A Comprehensive Comparison of Distance Measures for Time Series Classification

Tomasz Górecki and Paweł Piasecki

Abstract In the past two decades, interest in the area of time series has soared and many distance measures for time series have been proposed. The problem of pairwise similarity of time series is based on the underlying distance measure (which is not necessarily metric or even dissimilarity measure) and is common in many time series areas. To the best of our knowledge, there are over 40 distance measures already proposed in the literature. Thus, there is a need to decide which measure will be the most appropriate for our specific problem. The aim of our study is to give a comprehensive comparison of distance measures for time series classification enriched with extensive statistical analysis. We will follow a methodology that assumes evaluating the efficacy of distance measures by the prism of accuracy of 1NN classifier. Experimental results carried out on benchmark datasets originated from UCR Time Series Classification Archive are provided. We show that none of the distance measures is the best for all datasets, however there is a group performing statistically significantly better than the others.

Keywords Time series analysis · Classification · Distance measures · UCR archive

1 Introduction

Year after year, we see a remarkable increase of the interests in both collecting and mining of data. Typically, we differentiate time series problems from other data analysis tasks, because the attributes are ordered and we may look for a discriminatory feature that depends on the ordering [4]. In the past 20 years, interest in the area of time series has soared and many tasks have been deeply investigated, such as

T. Górecki · P. Piasecki (✉)
Adam Mickiewicz University, Poznań, Poland
e-mail: pawel.piasecki@amu.edu.pl

T. Górecki
e-mail: tomasz.gorecki@amu.edu.pl

A. Steland et al. (eds.), *Stochastic Models, Statistics and Their Applications*,
Springer Proceedings in Mathematics & Statistics 294,
https://doi.org/10.1007/978-3-030-28665-1_31

classification [4], clustering [29], indexing [26], prediction [50], anomaly detection [51], motif discovery [34] and more. In our opinion, there is a problem that appears throughout almost all of these topics: how to compare given two time series in the most appropriate way?

The problem of pairwise similarity of time series is based on the underlying distance measure (which are not necessarily metrics or even dissimilarity measures). To the best of our knowledge, there are about 40 distance measures proposed already in the literature. Some of them are based on certain feature of data, while the others use predictions, underlying models or some transformations. Such a variety may be confusing and makes it hard to find the most appropriate measure, especially for application-oriented scientists. Available research includes only 2 papers providing a partially comparison of selected distance measures.

Wang et al. [48] provide an extensive comparison of 9 different similarity measures and their 4 variants, which was carried on 38 time series datasets from UCR archive [13]. Authors of the paper conclude, that they did not find any measure, that is "universally better" at all datasets—some of them are better than the rest, while being worse on other datasets. However, dynamic time warping (DTW; [7])—slightly before some edit based measures: LCSS, EDR and ERP—seems to be superior to others. And it is in line with the widespread opinion that DTW is not always the best but in general hard to beat [45, 52]. From the other hand, the study points out that Euclidean distance remains a quick and efficient way of measuring distances between time series. Especially, when the training set increases, the accuracy of elastic measures converges to that of Euclidean distance.

Serrà et al. [44] compare 7 similarity measures on 45 datasets originated from UCR archive. Authors of the paper suggest that, in the set of investigated distances, there is a group of measures with no statistically significant differences: DTW, EDR and MJC. Another finding is that the TWED measure seems to consistently outperform all the considered distances. Euclidean distance is said to perform statistically worse than TWED, DTW, EDR, and MJC, and even its performance on large datasets was "not impressive". What is more, an interesting remark is made about various post-processing steps that may increase classification accuracy: the complexity-invariant correction [5], the hubness correction for time series classification [42], unsupervised clustering algorithms to prune nearest neighbor candidates [44]. For details see Serrà et al. [44].

Despite giving interesting results, both studies take into account only some distance measures, while nowadays, due to the very dynamic increase of interest in the time series area, there are about 40 measures available. As it is computationally expensive, in this paper we compare 30 of them, but we plan to develop our experiment in the nearest future. Our contribution is to give an extensive comparison, supported by deep statistical analysis. We would like to create a benchmark study, that could be used not only by researchers from different application fields, but as well by authors of new distance measures, to assess their effectiveness. We are going to give only basic descriptions of used similarity measures, provided along with some reference, as our intention is to not to develop distance measures itself, but rather to compare their efficacy.

2 Distances' Classification and Description

According to our best knowledge there exist about 40 distance measures, thus there is a strong need to classify them. Montero and Vilar [38] proposed to group measures into four categories: model-free measures, model-based measures, complexity-based measures, and prediction-based measures. Wang et al. [48] in their research named four groups of distance measures: lock-step measures, elastic measures, threshold-based measures, and pattern-based measures. In our opinion the most universal and covering almost all distances is categorization proposed by Esling and Agon [18]: shape-based measures, edit-based measures, feature-based measures and structure-based measures. We are going to follow the last classification. In this section, we list all 30 distance measures compared in this paper. We provide most important formulas, assuming we are given two time series: $\mathbf{X}_T = (x_1, x_2, \ldots, x_T)$, $\mathbf{Y}_T = (y_1, y_2, \ldots, y_T)$.

2.1 Shape-Based Distance Measures

This group of distance measures compares the overall shape of series looking mostly on the raw values.

The basic measures there are derived directly from L_p norms and we call them L_p distances: Manhattan distance, Minkowski distance, Euclidean distance, and Infinite norm distance. They are relatively simple in understanding and computation, but compare only time series of equal length and sometimes they perform poor and are highly influenced by outliers, noise, scaling or warping. For more information, we refer to Yi and Faloutsos [53], Antunes and Oliveira [2]. The basic formulas are given in Table 1.

Berndt and Clifford [7] proposed Dynamic Time Warping (DTW) distance, which not only solve most problems know from L_p distances, but due to its ability to deal with warping of the time axis became one of the most popular measure for time series. In practice, we compute DTW using dynamic programming with the following recurrence:

$$\Gamma(i, j) = D(i, j) + min\{\Gamma(i-1, j-1), \Gamma(i-1, j), \Gamma(i, j-1)\}$$

Table 1 L_p distances, $1 < p < \infty$

Distance	Formula
Manhattan	$\sum_{i=1}^{T} \lvert x_i - y_i \rvert$
Minkowski	$\sqrt[p]{\sum_{i=1}^{T} (x_i - y_i)^{\frac{1}{p}}}$
Euclidean	$\sqrt{\sum_{i=1}^{T} (x_i - y_i)^2}$
Infinite norm	$\max_{i=1,\ldots,T} \lvert x_i - y_i \rvert$

with initial conditions:

$$\Gamma(0,0) = 0, \quad \Gamma(0,i) = \infty, \quad \Gamma(i,0) = \infty \quad (i = 1, 2, \ldots, n),$$

where Γ is a cumulative distance matrix, $D(i,j) = d(x_i, y_j), d(x_i, y_j) = (x_i - y_j)^2$. The value of DTW at position (n,n) of the matrix Γ is calculated then as $DTW(x,y) = \sqrt{\Gamma(n,n)}$.

Because of the long computation time of basic DTW distance, several lower bounding and temporal constraints techniques have been proposed. In Sect. 4 we denote DTW with Sakoe–Chiba Band as "DTWc" and we use the window size as in Dau et al. [13]. For more details about DTW we refer to Bagnall et al. [4], Keogh and Ratanamahatana [30], Mori et al. [39]. We will also examine two distance measures expanding DTW with derivatives. Keogh and Pazzani [32] defined Derivative Dynamic Time Warping (DDTW) which is a DTW distance between the data transformed by the first (discrete) derivative. Górecki and Łuczak [22] proposed Parametric Derivative Dynamic Time Warping (DD_{DTW}) as a convex combination of the distances DTW and DDTW, what brought further performance improvements.

For irregularly spaced series, [37] proposed the Short Time Series (STS) distance given by

$$d_{STS}(\mathbf{X}_T, \mathbf{Y}_T) = \sqrt{\sum_{i=1}^{T-1} \left(\frac{y_{i+1} - y_i}{t_{i+1} - t_i} - \frac{x_{i+1} - x_i}{t'_{i+1} - t'_i} \right)},$$

where t and t' are the temporal indexes of series $\mathbf{X}_T$ and $\mathbf{Y}_T$ respectively. It is able to measure similarity of shapes formed by both the relative change of amplitude and the corresponding temporal information.

Another very important aspect of similarity measures is a tendency to put time series with high complexity level further apart than simple ones [5]. In order to fix this distortion, a Complexity-Invariant dissimilarity measure (CID) has been proposed by the authors. The general formula is as follows

$$d_{CID}(\mathbf{X}_T, \mathbf{Y}_T) = CF(\mathbf{X}_T, \mathbf{Y}_T) \cdot d(\mathbf{X}_T, \mathbf{Y}_T),$$

where $d(\mathbf{X}_T, \mathbf{Y}_T)$ is a distance which will be adjusted, $\mathrm{CF}(\mathbf{X}_T, \mathbf{Y}_T)$ is a complexity correction factor defined as

$$\mathrm{CF}(\mathbf{X}_T, \mathbf{Y}_T) = \frac{\max\{\mathrm{CE}(\mathbf{X}_T), \mathrm{CE}(\mathbf{Y}_T)\}}{\min\{\mathrm{CE}(\mathbf{X}_T), \mathrm{CE}(\mathbf{Y}_T)\}},$$

where $\mathrm{CE}(\mathbf{X}_T)$ stands for a complexity estimator of $\mathbf{X}_T$. Now, we can observe, that when the complexity of both time series is equal, we get $d_{CID}(\mathbf{X}_T, \mathbf{Y}_T) = d(\mathbf{X}_T, \mathbf{Y}_T)$ and, that increase of complexity difference results in increase of distance between time series. As a complexity estimator [5] proposed $\mathrm{CE}(\mathbf{X}_T) = \sqrt{\sum_{t=1}^{T-1} (X_t - X_{t+1})^2}$.

2.2 Edit-Based Distance Measures

Edit-based distances use the minimal number of operation necessary to transform one series into another. They were initially proposed to measure the similarity between two sequences of strings and use the minimal number of edit operations (delete, insert, replace) necessary to transform one series into another.

As edit-based distances may be computed for time series of different length, in this section we will assume we are given two time series: $\boldsymbol{X_N} = (x_1, x_2, \ldots, x_N)$ and $\boldsymbol{Y_M} = (y_1, y_2, \ldots, y_M)$. For clarification and simplicity, in all other sections the notation is as mentioned in the introduction to Sect. 2.

The LCSS distance was proposed by Vlachos et al. [47] and measures the similarity between time series in terms of the longest common subsequence, with the addition that gaps and unmatched regions are permitted. LCSS is robust to noise and we expect that it should be more accurate than DTW in the presence of outliers and noise. The measure has two constant parameters. The first one, δ controls the size of the window for matching given point from one series to a point in the other series. The second one, ϵ, is the matching threshold: two points are considered to match if their distance is less than ϵ. Given

$$L(i, j) = \begin{cases} 0 & \text{for } i = 0 \\ 0 & \text{for } j = 0 \\ 1 + L[i-1, j-1] & \text{for } |x_i - y_j| < \epsilon \\ & \text{and } |i - j| \leq \delta \\ \max(L[i-1, j], L[i, j-1]) & \text{in other cases,} \end{cases}$$

we can compute [43]

$$LCSS(\boldsymbol{X_N}, \boldsymbol{Y_M}) = \frac{N + M - 2L(N, M)}{N + M}.$$

The Edit Distance on Real Sequence (EDR) is an adaptation of the edit distance that finds the minimal number of edit operations to convert one series to another [11]. Similarly to LCSS, EDR permits gaps and unmatched regions, but penalizes such occurrences with a value equal to their length. Computation of the EDR measure can be converted into an iteration using dynamic programming as follows

$$\mathrm{EDR}(\mathrm{X_N}, \mathrm{Y_M}) = \begin{cases} N & \text{for } i = 0 \\ M & \text{for } j = 0 \\ \min\{(\mathrm{EDR}(\mathrm{Rest}(\mathrm{X_N}), \mathrm{Rest}(\mathrm{Y_M})) + \mathrm{d_{edr}}(\mathrm{x_1}, \mathrm{y_1}), & \\ \mathrm{EDR}(\mathrm{Rest}(\mathrm{X_N}), \mathrm{Y_M}) + 1, \mathrm{EDR}(\mathrm{X_N}, \mathrm{Rest}(\mathrm{Y_M})) + 1\} & \text{otherwise,} \end{cases}$$

where $\mathrm{Rest}((\boldsymbol{X_N})) = (x_2, x_3, \ldots, x_N)$ and d_{edr} stands for the distance between two points in the series computed along to the rule: if x_i and y_i are closer to each other in the absolute sense than ϵ, it is equal to 1. Otherwise, it is equal to 0.

The third variation of edit distance is the Edit Distance with a Real Penalty (ERP) [10] that may be considered as a combination of DTW and EDR. It uses the L_1

distance between elements of time series as the penalty for local shifting of time series. Penalization is carried out by setting a constant g and adding the euclidean distance of the unmatched points to g. The ERP measure is given by

$$\mathrm{ERP}(\mathrm{X_N}, \mathrm{Y_M}) = \begin{cases} \sum_{i=1}^{N} |y_1 - g| & \text{if } M - 1 = 0 \\ \sum_{i=1}^{M} |x_1 - g| & \text{if } N - 1 = 0 \\ \min\{(\mathrm{EDR}(\mathrm{Rest}(\mathrm{X_N}), \mathrm{Rest}(\mathrm{Y_M})) + \mathrm{d_{edr}}(\mathrm{x_1}, \mathrm{y_1}), \\ \mathrm{EDR}(\mathrm{Rest}(\mathrm{X_N}), \mathrm{Y_M}) + 1, \mathrm{EDR}(\mathrm{X_N}, \mathrm{Rest}(\mathrm{Y_M})) + 1\} & \text{otherwise.} \end{cases}$$

2.3 Feature-Based Distances

These distances look at some aspect of the time series by extracting certain feature. Then, based on it, a similarity measure is calculated.

Taking into account correlation in time series, we may define at least several measures. Golay et al. [21] defined distance based on Pearson's correlation coefficient as follows:

$$d_{\mathrm{PC}}(\mathbf{X}_T, \mathbf{Y}_T) = 2(1 - PC),$$

where PC denotes Pearson's correlation coefficient.

Warren Liao [49] proposed to use cross-correlation between two series and based on it, formulated

$$d_{\mathrm{CC}}(\mathbf{X}_T, \mathbf{Y}_T) = \sqrt{\frac{(1 - CC_0(X, Y))}{\sum_{k=1}^{\max} CC_k(X, Y)}},$$

where $\mathrm{CC}_k(X, Y)$ is the cross-correlation between two series at lag k.

Let $\hat{\boldsymbol{\rho}}_{X_T} = (\hat{\rho}_{1,X_T}, \ldots, \hat{\rho}_{L,X_T})^T$, $\hat{\boldsymbol{\rho}}_{Y_T} = (\hat{\rho}_{1,Y_T}, \ldots, \hat{\rho}_{L,Y_T})^T$ be the estimated autocorrelation vectors of $\mathbf{X}_T, \mathbf{Y}_T$ (respectively), for some L such that $\hat{\rho}_{i,X_T}, \hat{\rho}_{i,Y_T} \approx 0$ for $i > L$. Peña and Galeano [40] proposed the following distance:

$$d_{\mathrm{ACF}}(\mathbf{X}_T, \mathbf{Y}_T) = \sqrt{(\hat{\boldsymbol{\rho}}_{X_T} - \hat{\rho}_{Y_T})^T \boldsymbol{\Omega} (\hat{\boldsymbol{\rho}}_{X_T} - \hat{\rho}_{Y_T})},$$

where $\boldsymbol{\Omega}$ is a matrix of weights, which define the importance of correlation at different lags. Obviously, to emphasize slightly different aspect of the data, it is possible to replace autocorrelations by partial autocorrelations and obtain d_{PC}.

The first-order temporal correlation coefficient is defined by

$$\mathrm{CORT}(\mathbf{X}_T, \mathbf{Y}_T) = \frac{\sum_{i=1}^{T-1} (X_{t+1} - X_t)(Y_{t+1} - Y_t)}{\sqrt{\sum_{i=1}^{T-1} (X_{t+1} - X_t)^2} \sqrt{\sum_{i=1}^{T-1} (Y_{t+1} - Y_t)^2}}.$$

The CORT coefficient reflect the dynamic behaviors of the series [38]. The related dissimilarity measure was proposed by Chouakria and Nagabhushan [12] and it is defined as

$$d_{\text{CORT}}(\mathbf{X}_T, \mathbf{Y}_T) = \phi_k[\text{CORT}(\mathbf{X}_T, \mathbf{Y}_T)] \cdot d(\mathbf{X}_T, \mathbf{Y}_T),$$

where $\phi_k(\cdot)$ is an adaptive tuning function to automatically modulate a conventional data distance according to the temporal correlation. Chouakria and Nagabhushan proposed $\phi_k(u) = \frac{2}{1+exp(ku)}$, $k \geq 0$.

Another aspect of time series may be revealed by the Discrete Fourier Transform. Based on that we may compute Euclidean distance d_{FC} between the first n coefficients [1]:

$$\text{FC}(\mathbf{X}_T, \mathbf{Y}_T) = \sqrt{\sum_{i=0}^{n}((a_i - a_i^{'})^2 + (b_i - b_i^{'})^2)}.$$

There at least several distances based on the frequency domain of the time series. Caiado et al. [9] proposed the Euclidean distance d_{P} between the periodogram coordinates as follows:

$$d_{\text{P}}(\mathbf{X}_T, \mathbf{Y}_T) = \frac{1}{n}\sqrt{\sum_{k=1}^{n}(I_{X_T}(\lambda_k) - I_{Y_T}(\lambda_k))^2},$$

where $I_{X_T}(\lambda_k)$ and $I_{Y_T}(\lambda_k)$ for $k = 1, \ldots, n$ are periodograms of $\mathbf{X}_T$ and $\mathbf{Y}_T$ (respectively).

Alternatively, de Lucas [14] introduced distance measure based on integrated periodogram, arguing that—due to some properties of integrated periodogram—it presents several advantages over the previous one. The distance is defined as

$$d_{\text{IP}}(\mathbf{X}_T, \mathbf{Y}_T) = \int_{-\pi}^{\pi} |F_{\mathbf{X}_T}(\lambda) - F_{\mathbf{Y}_T}(\lambda)| d\lambda,$$

where $F_{\mathbf{X}_T}(\lambda_j) = C_{\mathbf{X}_T}^{-1} \sum_{i=1}^{j} I_{\mathbf{X}_T}(\lambda_i)$ and $F_{\mathbf{Y}_T}(\lambda_j) = C_{\mathbf{Y}_T}^{-1} \sum_{i=1}^{j} I_{\mathbf{Y}_T}(\lambda_i)$, with $C_{\mathbf{X}_T} = \sum_i I_{\mathbf{X}_T}(\lambda_i)$, $C_{\mathbf{Y}_T} = \sum_i I_{\mathbf{Y}_T}(\lambda_i)$.

Kakizawa et al. [24] proposed a general spectral disparity measure between two time series as

$$d_{\text{LLR}}(\mathbf{X}_T, \mathbf{Y}_T) = \int_{-\pi}^{\pi} \tilde{W}\left(\frac{f_{X_T}(\lambda)}{f_{Y_T}(\lambda)}\right) d\lambda,$$

where f_{X_T} and f_{Y_T} are spectral densities of $\mathbf{X}_T$ and $\mathbf{Y}_T$. $\tilde{W} = W(x) + W(x^{-1})$, $W(x) = \log(\alpha x + (1-\alpha)) - \alpha \log x$, with $0 < \alpha < 1$. $W(\cdot)$ is a divergence function satisfying regular quasi-distance conditions for d_{LLR}.

Alternatively, Díaz and Vilar [16] described the two following distances. The first one is defined as

$$d_{\mathrm{GLK}}(\mathbf{X}_T, \mathbf{Y}_T) = \sum_{k=1}^{n} \left[Z_k - \hat{\mu}(\lambda_k) - 2\log(1 + e^{Z_k - \hat{\mu}(\lambda_k)}) \right] - \\ - \sum_{k=1}^{n} \left[Z_k - 2\log(1 + e^{Z_k}) \right],$$

where $Z_k = \log(I_{X_T}(\lambda_k)) - \log(I_{Y_T}(\lambda_k))$ and $\hat{\mu}(\lambda_k)$ is the local maximum log-likelihood estimator of $\mu(\lambda_k) = \log(f_{X_T}(\lambda_k) - \log(f_{Y_T}(\lambda_k)$ computed by local linear fitting.

The second distance is given by

$$d_{\mathrm{ISD}}(\mathbf{X}_T, \mathbf{Y}_T) = \int_{-\pi}^{\pi} (\hat{m}_{X_T}(\lambda) - \hat{m}_{Y_T}(\lambda))^2 d\lambda,$$

where $\hat{m}_{X_T}(\lambda)$ and $\hat{m}_{Y_T}(\lambda)$ are local linear smoothers of the log-periodograms obtained with the maximum local likelihood criterion.

Moving on to an another characteristic, Aßfalg et al. [3] proposed a distance measure d_{TQ} based on Threshold Queries, using given τ parameter as a threshold in order to transform a time series into a sequence of time stamps, when the threshold is crossed. Let us denote the time stamps for a certain threshold τ as a sequence $(t_1, t_2, \ldots, t_n)$. For a time series $\boldsymbol{X_T}$ and a threshold τ we define the interval set $S(\boldsymbol{X_T}, \tau) = \{(t_1, t_2), (t_3, t_4), \ldots, (t_{n-1}, t_n)\}$. The distance between time series $\boldsymbol{X_T}$ and $\boldsymbol{Y_T}$, represented by the interval sets $S(\boldsymbol{X_T}, \tau)$ and $S(\boldsymbol{Y_T}, \tau)$ is given by

$$\mathrm{TQuest}(\boldsymbol{X_T}, \boldsymbol{Y_T}) = \frac{1}{|S(\boldsymbol{X_T}, \tau)|} \sum_{s \in S(\boldsymbol{X_T}, \tau)} \min_{s' \in S(\boldsymbol{X_T}, \tau)} d(s, s') + \\ + \frac{1}{|S(\boldsymbol{Y_T}, \tau)|} \sum_{s' \in S(\boldsymbol{Y_T}, \tau)} \min_{s \in S(\boldsymbol{Y_T}, \tau)} d(s', s),$$

where the distance between two intervals $s = (s_l, s_u)$ and $s' = (s'_l, s'_u)$ is computed as

$$d(s, s') = \sqrt{(s_l - s'_l)^2 + (s_u - s'_u)^2}.$$

The TQuest measure is based on an interesting feature extraction idea, but—in our opinion—it is highly dependent on user's specialist knowledge, as the τ parameter must be set.

The symbolic approximation representation (SAX) has been introduced by Lin et al. [33] and became one of the best symbolic representation for most time series problems [27]. The original data are first transformed into the piecewise aggregate approximation (PAA) representation [53] and then into a discrete string. For the full outline of MINDIST dissimilarity measure based on SAX representation see Lin et al. [35].

2.4 Structure-Based Distances

The last group of distance measures try to find some higher level structures and then compare time series on these basis. This category can be subdivided into two another groups: model-based—aiming to fit a model and then to compare coefficients thorough certain distance function and compression-based, which work by compression ratios.

The first category is represented by the distance described by Piccolo [41] as the Euclidean distance between coefficients derived from AR representation of processes:

$$d_{\mathrm{PIC}}(\mathbf{X}_T, \mathbf{Y}_T) = \sqrt{\sum_{j=1}^{k} (\hat{\pi}'_{j,X_T} - \hat{\pi}'_{j,Y_T})^2},$$

where the vectors of $AR(k_1)$ and $AR(k_2)$ for $\mathbf{X}_T$ and $\mathbf{Y}_T$ are denoted respectively by $\hat{\mathbf{\Pi}}_{X_T} = (\hat{\pi}_{1,X_T}, \ldots, \hat{\pi}_{k_1,X_T})$ and $\hat{\mathbf{\Pi}}_{Y_T} = (\hat{\pi}_{1,Y_T}, \ldots, \hat{\pi}_{k_2,Y_T})$, $k = \max(k_1, k_2)$, $\hat{\pi}'_{j,X_T} = \hat{\pi}_{j,X_T}$ if $j \leq k_1$ and $\hat{\pi}'_{j,X_T} = 0$ otherwise and analogously $\hat{\pi}'_{j,Y_T} = \hat{\pi}_{j,Y_T}$ if $j \leq k_2$ and $\hat{\pi}'_{j,Y_T} = 0$ otherwise. In case of nonstationary series, a differencing is carried out. To fit truncated AR(∞) model, a criterion such as BIC or AIC is used. There are at least two another distances (proposed by [25, 36]) based on the idea of fitting an ARIMA model to each series and then measure the dissimilarity between the models, but we will not use them due to implementation problems.

The distances from the second group compare levels of complexity of time series. Alternative to previous sections and paragraphs, complexity-based approaches do not rely on specific feature or knowledge of underlying models, but on evaluating the level of shared information by both series [38]. Keogh et al. [31] proposed compression-based dissimilarity measure defined as

$$d_{\mathrm{CDM}}(\mathbf{X}_T, \mathbf{Y}_T) = \frac{\mathrm{C}(\mathbf{X}_T, \mathbf{Y}_T)}{\mathrm{C}(\mathbf{X}_T)\mathrm{C}(\mathbf{Y}_T)}.$$

The CDM distance is descended from normalized compression distance (NCD) proposed by Lin et al. [34], using the compressed size of $\mathbf{X}_T$—$\mathrm{C}(\mathbf{X}_T)$—as an approximation of Kolmogorov complexity.

Dissimilarity measure based on permutation distribution clustering (PDC) uses permutation $\Pi(\mathbf{X}'_T)$ of m-dimensional embedding of $\mathbf{X}_T$. Dissimilarity between two time series $\mathbf{X}_T$ and $\mathbf{Y}_T$ is expressed in terms of divergence between distribution of these permutations, denoted by $P(\mathbf{X}_T)$, $P(\mathbf{Y}_T)$. Specifically, Brandmaier [8] proposed the α-divergence between $P(\mathbf{X}_T)$ and $P(\mathbf{Y}_T)$ as a dissimilarity between time series $\mathbf{X}_T$ and $\mathbf{Y}_T$.

3 Experimental Design

We performed experiments on 47 real time series that come from the UCR time series repository [13]. Each dataset from the database is split into training and testing subsets. Within the data, the number of classes ranges from 2 to 50, the number of time series per dataset go from 56 to 9236, and time series lengths ranges from 60 to 1882 samples. All time series instances are z-normalized.

In our paper, we will follow the methodology proposed by Keogh and Kasetty [28], which assumes evaluating the efficacy of distance measure by the prism of accuracy of 1NN classifier. While one should be aware that the proposed approach can not deliver us the overall evaluation of a distance measure, there seems to be more pros than cons of the chosen method. For example, Wang et al. [48] pointed out three aspects: the simplicity of implementation, performance directly dependent on distance choice and relatively (to other, often more complex classifiers) good performance. For more information we refer to Batista et al. [5], Ding et al. [17], Tan et al. [46], Xi et al. [52].

Specifically, for each dataset, we computed the classification error rate on a test subset. When a parameter to train the 1NN classifier was needed, we tried to use values proposed already in the literature (referred in the Sect. 2).

4 Results

The results are presented in Tables 2 and 3. We computed there the absolute error rates on the test subset with the 1NN classifier for each of 30 distance measures. In Fig. 1 we presented ranks for all considered distances.

If we look at the overall result, we can observe that none of the compared distances achieves the best performance for all, or even the most of datasets. In fact, the lowest error rates are computed for DD_{DTW} (15 wins), DDTW (9 wins), DTW (8 wins) ahead of ERP (6 wins), EDR (5 wins), LCSS (5 wins) and DTWc (5 wins). There are also CORT and ISD measures with 4 wins both, but the others are significantly worse. It may be the evidence for superiority of elastic measures and those connected with DTW distance over the rest. From the other hand, looking at average ranks, one may be surprised by the good performance of L_p norms: MAN, ED, and MIN. It is also worth mentioning about CID distance. It achieved better average rank than DTW, while in fact it only improves Euclidean distance by simple complexity correction factor.

Looking at certain datasets, we see, that some of them are almost perfectly classified (e.g., *Coffee, DiatSizeRed, GunPoint, Plane*), what could mean that their classes are relatively easy to recognize by the algorithm. Another interesting fact is, that there are datasets, which are better classified by some group of distances. For example, performance of L_p norms is relatively good for *MALLAT, SynthetCont*, while clearly worse for *CricketX, CricketY, Haptics*, what may indicate cases, where we

Table 2 Error rates (in%) of all considered distance measures on 1NN classifier. Best classifier for each dataset was bold. In the last but one row we computed number of wins for each distance and in the last one—average ranks

Dataset	DD_{DTW}	ERP	LCSS	DTWc	DDTW	EDR	CID	DTW	MAN	FC	COR	CORT	ED	MIN	IP
50words	20.7	25.1	**19.1**	24.2	23.5	21.1	33.6	31.0	33.2	36.9	36.9	38.7	36.9	39.8	66.6
Adiac	**29.9**	37.9	85.7	39.1	32.7	85.7	37.3	39.6	40.2	38.9	38.9	39.6	38.9	38.1	44.0
ArrowHead	18.9	20.6	20.0	20.0	21.1	18.9	**17.1**	29.7	20.6	20.0	20.0	21.7	20.0	18.9	21.1
Beef	33.3	40.0	33.3	33.3	33.3	**30.0**	36.7	36.7	36.7	33.3	33.3	33.3	33.3	33.3	53.3
BeetleFly	15.0	35.0	30.0	30.0	15.0	30.0	30.0	30.0	35.0	25.0	25.0	35.0	25.0	25.0	**10.0**
BirdCh	30.0	25.0	20.0	30.0	20.0	10.0	35.0	25.0	40.0	45.0	45.0	45.0	45.0	45.0	15.0
Car	21.7	28.3	**16.7**	23.3	21.7	25.0	26.7	26.7	28.3	26.7	26.7	25.0	26.7	26.7	31.7
CBF	1.7	**0.2**	0.4	0.4	45.4	4.6	1.6	0.3	11.1	15.1	14.8	20.6	14.8	19.9	38.0
CinCECG	6.7	3.8	11.9	7.0	6.3	13.1	8.4	34.9	5.6	10.3	10.3	9.3	10.3	12.9	36.4
Coffee	**0.0**	3.6	**0.0**	**0.0**	7.1	**0.0**	**0.0**	**0.0**	3.6	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
Comp	**29.6**	41.6	40.4	38.0	35.2	40.8	40.4	30.0	41.2	42.4	42.4	46.4	42.4	46.0	43.2
CricketX	**21.5**	25.9	28.7	22.8	40.5	29.2	37.2	24.6	36.9	42.6	42.3	47.4	42.3	50.0	49.7
CricketY	**20.0**	24.6	25.4	24.1	46.7	27.4	42.1	25.6	36.7	43.3	43.3	41.3	43.3	46.9	53.8
CricketZ	25.9	27.4	27.9	25.4	44.4	28.2	40.5	**24.6**	34.1	41.3	41.3	44.4	41.3	45.4	51.8
DiatSizeRed	6.5	7.2	6.2	6.5	11.1	6.2	6.5	**3.3**	7.2	6.5	6.5	6.2	6.5	5.6	6.5
DisOutCorr	20.8	23.0	22.2	27.5	21.5	**20.7**	25.5	28.3	23.8	24.7	24.8	26.2	28.3	25.5	26.2
DisOutGr	23.0	23.2	24.5	37.4	22.0	**21.2**	22.2	23.0	22.0	21.8	21.8	21.8	37.4	22.0	23.2
DisTW	27.8	**26.2**	27.0	36.7	28.8	27.8	26.5	41.0	28.2	27.2	27.2	27.8	36.7	28.2	28.5
Earthq	25.5	29.2	23.9	27.3	23.6	20.2	28.6	28.1	28.6	32.6	32.6	32.9	28.8	28.6	33.2
ECG200	**10.0**	**10.0**	12.0	12.0	14.0	11.0	11.0	23.0	11.0	12.0	12.0	12.0	12.0	11.0	20.0
ECGFive	20.3	20.7	14.6	20.3	32.6	17.2	21.8	23.2	21.5	20.2	20.3	23.8	20.3	20.7	3.0
FaceFour	11.4	6.8	5.7	11.4	38.6	**3.4**	19.3	17.1	15.9	21.6	21.6	22.7	21.6	27.3	40.9
FacesUCR	8.1	**4.3**	7.7	8.8	13.3	4.9	23.5	9.5	19.9	23.1	23.1	25.8	23.1	26.8	41.5
FISH	6.9	12.6	15.4	15.4	**6.3**	16.0	21.7	17.7	20.6	21.7	21.7	20.6	21.7	20.0	28.6

(continued)

Table 2 (continued)

Dataset	DD_{DTW}	ERP	LCSS	DTWc	DDTW	EDR	CID	DTW	MAN	FC	COR	CORT	ED	MIN	IP
GunPoint	2.0	4.7	5.3	8.7	**0.7**	6.7	7.3	9.3	4.7	8.7	8.7	7.3	8.7	12.0	8.7
Haptics	**57.5**	61.4	57.8	58.8	68.5	57.8	58.4	62.3	64.0	63.0	63.0	58.4	63.0	61.0	61.7
InlineSk	44.0	56.7	60.5	61.3	48.9	60.0	62.9	61.6	64.7	65.8	65.8	64.7	65.8	67.5	66.2
Light2	**13.1**	**13.1**	16.4	**13.1**	22.9	22.9	24.6	**13.1**	18.0	24.6	24.6	29.5	24.6	34.4	24.6
Light7	**23.3**	28.8	27.4	28.8	34.2	35.6	39.7	27.4	28.8	42.5	42.5	46.6	42.5	53.4	67.1
MALLAT	8.6	7.5	11.9	8.6	8.4	14.6	7.5	**6.6**	7.5	8.6	8.6	8.6	8.6	9.1	15.2
Medicallm	25.4	28.4	32.8	**25.3**	34.0	33.5	30.9	26.3	29.3	31.6	31.6	31.4	31.6	32.5	43.9
MiddleTW	42.1	43.4	45.4	49.4	42.9	45.6	42.6	49.4	42.4	43.9	43.9	43.1	48.7	42.4	42.9
MoteStrain	26.3	12.1	**10.8**	13.4	28.4	16.0	21.2	16.5	13.4	12.1	12.1	17.3	12.1	18.4	45.3
OliveOil	**13.3**	16.7	83.3	**13.3**	**13.3**	83.3	13.3	16.7	16.7	13.3	13.3	13.3	**13.3**	13.3	16.7
OSULeaf	**12.0**	39.3	21.1	38.8	**12.0**	24.0	43.8	40.9	45.5	47.9	47.9	46.3	47.9	48.3	35.5
Plane	**0.0**	**0.0**	0.9	**0.0**	**0.0**	0.9	3.8	**0.0**	3.8	3.8	3.8	3.8	3.8	3.8	1.9
SmallKit	33.3	32.8	62.9	32.8	32.8	68.0	60.8	35.7	60.3	65.6	65.6	69.3	65.9	67.5	56.3
SonySurf	30.4	31.3	29.9	30.4	25.5	30.9	18.5	27.4	31.3	30.4	30.4	20.8	30.4	28.8	22.5
SonySurfII	**10.5**	13.2	15.5	14.1	**10.5**	17.5	12.3	16.9	13.2	14.1	14.1	14.4	14.1	16.1	12.7
Strawberry	5.4	6.0	24.6	5.4	**5.2**	24.6	6.0	5.9	5.7	6.2	6.2	6.0	5.4	6.4	9.1
SwedLeaf	**9.3**	12.0	11.0	15.4	10.9	9.8	12.3	20.8	21.1	21.1	21.1	19.4	21.1	22.2	11.7
SynthCont	1.7	2.7	9.3	1.7	49.7	12.7	5.0	**0.7**	12.0	11.3	12.0	25.7	12.0	12.0	51.0
ToeSegm1	18.4	18.9	21.5	25.0	18.4	18.4	30.7	22.8	28.1	32.0	32.0	35.5	32.0	34.2	**13.2**
ToeSegm2	20.0	6.9	**6.2**	9.2	20.0	10.0	18.5	16.1	14.6	19.2	19.2	16.9	19.2	21.5	12.3
Trace	1.0	19.0	3.0	1.0	**0.0**	18.0	14.0	**0.0**	24.0	24.0	24.0	28.0	24.0	23.0	**0.0**
TwoLeadECG	4.6	13.3	13.2	13.2	**0.7**	18.7	23.2	9.6	26.5	25.2	25.3	22.1	25.3	25.6	7.1
WordsSyn	**23.7**	30.7	25.1	26.2	27.0	27.3	35.7	35.1	36.5	38.2	38.2	41.2	38.2	41.2	62.9
# wins	15	6	5	5	9	5	2	8	0	1	1	1	2	1	4
avg. rank	4.8	7.9	8.2	8.3	8.7	9.1	9.4	10.2	10.8	12.2	12.3	13.2	13.4	13.9	14.3

Table 3 Error rates (in%) of all considered distance measures on 1NN classifier. Best classifier for each dataset was bold. In the last but one row we computed number of wins for each distance and in the last one—average ranks

Dataset	CCOR	P	INF	ACF	LLR	ISD	TQ	PACF	STS	GLK	SAX	PDC	PIC	NCD	CDM
50words	39.6	60.0	47.5	72.7	80.4	74.1	42.6	67.5	48.4	80.4	43.1	89.9	74.3	93.8	94.7
Adiac	38.1	50.9	38.4	44.0	62.7	76.2	77.0	45.3	41.9	80.6	97.7	89.3	83.6	88.2	88.2
ArrowHead	19.4	23.4	17.7	28.6	36.6	41.1	44.0	41.1	40.6	40.6	49.1	48.6	57.7	58.3	58.9
Beef	36.7	56.7	**30.0**	56.7	60.0	63.3	56.7	46.7	**30.0**	63.3	63.3	66.7	53.3	76.7	80.0
BeetleFly	35.0	25.0	35.0	40.0	50.0	50.0	15.0	35.0	35.0	55.0	20.0	65.0	50.0	30.0	30.0
BirdCh	40.0	25.0	35.0	15.0	15.0	**5.0**	45.0	45.0	45.0	**5.0**	35.0	25.0	40.0	45.0	45.0
Car	26.7	45.0	28.3	61.7	56.7	41.7	35.0	66.7	33.3	51.7	40.0	55.0	70.0	75.0	76.7
CBF	25.9	31.2	57.8	31.6	41.8	43.3	32.0	58.8	66.0	44.9	21.0	60.7	49.9	43.1	44.2
CinCECG	12.5	18.7	20.3	48.6	26.7	11.3	30.7	15.4	21.4	12.7	30.7	**1.4**	37.8	46.5	47.9
Coffee	**0.0**	7.1	**0.0**	7.1	14.3	17.9	39.3	**0.0**	7.1	21.4	42.9	50.0	14.3	46.4	46.4
Comp	48.0	42.8	46.0	41.2	44.0	41.6	50.0	45.2	48.4	40.8	51.6	33.6	49.6	42.8	45.2
CricketX	58.2	40.8	66.9	53.1	63.8	68.5	50.0	68.5	85.9	68.7	46.7	87.2	75.4	83.6	83.6
CricketY	55.9	46.9	61.5	60.8	71.0	74.6	53.3	76.7	83.8	77.9	53.6	86.7	80.3	84.6	83.3
CricketZ	57.9	38.7	65.1	50.0	65.4	65.1	48.2	68.2	79.5	70.8	46.4	87.2	78.7	83.3	84.4
DiatSizeRed	6.9	6.5	3.6	8.2	44.4	22.2	12.1	11.8	11.1	25.8	69.3	33.7	39.2	61.8	61.4
DisOutCorr	25.2	27.5	26.8	26.0	24.5	26.2	37.2	32.2	25.2	33.8	44.8	40.8	34.3	41.2	45.0
DisOutGr	22.8	24.8	25.0	22.5	28.0	27.8	30.2	29.5	24.8	44.8	59.0	28.5	24.2	37.2	66.5
DisTW	27.5	31.2	30.2	30.5	28.8	33.0	35.0	32.5	28.8	45.2	75.8	34.8	33.2	39.0	79.2
Earthq	23.0	33.5	38.2	40.4	38.8	38.5	23.0	38.8	28.6	34.2	28.6	28.0	35.4	25.8	**19.3**
ECG200	11.0	22.0	13.0	26.0	16.0	15.0	22.0	20.0	15.0	20.0	16.0	38.0	24.0	32.0	33.0
ECGFive	17.2	**0.1**	28.9	0.6	17.4	25.6	22.6	22.9	38.3	27.6	46.0	38.3	31.6	38.8	39.6
FaceFour	30.7	34.1	45.5	46.6	52.3	51.1	12.5	55.7	47.7	58.0	21.6	55.7	69.3	56.8	69.3
FacesUCR	26.9	22.5	40.9	25.7	50.9	51.5	25.5	40.5	35.5	54.6	38.4	84.6	64.9	87.5	87.9
FISH	21.1	36.6	20.6	60.0	70.3	65.1	60.0	45.1	18.9	70.3	80.0	72.0	76.0	82.3	81.1

(continued)

Table 3 (continued)

Dataset	CCOR	P	INF	ACF	LLR	ISD	TQ	PACF	STS	GLK	SAX	PDC	PIC	NCD	CDM
GunPoint	8.7	10.7	14.7	8.0	10.0	22.7	19.3	4.7	8.7	22.0	28.7	34.0	32.7	30.7	33.3
Haptics	61.7	65.3	61.0	75.0	65.9	75.3	62.3	68.8	68.5	74.0	68.8	72.1	73.4	82.8	85.4
InlineSk	66.9	69.8	71.3	79.8	64.7	**17.8**	75.3	52.7	76.5	25.5	80.0	53.5	66.0	82.7	82.7
Light2	39.3	27.9	31.1	23.0	37.7	31.1	39.3	36.1	49.2	36.1	29.5	49.2	47.5	37.7	39.3
Light7	49.3	41.1	54.8	47.9	53.4	61.6	47.9	63.0	68.5	68.5	43.8	79.5	78.1	63.0	65.8
MALLAT	8.4	13.3	12.1	53.3	15.9	31.5	30.7	28.7	14.8	34.8	54.8	41.1	38.7	78.4	80.3
Medicallm	32.2	43.7	33.4	35.4	45.1	40.0	48.6	38.8	38.9	41.7	55.4	56.7	46.2	60.5	60.7
MiddleTW	42.6	42.6	43.4	42.9	47.6	46.6	44.4	43.4	**41.4**	51.1	78.2	47.9	44.9	55.6	80.2
MoteStrain	19.4	42.1	26.4	28.0	43.1	43.3	17.2	42.6	28.4	44.1	28.4	17.4	46.2	38.7	39.2
OliveOil	16.7	16.7	16.7	23.3	50.0	63.3	80.0	20.0	16.7	60.0	80.0	50.0	46.7	20.0	20.0
OSULeaf	46.7	47.5	47.1	52.1	34.7	32.2	50.4	47.5	60.7	35.5	50.4	57.4	66.1	70.7	69.8
Plane	3.8	3.8	3.8	3.8	1.0	4.8	4.8	2.9	2.9	11.4	20.0	50.5	1.9	65.7	65.7
SmallKit	64.3	60.0	61.1	50.7	52.8	50.9	45.6	58.1	66.4	68.0	67.5	**24.5**	56.8	39.5	48.8
SonySurf	50.1	39.4	31.3	27.0	17.0	**12.1**	38.8	31.6	31.6	15.1	36.8	49.1	27.0	40.8	43.6
SonySurfII	24.3	20.7	21.6	22.6	12.3	14.0	16.9	23.3	16.8	21.2	19.6	36.9	16.1	44.6	43.7
Strawberry	6.0	11.4	6.4	21.0	11.3	18.6	7.0	11.3	5.2	22.7	46.3	39.5	23.0	34.4	35.1
SwedLeaf	25.1	22.7	28.8	24.8	26.6	27.2	44.0	38.6	44.2	38.6	75.5	63.7	45.1	80.0	82.4
SynthCont	48.0	36.3	18.7	36.7	33.7	33.7	15.0	45.3	65.3	38.7	18.3	57.0	47.7	78.3	78.0
ToeSegm1	39.0	21.1	41.7	20.6	29.8	34.2	31.6	17.5	41.2	34.2	33.8	39.5	39.9	39.5	40.4
ToeSegm2	29.2	12.3	36.9	13.8	30.0	26.2	21.5	22.3	66.2	26.2	24.6	15.4	48.5	20.0	29.2
Trace	21.0	16.0	31.0	20.0	**0.0**	**0.0**	47.0	7.0	41.0	**0.0**	63.0	47.0	18.0	44.0	50.0
TwoLeadECG	24.1	21.2	27.2	15.1	23.1	31.1	27.2	27.0	15.3	27.4	45.4	37.7	36.3	36.1	41.4
WordsSyn	40.8	62.9	51.4	70.4	77.0	72.7	47.2	69.4	50.5	77.1	47.0	85.0	76.3	87.0	88.2
# wins	1	1	2	0	1	4	0	1	2	1	0	2	0	0	1
avg. rank	15.0	16.1	17.7	17.7	18.6	19.4	19.6	19.7	19.8	21.8	22.8	23.7	23.8	25.8	26.9

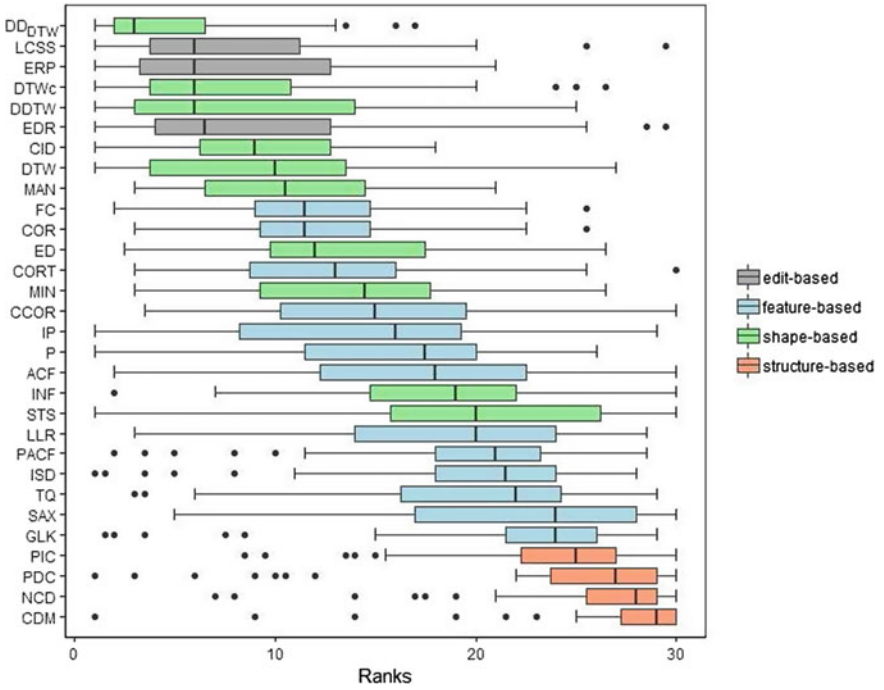

Fig. 1 Box plot of ranks of each measure across all datasets. Boxes are colored according to the category of a measure: shape-based (blue), edit-based (green), feature-based (orange), structure-based (gray)

should pay attention to shape (without editing) or not. Correlation-based distances (e.g. ACF, PACF, CCOR) may be considered as a good choice for datasets: *ECGFive, Trace*.

To assess the differences between examined methods, we performed a detailed statistical comparison. We tested the hypothesis that there are no differences between 1NN classifiers using different measures. Firstly, we employed the test proposed by Iman and Davenport [23], which is a less conservative variant of Friedman's ANOVA [19]. The test is recommended by Demšar [15] and Garcia and Herrera [20]. If the hypotheses is rejected, we can proceed with the post hoc test to provide all pairwise comparisons. In this way we can detect the statistically significant differences between certain classifiers. Garcia and Herrera [20] proved that the procedure presented in Bergmann and Hommel [6] is the most powerful post hoc comparison test. It is based on the idea of finding all elementary hypotheses, which cannot be rejected. However, finding all the possible exhaustive sets of hypotheses for a certain comparison is extremely computationally expensive. Thus, we are able to compare in a post hoc test up to 9 classifiers.

The p-value from the Iman and Davenport's test performed for all classifiers is equal to 0. We can, therefore, proceed with the post hoc tests. The results of

multiple comparisons are given in Table 5. We have chosen for the comparison 9 distance measures, which achieved best average ranks. The p-value from the Iman and Davenport's for these measures was equal to 0.

Based on the Fig. 1 and Table 4, we see that there is one measure that significantly outperform most of the rest—DD_{DTW}. In the group of 9 best classifiers, using p-values obtained from Bergmann–Hommel post hoc test, we can make a division of distances into 3 groups (Table 5). We observe, that there are not statistically significant differences between DD_{DTW} and DTWc as well as LCSS. MAN distance is the worst performing one in the group (taking into account mean ranks), but the post hoc test did not signalize differences with DTW—which is considered to be one of the most efficient measure—and EDR. Another interesting fact is, that CID distance may be treated as statistically equal to much more computationally expensive elastic measures such as DTW, DTWc, EDR, ERP, LCSS. In Fig. 2 we provided plot of critical differences (CD) from Bergmann–Hommel post hoc test, shown in Demšar [15].

Table 4 p-values in the Bergmann–Hommel post hoc test for best 9 measures (taking into account average ranks). Statistically significant differences ($p < 0.05$) are in bold

	ERP	LCSS	DTWc	DDTW	EDR	CID	DTW	MAN
DD_{DTW}	**0.021**	0.118	0.053	**0.016**	**0.005**	**0.000**	**0.000**	**0.000**
ERP		1.000	1.000	1.000	1.000	0.275	1.000	**0.034**
LCSS			1.000	1.000	1.000	0.118	1.000	**0.008**
DTWc				1.000	1.000	0.174	1.000	**0.015**
DDTW					1.000	0.334	1.000	**0.039**
EDR						0.879	1.000	0.118
CID							1.000	1.000
DTW								0.682

Table 5 Results of the Bergmann–Hommel post hoc test: division into groups

	Mean rank	Group 1	Group 2	Group 3
DD_{DTW}	4.82	*		
ERP	7.91	*		
LCSS	8.22	*	*	
DTWc	8.28		*	
DDTW	8.67		*	
EDR	9.13		*	*
CID	9.37		*	*
DTW	10.16		*	*
MAN	10.84			*

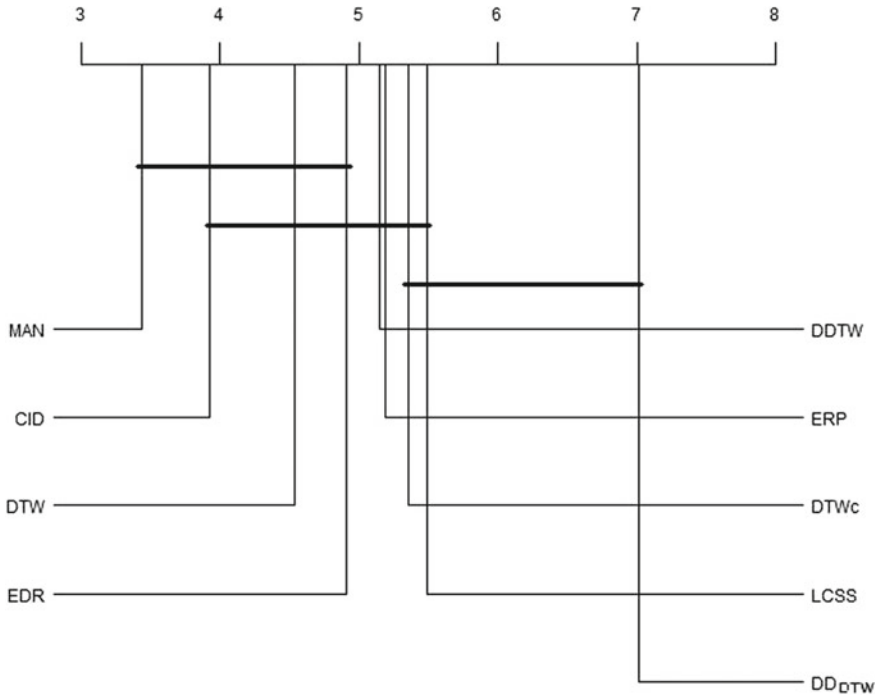

Fig. 2 Plot of critical differences from Bergmann–Hommel post hoc test. Groups of classifiers that are not statistically significantly different (at $p = 0.05$) are connected

We decided to provide comparisons of pairs of classifiers (Fig. 3). We can see, that DD_{DTW} is observably better than DTW and LCSS (most of the points is above the diagonal). Looking at DD_{DTW} and MAN, we see that there are some datasets classified better with the MAN distance, but it occurs extremely rarely. In most cases performance of DD_{DTW} is far better (points are distant to the diagonal). Comparing ERP with MAN and DTW we observed, that the edit-based measure achieves lower error rates than both shape-based distances. The plot of CID and ED shows, that adding a simple complexity correction factor results in a considerable increase of accuracy.

5 Conclusion

In this article, we have compared efficacy of 30 distance measures on 47 datasets, by the prism of 1NN classifier accuracy. Similarly to Serrà et al. [44], Wang et al. [48], we have observed, that there is no measure distinctly better than the others or appropriate for a majority of datasets. Thus, there is still a place for new ones, maybe

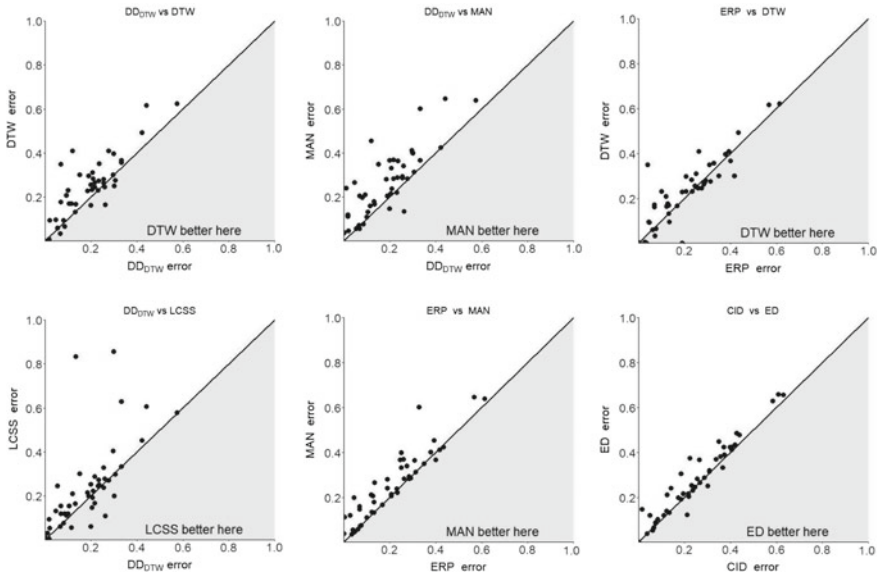

Fig. 3 Comparison of error rates

connecting some properties of already existing measures. From the other hand, best average ranks were achieved by modifications of DTW distance—DD_{DTW}, DDTW, DTWc and by edit-based distances—LCSS, ERP, EDR. Thus, we may draw two conclusions. First, processing shape of time series in a smart way may be a direction for future researches. Second, comparing time series by the mean of edit operations brings remarkable results. Finally, we have also observed, that there are some datasets that are classified better with some groups of measures. It would be highly desirable to find a set of metadata, which could help us to choose the most appropriate measure.

Since this study only discussed 30 of about 40 available distance measures, there is still potential to develop the presented comparison. We plan to cover all available distance measures in the nearest future and, as well, extend the number of datasets for testing them. It would be also interesting to confront conclusions made during these analyses with different time series mining tasks, e.g. with clustering.

Acknowledgements Resources supporting this work were provided by the Poznań Supercomputing and Networking Center.

References

1. Agrawal, R., Faloutsos, C., Swami, A.N.: Efficient similarity search in sequence databases. In: FODO, pp. 69–84 (1993)
2. Antunes, C., Oliveira, A.: Temporal data mining: an overview. KDD Workshop, pp. 1–13 (2001)

3. Aßfalg, J., Kriegel, H.P., Kröger, P., Kunath, P., Pryakhin, A., Renz, M.: Similarity search on time series based on threshold queries. Lecture Notes in Computer Science, pp. 276–294. Springer, Berlin (2006)
4. Bagnall, A., Lines, J., Bostrom, A., Large, J., Keogh, E.: The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Min. Knowl. Discov. **31**(3), 606–660 (2017)
5. Batista, G., Wang, X., Keogh, E.J.: A complexity-invariant distance measure for time series. In: SDM (2011)
6. Bergmann, B., Hommel, G.: Improvements of general multiple test procedures for redundant systems of hypotheses. Mult. Hypotheses Test. 100–115 (1988)
7. Berndt, D.J., Clifford, J.: Using dynamic time warping to find patterns in time series. In: KDD Workshop, AAAI Press, pp. 359–370 (1994)
8. Brandmaier, A.: Permutation distribution clustering and structural equation model trees. Ph.D. thesis, Saarland University (2012)
9. Caiado, J., Crato, N., Peña, D.: A periodogram-based metric for time series classification. Comput. Stat. Data Anal. **50**(10), 2668–2684 (2006)
10. Chen, L., Ng, R.: On the marriage of Lp-norms and edit distance. In: VLDB Endowment, VLDB'04, pp. 792–803 (2004)
11. Chen, L., Özsu, M.T., Oria, V.: Robust and fast similarity search for moving object trajectories. In: SIGMOD'05, pp. 491–502. ACM, New York (2005)
12. Chouakria, A.D., Nagabhushan, P.N.: Adaptive dissimilarity index for measuring time series proximity. Adv. Data Anal. Classif. **1**(1), 5–21 (2007)
13. Dau, H.A., Keogh, E., Kamgar, K., Yeh, C.C.M., Zhu, Y., Gharghabi, S., Ratanamahatana, C.A., Yanping, Hu.B., Begum, N., Bagnall, A., Mueen, A., Batista, G.: The UCR time series classification archive (2018). https://www.cs.ucr.edu/eamonn/time_series_data_2018/
14. de Lucas, D.C.: Classification techniques for time series and functional data. Ph.D. thesis, Universidad Carlos III de Madrid (2010)
15. Demšar, J.: Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. **7**, 1–30 (2006)
16. Díaz, S.P., Vilar, J.A.: Comparing several parametric and nonparametric approaches to time series clustering: a simulation study. J. Classif. **27**(3), 333–362 (2010)
17. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E.: Querying and mining of time series data: experimental comparison of representations and distance measures. PVLDB **1**(2), 1542–1552 (2008)
18. Esling, P., Agon, C.: Time-series data mining. CSUR **45**, 12 (2012)
19. Friedman, M.: A comparison of alternative test of significance for the problem of m rankings. Ann. Math. Stat. **11**, 86–92 (1940)
20. Garcia, S., Herrera, F.: An extension on "statistical comparisons of classifiers over multiple data sets" for all pairwise comparisons. J. Mach. Learn. Res. **9**, 2677–2694 (2008)
21. Golay, X., Kollias, S., Stoll, G., Meier, D., Valavanis, A., Boesiger, P.: A new correlation-based fuzzy logic clustering algorithm for FMRI. Magn. Reson. Med. **40**, 249–260 (1998)
22. Górecki, T., Łuczak, M.: Using derivatives in time series classification. Data Min. Knowl. Discov. **26**, 310–331 (2013)
23. Iman, R.L., Davenport, J.M.: Approximations of the critical region of the fbietkan statistic. Commun. Stat. Theory **9**(6), 571–595 (1980)
24. Kakizawa, Y., Shumway, R.H., Taniguchi, M.: Discrimination and clustering for multivariate time series. J. Am. Stat. Assoc. **93**(441), 328–340 (1998)
25. Kalpakis, K., Gada, D., Puttagunta, V.: Distance measures for effective clustering of ARIMA time-series. In: IEEE 2001, pp. 273–280 (2001)
26. Keogh, E.: A decade of progress in indexing and mining large time series databases. In: Proceedings of the International Conference on Very Large Databases (2006)
27. Keogh, E.: The SAX (symbolic aggregate approximation). http://www.cs.ucr.edu/~eamonn/SAX.htm. Accessed 26 Oct 2018

28. Keogh, E., Kasetty, S.: On the need for time series data mining benchmarks: a survey and empirical demonstration. Data Min. Knowl. Discov. **7**(4), 349–371 (2003)
29. Keogh, E., Lin, J.: Clustering of time-series subsequences is meaningless: implications for previous and future research. Knowl. Inf. Syst. **8**(2), 154–177 (2005)
30. Keogh, E., Ratanamahatana, C.A.: Exact indexing of dynamic time warping. Knowl. Inf. Syst. **7**(3), 358–386 (2005)
31. Keogh, E., Lonardi, S., Ratanamahatana, C.A.: Towards parameter-free data mining. In: SIGKDD, pp. 206–215. ACM (2004)
32. Keogh, E.J., Pazzani, M.J.: Derivative dynamic time warping. In: SDM (2001)
33. Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. In: SIGMOD, pp. 2–11 (2003)
34. Lin, J., Keogh, E., Lonardi, S., Lankford, J.P., Nystrom, D.M.: Visually mining and monitoring massive time series. In: SIGKDD, pp. 460–469. ACM, New York (2004)
35. Lin, J., Keogh, E., Wei, L., Lonardi, S.: Experiencing SAX: a novel symbolic representation of time series. Data Min. Knowl. Discov. **15**(2), 107–144 (2007)
36. Maharaj, E.A.: A significance test for classifying ARMA models. J. Stat. Comput. Simul. **54**(4), 305–331 (1996)
37. Möller-Levet, C.S., Klawonn, F., Cho, K.H., Wolkenhauer, O.: Fuzzy clustering of short time-series and unevenly distributed sampling points, pp. 330–340 (2003)
38. Montero, P., Vilar, J.A.: TSclust: an R package for time series clustering. J. Stat. Softw. **62**(1), 1–43 (2014)
39. Mori, U., Mendiburu, A., Lozano, J.A.: Distance measures for time series in R: the TSdist package. R J. **8**(2), 451–459 (2016)
40. Peña, D., Galeano, P.: Multivariate analysis in vector time series. Technical report (2000)
41. Piccolo, D.: A distance measure for classifying ARIMA models (corr: V11 p180). J. Time Ser. Anal. **11**, 153–164 (1990)
42. Radovanović, M., Nanopoulos, A., Ivanović, M.: Time-series classification in many intrinsic dimensions. SIAM Proc. S 677–688 (2010)
43. Ratanamahatana, C., Lin, J., Gunopulos, D., Keogh, E., Vlachos, M., Das, G.: Mining time series data, pp. 1049–1077 (2010)
44. Serrà, J., Zanin, M., Herrera, P., Serra, X.: Characterization and exploitation of community structure in cover song networks **33**(9), 1032–1041 (2012)
45. Spiegel, S., Jain, B., Albayrak, S.: Fast time series classification under lucky time warping distance. In: SAC (2014). https://doi.org/10.1145/2554850.2554885
46. Tan, P.N., Steinbach, M., Kumar, V.: Introduction to Data Mining, 1st edn. Addison-Wesley Longman Publishing Co., Inc., Boston (2005)
47. Vlachos, M., Kollios, G., Gunopulos, D.: Discovering similar multidimensional trajectories. In: Proceedings of the International Conference on Data, pp. 673–684 (2002)
48. Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., Keogh, E.: Experimental comparison of representation methods and distance measures for time series data. Data Min. Knowl. Discov. **26**(2), 275–309 (2013)
49. Warren Liao, T.: Clustering of time series data-a survey. Pattern Recognit. **38**(11), 1857–1874 (2005)
50. Weigend, A., Gershenfeld, N.: Time Series Prediction: Forecasting the Future and Understanding the Past. Addison-Wesley, Reading (1994)
51. Weiss, G.M.: Mining with rarity: a unifying framework. SIGKDD Explor. Newsl. **6**(1), 7–19 (2004)
52. Xi, X., Keogh, E., Shelton, C., Wei, L., Ratanamahatana, C.A.: Fast time series classification using numerosity reduction. In: ICML, pp. 1033–1040 (2006)
53. Yi, B.K., Faloutsos, C.: Fast time sequence indexing for arbitrary Lp norms. In: Proceedings of the International Conference on Very Large Databases, pp. 385–394 (2000)