The Performance of a Combined Distance Between Time Series

Margarida G. M. S. Cardoso and Ana Alexandra Martins

Abstract This paper presents the comparison of a proposed measure of dissimilarity between time series (COMB) with three baseline measures. COMB is a convex combination of Euclidean distance, a Pearson-correlation-based distance, a Periodogram-based measure and a distance between estimated autocorrelation structures. The comparison resorts to 1-Nearest Neighbour classifier (1NN) since the effectiveness of the dissimilarity measures is directly reflected on the performance of 1NN. Data considered is available in the University of California Riverside (UCR) Time-Series Archive which includes datasets from a wide variety of application domains and have been used in similar studies. The COMB measure shows promising results: a good trade-off performance-computation time when compared to the alternative distances considered.

Keywords Clustering · Distance measures · Time series

1 Introduction

The use of dissimilarity measures between time series is critical in several data analysis tasks which range from simple querying to classification, clustering and anomaly detection. The role of dissimilarity measures in these contexts has been acknowledged by several works, e.g. [1–3].

Recently, in [4], we proposed a new dissimilarity measure, COMB, a convex combination of four (normalized) distance measures which offer complementary perspectives on the differences between two time series: the Euclidean distance which

M. G. M. S. Cardoso (✉)
University Institute of Lisbon (ISCTE-IUL), Business Research Unit (BRU-IUL), Lisbon, Portugal
e-mail: margarida.cardoso@iscte-iul.pt

A. Alexandra Martins
CIMOSM, ISEL, Polytechnic of Lisbon, Lisbon, Portugal

R. Bispo et al. (eds.), *Recent Developments in Statistics and Data Science*,
Springer Proceedings in Mathematics & Statistics 398,
https://doi.org/10.1007/978-3-031-12766-3_6

captures differences in scale; a Pearson-correlation-based measure that takes into account linear increasing and decreasing trends over time; a Periodogram-based measure that expresses the dissimilarities between frequencies or cyclical components of the series and a distance between estimated autocorrelation structures, comparing the series in terms of their dependence on past observations.

COMB achieved quite good results when clustering electricity market prices time series in European regions and also when clustering electricity loads time series (Portuguese Transmission System Operator data)—[4, 5].

In this work, we conduct an experimental analysis to evaluate the comparative performance of the proposed combined distance measure.

The remainder is structured as follows: first we present the Methodology used to provide the comparison of COMB with alternative distance measures; then, the Data Analysis and Results section brings some insights regarding the comparative analysis and, finally, we end with Discussion and Future Research of the presented work.

2 Methodology

2.1 UCR Repository

We resort to the University of California Riverside (UCR) Time-Series Archive where we can find time series of diverse lengths and numbers of target classes, with corresponding test and train sets—[6]. The UCR time-series datasets are from a wide variety of application domains and have been used to study the comparative performance of time-series classifiers—e.g. [2]—and specifically used in comparative studies of dissimilarity measures between time series, e.g. [7].

We limit the datasets considered to 57 taking into account computational cost. This is a criterion that has been invoked in similar studies—e.g. [8]. In our study we found that, for example, the script routine, when referring to the analysis of the "GesturePebbleZ2" UCR dataset, took 18:45 h to run (using a PC with processor Intel(R) Core(TM) i7-10750H CPU @ 2.60 2.59 GHz with a RAM of 32 GB). Nevertheless, we tried to include dissimilar datasets, namely, in what regards the number of target classes: 28 datasets have 2 target classes while 29 have more than 2 target classes. The selected datasets are presented in the Appendix. As in previous studies—e.g. [7]—and although this can limit the analysis, z-standardization is adopted for fairness, since many of the UCR series are presented in their z-standardized form.

2.2 Using the 1NN Classifier

We follow a methodology suggested in previous studies that were conducted to compare several dissimilarity measures and their variants—e.g. [7]: we use one nearest neighbour (1NN) classifier on labelled data to evaluate the performance of the distance measures. In fact, since the distance measure used is critical to 1NN accuracy, this indicator directly reflects the effectiveness of the dissimilarity measure used. According to [7] p. 1890, *1NN classifiers are suitable methods for distance measure evaluation for several reasons:*

1. *resemble the problem solved in time-series similarity search;*
2. *are parameter-free and easy to implement;*
3. *are dependent on the choice of distance measure;*
4. *provide an easy-to-interpret (classification) accuracy measure which captures if the query and the nearest neighbour belong to the same class.*

2.3 Dissimilarity Measures

We compare COMB [4] with three alternative dissimilarity measures between time series. Comparisons are provided with three baseline measures: Euclidean distance, DTW (Dynamic Time-Warping with Sakoe-Chiba band [9] windowing considering 20% of the time-series length) and Complexity Invariance Distance (CID).

COMB Distance. Considering two time series x_t and y_t, $(t = 1, \ldots, T)$, the COMB distance is a convex combination of four distances: Euclidean (d_{Euclid}), a Pearson-correlation-based measure ($d_{Pearson}$), a Periodogram-based measure (d_{Period}) and an autocorrelation-based measure ($d_{Autocorr}$).

The Euclidean distance, d_{Eucl}, yields the sum of Euclidean distances corresponding to each pair (x_t, y_t) capturing the differences in scale:

$$d_{Eucl} = \left(\sum_{t=1}^{T} (x_t - y_t)^2 \right)^{\frac{1}{2}}. \tag{1}$$

The Pearson-correlation-based measure takes into account linear increasing and decreasing trends over time. The following measure was suggested by [10]:

$$d_{Pearson} = \sqrt{\frac{1 - r_{x_t, y_t}}{2}}, \tag{2}$$

where r_{x_t, y_t} represents the Pearson correlation.

The Periodogram-based measure [11] considers the Euclidean distances between the Periodograms expressing the contribution of the various frequencies or cyclical components to the variability of the series,

$$d_{Period} = \left(\sum_{j=1}^{[\frac{T}{2}]} \left(P_x\left(w_j\right) - P_y\left(w_j\right) \right)^2 \right)^{\frac{1}{2}}, \tag{3}$$

where $P_x\left(w_j\right)$ is the Periodogram of time series x_t at frequencies $w_j = 2\pi_j/n$, $j = 1, \ldots, [n/2]$ in the range 0 to π, being $[n/2]$ the largest integer less or equal to $n/2$,

$$P_x\left(w_j\right) = \left(\frac{1}{n}\right) \left| \sum_{t=1}^{T} x_t e^{-itw_j} \right|^2. \tag{4}$$

The autocorrelation-based distance [12] calculates Euclidean distances between autocorrelation structures, comparing the series in terms of their dependence on past observations

$$d_{Autocorr} = \left(\sum_{l=1}^{L} \left(r_l\left(x_t\right) - r_l\left(y_t\right) \right)^2 \right)^{\frac{1}{2}}, \tag{5}$$

where $r_l\left(x_t\right)$ and $r_l\left(y_t\right)$ represent the estimated autocorrelations of lag l of (x_t) and (y_t), respectively.

In this study, we specifically use an uniform convex combination, all four weights being equal.

Eucl—Euclidean Distance. The comparison with the performance of the Euclidean distance is unavoidable in all studies of this type. Even because, despite its simplicity, this distance can obtain surprisingly good results *especially if the size of the training set/database is relatively large*, [13], p. 281.

DTW—Dynamic Time-Warping. DTW is an elastic measure that computes the optimal alignment between two time series to minimize the sum of distances between aligned elements.

Considering two time series x_t and y_t, $(t = 1, \ldots, T)$, let M be the $T \times T$ matrix where each element is a dissimilarity $d_{i,j}$ (commonly the Euclidean distance is considered) between any pair of elements x_i and y_j $(i, j = 1, \ldots, T)$.
A warping path $P = ((i_1, j_1), (i_2, j_2))$ is a series of indexes of M defining a mapping from each element of one time series to one, or more than one, or even none, of the elements of the other time series. A valid path should satisfy several conditions, for example, $i_{k+1} \geq i_k$ ensures the path does not go back in time. For other step patterns constrains see, e.g. [14]. For each path P, through M, the total sum of the distances along it is

$$D(P) = \sum_{k=1}^{K} d_{i_k, j_k}. \tag{6}$$

For example, the Euclidean distance is the total distance along the diagonal of M. The goal of the DTW measure is to find a path P^* that minimizes the total distances $D(P)$:

$$P^* = \min_P D(P). \tag{7}$$

To improve the efficiency of the procedure, it is a common practice to limit the time distortion (e.g. considering 20% of the time-series length). For example, the Sakoe-Chiba band [9] limits the warping path to a band of size T_0 directly above and to the right of the diagonal of the matrix M, by enforcing the constraint $|i_k - j_k| < T_0$.

CID—Complexity Invariance Distance. CID measure was proposed by [15]. The time series' complexity is measured by stretching them and measuring the length of the resulting lines.

$$CID(x_t, y_t) = d_{Eucl}.CF(x_t, y_t), \tag{8}$$

where

$$CF(x_t, y_t) = \frac{max(CE(x_t), CE(y_t))}{min(CE(x_t), CE(y_t))} \tag{9}$$

is the Complexity Factor, and

$$CE(x_t) = \left(\sum_{t=1}^{T-1} (x_t - x_{t+1})^2 \right)^{\frac{1}{2}} \tag{10}$$

is the Complexity Estimate of time series x_t.

We resort to the R package "TSclust" [12] where the four distances that compose the COMB distance, the CID and the DTW (using the "dtw" package [14]) are implemented.

2.4 Evaluating the Classification Results

The evaluation of performance of the 1NN classifiers regards the test sets of the UCR time series considered. Balanced accuracy measure (average between sensitivity and specificity) when dealing with unbalanced sets is suggested by [6]. We propose using the Huberty index (HI)—e.g. [16], as a measure of classification performance:

$$HI = \sum_{k=1}^{K} \frac{p^c - p^{def}}{1 - p^{def}}, \tag{11}$$

where p^c and p^{def} are the proportion of observations correctly classified and the proportion of observations in the modal class, respectively. This measure is clearly useful for the evaluation of performance in unbalanced datasets. Furthermore, it provides a fair and interpretable view of the success of classification tasks which could be overestimated when high accuracy results are obtained in strongly unbalanced sets, e.g. a 90% accuracy result when a target class includes 95% of observations yields a negative Huberty index (one should do better by simply allocating all observations to the modal class). In addition, the computational time is also taken into account in the evaluation of the 1NN results referring to the four dissimilarity measures considered.

After the evaluation of aggregated results, comparisons referring to specific datasets are considered to get some dissimilarities' performance-related insights. On a "closer look to specific problems", [2] resorts to the selection of some time series from each target class, trying to capture the main differences between these classes on specific datasets. We propose using the medoids of each class as defined by each dissimilarity measure to obtain those insights. The medoid definition is the observations that minimize the sum of all distances to elements in the same class—[17].

3 Data Analysis and Results

3.1 General Comparisons

A brief exploratory data analysis leads to the conclusion that, in the datasets considered, DTW generally provides better classification results than the alternative distances, followed by COMB—Table 1. COMB comparative results are illustrated in Fig. 1. However, for time series with two target classes (K=2) only, COMB provides slightly better results—see Table 1. In what regards the computation time DTW clearly provides the worst results—see Table 2.

According to the Friedman test's results, there are no significant differences between the distributions of HI regarding the four dissimilarity measures (see Table 3). However, significant differences can be found when analysing data with more than two classes (K>2), which, after pairwise comparison of Dunn's test, can be referred to the significant difference between HI.Eucl and HI.DTW (see Table 4).

The differences between computation times regarding the four dissimilarity measures are all significant according to Friedman's test—see Table 5.

Table 1 All time-series results

	Hubert index				Computation time (seconds)			
	HI.Eucl	HI.DTW	HI.CID	HI.COMB	t.Eucl	t.DTW	t.CID	t.COMB
Mean	**0.581**	**0.631**	**0.599**	**0.603**	**0.01**	**7846.60**	**47.53**	**651.27**
Std. Dev.	0.279	0.268	0.273	0.261	0.01	15500.13	66.81	959.48
Coef. Var.	0.480	0.425	0.456	0.433	1	1.98	1.41	1.47
Perc. 25th	0.385	0.460	0.448	0.448	0.00	313.80	5.23	49.64
Perc. 50th	**0.633**	**0.679**	**0.609**	**0.636**	**0.00**	**1117.40**	**13.00**	**157.96**
Perc. 75th	0.799	0.842	0.830	0.791	0.02	7241.48	88.95	1047.15
IQR	0.414	0.382	0.382	0.343	0.02	6927.68	83.72	997.51

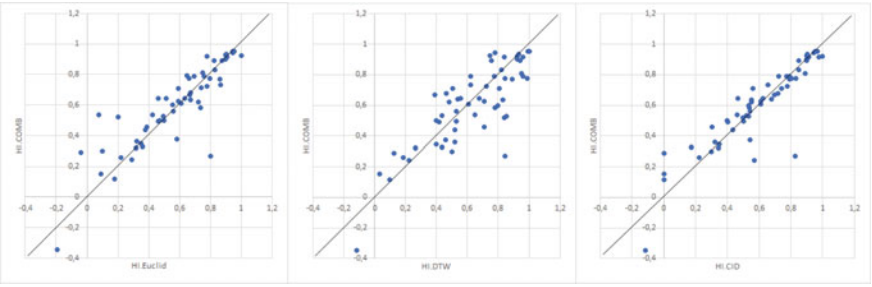

Fig. 1 Plot of Huberty index results: COMB versus Euclidean, DTW and CID

Table 2 Huberty index results: time series with two target classes versus more than two target classes

	Two target classes				More than two target classes			
	HI.Eucl	HI.DTW	HI.CID	HI.COMB	HI.Eucl	HI.DTW	HI.CID	HI.COMB
Mean	**0.516**	**0.547**	**0.544**	**0.562**	**0.642**	**0.712**	**0.651**	**0.643**
Std. Dev.	0.346	0.298	0.321	0.314	0.179	0.212	0.210	0.194
Coef. Var.	0.669	0.544	0.591	0.559	0.278	0.297	0.323	0.302
Perc. 25th	0.211	0.394	0.337	0.323	0.511	0.528	0.533	0.496
Perc. 50th	**0.546**	**0.579**	**0.551**	**0.612**	**0.662**	**0.722**	**0.622**	**0.636**
Perc. 75th	0.827	0.799	0.842	0.809	0.778	0.925	0.803	0.790
IQR	0.616	0.405	0.505	0.486	0.267	0.397	0.270	0.294

Table 3 Friedman test's results regarding Huberty index

	Test statistic (p-value)
All sample	7.062 (0.07)
K = 2	4.375 (0.228)
k > 2	12.761 (0.005)

Table 4 Dunn's pairwise comparison tests regarding Huberty index for data with more than two classes ("Adj. Sig" are p-values adjusted by Bonferroni correction)

	Test statistic	Sig.	Adj. Sig.
HI.Eucl-HI.CID	−0.328	0.334	1.000
HI.Eucl-HI.COMB	−0.414	0.222	1.000
HI.Eucl-HI.DTW	−1.121	0.001	0.006
HI.CID-HI.COMB	−0.086	0.799	1.000
HI.CID-HI.DTW	0.793	0.019	0.116
HI.COMB-HI.DTW	0.707	0.037	0.222

Table 5 Friedman test's results regarding computation time

	Test statistic (p-value)
All sample	171.0 (0.000)
K = 2	84.0 (0.000)
K > 2	87.0 (0.000)

3.2 COMB "Wins" and "Looses" Examples

In an attempt to understand the data conditions that could (un)favour COMB, we looked for some insights regarding a "COMB wins example" and a "COMB looses example": ToeSegmentation2 and Herring time series, respectively. ToeSegmentation2 was originated in the CMU Graphics Lab Motion Capture Data, referring to right toe movements, with target classes "Walk Normally" and "Walk Abnormally". Herring data refers to calcium carbonate structures from two classes of Herring: North sea or Thames. In Table 6, we present the details of data referring to these two datasets.

On the assumption that exploring the target classes in the test set could bring some insights into the performance of 1NN classifier, we obtained the medoids of target classes according to each of the four dissimilarity measures. The ToeSegmentation2 test set classes' medoids are depicted in Fig. 2. The COMB measure reveals not only scale differences between the medoids (as Euclidean distance does, with the poorest results) but it is also apparent (for example) how the medoids' tendencies diverge from each other, which, conjugated with the additional differences captured by COMB, results in its best performance, according to the HI.

Table 6 COMB "wins" and "looses" datasets

Name	Train	Test	Class	Length	HI.Eucl	HI.DTW	HI.CID	HI.COMB
ToeSegmentation2	36	130	2	343	−0.0416	0.1252	0.0001	0.2915
Herring	64	64	2	512	−0.192	−0.115	−0.115	−0.346

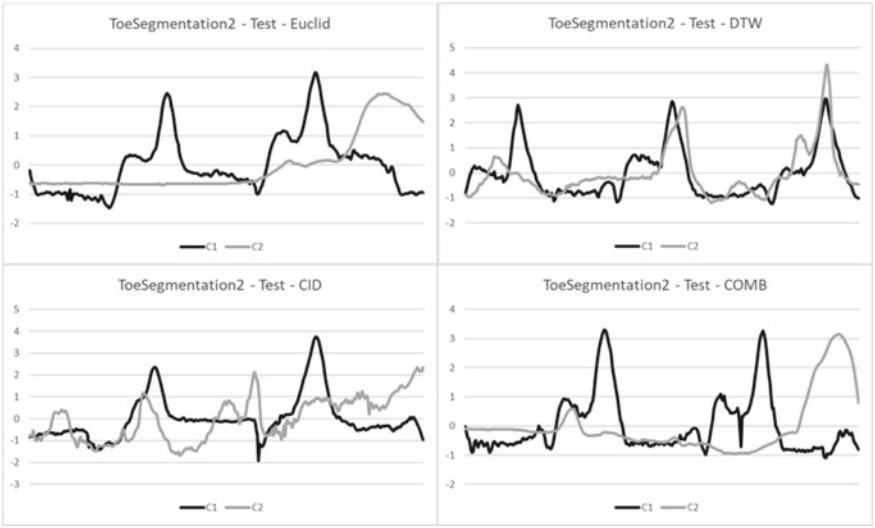

Fig. 2 Medoids of ToeSegmentation2 test set classes, according to dissimilarity measures Eucl, DTW, CID and COMB

The Herring test set classes' medoids coincide for all dissimilarity measures except DTW which presents slightly different medoids. Nevertheless, a negative HI was obtained for all measures (revisit Table 6).

In an attempt to explore the potential of COMB measure, in a worst-case scenario, we performed a brief sensitivity analysis manipulating the COMB's weights. After some trials, when considering the COMB weights regarding d_{Period} and $d_{Autocorr}$ as nine times the weights regarding d_{Euclid} and $d_{Pearson}$, we managed to cross the "waterline", obtaining a HI slightly positive which the alternative measures could not. Note, however, that a customized parametrization of DTW could eventually obtain better results also, but we believe that it would also bring a relevant increase in computation time.

4 Discussion and Future Research

We conducted experiments on 57 time-series datasets from diverse application domains to compare the proposed dissimilarity measure, COMB, with three baseline alternative measures: Euclidean, Dynamic Time-Warping and Complexity Invariance Distance. We resorted to the 1-Nearest Neighbour classifier, using the four dissimilarities, to compare their effectiveness. Huberty index was used as a classification metric providing more informative analysis results than the simple Accuracy measure,

adopted in previous studies to evaluate performance (ignoring prevalence). Experimental results obtained indicate that there are no significant differences between the classification performance (Huberty index measures) of the four dissimilarity measures. Nevertheless, it appears COMB can produce better results regarding time series with two target classes. Furthermore, there is also the potential to improve the results obtained with COMB by changing the weights in the convex combination: an example was provided for the Herring dataset where the COMB with uniform weights provided the worst classification results, while COMB with tuned weights was able to provide the best results. In what regards the computation time, Dynamic Time-Warping, which appears to be the most direct COMB competitor regarding classification performance, presented the (significantly) worst results. Considering the classification performance-runtime results we conclude that the proposed combined measure can be seen as competitive in several settings.

In future research, we aim to extend the present analysis to all (128) UCR datasets which will require to explore hardware-aware implementations and/or algorithmic solutions to turn the measures' implementation the most efficient. We also think the Complexity Invariance Distance, which revealed to be a competitive measure, should definitely play a role in future similar studies (along with the unavoidable Euclidean and Dynamic Time-Warping dissimilarities and other eventual baseline measures). An investigation of the process to determine COMB weights should also be considered. Finally, the experimental design should include additional characteristics of the time-series data, besides the number of target classes, namely, we think that the inclusion of a measure of separation between classes should be considered.

Acknowledgements This work was supported by Fundação para a Ciência e a Tecnologia, grant UIDB/00315/2020.

5 Appendix: The Datasets

The characteristics of the 57 datasets used in this work are presented in Tables 7 and 8. Several time series have missing values which were treated with linear interpolation. In order to make all time series of the same dataset with equal length, low-amplitude random noise was imputed to the end of time series with smallest length. For more details, see web page https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.

Table 7 The datasets' dimensions

Name	Train	Test	No. Classes	Length
ArrowHead	36	175	3	251
Beef	30	30	5	470
BeetleFly	20	20	2	512
BirdChicken	20	20	2	512
BME	30	150	3	128
Car	60	60	4	577
CBF	30	900	3	128
Chinatown	20	343	2	24
Coffee	28	28	2	286
DiatomSizeReduction	16	306	4	345
DodgerLoopDay	78	80	7	288
DodgerLoopGame	20	138	2	288
DodgerLoopWeekend	20	138	2	288
ECG200	100	100	2	96
ECGFiveDays	23	861	2	136
FaceFour	24	88	4	350
Fish	175	175	7	463
FreezerSmallTrain	28	2850	2	301
Fungi	18	186	18	201
GestureMidAirD1	208	130	26	360
GestureMidAirD2	208	130	26	360
GestureMidAirD3	208	130	26	360
GesturePebbleZ1	132	172	6	455
GesturePebbleZ2	146	158	6	455
GunPoint	50	150	2	150
GunPointAgeSpan	135	316	2	150
GunPointMaleVersusFemale	135	316	2	150
GunPointOldVersusYoung	136	315	2	150
Ham	109	105	2	431
Herring	64	64	2	512
HouseTwenty	40	119	2	2000
InsectEPGRegularTrain	62	249	3	601

Table 8 The datasets' dimensions (continuation)

Name	Train	Test	No. Classes	Length
InsectEPGSmallTrain	17	249	3	601
ItalyPowerDemand	67	1029	2	24
Lightning2	60	61	2	637
Lightning7	70	73	7	319
Meat	60	60	3	448
MoteStrain	20	1252	2	84
OliveOil	30	30	4	570
OSULeaf	200	242	6	427
PickupGestureWiimoteZ	50	50	10	361
Plane	105	105	7	144
PowerCons	180	180	2	144
Rock	20	50	4	2844
ShakeGestureWiimoteZ	50	50	10	385
ShapeletSim	20	180	2	500
SmoothSubspace	150	150	3	15
SonyAIBORobotSurface1	20	601	2	70
SonyAIBORobotSurface2	27	953	2	65
Symbols	25	995	6	398
ToeSegmentation1	40	228	2	277
ToeSegmentation2	36	130	2	343
Trace	100	100	4	275
TwoLeadECG	23	1139	2	82
UMD	36	144	3	150
Wine	57	54	2	234
WormsTwoClass	181	77	2	900

References

1. Ding, H., Trajcevski, G., Scheuermann, P., Wang, X., Keogh, E.: Querying and mining of time series data: experimental comparison of representations and distance measures. In: Proceedings of the VLDB Endowment, vol. 1, pp. 1542–1552 (2008)
2. 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. Disc. **31**(3), 606–660 (2017)
3. Javed, A., Lee, B.S., Rizzo, D.M.: A benchmark study on time series clustering. In: Machine Learning with Applications, vol. 1, p. 100001 (2020)
4. Cardoso, M., Martins, A., Lagarto, J.: Combining various dissimilarity measures for clustering electricity market prices. In: Milheiro, P., Pacheco, A., de Sousa, B., Alves, I.F., Pereira, I., Polidoro, M.J., Ramos, S. (eds.) Estatística: Desafios Transversais ás Ciências dos Dados—Atas do XXIV Congresso da Sociedade Portuguesa de Estatística (), Edições SPE, pp. 197–212 (2021)

5. Martins, A., Lagarto, J., Canacsinh, H., Reis, F., Cardoso, M.: Short-term load forecasting using time series clustering. In: Proceedings of 16th Conference on Sustainable Development of Energy, Water and Environment Systems (2021). ISSN: 1847-7178
6. Dau, H.A., Bagnall, A., Kamgar, K., Yeh, C.-C.M., Zhu, Y., et al.: The UCR time series archive. IEEE/CAA J. Autom. Sin. **6**(6), 1293–1305 (2019)
7. Paparrizos, J., Liu, C., Elmore, A.J., Franklin, M.J.: Debunking four long-standing misconceptions of time-series distance measures. In: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data. ACM (2020)
8. Hills, J., Lines, J., Baranauskas, E., Mapp, J., Bagnall, A.: Classification of time series by shapelet transformation. Data Min. Knowl. Disc. **28**(4), 851–881 (2013)
9. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Signal Process. **26**(1), 43–49 (1978)
10. Rodrigues, P., Gama, J., Pedroso, J.: Hierarchical clustering of time-series data streams. IEEE Trans. Knowl. Data Eng. **20**(5), 615–627 (2008)
11. Caiado, J., Crato, N., Peña, D.: A periodogram-based metric for time series classification. Comput. Stat. Data Anal. **50**(10), 2668–2684 (2006)
12. Montero, P., Vilar, J.A.: TSclust: an R package for time series clustering. J. Stat. Softw. **62**(1) (2014)
13. Wang, X., Mueen, A., Ding, H., Trajcevski, G., Scheuermann, P., et al.: Experimental comparison of representation methods and distance measures for time series data. Data Min. Knowl. Disc. **26**(2), 275–309 (2012)
14. Giorgino, T.: Computing and visualizing dynamic time warping alignments in R: the `dtw` package. J. Stat. Softw. **31**(7) (2009)
15. Batista, G., Wang, X., Keogh, E.: A complexity-invariant distance measure for time series. In: Proceedings of the 2011 SIAM International Conference on Data Mining, pp. 699–710 (2011)
16. Sharma, S.: Applied Multivariate Techniques. Wiley, New York (1996)
17. Kaufman, L., Rousseeuw, P.J.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (2009)