An Empirical Evaluation of Similarity Measures for Time Series Classification

Joan Serrà, Josep Ll. Arcos

Artificial Intelligence Research Institute (IIIA-CSIC),
Spanish National Research Council,
08193 Bellaterra, Barcelona, Spain.

Abstract

Time series are ubiquitous, and a measure to assess their similarity is a core part of many computational systems. In particular, the similarity measure is the most essential ingredient of time series clustering and classification systems. Because of this importance, countless approaches to estimate time series similarity have been proposed. However, there is a lack of comparative studies using empirical, rigorous, quantitative, and large-scale assessment strategies. In this article, we provide an extensive evaluation of similarity measures for time series classification following the aforementioned principles. We consider 7 different measures coming from alternative measure 'families', and 45 publicly-available time series data sets coming from a wide variety of scientific domains. We focus on out-of-sample classification accuracy, but in-sample accuracies and parameter choices are also discussed. Our work is based on rigorous evaluation methodologies and includes the use of powerful statistical significance tests to derive meaningful conclusions. The obtained results show the equivalence, in terms of accuracy, of a number of measures, but with one single candidate outperforming the rest. Such findings, together with the followed methodology, invite researchers on the field to adopt a more consistent evaluation criteria and a more informed decision regarding the baseline measures to which new developments should be compared.

Keywords: Time Series, Similarity, Classification, Evaluation

1. Introduction

Data in the form of time series pervades a large number of scientific domains (Keogh, 2011; Keogh et al., 2011). Observations that unfold over time usually represent valuable information subject to analysis, classification, indexing, prediction, or interpretation (Kantz and Schreiber, 2004; Han and Kamber, 2005; Liao, 2005; Fu, 2011). Real-world examples include financial data (e.g., stock market fluctuations), medical data (e.g., electrocardiograms), computer data (e.g., log sequences), or motion data (e.g,. location of moving objects). Even object shapes or handwriting can be effectively transformed into time series, facilitating their analysis and retrieval (Keogh et al., 2009, 2011).

A core issue when dealing with time series is determining their pairwise similarity, i.e., the degree to which a given time series resembles another. In fact, a time series similarity (or dissimilarity) measure is central to many mining, retrieval, clustering, and classification tasks (Han and Kamber, 2005; Liao, 2005; Fu, 2011; Keogh and Kasetty, 2003). Furthermore, there is evidence that simple approaches to such tasks exploiting generic time series similarity measures usually outperform more elaborate, sometimes specifically-targeted strategies. This is the case, for instance, with time series classification, where a one-nearest neighbor approach using a well-known time series similarity measure was found to outperform an exhaustive list of alternatives (Xi et al., 2006), including decision trees, multi-scale histograms, multi-layer perceptron neural networks, order logic rules with boosting, or multiple classifier systems.

Deriving a measure that correctly reflects time series similarities is not straightforward. Apart from dealing with high dimensionality (time series can be roughly considered as multi-dimensional data), the calculation of such measures needs to be fast and efficient (Keogh and Kasetty, 2003). Indeed, with better information gathering tools, the size of time series data sets may continue to increase in the future. Moreover, there is the need for generic/multi-purpose similarity measures, so that they can be readily applied to any data set, whether this application is the final goal or just an initial approach to a given task. This last aspect highlights another desirable quality for time series similarity measures: their robustness to different types of data (cf. Keogh and Kasetty, 2003; Wang et al., 2012).

Over the years, several time series similarity measures have been proposed (for pointers to such measures see, e.g., Liao, 2005; Fu, 2011; Wang et al., 2012). Nevertheless, few quantitative comparisons have been made in order to evaluate their efficacy in a multiple-data framework. Apart from being an interesting and important task by itself, and as opposed to clustering, time series classification offers the possibility to straightforwardly assess the merit of time series similarity measures under a controlled, objective, and quantitative framework (Keogh and Kasetty, 2003).

In a recent study, Wang et al. (2012) perform an extensive comparison of classification accuracies for 9 measures (plus 4 variants) across 38 data sets coming from various scientific domains. One of the main conclusions of the study is that, even though the newly proposed measures can be theoretically attractive, the efficacy of some common and well-established measures is, in the vast majority of cases, very difficult to beat. Specifically, dynamic time warping (DTW; Berndt and Clifford, 1994) is found to be consistently superior to the other studied measures (or, at worst, for a few data sets, equivalent). In addition, the authors emphasize that the Euclidean distance remains a quite accurate, robust, simple, and efficient way of measuring the similarity between two time series. Finally, by looking in detail at the results presented by Wang et al. (2012), we can spot a group of time series similarity measures that seems to have an efficacy comparable to DTW: those based on edit distances. In particular, the edit distance for real sequences (EDR; Chen et al., 2005) seems to be very competitive, if not slightly better than DTW. Interestingly, none of the three measures above was initially targeted to generic time series data, but were introduced with hindsight (Agrawal et al., 1993; Berndt and Clifford, 1994; Chen et al., 2005). The intuition behind Euclidean distance relates to spatial proximity, DTW was initially devised for the specific task of spoken word recognition (Sakoe and Chiba, 1978), and edit distances were introduced for measuring the dissimilarity between two strings (Levenshtein, 1966).

The study by Wang et al. (2012) is, to the best of our knowledge, the only comparative study dealing with time series classification using multiple similarity measures and a large collection of data. In general, the studies introducing a new measure only compare against a few other measures[1], and usually using a reduced data set corpus (cf. Keogh and Kasetty, 2003). Furthermore, there is a lack of agreement in the literature regarding evaluation methodologies. Besides, statistical significance is usually not studied or, at best, improperly evaluated. This is very inconvenient, as robust evaluation methodologies and statistical significance are the principal tools by which we can establish, in a formal and rigorous way, differences across the considered measures (Salzberg, 1997; Hollander and Wolfe, 1999; Demšar, 2006). In addition, the chosen parameter values for every measure are rarely discussed. All these issues impact the scientific development of the field as one is never sure, e.g., of which measure should be used as a baseline for future developments, or of which parameters are the most sensible choice.

[1]In the majority of cases, as our results will show, not the most appropriate ones.

In this work, we perform an empirical evaluation of similarity measures for time series classification. We follow the initiative by Wang et al. (2012), and consider a big pool of publicly-available time series data sets (45 in our case). However, instead of additionally focusing on representation methods, computational/storage demands, or more theoretical issues, we here take a pragmatic approach and restrict ourselves to classification accuracy. We believe that this is the most important aspect to be considered in a first stage and that, in contrast to the other aforementioned issues, it is not sufficiently well-covered in the existing literature. As for the considered measures, we decide to include DTW and EDR, as these were found to generally achieve the highest accuracies among all measures compared in Wang et al. (2012). Apart from these two, we choose the Euclidean distance plus 4 different measures not considered in such study, making up to a total of 7. Further important contributions that differentiate the current work from previous studies include (a) an extensive summary and background of the considered measures, with basic formulations, applications, and references, (b) the formalization of a robust evaluation methodology, exploiting standard out-of-sample cross-validation strategies, (c) the use of rigorous statistical significance tests in order to assess the superiority of a given measure, (d) the evaluation of both train and test accuracies, and (e) the assessment of the chosen parameters for each measure and data set.

The rest of the paper is organized as follows. Firstly, we provide the background on time series similarity measures, outline some of their applications, and detail their calculation (Sec. 2). Next, we explain the proposed evaluation methodology (Sec. 3). Subsequently, we report the obtained results (Sec. 4). A conclusion section ends the paper (Sec. 5).

2. Time series similarity measures

The list of approaches for dealing with time series similarity is vast, and a comprehensive enumeration of them all is beyond the scope of the present work (for that, the interested reader is referred to Gusfield, 1997; Wang et al., 2012; Han and Kamber, 2005; Liao, 2005; Marteau, 2009; Fu, 2011). In this section, we present several representative examples of different 'families' of time series similarity measures: lock-step measures (Euclidean distance), feature-based measures (Fourier coefficients), model-based measures (auto-regressive), and elastic measures (DTW, EDR, TWED, and MJC). An effort has been made in selecting the most standard measures of each group, emphasizing the approaches that are reported to have good performance. We also try to avoid measures with too many parameters, since such parameters may be difficult to learn in small training data sets and, furthermore, could lead to over-fitting. Alternative measures found to be consistently less accurate than DTW or EDR are not considered (see Wang et al., 2012). Apart from all the aforementioned measures, we also include a random dissimilarity measure, consisting of a uniformly distributed random number between 0 and 1. This will act as our random baseline, informing us of the error rates we can expect by chance. By comparing its accuracy to the one achieved by other measures, it also gives us qualitative information regarding their 'usefulness' or improved capacity for classification.

2.1. Euclidean distance

The simplest way to estimate the dissimilarity between two time series is to use any L_n norm such that

$$d_{\mathrm{L}_n}(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{M}(x_i - y_i)^n\right)^{\frac{1}{n}}, \tag{1}$$

where n is a positive integer, M is the length of the time series, and x_i and y_i are the i-th element of time series $\mathbf{x}$ and $\mathbf{y}$, respectively. Measures based on L_n norms correspond to the group of so-called lock-step measures (Wang et al., 2012), which compare samples that are at exactly the same temporal location (Fig. 1, top). Notice that in case the time series $\mathbf{x}$ and $\mathbf{y}$ not being of the same length, one can always re-sample one to the length of the other, an approach that works well for a number of data sources (Keogh and Kasetty, 2003).

Using Eq. 1 with $n = 2$ we obtain the Euclidean distance, one of the most used time series dissimilarity measures, favored by its computational simplicity and indexing capabilities. Applications range from early classification of time series (Xing et al., 2011) to rule discovery in economic, communications, and ecological time series (Das et al., 1998). Some authors state that the accuracy of the Euclidean distance can be very difficult to beat, specially for large data sets containing many time series (cf. Wang et al., 2012). To the best of our knowledge, these claims are only quantitatively supported by one-nearest neighbor classification experiments using two artificially-generated/synthetic data sets (Geurts, 2002). We believe that such claims need to be carefully assessed with extensive experiments and under broader conditions, considering multiple measures, different distance-exploiting algorithms, and real-world data sets.

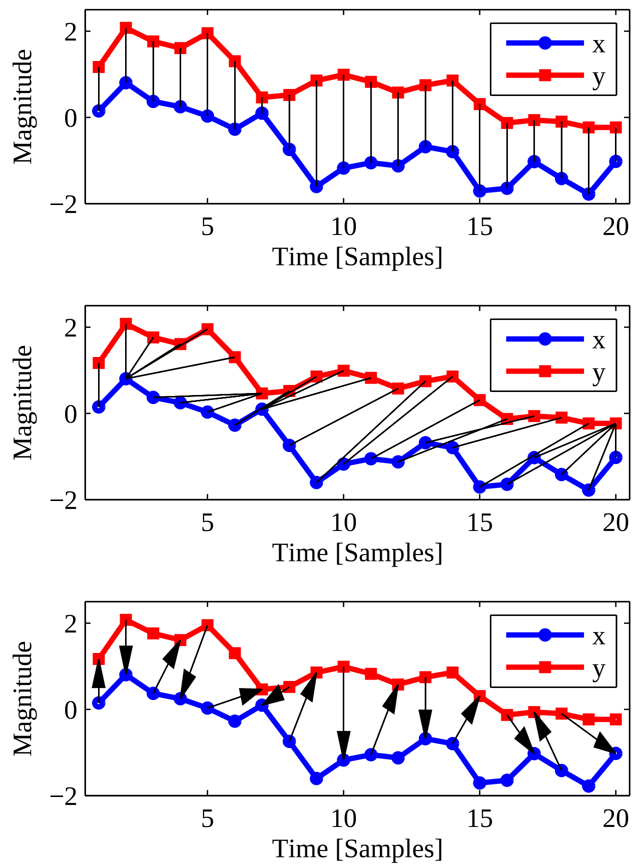

Figure 1: Examples of dissimilarity calculations between time series $\mathbf{x}$ and $\mathbf{y}$: Euclidean distance (top), DTW alignment (center), and MJC (bottom). See text for details.

2.2. Fourier coefficients

A simple extension of the Euclidean distance is not to compute it directly using the raw time series, but using features extracted from it. For instance, by first representing the time series by their Fourier coefficients (FC), one uses

$$d_{\mathrm{FC}}(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{\theta} (\hat{x}_i - \hat{y}_i)^2 \right)^{\frac{1}{2}}, \tag{2}$$

where $\hat{x}_i$ and $\hat{y}_i$ are complex value pairs denoting the i-th Fourier coefficient of $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$, the discrete Fourier transforms (DFT) of the raw time series (Oppenheim et al., 1999). Notice that in Eq. 2 we introduce the parameter θ, the actual number of considered coefficients. Because of the symmetry of the DFT, the sum only needs to be performed, at most, over half of the coefficients, so that $\theta = M/2$. Notice that, by the Parseval theorem (Oppenheim et al., 1999), the Euclidean distance between FCs is equivalent to the standard Euclidean distance between the raw time series (see, e.g., Agrawal et al., 1993). However, having parameter θ, one usually takes the opportunity to filter out high-frequency coefficients, i.e., coefficients $\hat{x}_i$ and $\hat{y}_i$ whose i is close to $M/2$. This has the (sometimes desired) effect of removing rapidly-fluctuating components of the signal. Hence, if high frequencies are not relevant for the intended analysis or we have some high-frequency noise, this operation will usually carry some increase in accuracy. Furthermore, if θ is relatively small, similarity computations can be substantially accelerated.

Computing the Euclidean distance on a reduced set of features is an extremely common approach in literature. FCs are the standard choice for efficient time series retrieval, exploiting the aforementioned acceleration capabilities. Pioneering work includes Agrawal et al. (1993) and Faloutsos et al. (1994) dealing with synthetic and financial data. More recent works use FCs with data from other domains. For instance, the case-based reasoning system of Montani et al. (2006) uses FCs to compare medical time series. Apart from FCs, wavelet coefficients have been extensively used (Chan and Fu, 1999). For instance, Olsson et al. (2004) use a wavelet analysis to remove noise and extract features in their system of fault diagnosis in industrial equipment. Research suggests that, although they provide some advantages, wavelet coefficients do not generally outperform FCs for the considered task (Wu et al., 2000). Comparatively less used time series features are based on singular value decomposition (Wu et al., 1996), piece-wise aggregate approximations (Keogh et al., 2001), or the coefficients of fitted polynomials (Cai and Ng, 2004) among others.

2.3. Auto-regressive models

A further option for computing similarities between time series using features extracted from them is to employ time series models (Liao, 2005; Fu, 2011). The main idea behind model-based measures is to learn a model of the two time series and then use its parameters for computing a similarity value. In the literature, several approaches follow this idea. For instance, Maharaj (2000) uses the p-value of a chi-square statistic to cluster auto-regressive coefficients representing stationary time series. Ramoni et al. (2002) present a Bayesian algorithm for clustering time series. They transform each series into a Markov chain and then cluster similar chains to discover the most probable set of generating processes. Povinelli et al. (2004) use Gaussian mixture models of reconstructed phase spaces to classify time series of different sources. Serrà et al. (2012a) study the use of the error of several learned models to identify similar time series corresponding to musical information.

In the present study we consider the use of auto-regressive (AR) models for time series feature extraction. Given an AR model of the form

$$x_i = a_0 + \sum_{j=1}^{\eta} a_j x_{i-j}, \tag{3}$$

where a_j denotes the j-th regression coefficient and η is the order of the model, we can estimate its coefficients, e.g., by the Yule-Walker function (Marple, 1987). Then, the dissimilarity between two time series can be calculated, for instance, using the Euclidean distance between their estimated coefficients, analogously as in Eq. 2 (Piccolo, 1990). The number of AR coefficients is controlled by the parameter η which, similarly to θ with FCs, directly affects the final speed of similarity calculations (AR and FCs are usually estimated offline, prior to similarity calculations).

2.4. Dynamic time warping

Dynamic time warping (DTW; Sakoe and Chiba, 1978; Berndt and Clifford, 1994) is a classic approach for computing the dissimilarity between two time series. It has been exploited in countless works: to construct decision trees (Rodríguez and Alonso, 2004), to retrieve similar shapes from large image databases (Bartolini et al., 2005), to match incomplete time series in medical applications (Tormene et al., 2009), to align signatures in an identity authentication task (Kholmatov and Yanikoglu, 2005), etc. In addition, several extensions for speeding up its calculations exist (Keogh and Ratanamahatana, 2005; Salvador and Chan, 2007; Lemire, 2009).

DTW belongs to the group of so-called elastic dissimilarity measures (Wang et al., 2012), and works by optimally aligning (or 'warping') the time series in the temporal domain so that the accumulated cost of this alignment is minimal (Fig. 1, center). In its canonical form, this accumulated cost can be obtained by dynamic programming, recursively applying

$$D_{i,j} = f(x_i, y_j) + \min\{D_{i,j-1}, D_{i-1,j}, D_{i-1,j-1}\} \tag{4}$$

for $i = 1, \ldots, M$ and $j = 1, \ldots, N$, being M and N the lengths of time series $\mathbf{x}$ and $\mathbf{y}$, respectively. Except for the first cell, which is initialized to

$D_{0,0} = 0$, the matrix D is initialized to $D_{i,j} = \infty$ for $i = 0, 1, \ldots, M$ and $j = 0, 1, \ldots, N$. In the case of dealing with uni-dimensional time series, the local cost function $f()$, also called sample dissimilarity function, is usually taken to be the square of the difference between x_i and y_j (Berndt and Clifford, 1994), i.e., $f(x_i, y_j) = (x_i - y_j)^2$. In the case of dealing with multidimensional time series or having some domain-specific knowledge, the local cost function $f()$ must be chosen appropriately, although the Euclidean distance is often used. The final DTW dissimilarity measure typically corresponds to the total accumulated cost, i.e., $d_{\text{DTW}}(\mathbf{x}, \mathbf{y}) = D_{M,N}$. A normalization of d_{DTW} can be performed on the basis of the alignment of the two time series, which is found by backtracking from $D_{M,N}$ to $D_{0,0}$ (Rabiner and Juang, 1993). However, in preliminary analysis we found the normalized variant to be equivalent, or sensibly less accurate, than the unnormalized one.

The canonical form of DTW presented in Eq. 4 can incorporate many variants. In particular, several constraints can be applied to the computation of D. A common constraint (Sakoe and Chiba, 1978) is to introduce a window parameter $\omega \in [0, N]$, such that the recursive formula of Eq. 4 is only applied for $i = 1, \ldots, M$ and

$$j = \max\{1, i' - \omega\}, \ldots, \min\{N, i' + \omega\}, \tag{5}$$

where i' is progressively adjusted for dealing with different time series lengths, i.e., $i' = \lfloor iN/M \rceil$, using $\lfloor\ \rceil$ as the round-to-the-nearest-integer operator. Notice that if $\omega = 0$ and $N = M$, d_{DTW} will correspond to the squared Euclidean distance (the value in $D_{M,N}$ will be the sum of the squared differences, see Eqs. 1 and 4). Notice furthermore that, when $\omega = N$, we are using the unconstrained version of DTW (the constraints in Eq. 5 have no effect). Thus, we include two DTW variants in a single formulation. In general, the introduction of constraints, and specially of the window parameter ω, carries some advantages (Keogh and Kasetty, 2003; Rabiner and Juang, 1993; Wang et al., 2012). For instance, constraints prevent 'pathological alignments' and, therefore, usually provide better similarity estimates (pathological alignments typically go beyond the main diagonal of D). Moreover, constraints allow for reduced computational costs, since only a percentage of the cells in D needs to be examined (Sakoe and Chiba, 1978; Rabiner and Juang, 1993).

DTW currently stands as the main benchmark against which new similarity measures need to be compared (Xi et al., 2006; Wang et al., 2012). Very few measures have been proposed that systematically outperform DTW for a number of different data sources. These measures are usually more complex than DTW, sometimes requiring extensive tuning of one or more parameters. Additionally, it is often the case that no careful, rigorous, and extensive evaluation of the accuracy of such measures is done, and further studies fail to assess the statistical significance of their improvement. Thus we could say that the superiority of such measures is, at best, unclear. In this paper, we pay special attention to all these aspects in order to formally assess the considered measures under a common framework. As it will be shown, there exists a similarity measure outperforming DTW for a statistically significant margin (Sec. 4).

2.5. Edit distance on real sequences

Turning to previous evidence (Wang et al., 2012), we observe that perhaps the only measure able to seriously challenge DTW is the edit distance on real sequences (EDR; Chen et al., 2005). The EDR corresponds to the extension of the original edit or Levensthein distance (Levenshtein, 1966) to real-valued time series. Such extensions are not commonplace, but recent research is starting to focus on them (Morse and Patel, 2007; Marteau, 2009). As noted by Chen et al. (2005), EDR outperformed previous edit distance variants for time series similarity.

The computation of the EDR can be formalized by a dynamic programming approach. Specifically, we compute

$$D_{i,j} = \begin{cases} D_{i-1,j-1} & \text{if } m(x_i, y_j) = 1 \\ 1 + \min\left\{D_{i,j-1}, D_{i-1,j}, D_{i-1,j-1}\right\} & \\ & \text{if } m(x_i, y_j) = 0, \end{cases} \tag{6}$$

for $i = 1, \ldots, M$ and $j = 1, \ldots, N$. The match function used is

$$m(x_i, y_j) = \Theta\left(\varepsilon - f\left(x_i, y_j\right)\right), \tag{7}$$

where $\Theta()$ is the Heaviside step function such that $\Theta(z) = 1$ if $z \geq 0$ and 0 otherwise, and $\varepsilon \in [0, \infty)$ is a suitably chosen threshold parameter that controls the degree of resemblance between two time series samples being considered as a match. The first row of D is initialized to $D_{i,0} = i$ for $i = 0, 1, \ldots, M$ and the first column of D to $D_{0,j} = j$ for $j = 0, 1, \ldots, N$. Following Chen et al. (2005), who initially reported some accuracy improvements of EDR over DTW, we set the local cost function $f()$ to the absolute difference between the sample values, i.e., $f(x_i, y_j) = |x_i - y_j|$. This has the additional advantage that we can easily relate ε to the standard deviation of the time series (Sec. 3.5).

2.6. Time-warped edit distance

The time-warped edit distance (TWED; Marteau, 2009) is perhaps the most interesting extension of dynamic programming algorithms like DTW and EDR. In a sense, it is a combination of these two. Like edit distances, TWED comprises a mismatch penalty λ and, like dynamic time warping, it introduces a so-called stiffness parameter ν, controlling its 'elasticity' (Marteau, 2009). For uniformly-sampled time series, the formulation of TWED corresponds to

$$D_{i,j} = \min\left\{D_{i,j} + \Gamma_{\mathbf{xy}}, D_{i-1,j} + \Gamma_{\mathbf{x}}, D_{i,j-1} + \Gamma_{\mathbf{y}}\right\}, \tag{8}$$

for $i = 1, \ldots, M$ and $j = 1, \ldots, N$, with

$$\begin{array}{rcl} \Gamma_{\mathbf{xy}} & = & f(x_i, y_j) + f(x_{i-1}, y_{j-1}) + 2\nu|i-j|, \\ \Gamma_{\mathbf{x}} & = & f(x_i, x_{i-1}) + \nu + \lambda, \\ \Gamma_{\mathbf{y}} & = & f(y_j, y_{j-1}) + \nu + \lambda, \end{array} \tag{9}$$

where $f()$ can be any $\mathrm{L_n}$ metric (Eq. 1). Following Marteau (2009), and as done for EDR as well, we choose $f(x_i, y_j) = |x_i - y_j|$. Together with DTW and EDR, the final dissimilarity value is taken to be $d_{\mathrm{TWED}}(\mathbf{x}, \mathbf{y}) = D_{M,N}$.

An interesting aspect of TWED is that, in its original formulation (Marteau, 2009), it takes time stamp differences into account. Therefore, it is able to cope with time series of different sampling rates, including down-sampled time series. A further interesting aspect, and contrasting to DTW and other measures, is that TWED is a metric (Marteau, 2009). Thus, one can exploit the triangular inequality to speed up the search in the metric space. Finally, it is worth mentioning that the combination of the two previous characteristics results in a lower bound of the TWED dissimilarity, which can be used to speed up nearest neighbor retrieval.

2.7. Minimum jump costs dissimilarity

The main idea behind the minimum jump costs dissimilarity measure (MJC; Serrà and Arcos, 2012) is that, if a given time series $\mathbf{x}$ resembles $\mathbf{y}$, the cumulative cost of iteratively 'jumping' between their samples should be small[2] (Fig. 1, bottom). Supposing that for the i-th jump we are at time step t_x

[2] An implementation of MJC is made available online by the authors: `http://www.iiia.csic.es/~jserra/downloads/2012_SerraArcos_MJC-Dissim.tar.gz` (last accessed on September 15, 2013).

of time series $\mathbf{x}$, and that we previously visited time step $t_y - 1$ of $\mathbf{y}$, the minimum jump cost is expressed as

$$c_{\min}^{(i)} = \min \left\{ c_{t_x}^{t_y}, c_{t_x}^{t_y+1}, c_{t_x}^{t_y+2}, \dots \right\}, \quad (10)$$

where $c_{t_x}^{t_y+\Delta}$ is the cost of jumping from x_{t_x} to $y_{t_y+\Delta}$ and $\Delta = 0, 1, 2, \dots$ is an integer time step increment such that $t_y + \Delta \leq N$. After a jump is made, t_x and t_y are updated accordingly: t_x becomes $t_x + 1$ and t_y becomes $t_y + \Delta + 1$. In case we want to jump from $\mathbf{y}$ to $\mathbf{x}$, only t_x and t_y need to be swapped (Serrà and Arcos, 2012).

To define a jump cost $c_{t_x}^{t_y+\Delta}$, the temporal and the magnitude dimensions of the time series are considered:

$$c_{t_x}^{t_y+\Delta} = (\phi\Delta)^2 + f(x_{t_x}, y_{t_y+\Delta}), \quad (11)$$

where ϕ represents the cost of advancing in time and $f()$ is the local cost function, which we take to be $f(x_{t_x}, y_{t_y+\Delta}) = (x_{t_x} - y_{t_y+\Delta})^2$, similarly to what is done with DTW (Eq. 4). Notice that, akin to the general formulation of TWED, the term $(\phi\Delta)^2$ introduces a nonlinear penalty that depends on the temporal gap. Here, the value of ϕ is set proportional to the standard deviation σ expected for the time series and, at the same time, proportional to the real-valued parameter $\beta \in [0, \infty)$, which controls how difficult is to advance in time (for more details see Serrà and Arcos, 2012). To obtain a symmetric dissimilarity measure, $d_{\text{MJC}}(\mathbf{x}, \mathbf{y}) = \min\{d_{\text{XY}}, d_{\text{YX}}\}$ can be used, where d_{XY} and d_{YX} are the cumulative MJCs obtained by starting at x_1 and y_1, respectively.

3. Evaluation methodology

3.1. Classification scheme

The efficacy of a time series similarity measure is commonly evaluated by the classification accuracy it achieves (Keogh and Kasetty, 2003; Wang et al., 2012). For that, the error ratio of a distance-based classifier is calculated for a given labeled data set, understanding the error ratio as the number of wrongly classified items divided by the total number of tested items. The standard choice for the classifier is the one-nearest neighbor (1NN) classifier. Following Wang et al. (2012), we can enumerate several advantages of using this approach. First, the error of the 1NN classifier critically depends on the similarity measure used. Second, the 1NN classifier is parameter-free and easy to implement. Third, there are theoretical results relating the error of an 1NN classifier to errors obtained with other classification schemes. Fourth, some works suggest that the best results for time series classification come from simple nearest neighbor methods. For more details on these aspects we refer to Mitchell (1997); Hastie et al. (2009), and the references provided by Wang et al. (2012).

3.2. Data sets

We perform experiments with 45 publicly-available time series data sets from the UCR time series repository (Keogh et al., 2011). This is the world's biggest time series repository, and some authors estimate that it makes up to more than 90% of all publicly-available, labeled data sets (Wang et al., 2012). The repository comprises synthetic, as well as real-world data sets, and also includes one-dimensional time series extracted from two-dimensional shapes (Keogh et al., 2011). The 45 data sets considered here correspond to the totality of the UCR repository, as by March 2013. Within such data sets, the number of classes ranges from 2 to 50, the number of time series per data set ranges from 56 to 9,236, and time series lengths go from 24 to 1,882 samples. For further details on these data sets we refer to (Keogh et al., 2011).

3.3. Cross-validation

To properly assess a classifier's error, out-of-sample validation needs to be done (Salzberg, 1997). In our experiments, we follow a standard 3-fold cross-validation scheme using balanced data sets (Mitchell, 1997; Hastie et al., 2009), i.e., using the same number of items per class. We repeat the validation 20 times and report average error ratios. Balancing the data sets allows for balanced error estimations regarding the class distribution, and repeating cross-fold validation several times allows for more precise estimations (Mitchell, 1997; Hastie et al., 2009). The use of a cross-fold validation scheme is essential for avoiding the bias that a particular split of the data could introduce (Salzberg, 1997; Hastie et al., 2009).

We also computed error ratios for the original splits provided in the UCR time series repository (Keogh et al., 2011). This allowed us to confirm that the 1NN error ratios from our implementations of DTW and Euclidean distance agree with the values reported there. In addition, we observed that the error ratios obtained by such splits were substantially different from the ones obtained by cross-validation, up to the point of even modifying the ranking of some algorithms with respect to those error ratios in some data sets. This indicates a potential bias in such individual splits, an aspect that is well-known in the machine learning community (Salzberg, 1997; Mitchell, 1997; Hastie et al., 2009). We refer the interested reader to any machine learning textbook for a more in-depth discussion of cross-fold validation schemes and their appropriateness over individual splits. Besides, individual splits difficult statistical significance assessment (see below). A full account of the raw error ratios for all measures and data sets is available online[3], including the error ratios for the aforementioned original splits.

3.4. Statistical significance

To assess the statistical significance of the difference between two error ratios we employ the well-known Wilcoxon signed-rank test (Hollander and Wolfe, 1999). The Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two repeated measurements (or related samples, or matched samples) in order to assess whether their population mean ranks differ. It is the natural alternative to the Student's t-test for dependent samples when the population distribution cannot be assumed to be normal (Hollander and Wolfe, 1999). For a given data set, we use as input the 20×3 accuracy values obtained for each classifier (i.e., the test fold accuracies). Besides, for comparing similarity measures on a more global basis using all data sets, we employ as input the 45 average accuracy values obtained for each data set. Following common practice (Salzberg, 1997; Hollander and Wolfe, 1999), the threshold significance level is set to 5%. Additionally, to compensate for multiple pairwise comparisons, we apply the Holm-Bonferroni method (Holm, 1979), a post-hoc statistical analysis method controlling the so-called family-wise error rate that is more powerful than the usual Bonferroni correction (Demšar, 2006).

3.5. Parameter choices

Before performing the experiments, all time series from all data sets were z-normalized so that each individual time series had zero mean and unit variance. Furthermore, we optimized the measures' parameters in the training phase of our cross-validation. This optimization step consisted of a grid search within a suitable range of parameter values, forcing the same number of parameter combinations per algorithm (Table 1). The values of the grid are chosen according to common practice and the specifications given in the papers introducing each measure (Sec. 2). Specifically, for FC we used 25 linearly-spaced integer values of $\theta \in [2, N/2]$. For AR we used 25 linearly-spaced integer values of $\eta \in [1, 0.25N]$ (because of the z-normalization, we

[3] `http://www.iiia.csic.es/~jserra/downloads/2013_SerraArcos_AnEmpiricalEvaluation.tar.gz` (last accessed on September 15, 2013).

Measure	Parameter	Minimum value	Maximum value	Number of steps	Extra value
FC	θ	2	$0.5N$	25	-
AR	η	1	$0.25N$	25	-
DTW	ω	0	$0.25N$	24	N
EDR	ε	0.02σ	σ	25	-
TWED	ν	10^{-5}	1	5	-
TWED	λ	0	1	5	-
MJC	β	0	25	24	10^{10}

Table 1: Parameter grid for the considered similarity measures (recall that N corresponds to the length of the time series and, since we z-normalize all time series, $\sigma = 1$). For DTW and MJC we consider an extra value corresponding to unconstrained DTW and to the Euclidean configuration of MJC, respectively. All parameter values were linearly spaced except ν, which was logarithmically spaced.

remove a_0 in Eq. 3). For DTW we used 24 linearly-spaced integer values of $\omega \in [0, 0.25N]$ plus $w = N$, the unconstrained DTW variant (we also considered $\omega \in [0, 0.1N]$ and $\omega \in [0, 0.15N]$, but obtained no statistically significant differences from $\omega \in [0, 0.25N]$ and none of the overall results changed; considering $\omega \in [0, 0.05N]$ made DTW closely approach the results of the Euclidean distance). For EDR we used 25 linearly-spaced real values of $\varepsilon \in [0.02\sigma, \sigma]$, σ being the standard deviation of the time series (because of the z-normalization $\sigma = 1$). For TWED we used all possible 25 combinations for $\nu = [10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1]$ and $\lambda = [0, 0.25, 0.5, 0.75, 1]$. For MJC we used 24 linearly-spaced real values of $\beta \in [0, 25]$ plus $\beta = 10^{10}$ (in practice corresponding to the squared Euclidean distance variant, Eq. 11). After the grid search, the parameter value yielding to the lowest leave-one-out error ratio for the training set was kept for out-of-sample testing.

4. Results

4.1. Classification performance: test

If we look at the overall results, we see that all considered measures clearly outperform the random baseline for practically all the 45 data sets (Table 2). Furthermore, we see that some of them achieve near-perfect accuracies for a number of data sets (e.g., *CBF*, *CinC_ECG_torso*, *ECGFiveDays*, *Two_Patterns*, or *TwoLeadECG*). However, no single measure achieves the best performance for all the data sets. The Euclidean distance is found to be the best-performing measure in 2 data sets, FC is the best-performing in 4 data sets, AR in 1, DTW in 6, EDR in 7, TWED in 20, and MJC in 5. If we count only the data sets where one measure statistically significantly outperforms the rest, the numbers reduce to 0 for Euclidean, 2 for FC, 1 for AR, 2 for DTW, 2 for EDR, 6 for TWED, and 0 for MJC. Thus, interestingly, there are some data sets where choosing a specific similarity measure can make a difference.

Beyond accuracies, this latter aspect can potentially highlight inherent data set qualities. For instance, the fact that a feature/model-based measure clearly outperforms the others for a particular data set indicates that such time series may be very well characterized by the extracted features/fitted model (e.g., FC with *Adiac* for features and AR with *ChlorineConcentration* for models). In addition, the good or bad performance of Euclidean and elastic measures gives us an intuition of the importance of alignments, warping, or sample correspondences (e.g., these may be very important for *Trace* and the three *Face* data sets, where there is an order of magnitude difference between Euclidean and warping-based measures, but not much for *DiatomSizeReduction* or *NonInvasiveFetalECG2*, where Euclidean gets numbers that are very close, or even better than the ones obtained by the warping-based measures).

In general, we see that TWED outperforms the other measures in several data sets, with an average rank of 2.29 (Table 2). In fact, if we compare the considered measures on a more global scale, taking the matched error ratios across data sets (Sec. 3.4), we obtain that TWED is statistically significantly superior to the rest (Fig. 2). Next, we see that DTW, MJC, and EDR form a group of equivalent measures, with no statistically significant difference between them. The performed statistical analysis also separates the remaining measures from these and also between themselves. Apart from this more global analysis, further pairwise comparisons can be made, confirming the aforementioned global tendencies (Fig. 3).

4.2. Classification performance: test vs. train

For choosing the parameters for a given measure and data set we solely dispose of the training data. Hence, it is important to know whether the error ratios for training and testing sets are similar, otherwise one could be incurring into the so-called "Texas sharpshooter fallacy" (Batista et al., 2011), i.e., one could not predict a measure's utility ahead of time by just looking at training data. For comparing train and test error ratios, we can compute an error gain value for a couple of measures on each data set and check whether such values for train and test agree. To do so, a kind of real-valued contingency table can be plotted, called the "Texas sharpshooter plot" by Batista et al. (2011). Due to space reasons, we here only show such contingency tables for TWED against DTW and Euclidean distance (Fig. 4).

#	Data set	Random	Euc	FC	AR	DTW	EDR	TWED	MJC
1	*50words*	0.969	0.503	0.685	0.867	0.332	0.289	**0.237***	0.319
2	*Adiac*	0.970	0.345	**0.266***	0.725	0.355	0.423	0.335	0.346
3	*Beef*	0.763	0.417	**0.390**	0.504	0.472	0.439	0.506	0.448
4	*CBF*	0.655	0.013	0.358	0.432	0.000	0.002	**0.000**	0.001
5	*ChlorineConcentration*	0.673	0.071	0.063	**0.038***	0.072	0.094	0.093	0.070
6	*CinC_ECG_torso*	0.749	0.002	0.008	0.102	0.001	**0.000***	0.001	0.002
7	*Coffee*	0.394	0.019	0.024	0.139	**0.014**	0.031	0.021	0.023
8	*Cricket_X*	0.913	0.378	0.348	0.713	0.209	0.237	**0.190***	0.253
9	*Cricket_Y*	0.928	0.423	0.411	0.814	0.222	0.224	**0.209**	0.267
10	*Cricket_Z*	0.920	0.380	0.353	0.731	0.212	0.235	**0.194***	0.254
11	*DiatomSizeReduction*	0.744	0.008	0.011	0.222	0.010	0.016	0.012	**0.007**
12	*ECG200*	0.515	0.130	0.145	0.227	0.139	0.148	**0.109**	0.130
13	*ECGFiveDays*	0.505	0.007	**0.000**	0.072	0.003	0.003	0.005	0.001
14	*FaceAll*	0.931	0.139	0.152	0.649	0.053	0.019	**0.019**	0.034
15	*FaceFour*	0.679	0.111	0.149	0.545	0.069	0.028	0.025	**0.024**
16	*FacesUCR*	0.929	0.138	0.148	0.648	0.052	0.019	**0.018**	0.041
17	*Fish*	0.871	0.183	0.234	0.617	0.184	**0.084**	0.094	0.114
18	*Gun_Point*	0.506	0.058	0.031	0.149	0.023	**0.010**	0.017	0.014
19	*Haptics*	0.793	0.604	0.610	0.678	0.554	0.611	**0.544**	0.563
20	*InlineSkate*	0.862	0.524	0.601	0.497	0.462	0.456	0.416	**0.411**
21	*ItalyPowerDemand*	0.489	0.035	0.083	0.261	**0.033**	0.042	0.036	0.034
22	*Lighting2*	0.488	0.297	0.281	0.450	0.162	0.220	**0.161**	0.254
23	*Lighting7*	0.817	0.371	0.463	0.707	**0.252**	0.362	0.256	0.336
24	*Mallat*	0.870	0.018	0.020	0.058	0.015	0.006	**0.006**	0.014
25	*MedicalImages*	0.912	0.313	0.455	0.458	0.247	0.330	**0.228**	0.305
26	*MoteStrain*	0.513	0.087	0.162	0.336	0.058	0.024	**0.021**	0.034
27	*NonInvasiveFetalECG1*	0.978	0.171	0.213	0.401	0.175	0.186	0.182	**0.169**
28	*NonInvasiveFetalECG2*	0.975	**0.106**	0.146	0.296	0.107	0.118	0.108	0.110
29	*OliveOil*	0.644	**0.104**	0.185	0.663	0.154	0.194	0.146	0.127
30	*OSULeaf*	0.832	0.409	0.306	0.617	0.359	**0.191***	0.232	0.256
31	*SonyAIBORobotSurface*	0.510	0.017	0.040	0.079	0.018	0.026	0.017	**0.015**
32	*SonyAIBORobotSurfaceII*	0.489	0.018	0.032	0.113	0.021	0.023	**0.016**	0.019
33	*StarLightCurves*	0.671	0.124	**0.070***	0.274	0.083	0.107	0.097	0.109
34	*SwedishLeaf*	0.932	0.196	0.142	0.376	0.129	0.101	**0.094**	0.100
35	*Symbols*	0.838	0.038	0.074	0.260	0.019	**0.015**	0.016	0.018
36	*Synthetic_control*	0.834	0.087	0.393	0.511	**0.009***	0.047	0.014	0.034
37	*Trace*	0.757	0.169	0.117	0.117	**0.000***	0.034	0.011	0.038
38	*Two_Patterns*	0.743	0.020	0.491	0.724	0.000	0.000	**0.000**	0.001
39	*TwoLeadECG*	0.507	0.006	0.012	0.202	0.001	0.002	**0.001**	0.003
40	*UWaveGestureLibrary_X*	0.872	0.234	0.566	0.694	0.199	0.214	**0.192***	0.203
41	*UWaveGestureLibrary_Y*	0.876	0.288	0.631	0.645	**0.263**	0.280	0.265	0.267
42	*UWaveGestureLibrary_Z*	0.879	0.298	0.546	0.678	0.265	0.271	**0.250***	0.261
43	*Wafer*	0.497	0.004	0.003	0.013	0.005	**0.002**	0.003	0.005
44	*WordsSynonyms*	0.960	0.496	0.675	0.855	0.327	0.304	**0.251***	0.310
45	*Yoga*	0.500	0.070	0.108	0.333	0.061	**0.034**	0.037	0.047
	Average rank	7.99	4.40	5.07	6.80	3.00	3.42	**2.29**	3.04

Table 2: Error ratios for all considered measures and data sets. The symbol * denotes a statistically significant difference with respect to the other measures for a given data set ($p < 0.05$, Sec. 3.4). The last row contains the average rank of each measure across all data sets (i.e., the average position after sorting the errors for a given data set in ascending order).

The results show that error gains between TWED and DTW/Euclidean mostly agree between training and testing. As mentioned in Sec. 3.3, a full, raw account of train and test errors is available online. Having a close look

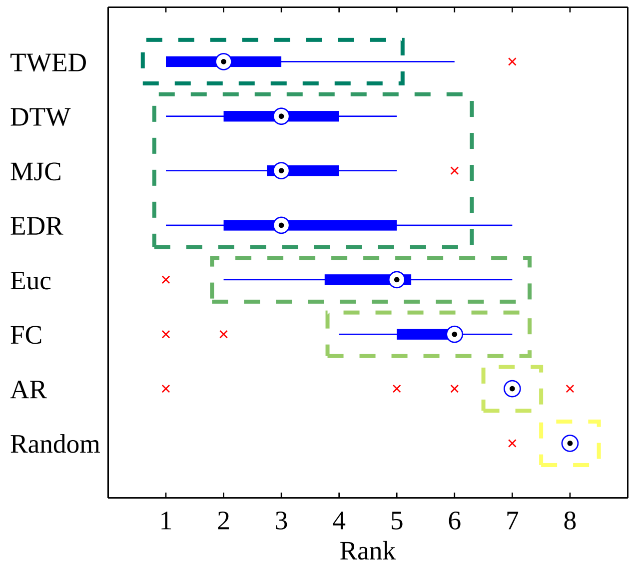

Figure 2: Box plot for the distribution of performance ranks of each measure across data sets. The dashed lines denote statistically significantly equivalent groups of measures ($p < 0.05$, Sec. 3.4).

at those full results, we can see that, in general, the best-performing measure at the training stage is also the best-performing measure at the testing stage. The few exceptions can be easily listed (Table 3). The relative rankings for the measures that do not perform best also mostly agree between train and test.

4.3. Parameter assessment

We finally report on the parameters chosen for each measure after training with 66% of balanced data (Fig. 5). Firstly, we observe that, in the vast majority of cases, a specific value for a given parameter is consistently chosen across the 20×3 performed iterations (we see clear peaks in the distributions of Fig. 5). Among these consistent choices, perhaps TWED and MJC present the most spread distributions. Such aspect, together with the fairly good accuracies obtained for these two specific measures (Sec. 4.1), indicates a certain degree of robustness against specific parameter choices. This is a very desirable quality of a time series similarity measure, even more if we have to train a classifier with a potentially incomplete set of training instances.

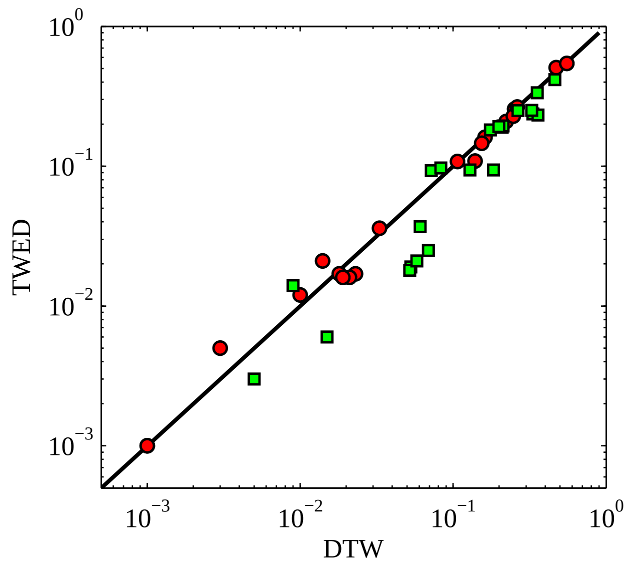

Figure 3: Error ratios comparison between DTW and TWED (notice the logarithmic axes). The lower-right triangular part corresponds to TWED outperforming DTW, whereas the upper-left part corresponds to the opposite case. The green squares indicate statistically significant performance differences ($p < 0.05$, Sec. 3.4).

Next, we see that the selected parameters are generally not in the borders of the specified ranges, thus indicating that a reasonable choice for these has been made (Fig 5). This is particularly true for DTW and EDR. Moreover, in the case of DTW, we see that ω values generally coincide with the ones suggested in the original data source (Keogh et al., 2011). In a total of 45 data sets we see 20 coincidences within ± 0.02 and 32 coincidences within ± 0.04. The only measure that could potentially benefit from reconsidering the parameters' range is TWED. As it can be seen, ν and λ seem to be consistently chosen in the lower and upper parts of the specified ranges, respectively. This suggests that the best combination for some data sets could lie outside the parameter space outlined by Marteau (2009), i.e., in $0 < \nu < 10^{-4}$ and/or $\lambda > 1$. If that was the case, TWED could potentially achieve even much higher accuracies. Interestingly, TWED is not the best-performing measure for some of the data sets where 'border' parameter values are chosen (e.g., *CBF*, *Fish*, *StarLightCurves*, *TwoPatterns*).

Finally, we can comment on the particularities of some data sets with relation to classification. For instance, we see that a relatively large window

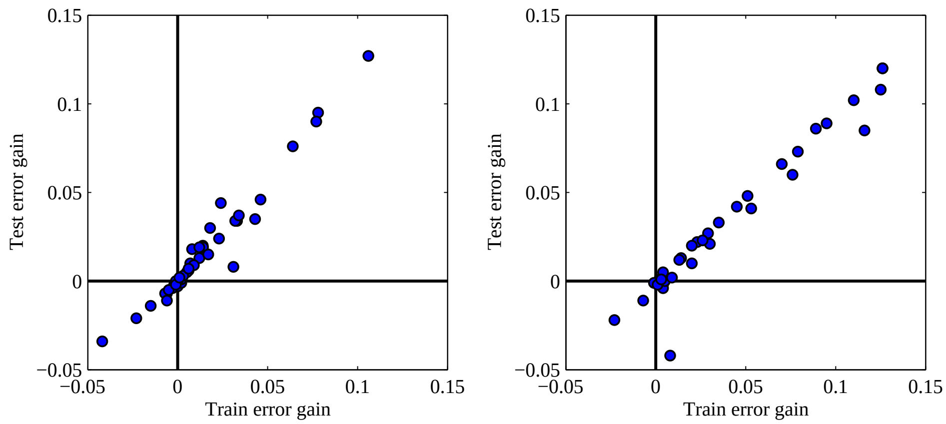

Figure 4: Texas sharpshooter plots for TWED against DTW (left) and Euclidean distance (right). Here, error gain is measured by subtracting the TWED error ratio from the one of DTW/Euclidean. Dots around the diagonal indicate agreement of error gain for train and test. False positives, i.e., dots in the lower-right quadrant, indicate that TWED, being the best measure after training, does not reach the lowest error at testing. For instance, in the case of TWED vs. Euclidean (right), the *OliveOil* data set false positive stands out at coordinates $(0.008, -0.042)$ (see also Table 3). For further details on the construction of Texas sharpshooter plots we refer to Batista et al. (2011).

parameter ω (DTW) is chosen for data sets 36 to 39 (i.e., *Synthetic_control*, *Trace*, *Two_Patterns*, and *TwoLeadECG*). This denotes that tracking alignments or warping paths beyond the main diagonal of D (Eq. 4) might be advantageous for classification in these data sets. In fact, when we re-ran the same experiment restricting ω to be between 0 and 0.1 we obtained the same or worse error rates, an effect that can also be observed by comparing the results obtained for the UCR splits (Keogh et al., 2011) which, as mentioned in Sec. 3.3, are available online. The stiffness parameter ν (TWED), which accounts for a similar but opposite concept (Sec. 2.6), takes relatively small values. Such agreement across different measures reinforces the hypothesis that tracking intricate alignments or strongly warped paths may be advantageous for these data sets. Analogous and complementary conclusions can be derived for other data sets. For instance, in data sets 11 (*DiatomSizeReduction*) and 13 (*ECGFiveDays*), a small number of both FCs θ and AR coefficients η is chosen. As FC and AR achieve competitive accuracies in those specific data sets, we could suspect that low-frequency components are important for correctly classifying the instances in those data sets (Secs. 2.2 and 2.3).

#	Data set	Measure	Outperf. by	Gain
3	*Beef*	FC	EDR	0.049
4	*CBF*	TWED	DTW	<0.001
7	*Coffee*	DTW	FC	0.004
12	*ECG200*	TWED	MJC	0.002
15	*FaceFour*	MJC	EDR	0.007
18	*Gun_Point*	EDR	MJC	0.004
19	*Haptics*	TWED	MJC	0.009
21	*ItalyPowerDemand*	DTW	EDR	0.001
28	*NonInvasiveFetalECG2*	Euclidean	TWED	0.001
29	*OliveOil*	Euclidean	TWED	0.008
39	*TwoLeadECG*	TWED	DTW	<0.001

Table 3: List of best-performing measures in testing (the column "Measure") but actually outperformed by others in training (the column "Outperf. by"). The column "Gain" corresponds to the absolute value of the train error gain, i.e., the absolute difference between error ratios at training stage (see also Fig. 4).

5. Conclusion

From a general perspective, the obtained results show that there is a group of equivalent similarity measures, with no statistically significant differences among them (DTW, EDR, and MJC). The existing literature suggests that some longest common sub-sequence approaches, together with alternative variants of DTW and EDR, could potentially join this group (Marteau, 2009; Wang et al., 2012). However, according to the results reported here, the TWED measure originally proposed by Marteau (2009) seems to consistently outperform all the considered distances, including DTW, EDR, and MJC. Thus, we believe this often unconsidered measure should take a baseline role in future evaluations of time series similarity measures (beyond accuracy, additional properties enumerated in Sec. 2.6 make it also very attractive). The Euclidean distance, although somehow competitive, generally performs statistically significantly worse than TWED, DTW, MJC, and EDR. Its accuracy on large data sets was also not very impressive. Below Euclidean distance, but statistically significantly above the random baseline, we find FC and AR measures. Of course, the general statements above do not exclude the possibility that a particular measure or variant could be very well-suited for a specific data set and statistically significantly outperform the rest (cf. Keogh and Kasetty, 2003). In Sec. 4.1 have enumerated several

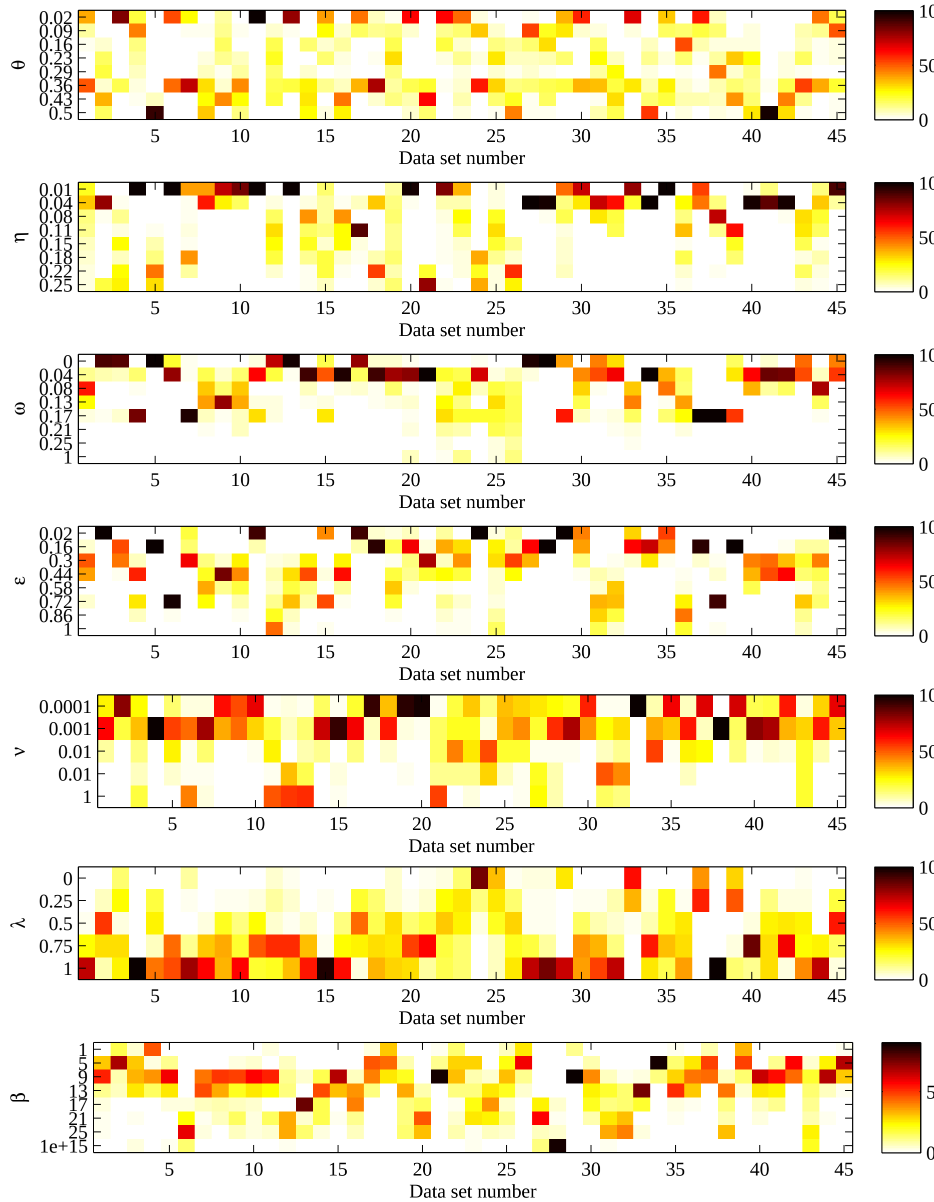

Figure 5: Percentage of times (color code) that a given parameter value (vertical axis) is chosen for each data set (horizontal axis; for the names behind each number see Table 2). From top to bottom, the plots correspond to FC (θ), AR (η), DTW (ω), EDR (ε), TWED (ν), TWED (λ), and MJC (β).

examples of that.

When comparing train and test errors, we have seen that these mostly agree, with train errors generally providing a good guess of the test errors on unseen data. We have listed some notable exceptions to this rule and used Texas sharpshooter plots to further assess this aspect for TWED vs. DTW and Euclidean. When assessing the best parameter choices for each measure, we have seen that the considered ranges are typically suitable for the task at hand. We have also discussed some particularities regarding parameter choices and the nature of a few data sets.

The similarity measure is a crucial step in computational approaches dealing with time series. However, there are some additional issues worth mentioning, in particular with regard to post-processing steps focused on improving similarity assessments (pre-processing steps are sufficiently well-discussed in the existing literature, see, e.g., Keogh and Kasetty (2003); Han and Kamber (2005); Wang et al. (2012) and references therein). A very interesting post-processing step is the complexity-invariant correction factor introduced by Batista et al. (2011). Such correction factor prevents from assigning low dissimilarity values to time series of different complexity, thus preventing the inclusion of time series of different nature in the same cluster. The way to assess complexity depends on the situation, but Batista et al. (2011) introduce a quite straightforward way: the L_2 norm of the sample-based derivative of a time series. Overall, considering different types of 'invariance' is a sensible approach (Batista et al., 2011, provide a good overview). Here, we have already implicitly considered a number of them, although more as a pre-processing or method-specific strategy: global amplitude and scale invariance (z-normalization), warping invariance (any elastic measure, in our case DTW, EDR, TWED, and MJC), phase invariance (AR[4]), and occlusion invariance (EDR and TWED).

Another interesting post-processing step is the hubness correction for time series classification introduced by Radovanovič et al. (2010). Based on the finding that some instances in high-dimensional spaces tend to become hubs by being unexpectedly (and usually wrongly) considered nearest neighbors of several other instances, a correction factor can be introduced. This usually does not harm classification accuracy and can definitely improve performance for some data sets (Radovanovič et al., 2010). A further strategy for enhancing time series similarity and potentially reducing hubness is the use of unsupervised clustering algorithms to prune nearest neighbor candi-

[4]For FC we use both phase and magnitude (Sec. 2.2).

dates (Serrà et al., 2012b). Future work should focus on the real quantitative impact of strategies for enhancing time series similarity like the ones above, with a special emphasis on its impact to different measures and classification schemes.

The empirical comparison of multiple approaches across a large-scale case basis is an important and necessary step towards any mature research field. Besides getting a more global picture and highlighting relevant approaches, it pushes towards unified validation procedures and analysis tools. It is hoped that this article will serve as a steppingstone for those interested in advancing in time series similarity, clustering, and classification.

Acknowledgements

We thank the people who made available or contributed to the UCR time series repository. This research has been funded by 2009-SGR-1434 from Generalitat de Catalunya, JAEDOC069/2010 from Consejo Superior de Investigaciones Científicas, and TIN2009-13692-C03-01 and TIN2012-38450-C03-03 from the Spanish Government, and EU Feder funds.

References

Agrawal, R., Faloutsos, C., Swami, A., 1993. Efficient similarity search in sequence databases, in: Proc. of the Int. Conf. on Foundations of Data Organization and Algorithms, pp. 69–84.

Bartolini, I., Ciaccia, P., Patella, M., 2005. WARP: accurate retrieval of shapes using phase of Fourier descriptors and time warping distance. IEEE Trans. on Pattern Analysis and Machine Intelligence 27, 142–147.

Batista, G.E.A.P.A., Wang, X., Keogh, E.J., 2011. A complexity-invariant distance measure for time series, in: Proc. of the SIAM Int. Conf. on Data Mining, pp. 699–710.

Berndt, D.J., Clifford, J., 1994. Using dynamic time warping to find patterns in time series, in: Proc. of the AAAI Workshop on Knowledge Discovery in Databases, pp. 359–370.

Cai, Y., Ng, R., 2004. Indexing spatio-temporal trajectories with Chebyshev polynomials, in: Proc. of the ACM SIGMOD Int. Conf. on Management of Data, pp. 599–610.

Chan, K.P., Fu, A.C., 1999. Efficient time series matching by wavelets, in: Proc. of the IEEE Int. Conf. on Data Engineering, pp. 126–133.

Chen, L., Öszu, M.T., Oria, V., 2005. Robust and fast similarity search for moving object trajectories, in: Proc. of the ACM SIGMOD Int. Conf. on Management of Data, pp. 491–502.

Das, G., Lin, K.I., Mannila, H., Renganathan, G., Smyth, P., 1998. Rule discovery from time series, in: Proc. of the AAAI Int. Conf. on Knowledge Discovery and Data Mining, pp. 16–22.

Demšar, J., 2006. Statistical comparison of classifiers over multiple data sets. Journal of Machine Learning Research 7, 1–30.

Faloutsos, C., Ranganathan, M., Manolopoulos, Y., 1994. Fast subsequence matching in time-series databases, in: Proc. of the ACM SIGMOD Int. Conf. on Management of Data, pp. 419–429.

Fu, T.C., 2011. A review on time series data mining. Engineering Applications of Artificial Intelligence 24, 164–181.

Geurts, P., 2002. Contributions to decision tree induction: bias/variance tradeoff and time series classification. Ph.D. thesis. University of Liège, Liège, Belgium.

Gusfield, D., 1997. Algorithms on strings, trees, and sequences: computer science and computational biology. Cambridge University Press, Cambridge, UK.

Han, J., Kamber, M., 2005. Data mining: concepts and techniques. Morgan Kaufmann, Waltham, USA.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning. 2nd ed., Springer, Berlin, Germany.

Hollander, M., Wolfe, D.A., 1999. Nonparametric statistical methods. 2nd ed., Wiley, New York, USA.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics 6, 65–70.

Kantz, H., Schreiber, T., 2004. Nonlinear time series analysis. Cambridge University Press, Cambridge, UK.

Keogh, E.J., 2011. Machine learning in time series databases (and everything is a time series!). Tutorial at the AAAI Int. Conf. on Artificial Intelligence.

Keogh, E.J., Chakrabarti, K., Pazzani, M., Mehrotra, S., 2001. Dimensionality reduction for fast similarity search in large time series databases. Knowledge and Information Systems 3, 263–286.

Keogh, E.J., Kasetty, S., 2003. On the need for time series data mining benchmarks: a survey and empirical demonstration. Data Mining and Knowledge Discovery 7, 349–371.

Keogh, E.J., Ratanamahatana, C.A., 2005. Exact indexing of dynamic time warping. Knowledge and Information Systems 7, 358–386.

Keogh, E.J., Xi, X., Wei, L., Ratanamahatana, C.A., 2009. Supporting exact indexing of arbitrarily rotated shapes and periodic time series under Euclidean and warping distance measures. VLDB Journal 11, 611–630.

Keogh, E.J., Zhu, Q., Hu, B., Hao, Y., Xi, X., Wei, L., Ratanamahatana, C.A., 2011. The UCR time series classification/clustering homepage. URL: `http://www.cs.ucr.edu/\%7eeamonn/time\_series\_data`.

Kholmatov, A., Yanikoglu, B., 2005. Identity authentication using improved online signature verification method. Pattern Recognition Letters 26, 2400–2408.

Lemire, D., 2009. Faster retrieval with a two-pass dynamic time warping lower bound. Pattern Recognition 42, 2169–2180.

Levenshtein, V.I., 1966. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10, 707–710.

Liao, T.W., 2005. Clustering of time series data: a survey. Pattern Recognition 38, 1857–1874.

Maharaj, E.A., 2000. Clusters of time series. Journal of Classification 17, 297–314.

Marple, S.L., 1987. Digital spectral estimation. Prentice-Hall, Englewood Cliffs, USA.

Marteau, R.F., 2009. Time warp edit distance with stiffness adjustment for time series matching. IEEE Trans. on Pattern Analysis and Machine Intelligence 31, 306–318.

Mitchell, T.M., 1997. Machine Learning. McGraw-Hill, New York, USA.

Montani, S., Portinale, L., Leonardi, G., Bellazzi, R., 2006. Case-based retrieval to support the treatment of end stage renal failure patients. Artificial Intelligence in Medicine 37, 31–42.

Morse, M.D., Patel, J.M., 2007. An efficient and accurate method for evaluating time series similarity, in: Proc. of the ACM SIGMOD Int. Conf. on Management of Data, pp. 569–580.

Olsson, E., Funk, P., Xiong, N., 2004. Fault diagnosis in industry using sensor readings and case-based reasoning. Journal of Intelligent Fuzzy Systems 15, 41–46.

Oppenheim, A.V., Schafer, R.W., Buck, J.R., 1999. Discrete-time signal processing. 2nd ed., Prentice-Hall, Upper Saddle River, USA.

Piccolo, D., 1990. A distance measure for classifying ARMA models. Journal of Time Series Analysis 11, 153–163.

Povinelli, R.J., Johnson, M.T., Lindgren, A.C., Ye, J., 2004. Time series classification using Gaussian mixture models of reconstructed phase spaces. IEEE Trans. on Knowledge Discovery and Data Engineering 16, 779–783.

Rabiner, L.R., Juang, B., 1993. Fundamentals of speech recognition. Prentice-Hall, Upper Saddle River, USA.

Radovanovič, M., Nanopoulos, A., Ivanovic, M., 2010. Time-series classification in many intrinsic dimensions, in: Proc. of the SIAM Int. Conf. on Data Mining, pp. 677–688.

Ramoni, M., Sebastiani, P., Cohen, P., 2002. Bayesian clustering by dynamics. Machine Learning 47, 91–121.

Rodríguez, J.J., Alonso, C., 2004. Interval and dynamic time warping-based decision trees, in: Proc. of the ACM Symp. on Applied Computing (SAC), pp. 548–552.

Sakoe, H., Chiba, S., 1978. Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. on Acoustics, Speech, and Language Processing 26, 43–50.

Salvador, S., Chan, P., 2007. Toward accurate dynamic time warping in linear time and space. Intelligent Data Analysis 11, 561–580.

Salzberg, S.L., 1997. On comparing classifiers: pitfalls to avoid and a recommended approach. Data Mining and Knowledge Discovery 1, 317–328.

Serrà, J., Arcos, J.L., 2012. A competitive measure to assess the similarity between two time series, in: Proc. of the Int. Conf. on Case-Based Reasoning (ICCBR), pp. 414–427.

Serrà, J., Kantz, H., Serra, X., Andrzejak, R.G., 2012a. Predictability of music descriptor time series and its application to cover song detection. IEEE Trans. on Audio, Speech and Language Processing 20, 514–525.

Serrà, J., Zanin, M., Herrera, P., Serra, X., 2012b. Characterization and exploitation of community structure in cover song networks. Pattern Recognition Letters 33, 1032–1041.

Tormene, P., Giorgino, T., Quaglini, S., Stefanelli, M., 2009. Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation. Artificial Intelligence in Medicine 45, 11–34.

Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., Keogh, E.J., 2012. Experimental comparison of representation methods and distance measures for time series data. Data Mining and Knowledge Discovery In press. URL: http://dx.doi.org/10.1007/s10618-012-0250-5.

Wu, D., Agrawal, D., El Abbadi, A., Singh, A., Smith, T.R., 1996. Efficient retrieval for browsing large image databases, in: Proc. of the Int. Conf. on Knowledge Information, pp. 11–18.

Wu, Y.L., Agrawal, D., El Abbadi, A., 2000. A comparison of DFT and DWT based similarity search in time-series databases, in: Proc. of the Int. Conf. on Information and Knowledge Management (CIKM), pp. 488–495.

Xi, X., Keogh, E.J., Shelton, C.R., Wei, L., Ratanamahatana, C.A., 2006. Fast time series classification using numerosity reduction, in: Proc. of the Int. Conf. on Machine Learning, pp. 1033–1040.

Xing, Z., Pei, J., Yu, P.S., 2011. Early classification on time series. Knowledge and Information Systems 31, 105–127.