HAL
open science

Time Adaptive Optimal Transport: A Framework of Time Series Similarity Measure

Zheng Zhang, Ping Tang, Thomas Corpetti

To cite this version:

Zheng Zhang, Ping Tang, Thomas Corpetti. Time Adaptive Optimal Transport: A Framework of Time Series Similarity Measure. IEEE Access, 2020, 8 (24), pp.149764-149774. 10.1109/ACCESS.2020.3016529 . hal-03045373

HAL Id: hal-03045373
https://hal.science/hal-03045373

Submitted on 14 Dec 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3016529, IEEE Access

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2020.DOI

Time Adaptive Optimal Transport : A Framework of Time Series Similarity Measure

ZHENG ZHANG[1], PING TANG[1], AND THOMAS CORPETTI[2]
[1]Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, 100101, China (e-mail: zhangzheng,tangping@aircas.ac.cn)
[2]LETG UMR CNRS 6554, University of Rennes 2, place du recteur Henri Le Moal, 35000 Rennes, France (e-mail: thomas.corpetti@univ-rennes2.fr)

Corresponding author: Zheng Zhang (e-mail: zhangzheng2035@163.com).

This work was supported by the National Natural Science Foundation of China (Grant No.41701399), and the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No.XDA19080301).

ABSTRACT Similarity measure is a critical tool for time series analysis. However, currently established methods, for instance, dynamic time warping (DTW) and its variants, are still facing some issues such as non-maximum-to-maximum alignment and pathological alignment, etc. Despite many attempts to improve, these issues remain stubborn because they are directly caused by the intrinsic mechanism of DTW. Thinking out of the context of DTW based methods, we propose in this paper a new time series similarity measure framework which we call Time Adaptive Optimal Transport (TAOT). As its name implies, TAOT is based on optimal transport, a powerful distance measure for histograms and probability distributions, and TAOT inherits several promising properties from optimal transport to tackle the problems of classic DTW based methods. We make optimal transport capable of handling time series data by considering both observed values and their corresponding time coordinates simultaneously. TAOT can generate a many-to-many alignment between time series that further releases the search space for a more correct result. Experimental results show that TAOT can outperform other widely used similarity measures on classification tasks on multiple datasets. We also introduce the parameter extracting and visualization strategies of TAOT in this paper.

INDEX TERMS Optimal transport, time series, similarity measure, Sinkhorn distance, classification.

I. INTRODUCTION

SIMILARITY measure is fundamental to many machine learning and data mining tasks, such as classification [1]–[4], clustering [5]–[7], indexing [8]–[11], etc. In the context of time series analytics, similarity measure has long been a research hotspot and many methods have been proposed [2], [12]–[16]. These similarity measures can be roughly categorized into: time-rigid measures (Euclidean distance), time-flexible measures (dynamic time warping) [15], [17], feature-based measures (Fourier coefficients) [18]–[20], and model-based measures (auto-regression and moving average model) [14], [21]. Among these methods, dynamic time warping (DTW) [17] and its variants [22]–[27] are probably the most popular and established ones [12], [21], [28]. However, this kind of DTW-based methods still exhibit several intrinsic alignment problems:

a) Pathological alignment: A critical job for similarity measure of time series is to tackle time distortions. DTW screens most time distortions by realigning data points in the two time series to be compared, and that helps DTW outperform Euclidean distance in many scenarios [28]. However, the intrinsic rules used by DTW to generate the new alignment can lead to pathological alignment, where a single point in one time series links to a large subsection of another time series [23], as shown in Fig. 1(a). In order to avoid this undesirable alignment, various constraints for DTW [29]–[32] have been proposed, for instance, Sakoe-Chiba band and Itakura parallelogram. But most of these constraints are rigid such that they take a risk of preventing the correct alignment from being generated.

b) Pairing of maximum values: In some applications, the maximum value of a time series is of overwhelming importance, and thus it is decisive to guarantee that the maximum value of one time series aligns to the maximum value of another time series. Unfortunately, DTW and its existing variants provide no such guarantee, and to some extent this

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3016529, IEEE Access

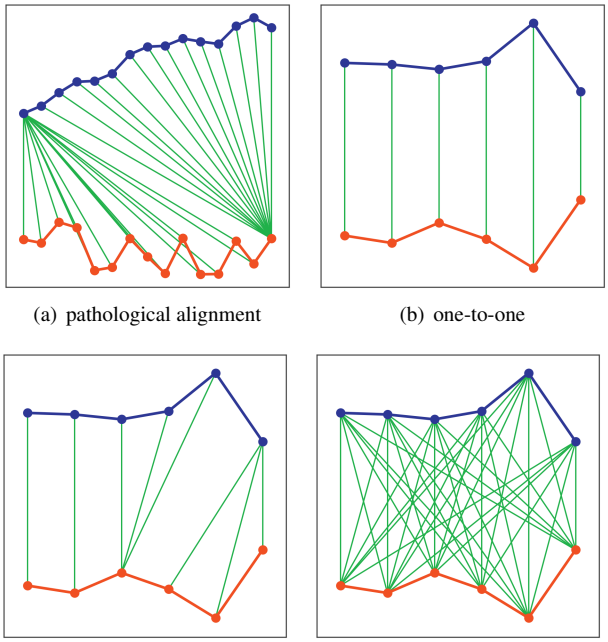
requirement even conflicts with the underlying mechanism of DTW. A mandatory pairing of maximum values can break the time-monotonicity of DTW and prevent DTW from finding the optimal alignment with the lowest global cost. Other common similarity measures, such as Euclidean distance and Minkowski distance, can only succeed in a certain situation when two peaks happen at the same time point, because they obey a time-rigid alignment rule.

c) Many-to-many alignment: From the perspective of typology, there should be three categories of alignments between time series, one-to-one, one-to-many, and many-to-many, as illustrated in Fig. 1(b-c-d) respectively. Among the three categories, the potential of one-to-one alignment and one-to-many alignment has already been exploited; for instance, time-rigid methods employ one-to-one alignment and DTW-based methods employ one-to-many alignment. However, the use of many-to-many alignment is still an open direction. In theory, many-to-many alignment has a larger capacity and a higher flexibility. Therefore, many-to-many alignment based methods deserve to be developed.

d) Privilege of time ordering: The similarity between time series is measured mainly from two aspects: observed values and time ordering. To some extent existing methods always consider time ordering in the first place. For instance, DTW first sets a rule of time ordering (continuity, monotonicity, etc), and then finds the minimum accumulated cost of observed values under the given time ordering rule. Euclidean distance simply follows a time-rigid ordering. This privilege of time ordering narrows the search space and as a result forfeits the opportunity to find a more correct alignment. Instead, a new method that considers observed values and time ordering simultaneously may further release the capacity and lead to a more accurate similarity measure.

A review of literature shows a lack of method effectively responding to the aforementioned issues, and one reason is that these alignment problems are closely related to the core mechanism of DTW. Thinking outside the context of classic DTW-based methods, we find Optimal Transport (OT) [33], also known as the Earth Mover's Distance (EMD) [34] or Wasserstein Distance, is a successful method to compare probabilities or histograms [35]–[37], and theoretically it has several promising properties [38] to solve the aforementioned alignment problems of DTW. For example, OT does not rely on the time order of data points and it always aligns maximum values with each other. In the past the utility of OT was greatly limited by its high computational complexity, but fortunately this issue has been largely alleviated by one of its variants, Sinkhorn distance [38], which transfers OT into a problem with a fast solution by adding an entropic regularization term and ends up making the computation several orders of magnitude faster. Another barrier to the use of OT on time series data is that OT only considers the observed values while ignores the corresponding time coordinates.

In this paper, we propose a time-adaptive version of OT (TAOT) to serve as a new similarity measure framework between time series. We make Sinkhorn distance capable of handling time series data by considering both observed values and their corresponding time coordinates simultaneously. Classification experiments on multiple real-world and synthetic datasets are conducted to show the accuracy of TAOT and how the aforementioned alignment problems of DTW are coped with. Other closely related topics such as the parameter extracting strategies and the visualization of TAOT will also be discussed.



(a) pathological alignment (b) one-to-one (c) one-to-many (d) many-to-many

FIGURE 1. Examples of different categories of alignments between time series.

The rest of the paper is structured as follows. Section II briefly revisits related works on dynamic time warping (DTW), optimal transport (OT), and their major variants. Section III details the proposed new similarity measure, namely TAOT. Section IV evaluates the performance of TAOT with classification experiments on UCR time series datasets and discusses several practical issues. Finally, Section V summarizes our main conclusions.

II. BACKGROUND

A. DYNAMIC TIME WARPING

DTW was first introduced in the field of speech recognition in the influential work by [17] and soon become widely used because it can cope with time distortions effectively. DTW aims at finding the optimal alignment between time series that can achieve the minimum accumulated cost. Given two time series, $A = a_1, a_2, ..., a_i, ..., a_m$ and $B = b_1, b_2, ..., b_j, ..., b_n$, DTW first constructs an m-by-n cost matrix D. Each matrix cell (i, j) contains the distance $d(i, j)$ between the two data points a_i and b_j. Squared Euclidean distance is normally used when calculating the cost matrix, and thus $d(a_i, b_j) =$

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3016529, IEEE Access

$(a_i - b_j)^2$. Then DTW tries to find the optimal alignment that leads to the minimum accumulated cost. The alignment is represented by a warping path $W = w_1, w_2, ..., w_k, ..., w_K$ that consists of a continuous set of matrix cells. Each matrix cell, for instance, $w_k = (i, j)$ corresponds to a pairing of points a_i and b_j, and the cost between them $d(i, j)$ is added to the final accumulated cost. Fig. 2 shows an example of the cost matrix and warping path. Eq. 1 defines the above-described DTW distance:

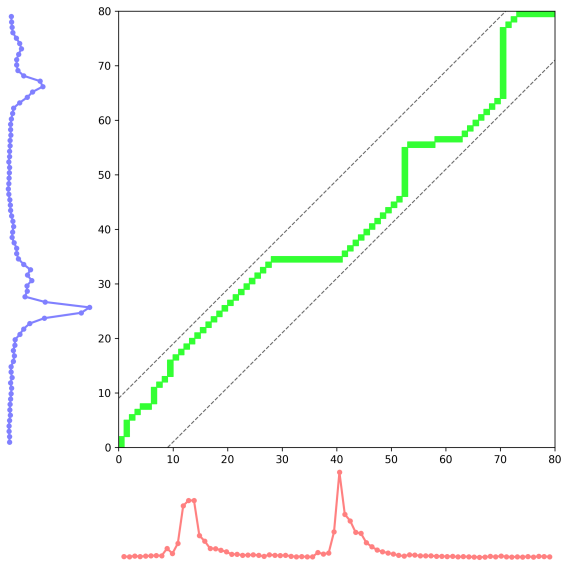


FIGURE 2. Warping path and cost matrix of DTW.

$$dtw(A, B) = \min_{W} \sum_{k=1}^{K} d(w_k) \tag{1}$$

DTW satisfies three time ordering rules: boundary condition, continuity, and monotonicity [23]. The boundary condition restricts the warping path to start at the lower-left corner of the warping matrix $w_1 = (1, 1)$ and finish at the upper-right corner $w_K = (m, n)$. The continuity restricts the allowable steps to adjacent cells. The monotonicity prevents all steps in the warping path from turning backward in any circumstances.

In practice, DTW distance can be efficiently calculated by the following recursive formula: $dtw(i, j) = d(i, j) + min\{dtw(i-1, j-1), dtw(i, j-1), dtw(i-1, j)\}$, where $dtw(i, j)$ is the accumulated cost found in current cell (i, j) and the final DTW distance is $dtw(m, n)$.

B. VARIANTS OF DTW

Despite its popularity, DTW was proposed decades ago and its pathological alignment problem has long been noted. In order to alleviate the problem, different variants of DTW have been proposed and we briefly review them here. These methods can be classified into two major categories.

The first category sets constraints on DTW. The most common and straightforward constraint is windowing [29]–[32], where allowable elements of warping path must be located within a warping window, as illustrated by the dashed line in Fig. 2. Different warping windows have been proposed, among which Sakoe-Chiba band and Itakura parallelogram are two most commonly used ones. In recent years, some researchers suggest to learn adaptive-shaped windows [39], [40] from the data instead of fixed-shaped windows. Another frequently used type of constraints is weighting. For example, Weighted DTW (WDTW) [24] is a representative method of weighting which applies different weights to temporally adjacent points when computing DTW. By penalizing further points, WDTW prevents minimum distance distortions caused by outliers and enhances the detection of similarity between two time series. Besides windowing and weighting, recently proposed LDTW [27] constrains the maximum allowable length of the warping path based on the observation that pathological alignment always happens concurrently with an unusually long warping path. SP-DTW and its kernelization version SP-KrDTW [41] address the sparsification of the alignment path search space for DTW-like measures to improve their efficiency without loosing accuracy.

The second category replaces the feature DTW considers. For example, Piecewise DTW (PDTW) [42], [43] proposes to use a compact abstraction instead of the raw data to compute DTW, with the aim of avoiding the impact of outliers and accelerating the computation. PDTW first splits time series into fixed-length segments and compute the mean of each segment. Then these mean values are used instead of raw data points. A challenge of PDTW is the choice of the optimal size of segments. To avoid brute-force search, Parameter Free Piecewise DTW (FDTW) [26] proposes a heuristic search of the size of segments for PDTW on the basis of classification accuracy. Derivative DTW (DDTW) [23] considers the local derivative of each data point rather than the raw value. Many variants of DTW can have a derivative version, for instance, WDTW and Weighted Derivative DTW (WDDTW) [24]. SC-DTW [25] views time series as 2D contours and employs shape context, a rich shape descriptor, to compute DTW. Local feature based DTW (LFDTW) [20] proposes a general framework to use different type of local features to generate the warping path.

C. OPTIMAL TRANSPORT

In the context of machine learning, OT [33], [34](also known as Earth Mover's Distance (EMD) or Wasserstein Distance) has long been a powerful tool to compare probabilities or histograms [37], [44], [45]. OT is modeled as the solution to the transportation problem. Suppose there are a collection of mines mining iron ores, and a collection of factories that consume the iron ores. Given the amount of supply and demand of each mine and factory, and the shipment cost from each mine to each factory, OT can find the optimal allocation plan with a minimum total shipment cost for resolving the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3016529, IEEE Access

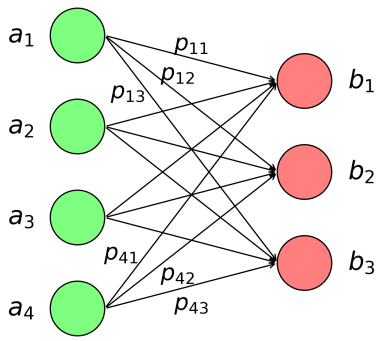


FIGURE 3. An example of optimal transport. We can observe that OT generates a many-to-many alignment.

supply-demand transports.

Given two probability distributions denoted as:

$$\begin{aligned}(A|p_A) &= \{(a_1|p_{a1}), (a_2|p_{a2}), ..., (a_d|p_{ad})\} \\ (B|p_B) &= \{(b_1|p_{b1}), (b_2|p_{b2}), ..., (b_d|p_{bd})\}\end{aligned} \tag{2}$$

where a_i or b_i is the i-th observed value in the respective distributions and p is its corresponding probability value. Let $M = \{m_{ij}\}$ be the cost matrix, where typically $m_{ij} = (a_i - b_j)^2$, then the OT distance can be defined as:

$$d_M(A, B) := \min_{P \in U(A,B)} \sum_{i,j=1}^{d} p_{ij} m_{ij} \tag{3}$$

$$U(A, B) := \{P \in \mathbb{R}_+^{d \times d} \mid P1_d = p_A, P^T 1_d = p_B\} \tag{4}$$

where 1_d is the vector of ones with a dimension of d, and $U(A, B)$ contains all possible joint probabilities of A and B, whose row and column sums to p_A and p_B respectively. Fig. 3 illustrates an example of OT. The optimal transport plan $P^\star$, is thus defined as:

$$P^\star := \arg\min_{P \in U(A,B)} \sum_{i,j=1}^{d} p_{ij} m_{ij} \tag{5}$$

OT is a special case of the linear programming (LP) problem whose worst case time complexity is $O(d^3 logd)$ [46]. Although the acceleration of OT has attracted a large amount of research interest and various algorithms have been proposed, the computational overhead remains high (super-cubic). Recently a variant of OT, Sinkhorn distance [38], successfully makes OT several orders of magnitude faster, which greatly enhances the utility of OT. Sinkhorn distance adds an entropic regularization to the classic OT formula and then enforces a simple structure on the optimal transport matrix, as defined by the following equation:

$$d_M^\lambda(A, B) := \min_{P \in U(A,B)} \left[\sum_{i,j=1}^{d} p_{ij} m_{ij} + \frac{1}{\lambda} \sum_{i,j=1}^{d} p_{ij} \log p_{ij} \right] \tag{6}$$

where λ is the regularization coefficient. As λ increases Sinkhorn distance converges to the classic OT distance. When M itself is a metric matrix, Sinkhorn distance satisfies all the three distance axioms, including the symmetry and the triangle inequality. As its name implies, Sinkhorn distance can be computed by SinkhornâĂŹs fixed point iteration [47], [48]. To converge to the optimal transport matrix P^λ, one only needs to iterate SinkhornâĂŹs update a sufficient number of times. The algorithm can be even faster with GPUs because it supports 1-vs-N mode where the distances against multiple targets are computed simultaneously. The 1-vs-N mode turns vector-by-matrix operations into matrix-by-matrix operations, which leads to a higher computing density and makes the algorithm more suitable for parallel computing. A flaw of Sinkhorn distance is that sometimes it will be constrained by machine-precision limit when λ increases beyond a problem-dependent value λ_{max} beyond which some elements of $e^{-\lambda M}$ are represented as zeroes.

III. TIME ADAPTIVE OPTIMAL TRANSPORT (TAOT)

The motivation of TAOT is to migrate OT from probability distributions (one-dimensional) to time series data (two-dimensional), which means except for the difference in observed values, TAOT also need to capture the difference in time coordinates. An intuitive solution is to consider both observed values and their corresponding time coordinates when calculating the cost between data points in respective time series. In this setting, the one dimensional data point, for example a_i, is extended by another time dimension, (a_i, t_i), and then a straightforward cost matrix can be defined as $M(i, j) = (a_i - b_j)^2 + w * (t_i - t_j)^2$, where w is a weight parameter to balance the two parts.

Originally the OT algorithm copes with probability distributions and each input value must have its corresponding probability. Here we assume all observed values of a time series share an equal probability based on the fact that these values come from independent observations. As a result, the input time series can be denoted by:

$$\begin{aligned}(A|p_A) &= \{((a_1, t_1)|\frac{1}{d}), ((a_2, t_2)|\frac{1}{d}), ..., ((a_d, t_d)|\frac{1}{d})\} \\ (B|p_B) &= \{((b_1, t_1)|\frac{1}{d}), ((b_2, t_2)|\frac{1}{d}), ..., ((b_d, t_d)|\frac{1}{d})\}\end{aligned} \tag{7}$$

Since each observed value shares an equal probability, the Sinkhorn iteration [38] can be further simplified. Algorithm 1 summarizes the simplified TAOT algorithm. Basically it contains two phases, where it first prepares the cost matrix and then iterates the simplified Sinkhorn update until the stopping criterion or the maximum iteration number is reached. Note that the time coordinates are normalized into z-score before calculating the cost matrix M. We can get not only the TAOT distance from the algorithm, but also the optimal transport plan, which is significant to the visualization of the results and the execution of other analysises.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3016529, IEEE Access

Algorithm 1 Computation of Time Adaptive Optimal Transport

Input: $A, B, \lambda, w, maxIter = 5000, tolerance = 0.005$
Output: $distance, plan$

$d = |A|$
$t = zscore(linspace(1, d, d))$
for $i = 1 : |A|$ **do**
  for $j = 1 : |B|$ **do**
    $M(i, j) = (a_i - b_j)^2 + w * (t_i - t_j)^2$
  end for
end for
$M = M/median(M(:))$
$K = exp(-\lambda * M)$
$curIter = 0$
$u = ones(d, 1)/d$
while $curIter < maxIter$ **do**
  $v = 1./(d * K' * u)$
  $u = 1./(d * K * v)$
  $curIter = curIter + 1$
  if $mod(curIter, 20) == 1$ **then**
    $criterion = sum(abs(v. * (K' * u) - 1/d))$
    if $criterion < tolerance$ **then**
      $break$
    end if
  end if
end while
$distance = sum(u. * ((K. * M) * v))$
$plan = bsxfun(@times, v', (bsxfun(@times, u, K)))$

The general idea of TAOT is to integrate temporal dissimilarity between time series to the final distance measure. For Sinkhorn distance based algorithm, the temporal dissimilarity is usually contained in how we construct the cost matrix, and we can observe from Algorithm 1 that any kind of cost matrix can easily fit into this algorithm framework and lead to a new distance measure. In this paper we just employ the most straightforward way to construct our cost matrix. Other definitions of cost matrix is still an open direction for future work.

As for the computational efficiency, given two time series of length N, the time complexity of Euclidean distance is $O(N)$. The time complexity of DTW-based methods, for example, DTW, derivative DTW, or weighted DTW, is $O(N^2)$. The time complexity of TAOT is relatively tricky because it is not as straightforward as the Euclidean distance or most DTW-based methods. TAOT involves a stopping criterion for its iteration and the criterion is calculated based on the current state of the iteration. According to a study on the complexity of multimarginal optimal transport [49], the time complexity of TAOT is approximately $O(N^2 \log N)$. Note that TAOT is based on Sinkhorn distance [38], which is a faster version of optimal transport. The naive optimal transport is a computational heavy method, whose cost of computing scales at least in $O(N^3 \log N)$.

IV. EXPERIMENT

A. 1NN ACCURACY REPORT

In this section, we evaluate the classification accuracy of TAOT on 63 datasets of the UCR time series classification archive [50], which has the largest and most widely used collection of public benchmark datasets of time series. These datasets include both real-world time series and synthetic time series from various application domains. The length of time series varies between 60 (Synthetic Control) and 1639 (CinC_ECG_torso), and the number of classes varies between 2 (Gun-Point) and 60 (ShapesAll).

In order to ensure fair comparisons between different similarity measures, we employ the one nearest neighbor classifier (1NN), because the classifier itself does not involve any parameter, and thus the accuracy depends only on the similarity measure. Six most well-established methods, Euclidean distance, DTW, DTW with Sakoe-Chiba constraint, DDTW, WDTW, and FDTW are selected for comparison.

Table 1 summarizes the 1NN classification error rate of TAOT and other competing methods on all datasets. If any parameter is involved, for instance, the width of Sakoe-Chiba band, we choose the optimal one found by cross validation on the training set. From these error rates we can observe that in general TAOT outperforms all the other measures on 28 datasets (Synthetic Control, Swedish Leaf, 50Words, ECG200, Adiac, etc), and achieves leading accuracy together with one or more other measures on another 7 datasets (Wafer, Face(four), Lightning-7, Plane, Coffee, ECGFiveDays, and BirdChicken).

TAOT encounters the aforementioned machine precision limit problem on four datasets (Face(all), Lighting2, Fish, LargeKitchenAppliances) of the original UCR time series classification archive, and therefore there is no accuracy report on these datasets. This is harmless for the classification task since we can find the problem during the early training phase and then change for another alternative method.

Fig. 4(a-b-c-d-e-f) shows the pairwise comparisons between TAOT and the six competing methods, respectively. In Fig. 4, each dot represents a dataset, whose x-coordinate and y-coordinate are respectively the accuracy generated by the competing method and TAOT. In this setting, a dot falling above, on, or below the diagonal indicates that the proposed TAOT outperforms, ties with, or lose to the competing method. The numbers of dots in each different region are labeled on these figures. We can observe that TAOT achieves a better performance in all of the six pairwise competitions. To demonstrate whether TAOT is statistically significant different from other methods, the p-value generated from Friedman test is shown the in lower-right corner of each figure. The p-value ranges from 0 to 1, and a smaller p-value indicates a more significant difference between methods. Note that all these p-values are smaller than 0.05 and it proves that TAOT performs significantly better than the six competing methods.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3016529, IEEE Access

TABLE 1. Comparisons on 1NN Error Rate

Dataset	Euclidean distance	DTW	DTW ($r\%$)[a]	DDTW	WDTW (g)[b]	FDTW	TAOT (λ and w)[c]
Synthetic Control	0.120	0.007	0.017 (6)	0.433	0.010 (0.30)	0.007	**0.003 (195 4)**
Gun-Point	0.087	0.093	0.087 (0)	**0.007**	0.027 (0.20)	0.020	0.027 (50 0.3)
CBF	0.148	0.003	0.004 (11)	0.408	0.009 (0.08)	**0.001**	0.006 (145 1)
OSU Leaf	0.479	0.409	0.388 (7)	**0.120**	0.479 (0.60)	0.409	0.450 (160 3)
Swedish Leaf	0.211	0.208	0.154 (2)	0.115	0.173 (0.03)	0.208	**0.107 (175 0.9)**
50Words	0.369	0.310	0.242 (6)	0.308	0.253 (0.10)	0.268	**0.233 (150 2)**
Trace	0.240	**0.000**	0.010 (3)	**0.000**	**0.000 (0.01)**	**0.000**	0.020 (200 0.3)
Two Patterns	0.090	**0.000**	0.002 (4)	0.003	**0.000 (0.01)**	**0.000**	0.010 (5 6)
Wafer	0.005	0.020	0.005 (1)	0.022	**0.003 (0.30)**	0.008	**0.003 (70 8)**
Face(four)	0.216	0.170	0.114 (2)	0.375	0.125 (0.10)	**0.102**	**0.102 (40 5)**
Lightning-7	0.425	**0.274**	0.288 (5)	0.425	**0.274 (0.10)**	0.301	**0.274 (10 0.9)**
ECG200	0.120	0.230	0.120 (0)	0.170	0.130 (0.50)	0.180	**0.110 (130 3)**
Adiac	0.389	0.396	0.391 (3)	0.414	0.366 (0.10)	0.414	**0.289 (185 0.1)**
Yoga	0.170	0.164	0.155 (2)	0.180	**0.153 (0.10)**	0.154	0.159 (75 0.2)
Plane	0.038	**0.000**	**0.000 (6)**	**0.000**	**0.000 (0.01)**	**0.000**	**0.000 (125 0.5)**
Car	0.267	0.267	0.233 (1)	0.267	**0.217 (0.19)**	0.367	0.283 (160 0.8)
Beef	0.333	0.367	0.333 (0)	0.333	**0.300 (0.20)**	0.367	0.333 (100 6)
Coffee	**0.000**	**0.000**	**0.000 (0)**	0.071	**0.000 (0.01)**	**0.000**	**0.000 (85 2)**
OliveOil	0.133	0.167	0.133 (0)	0.133	0.167 (0.01)	**0.100**	0.133 (200 0.6)
CinC_ECG_torso	0.103	0.349	**0.070 (1)**	0.289	0.075 (0.08)	0.317	0.084 (195 10)
DiatomSizeReduction	0.065	0.033	0.065 (0)	0.065	0.036 (0.10)	0.033	**0.020 (10 0.2)**
ECGFiveDays	0.203	0.232	0.203 (0)	0.314	0.138 (0.60)	**0.117**	**0.117 (20 5)**
FacesUCR	0.231	0.095	0.088 (12)	0.157	0.078 (0.03)	0.095	**0.067 (30 3)**
ItalyPowerDemand	0.045	0.050	0.045 (0)	0.086	0.043 (0.10)	**0.033**	0.038 (15 7)
MedicalImages	0.316	0.263	**0.253 (20)**	0.337	0.263 (0.08)	0.280	0.296 (40 4)
MoteStrain	0.121	0.165	0.134 (1)	0.284	0.142 (0.30)	0.165	**0.107 (20 1)**
SonyAIBORobotSurface	0.305	0.275	0.305 (0)	0.270	0.255 (0.50)	0.304	**0.166 (95 2)**
SonyAIBORobotSurfaceII	0.141	0.169	0.141 (0)	0.142	0.154 (0.05)	0.178	**0.134 (105 10)**
Symbols	0.101	0.050	0.062 (8)	**0.029**	0.049 (0.03)	0.060	0.062 (115 0.8)
TwoLeadECG	0.253	0.096	0.132 (4)	**0.005**	0.111 (0.05)	0.112	0.070 (35 0.1)
Cricket_X	0.423	0.246	0.228 (10)	0.369	**0.210 (0.03)**	0.269	0.267 (30 3)
Cricket_Y	0.433	0.256	0.241 (17)	0.441	**0.238 (0.01)**	0.244	0.277 (145 3)
Cricket_Z	0.413	0.246	0.254 (5)	0.444	0.246 (0.05)	**0.233**	0.287 (105 5)
InsectWingbeatSound	0.438	0.645	**0.415 (1)**	0.757	0.431 (0.20)	0.591	0.429 (75 10)
ArrowHead	0.200	0.297	0.200 (0)	0.223	0.183 (0.50)	0.223	**0.177 (150 3)**
BeetleFly	0.250	0.300	0.300 (7)	0.250	0.300 (0.01)	0.350	**0.100 (10 0.3)**
BirdChicken	0.450	0.250	0.300 (6)	0.150	0.250 (0.01)	0.300	**0.150 (45 0.1)**
Ham	0.400	0.533	0.400 (0)	0.524	0.429 (0.20)	0.432	**0.371 (40 0.7)**
Herring	0.484	0.469	0.469 (5)	0.500	0.453 (0.01)	0.562	**0.297 (60 0.2)**
ProximalPhalanxOAG	0.215	0.195	0.215 (0)	**0.180**	0.195 (0.01)	0.195	0.185 (15 0.1)
ProximalPhalanxOC	0.192	0.217	0.210 (1)	**0.182**	0.213 (0.10)	0.216	0.196 (40 0.7)
ProximalPhalanxTW	0.293	0.244	0.244 (2)	0.270	0.260 (0.03)	0.288	**0.215 (15 0.7)**
ToeSegmentation1	0.320	0.228	0.250 (8)	0.215	0.219 (0.01)	0.276	**0.171 (35 0.1)**
ToeSegmentation2	0.192	0.162	0.092 (5)	0.315	0.115 (0.03)	0.154	**0.077 (5 0.8)**
DistalPhalanxOAG	0.374	0.230	0.374 (0)	0.240	0.225 (0.40)	0.223	**0.185 (5 1)**
DistalPhalanxOC	0.283	0.283	0.275 (1)	0.220	0.237 (0.03)	0.238	**0.213 (45 0.4)**
DistalPhalanxTW	0.367	0.410	0.367 (0)	0.273	0.268 (0.30)	0.278	**0.245 (5 0.5)**
Earthquakes	0.288	0.281	0.273 (6)	0.276	0.292 (0.03)	0.276	**0.174 (5 7)**
MiddlePhalanxOAG	0.481	0.500	0.481 (0)	0.255	0.260 (0.05)	0.283	**0.238 (60 0.2)**
MiddlePhalanxOC	0.234	0.302	0.234 (0)	0.278	0.292 (0.30)	0.257	**0.227 (20 0.5)**
MiddlePhalanxTW	0.487	0.494	0.494 (3)	0.444	0.414 (0.08)	0.416	**0.381 (70 0.4)**
ShapeletSim	0.461	0.350	0.300 (3)	0.461	0.244 (0.03)	**0.122**	0.439 (60 2)
Wine	0.389	0.426	0.389 (0)	0.481	0.426 (0.10)	0.370	**0.296 (55 9)**

[a] r is the radius of best Sakoe-Chiba band, measured in percentages of time series length. Optimal values of r are published along with the datasets.

[b] g is a parameter of WDTW, which controls the level of penalization for the points with larger time gap. Optimal values of g are reported by the paper proposing WDTW [24]

[c] Optimal values of λ and w are extracted from cross validations, the details of which will be introduced in Section IV-B.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3016529, IEEE Access

TABLE 2. Comparisons on 1NN Error Rate (Continued)

Dataset	Euclidean distance	DTW	DTW ($r\%$)[a]	DDTW	WDTW (g)[b]	FDTW	TAOT (λ and w)[c]
WordsSynonyms	0.382	0.351	0.262 (9)	0.320	**0.249 (0.05)**	0.343	0.306 (55 3)
Computers	0.424	**0.300**	0.380 (12)	0.332	0.416 (0.01)	0.348	0.340 (15 0.6)
Meat	0.067	0.067	0.067 (0)	0.333	0.067 (0.10)	0.067	**0.017 (95 0.9)**
RefrigerationDevices	0.605	0.536	0.560 (8)	0.592	0.592 (0.03)	0.536	**0.469 (15 0.2)**
ScreenType	0.640	0.603	0.589 (17)	0.573	0.589 (0.01)	0.603	**0.565 (5 0.2)**
ShapesAll	0.248	0.232	0.198 (4)	**0.165**	0.192 (0.05)	0.232	0.193 (165 0.8)
SmallKitchenAppliances	0.659	0.357	**0.328 (15)**	0.347	0.347 (0.01)	0.333	0.336 (35 4)
Strawberry	**0.054**	0.059	**0.054 (0)**	0.059	0.062 (0.40)	**0.054**	0.060 (95 0.2)
Worms	0.545	**0.416**	0.468 (9)	0.497	0.552 (0.03)	0.514	0.575 (35 1)
WormsTwoClass	0.390	0.377	0.416 (7)	**0.298**	0.376 (0.05)	0.354	0.359 (25 8)

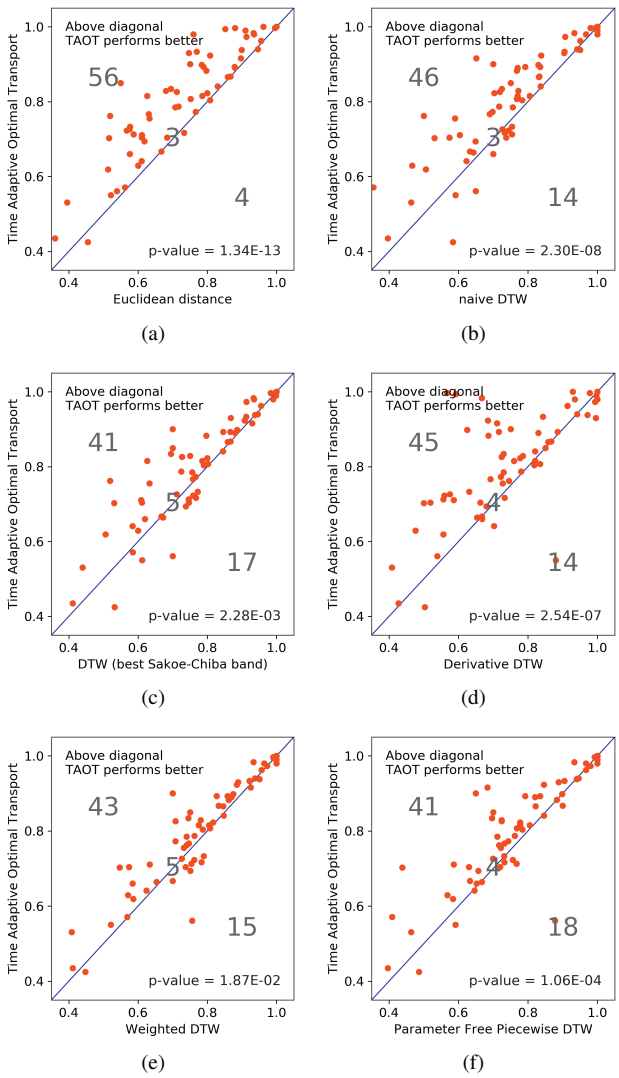


FIGURE 4. 1NN accuracy of TAOT versus the six competitors respectively.

B. EXTRACTING RELIABLE PARAMETERS FOR TAOT

TAOT involves two parameters: the regularization coefficient λ and the time weight w. To find a reliable combination of λ and w, we employ grid-search based leave-one-out cross validation on the training set. We choose leave-one-out cross validation over hold-out or k-fold validation because sometimes the training set is too small to further divide. Typically, grid-search is considered to be time-consuming, but for TAOT the search spaces of λ and w are both small and thus the efficiency is acceptable.

For w, if it is too large, for example $w > 10$, the difference in time coordinates will predominate. Contrarily, if w is too small, for example $w < 0.1$, the difference in observed values will predominate. Either way we lose the balance. Therefore we set the search space of w to a moderate range, $[0.1, 10]$. Given a uniform step size to the integer and the fraction part respectively, our candidates for w are $0.1, 0.2, ..., 0.9$ and $1, 2, ..., 10$.

For λ, recall that as it increases, Sinkhorn distance converges to the classic OT distance. Empirically, 200 is large enough to make Sinkhorn distance close to the classic OT distance, and 5 is strict enough for regularizing. So we set the search space of λ to $[5, 200]$ with a fixed step size of 5.

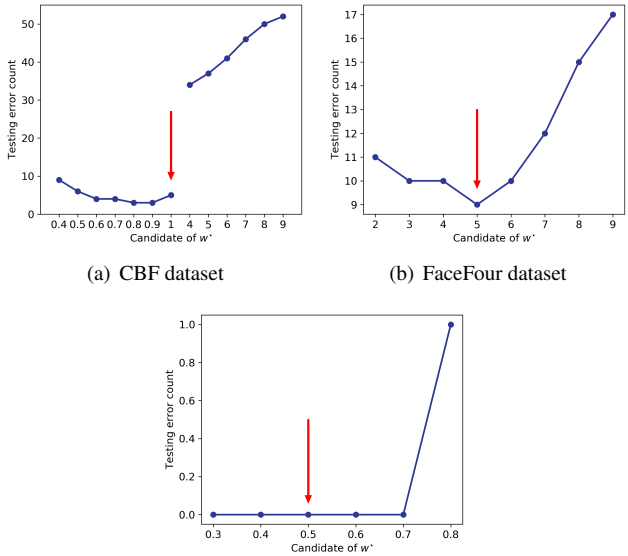
During grid-search, each pair of λ and w results in an error count, $error(\lambda, w)$. There can often be multiple pairs of parameters that all achieve the lowest error count, and these optimal parameters on the training set may not perform as well on the testing set. Therefore we develop another strategy to decide the sole and most adaptable pair of parameters. For λ, the most adaptable one should achieve the lowest cumulative error count given every different w, as defined by the following Eq. 8. If there are ties, we choose the smallest λ because it leads to a faster computation.

$$\lambda^\star = \min\left\{\arg\min_{\lambda} \sum_{n=1}^{N} error(\lambda, w_n)\right\} \tag{8}$$

After $\lambda^\star$ is fixed, the best w should be the one to have the lowest error count when pairing with the $\lambda^\star$, as defined by the following Eq. 9. If there are ties, we choose the median. Fig. 5 shows the real testing error counts of different $w^\star$

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3016529, IEEE Access



(a) CBF dataset (b) FaceFour dataset

(c) Plane dataset

FIGURE 5. Performance of different candidates of $w^\star$ on testing sets.

candidates that all achieve the lowest training error count. We can observe that the medians of those $w^\star$ candidates achieve the most stable performance in general on the three datasets.

$$w^\star = median\left\{\underset{w\in\{w_1,\ldots,w_N\}}{\arg\min} \quad error(\lambda^\star, w)\right\} \tag{9}$$

C. THE TEXAS SHARPSHOOTER FALLACY

To further demonstrate the capability of TAOT, we will test whether TAOT is capable of overcoming the Texas Sharpshooter Fallacy. The Texas Sharpshooter Fallacy is a common logic error that happens when comparing methods on multiple datasets. A pervasive scenario is that as long as a method can win on some datasets, the author then claim the method is valuable. Because the author think since the method can win on the datasets from some domains, it could be useful in those domains. However, it is not enough to have a method that can be more accurate on some datasets unless you can tell ahead of time that on which datasets it will be more accurate [51].

One way to show whether a method can tell in advance that it can be more accurate on a certain dataset is to test whether the expected accuracy gain over another competing method coincides with the actual accuracy gain. The expected accuracy gain and the actual accuracy gain are defined by Eq. 10 and Eq. 11, respectively. In our setting, the expected accuracy gain is based on the best training accuracy during leave-one-out cross validation, and the actual accuracy gain is based on the best testing accuracy. An expected accuracy gain larger than one indicates that we predict TAOT will perform better, while an actual accuracy gain larger than one indicates that TAOT indeed performs better.

$$expected\ gain = \frac{training\ accuracy(TAOT)}{training\ accuracy(competing\ method)} \tag{10}$$

$$actual\ gain = \frac{testing\ accuracy(TAOT)}{testing\ accuracy(competing\ method)} \tag{11}$$

As shown in Fig. 6, Texas sharpshooter plot is a convenient tool to visualize the comparison between the expected accuracy gain and the actual accuracy gain on multiple datasets. Each point represents a dataset to test and each dataset falls into one of the following four possibilities:

- TP(True Positive): In this region we predicted TAOT would increase accuracy, and TAOT did. Obviously, this is the most beneficial situation for TAOT and the majority of points fall into this region.
- TN(True Negative): In this region we predicted TAOT would decrease accuracy, and TAOT did. This is not a bad case. Because if we know ahead of time that TAOT will do worse, we can choose another method to avoid the loss of accuracy.
- FN(False Negative): In this region we predicted TAOT would decrease accuracy, but the accuracy actually increased. This is also not a bad case. We might miss the opportunity to improve, but we will not do worse.
- FP(False Positive): In this region we predicted TAOT would increase accuracy, but the accuracy actually decreased. This is the truly bad case. But not many points fall into this region.

D. THE PAIRING OF MAXIMUM VALUES

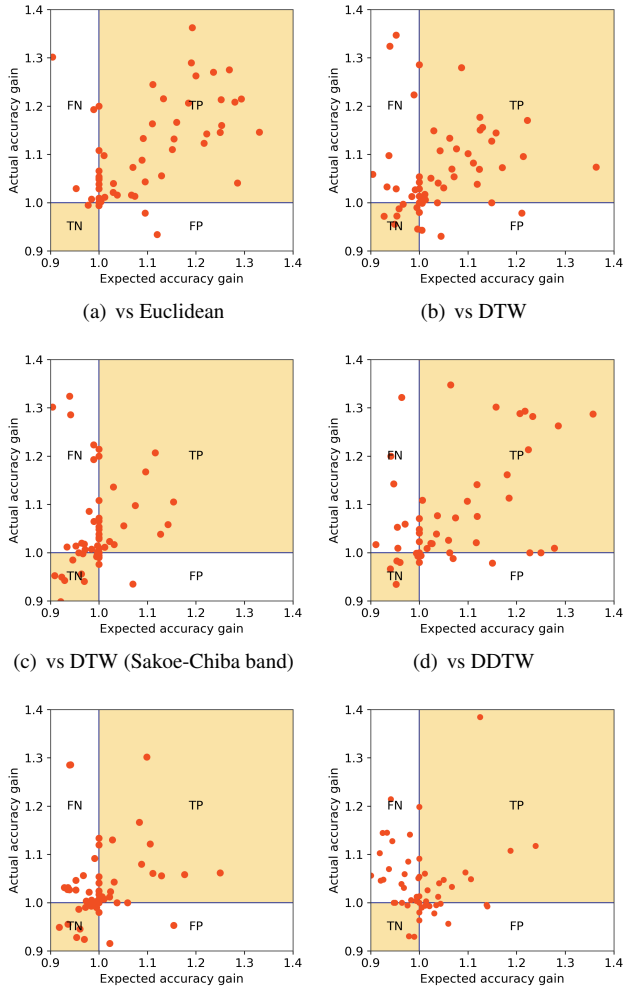
One motivation of TAOT is that we want global maximum values to pair with each other if such a pairing is critical for the current problem. Classic OT, which is a special case of TAOT when $w = 0$, can guarantee the pairing of global maximum values in theory. As for TAOT, if this pairing is decisive, then normally we will get a relatively small w from the training phase, and that will lead to a high possibility of generating a maximum-to-maximum alignment. Fig. 7 gives an example to visually demonstrate that TAOT is more likely to generate the required maximum-to-maximum alignment than the other currently established methods.

E. VISUALIZATION OF ALIGNMENTS GENERATED BY TAOT

The many-to-many alignment generated by TAOT, namely the transport plan, is usually represented by a weight matrix. As the most established and popular similarity measure of time series, DTW visualizes its output alignment with a warping path in a warping matrix, as shown in the previous Fig. 2. In order to keep consistency with DTW, the visualization of TAOT is still based on a warping matrix, but with a fuzzy warping path or a heat map of the warping path. Fig. 8 illustrates the fuzzy version of a warping path where each

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3016529, IEEE Access

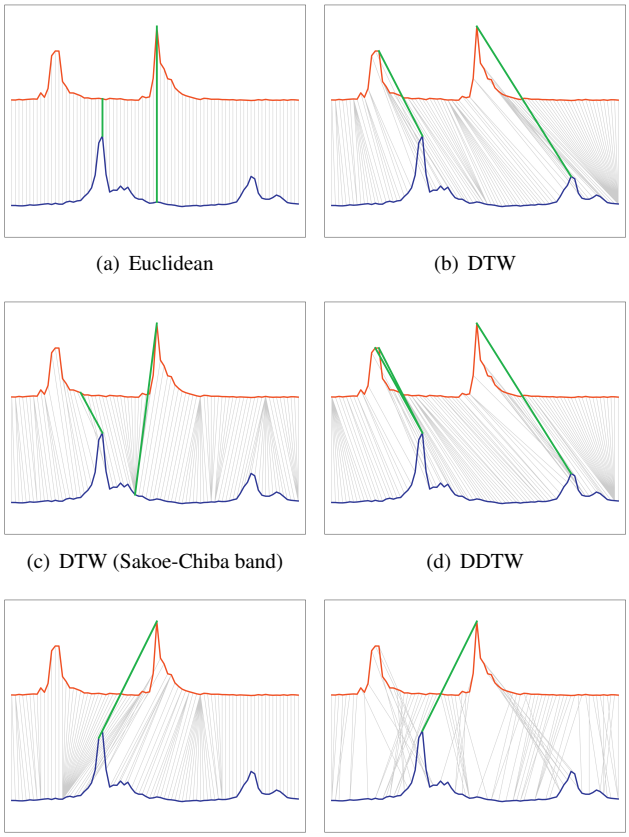


(a) vs Euclidean (b) vs DTW

(c) vs DTW (Sakoe-Chiba band) (d) vs DDTW

(e) vs WDTW (f) vs FDTW

FIGURE 6. Texas sharpshooter plot of TAOT versus the other six competing methods respectively.



(a) Euclidean (b) DTW

(c) DTW (Sakoe-Chiba band) (d) DDTW

(e) WDTW (f) TAOT

FIGURE 7. An example where only TAOT generates the required maximum-to-maximum alignment while the other methods fail.

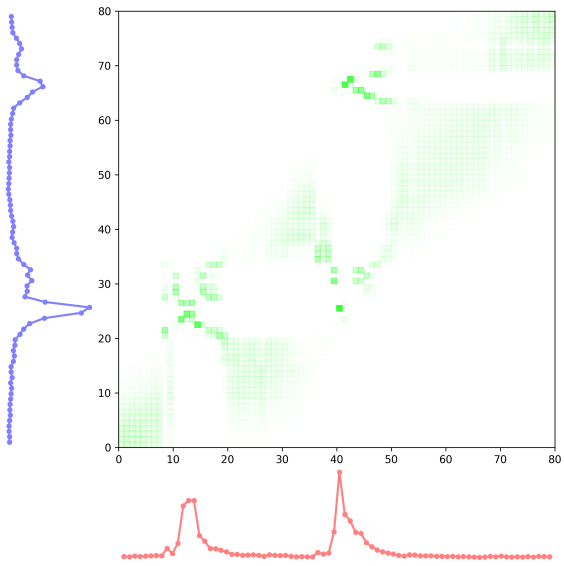


FIGURE 8. A fuzzy warping path generated by TAOT.

matrix cell corresponds to a pairing between the two points respectively indicated by the x-coordinate and y-coordinate of the cell, and we use the transparency of a matrix cell to represent the weight of the pairing. In this setting, a more conspicuous area indicates more intense similarity and vice versa. Note that although we use an alignment path matrix here, TAOT does not have a real alignment path like other DTW-based methods. TAOT is based on optimal transport and the values of all transport matrix cells are updated simultaneously during the calculation instead of a step-by-step motion.

V. CONCLUSION

In this paper, we proposed a new time series similarity measure framework entitled Time Adaptive Optimal Transport (TAOT). It inherits the advantages of several promising properties from the optimal transport (OT) that can greatly alleviate some long-suffered issues with currently established time series similarity measures. To make OT capable of handling time series data, TAOT takes into consideration not only the difference in observed values but also the differ-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3016529, IEEE Access

ence in time coordinates. To make the efficiency of TAOT acceptable, the calculation of TAOT is based on Sinkhorn distance, a very fast variant of OT. TAOT further simplifies the Sinkhorn iteration by assuming that all observed values of a time series share a equal probability. The performance of TAOT was demonstrated by a series of one nearest neighbor classification experiments on multiple public datasets. Compared with other currently established methods, TAOT exhibits a better accuracy and it even has the ability to predict whether there will be an accuracy improvement on the majority of experimental datasets. We also introduced a grid-search based parameter extracting strategy and a fuzzy warping path based visualization method for TAOT. TAOT is not just a single distance measure, but also an algorithm framework where other kinds of cost matrix can easily fit into and each creates a new distance measure. We believe this may lead to some potential future work.

REFERENCES

[1] Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti, "Similarity-based classification: Concepts and algorithms," Journal of Machine Learning Research, vol. 10, no. Mar, pp. 747–776, 2009.

[2] S. Lhermitte, J. Verbesselt, W. Verstraeten, and P. Coppin, "A comparison of time series similarity measures for classification and change detection of ecosystem dynamics," Remote Sensing of Environment, vol. 115, no. 12, pp. 3129–3152, 2011.

[3] Y.-S. Lin, J.-Y. Jiang, and S.-J. Lee, "A similarity measure for text classification and clustering," IEEE transactions on knowledge and data engineering, vol. 26, no. 7, pp. 1575–1590, 2014.

[4] P. Wan, Y. Zhan, and W. Jiang, "Study on the satellite telemetry data classification based on self-learning," IEEE Access, vol. 8, pp. 2656–2669, 2020.

[5] T. W. Liao, "Clustering of time series dataąła survey," Pattern recognition, vol. 38, no. 11, pp. 1857–1874, 2005.

[6] F. Petitjean, J. Inglada, and P. Gançarski, "Satellite image time series analysis under time warping," IEEE transactions on geoscience and remote sensing, vol. 50, no. 8, pp. 3081–3095, 2012.

[7] D. Li, Y. Zhao, and Y. Li, "Time-series representation and clustering approaches for sharing bike usage mining," IEEE Access, vol. 7, pp. 177 856–177 863, 2019.

[8] I. Bartolini, P. Ciaccia, and M. Patella, "Warp: Accurate retrieval of shapes using phase of fourier descriptors and time warping distance," IEEE transactions on pattern analysis and machine intelligence, vol. 27, no. 1, pp. 142–147, 2005.

[9] E. Keogh and C. A. Ratanamahatana, "Exact indexing of dynamic time warping," Knowledge and information systems, vol. 7, no. 3, pp. 358–386, 2005.

[10] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh, "Searching and mining trillions of time series subsequences under dynamic time warping," in Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2012, pp. 262–270.

[11] Z. Li, J. Guo, H. Li, T. Wu, S. Mao, and F. Nie, "Speed up similarity search of time series under dynamic time warping," IEEE Access, vol. 7, pp. 163 644–163 653, 2019.

[12] X. Wang, A. Mueen, H. Ding, G. Trajcevski, P. Scheuermann, and E. J. Keogh, "Experimental comparison of representation methods and distance measures for time series data," Data Mining and Knowledge Discovery, vol. 26, no. 2, pp. 275–309, 2013.

[13] H. Pree, B. Herwig, T. Gruber, B. Sick, K. David, and P. Lukowicz, "On general purpose time series similarity measures and their use as kernel functions in support vector machines," Information Sciences, vol. 281, pp. 478–495, 2014.

[14] J. Serri and J. L. Arcos, "An empirical evaluation of similarity measures for time series classification," Knowledge Based Systems, vol. 67, pp. 305–314, 2014.

[15] D. Folgado, M. Barandas, R. Matias, R. S. Martins, M. F. Carvalho, and H. Gamboa, "Time alignment measurement for time series," Pattern Recognition, vol. 81, pp. 268–279, 2018.

[16] N. S. Savaş, F. Bakkal, S. Eken, and A. Sayar, "Evaluation of different algorithms for measuring the similarities of trajectory datasets," in 2017 25th Signal Processing and Communications Applications Conference (SIU), 2017, pp. 1–4.

[17] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," IEEE transactions on acoustics, speech, and signal processing, vol. 26, no. 1, pp. 43–49, 1978.

[18] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, Discrete-time signal processing (2nd ed.), 1999.

[19] Y. leh Wu, D. Agrawal, and A. E. Abbadi, "A comparison of dft and dwt based similarity search in time-series databases," in In Proceedings of the 9 th International Conference on Information and Knowledge Management, 2000, pp. 488–495.

[20] Z. Zhang, L. Tang, and P. Tang, "Local feature based dynamic time warping," in IEEE International Conference on Data Science and Advanced Analytics (DSAA), 2014. IEEE, 2014, pp. 425–429.

[21] T.-c. Fu, "A review on time series data mining," Engineering Applications of Artificial Intelligence, vol. 24, no. 1, pp. 164–181, 2011.

[22] H. Li and C. Wang, "Similarity measure based on incremental warping window for time series data mining," IEEE Access, vol. 7, pp. 3909–3917, 2019.

[23] E. J. Keogh and M. J. Pazzani, "Derivative dynamic time warping," in Proceedings of the 2001 SIAM International Conference on Data Mining. SIAM, 2001, pp. 1–11.

[24] Y.-S. Jeong, M. K. Jeong, and O. A. Omitaomu, "Weighted dynamic time warping for time series classification," Pattern Recognition, vol. 44, no. 9, pp. 2231–2240, 2011.

[25] Z. Zhang, P. Tang, and R. Duan, "Dynamic time warping under pointwise shape context," Information Sciences, vol. 315, pp. 88–101, 2015.

[26] V. S. Siyou Fotso, E. Mephu Nguifo, and P. Vaslin, "Parameter Free Piecewise Dynamic Time Warping for time series classification," Dec. 2016, working paper or preprint. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01408944

[27] Z. Zhang, R. Tavenard, A. Bailly, X. Tang, P. Tang, and T. Corpetti, "Dynamic time warping under limited warping path length," Information Sciences, vol. 393, pp. 91–107, 2017.

[28] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh, "Querying and mining of time series data: experimental comparison of representations and distance measures," Proceedings of the VLDB Endowment, vol. 1, no. 2, pp. 1542–1552, 2008.

[29] F. Itakura, "Minimum prediction residual principle applied to speech recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 23, no. 1, pp. 67–72, 1975.

[30] L. Rabiner, A. Rosenberg, and S. Levinson, "Considerations in dynamic time warping algorithms for discrete word recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, no. 6, pp. 575–582, 1978.

[31] C. Myers, L. Rabiner, and A. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 6, pp. 623–635, 1980.

[32] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series." in KDD workshop, vol. 10, no. 16. Seattle, WA, 1994, pp. 359–370.

[33] C. Villani, Optimal transport: old and new. Springer Science & Business Media, 2008, vol. 338.

[34] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in Computer Vision, 1998. Sixth International Conference on. IEEE, 1998, pp. 59–66.

[35] H. Ling and K. Okada, "An efficient earth mover's distance algorithm for robust histogram comparison," IEEE transactions on pattern analysis and machine intelligence, vol. 29, no. 5, pp. 840–853, 2007.

[36] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," International journal of computer vision, vol. 40, no. 2, pp. 99–121, 2000.

[37] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy, "Optimal transport for domain adaptation," IEEE transactions on pattern analysis and machine intelligence, vol. 39, no. 9, pp. 1853–1865, 2017.

[38] M. Cuturi, "Sinkhorn distances: Lightspeed computation of optimal transport," in Advances in neural information processing systems, 2013, pp. 2292–2300.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3016529, IEEE Access

[39] D. Yu, X. Yu, Q. Hu, J. Liu, and A. Wu, "Dynamic time warping constraint learning for large margin nearest neighbor classification," Information Sciences, vol. 181, no. 13, pp. 2787–2796, 2011.

[40] K. S. Candan, R. Rossini, X. Wang, and M. L. Sapino, "sdtw: computing dtw distances using locally relevant constraints based on salient feature alignments," Proceedings of the VLDB Endowment, vol. 5, no. 11, pp. 1519–1530, 2012.

[41] S. Soheily-Khah and P.-F. Marteau, "Sparsification of the alignment path search space in dynamic time warping," Applied Soft Computing, vol. 78, pp. 630–640, 2019.

[42] E. J. Keogh and M. J. Pazzani, "Scaling up dynamic time warping for datamining applications," in Proceedings of the sixth ACM SIGKDD international conference on Knowledge Discovery and Data Mining,, 2000, pp. 285–289.

[43] Q. Cai, L. Chen, and J. Sun, "Piecewise statistic approximation based similarity measure for time series," Knowledge Based Systems, vol. 85, pp. 181–195, 2015.

[44] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde, "Optimal mass transport: Signal processing and machine-learning applications," IEEE Signal Processing Magazine, vol. 34, no. 4, pp. 43–59, 2017.

[45] Y. Robin, P. Yiou, and P. Naveau, "Detecting changes in forced climate attractors with wasserstein distance," Nonlinear Processes in Geophysics, vol. 24, no. 3, p. 393, 2017.

[46] O. Pele and M. Werman, "Fast and robust earth mover's distances," in Computer vision, 2009 IEEE 12th international conference on. IEEE, 2009, pp. 460–467.

[47] J. Franklin and J. Lorenz, "On the scaling of multidimensional matrices," Linear Algebra and its applications, vol. 114, pp. 717–735, 1989.

[48] P. A. Knight, "The sinkhorn–knopp algorithm: convergence and applications," SIAM Journal on Matrix Analysis and Applications, vol. 30, no. 1, pp. 261–275, 2008.

[49] T. Lin, N. Ho, M. Cuturi, and M. I. Jordan, "On the complexity of approximating multimarginal optimal transport." arXiv: Machine Learning, 2019.

[50] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista, "The UCR Time Series Classification Archive," July 2015, www.cs.ucr.edu/∼eamonn/time_series_data/.

[51] G. E. Batista, X. Wang, and E. J. Keogh, "A complexity-invariant distance measure for time series," in Proceedings of the 2011 SIAM International Conference on Data Mining, vol. 11. SIAM, 2011, pp. 699–710.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3016529, IEEE Access

ZHENG ZHANG received the B.S. degree in spatial information and digital technology from Wuhan University, in 2011 and the Ph.D. degree in signal and information processing from Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China, in 2016. He has also been a visiting scholar in LETG UMR CNRS 6554, University of Rennes 2, France, in 2019.

He is currently an Associate Professor with Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing, China. His research interests include remote sensing image processing, time series analytics, deep learning, and artificial intelligence.

PING TANG received the B.S. degree in mathematics from Ningxia University, Yinchuan, China, in 1986, and the M.S. and Ph.D. degrees in mathematics from Beijing Normal University, Beijing, China, in 1993 and 1996, respectively. In 1998, she was a Professor with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing, China.

Currently, she is serving as a Team Leader and leading a project of multispectral imagery radiometric and geometric correction for large volume of image data at global scale for higher resolution global land-cover mapping. She has significant experience in managing and designing software systems for satellite image processing and applications. Her research interests include the use of mathematical theories to development algorithms related to image processing and analysis.

THOMAS CORPETTI received an engineer degree in electrical engineering and a master degree in computer vision, in 1999. He obtained his Ph.D. and habilitation degree in 2002 and 2011, in the field of Computer Vision and Applied Mathematics with University Rennes I, France. After Ph.D. he spent one year as an assistant professor in Rennes and one year as a post-doc with IRSTEA (environment institute). He then obtained a permanent researcher position at CNRS in 2004 (French National Institute for Scientific Research) on the analysis of remote sensing image sequences for environmental applications. From 2009 to 2012, he has been with LIAMA, a sino-french lab in computer sciences, automatics and applied mathematics at Beijing, China, where he headed the TIPE group (Turbulence, Images, Physics and Environment).

He is now with OSUR (Observatory for Universe Sciences of Rennes, France) and LETG (Littoral, Environnement, Télédétection, Géomatique) UMR 6554 as a director of research, CNRS. His main research interests are concerned with the definition of computer vision tools for the analysis of remote sensing data (low and high resolution) for environmental applications.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/.