Context-aware Distance for Time Series

Zhihui Wang, Changlian Tan
School of Computer Science, Fudan University, Shanghai, China
Shanghai Key Laboratory of Data Science, Shanghai, China
zhhwang@fudan.edu.cn

Abstract

Dynamic Time Warping (DTW) is a widely used elastic distance measure for time series. It can warp the time axis to cope with local time shift, but it also causes singularities due to its warping ability. The singularities indicate pathological warping. One kind of approaches for limiting the singularities is to mechanically limit the warping ability without considering the values of time series, and these approaches are likely to miss a good warping path. Another kind of approaches is to use the derivative of the values of time series to limit the singularities, but the derivative is easily affected by noise, and the derivation is also de-informatized. This kind of approaches tends to miss a good warping path, too. The ideal situation is to achieve a balance between limiting the singularities and finding a good warping path. To this end, we propose Context-aware DTW (CDTW), which uses the context information of the current point in the time series, and can find the right warping path while limiting singularities. For illustrating that our idea can be easily applied to other elastic distances, we also introduce context information in MSM and propose Context-aware MSM (CMSM). The experiments on the UCR datasets show that our CDTW and CMSM can achieve better accuracy than the original DTW and MSM respectively, and demonstrate the effectiveness of our context-aware distance analysis approach for time series.

1 Introduction

Time series are one of the most common type of data in data science. In many cases, the elastic distance between two time series needs to be calculated for classification tasks, such as EE [Lines and Bagnall, 2015], COTE [Bagnall *et al.*, 2015], HIVE-COTE [Lines *et al.*, 2016], DTWF [Kate, 2016], etc, or for clustering tasks [Tim, 2015]. One category of elastic distances is DTW and its variants, such as DDTW [Keogh and Pazzani, 2001], WDTW [WDT, 2011], DD_{DTW} [Górecki and Łuczak, 2013], MVM [Latecki *et al.*, 2005], etc. Another category is edit distance based elastic distances, such as LCSS [HIRSCHBERG, 1977], EDR [Chen *et al.*, 2005], ERP [Chen and Ng, 2004], TWE [Marteau, 2009], MSM [Stefan *et al.*, 2013], etc.

DTW is a commonly used approach for distance measurement of the time series. It is also widely used in data mining [Dau *et al.*, 2018b; Holder *et al.*, 2024], gesture recognition [Ding and Chang, 2016], and speech processing [Yadav and Alam, 2018]. Since DTW can warp the time axis, it has good performance on time series. However, its strong warping ability is also accompanied by the problem of singularity. A point on one sequence that maps onto a large subsection of another sequence in the warping path found by DTW is defined as singularity [Hsu *et al.*, 2011; Keogh and Pazzani, 2001]. Singularities indicate pathological warping and are generated by the extreme warping of the time axis to explain the change of time series values [Keogh and Pazzani, 2001]. Too many singularities will seriously damage the alignment of two time series. One kind of approaches to limit singularities is to mechanically put some restrictions on the warping path, including limiting the path window [Chen *et al.*, 2012; Keogh and Pazzani, 2001], restricting the step pattern [Giorgino and others, 2009; Keogh and Ratanamahatana, 2005], setting slope weights [Giorgino and others, 2009], and punishing phase difference [WDT, 2011]. The defect of such approaches is not taking time series values into consideration, making it easy to miss a good warping path. Another kind of approaches is to calculate the derivative of time series and use it as a feature for DTW, such as DDTW [Keogh and Pazzani, 2001]. Such approaches do use the time series values, but due to the fact that derivative only represents the information from the current point and two points around, it is easily affected by noise, and thus the original value is lost in derivation. Therefore, this kind of approaches is difficult to find a good warping path, too.

To achieve a balance between limiting the singularities and finding a good warping path, we propose a novel approach, called Context-ware DTW (CDTW), which extracts information from the context of the current point as additional features, and calculates the point pair distance by combining the value and additional features of the current point. Although the additional features make the points more informative and more scattered, the points with similar contexts are still adjacent. In this way, our CDTW can inhibit the appearing of singularities. CDTW also retains the ability to find a good warping path, which is manifested in two aspects. First, CDTW keeps the warping ability so that the contextually similar segments of two time series can be aligned correctly. Second, CDTW is immune to noise. We also find that constructing additional features from context information not only works on the DTW, but also is suitable for all elastic distance based on point pair distances, including DDTW, WDTW, DD_{DTW}, EDR, ERP, LCSS, TWE, MSM, etc. In Section 4, we extend our work and apply the context information on MSM and then propose the CMSM.

The rest of the paper is organized as follows. Section 2 gives related works. Section 3 gives a generalized form of elastic distances. Section 4 proposes context elastic distances and describes CDTW's properties. Experimental results are presented and analyzed in Section 5. Section 6 concludes the paper.

2 Related Work

2.1 DTW-based Elastic Distance

Singularities were noted at least as early as 1978 [Sakoe and Chiba, 1978] and lots of research had been done to limit the warping ability of DTW. [Chen *et al.*, 2012] considered to set the window for paths, only paths in the window were allowed. [Zhang *et al.*, 2015] had experimented with different shaped windows. [Sakoe and Chiba, 1978] multiplied the steps parallel to the coordinate axis by a slope weight. [Keogh and Ratanamahatana, 2005] considered to restrict the step pattern. WDTW [WDT, 2011] calculated the distance of point pair multiplied by a weight that was positively related to the phase difference. DDTW [Keogh and Pazzani, 2001] used the derivative of time series values to limit singularities. DD_{DTW} [Górecki and Łuczak, 2013] proposed a weighted average of DTW distance and DDTW distance. In addition, MVM [Latecki *et al.*, 2005] proposed an approach that maps a subsequence of one sequence to the full sequence of another sequence. [Yurtman *et al.*, 2023] studied how to estimate DTW distance between time series with missing data.

2.2 Edit Distance-based Elastic Distance

LCSS [HIRSCHBERG, 1977] was applied to time series by defining how two real values are the same. Two real numbers were considered to be the same when the absolute value of their difference was less than a threshold. Similarly, Edit Distance could also set a threshold to measure the distance of time series, such as EDR [Chen *et al.*, 2005]. Neither DTW nor EDR was metric. That was, they did not meet the triangle inequality. ERP [Chen and Ng, 2004], MSM [Stefan *et al.*, 2013; Holznigenkemper *et al.*, 2023], TWE [Marteau, 2009] were metric distances. The point pair distance of EDR was set to 0 or 1, while the point pair distance of ERP was related to the value of points. MSM introduced Move, Split, and Merge operations similar to Change, Add, and Delete in EDR or ERP. TWE used a stiffness coefficient that penalized phase differences.

3 Preliminaries

In this paper, we assume that there are time series Q and O with lengths n and m, respectively. They are described as follows.

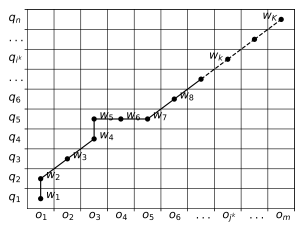


Figure 1: Warping path

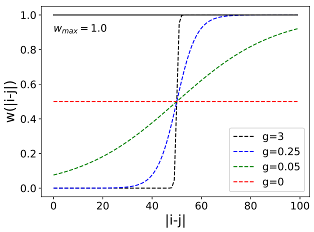


Figure 2: WDTW weight

$$\begin{aligned} Q &= q_1, q_2, q_3, ..., q_i, ..., q_n \\ O &= o_1, o_2, o_3, ..., o_j, ..., o_m \end{aligned} \tag{1}$$

To calculate the distance between two time series, a n-by-m distance matrix M is constructed. The (i, j) element of the matrix represents the distance between two points q_i and o_j, denoted by $\vec{d_{i,j}}$. The process of elastic distance calculating the distance between two time series is to choose a path W with the smallest cumulative distance on the matrix that satisfies the Boundary, Continuity, and Monotonicity conditions [WDT, 2011; Keogh and Pazzani, 2001]. This is a path starting at the point $(1, 1)$ and ending at point (n, m) of length K, shown in Figure 1. It can be described as the following equation.

$$\begin{aligned} W &= w_1, w_2, w_3, ..., w_k, ..., w_K \\ &= (q_1, o_1), ..., (q_{i^k}, o_{j^k}), ..., (q_n, o_m) \\ &\max(m, n) \leq K \leq m + n - 1 \end{aligned} \tag{2}$$

Here w_k corresponds to the point pair (q_{i^k}, o_{j^k}), (w_{k-1}, w_k) is called a step, and there are three step directions according to w_{k-1}. When $w_{k-1} = (q_i, o_{j-1})$, the direction is horizontal, denoted by $d^h_{i,j}$; when $w_{k-1} = (q_{i-1}, o_j)$, the direction is vertical, denoted by $d^v_{i,j}$; and when $w_{k-1} = (q_{i-1}, o_{j-1})$, the direction is diagonal, denoted by $d^d_{i,j}$. We use $d^{ud}_{i,j}$ for the non-directional case, and set $\vec{d_{i,j}} = [d^h_{i,j}, d^v_{i,j}, d^d_{i,j}, d^{ud}_{i,j}]$. There are the elastic distances in which the four components of $\vec{d_{i,j}}$ are different, such as EDR, ERP, TWE, and MSM. There are also the elastic distances in which the four components are the same, such as DTW-based elastic distance and LCSS. For convenience, the elastic distance (ELD) corresponding to the warping path is denoted as follows.

$$ELD(Q, O) = \sum_{k=1}^{K} \tilde{d}_{i^k, j^k} \tag{3}$$

Here K is the length of path W, and i^k, j^k are the subscript of point pairs (q_{i^k}, o_{j^k}) corresponding to path point w_k, and $\tilde{d}_{i^k, j^k}$ is one element of $[d^h_{i,j}, d^v_{i,j}, d^d_{i,j}, d^{ud}_{i,j}]$ according to the direction of (w_{k-1}, w_k). ELD can also be generalized to the following equation.

$$ELD_p(Q,O) = \sqrt[p]{\gamma(n,m)}$$

$$\gamma(i,j) = \begin{cases} d_{i,j}^{ud}, & if\ i = 1\ and\ j = 1; \\ \gamma(i-1,j) + d_{i,j}^{v}, & if\ n \geq i > 1\ and\ j = 1; \\ \gamma(i,j-1) + d_{i,j}^{h}, & if\ i = 1\ and\ m \geq j > 1; \\ \min\{\gamma(i-1,j-1) + d_{i,j}^{d}, & \\ \quad \gamma(i-1,j) + d_{i,j}^{v}, & \\ \quad \gamma(i,j-1) + d_{i,j}^{h}\}, & if\ n \geq i > 1 \\ & and\ m \geq j > 1. \end{cases} \tag{4}$$

Here q_i is the ith point of time series Q, o_j is jth point of O, n and m are the lengths of time series Q and O, respectively. The parameter p means that $\vec{d_{i,j}}$ uses L_p norm distance.

For DTW, the elements of $\vec{d_{i,j}}$ are the same, which can be denoted as follows.

$$d_{i,j}^{h} = d_{i,j}^{v} = d_{i,j}^{d} = d_{i,j}^{ud} = |q_i - o_j|^p \tag{5}$$

For DDTW, the elements of $\vec{d_{i,j}}$ are aslo the same, which can be denoted as follows.

$$d_{i,j}^{h} = d_{i,j}^{v} = d_{i,j}^{d} = d_{i,j}^{ud} = |q_i^{'} - o_j^{'}|^p \tag{6}$$

Here the calculation of first derivatives $q_i^{'}$ and $o_j^{'}$ can be found in [Keogh and Pazzani, 2001].

For WDTW, the elements of $\vec{d_{i,j}}$ are also the same, which can be denoted as follows.

$$d_{i,j}^{h} = d_{i,j}^{v} = d_{i,j}^{d} = d_{i,j}^{ud} = (w(|i-j|) * |q_i - o_j|)^p \tag{7}$$

Here $w(|i-j|)$ is a penalty coefficient for phase difference $|i-j|$. $w(|i-j|)$ is different according to parameter g [WDT, 2011] as shown in Figure 2.

For MSM, the elements of $\vec{d_{i,j}}$ are as the follows.

$$\begin{aligned} d_{i,j}^{v} &= C(q_i, q_{i-1}, o_j) \\ d_{i,j}^{h} &= C(o_j, q_i, o_{j-1}) \\ d_{i,j}^{d} &= d_{i,j}^{ud} = d(q_i, o_j) \end{aligned} \tag{8}$$

The C function in Equation (8) can be calculated as follows.

$$\begin{aligned} &C(z,r,t) = \begin{cases} c, & if\ r \leq z \leq t\ or\ r \geq z \geq t; \\ c + \min(d(z,r), d(z,t)), & otherwise. \end{cases} \\ &d(a_x, b_y) = |a_x - b_y| \end{aligned} \tag{9}$$

Here a_x, b_y represents the time series point passed into the d function, x, y represents the time subscript of a_x, b_y in the respective time series, and c is a parameter.

4 Context-aware Elastic Distance

4.1 Context-aware Dynamic Time Warping (CDTW)

For the time series Q and O described by Equation (1), the context of one point means a collection of points in the front and back of this point, we choose $2L$ points from the context of q_i, o_j to construct multi-dimensional features $\hat{q}_i$, $\hat{o}_j$,

$$\begin{aligned} \hat{q}_i &= (.., q_{max(i-l*s*n_c,1)}, \cdots, q_i, .., q_{min(i+l*s*n_c,n)}, \cdots) \\ \hat{o}_j &= (.., o_{max(j-l*s*n_c,1)}, \cdots, o_j, .., o_{min(j+l*s*n_c,m)}, \cdots) \\ & \qquad l = 1, 2, ..., L \end{aligned} \tag{10}$$

Here L indicates that L points are selected from the front and back of current point respectively. The parameter s implies the relative time interval between selecting points, and $n_c = (m+n)/2$ means that the subscript of point should be no less than 1 and no more than n or m. Through the discussion in Section 3, we find that the difference between DTW, DDTW, and WDTW can be summarized as the difference on $\vec{d_{i,j}}$. For our CDTW, the multi-dimensional features $\hat{q}_i$ and $\hat{o}_j$ are used to calculate $\vec{d_{i,j}}$, so the points with similar context have smaller $\vec{d_{i,j}}$.

Context Point Pair Distance

Our CDTW changes Equation (5), and uses the multidimensional features $\hat{q}_i$, $\hat{o}_j$ to calculate $\vec{d_{i,j}}$, which are shown as follows.

$$\begin{aligned} d_{i,j}^{h} &= d_{i,j}^{v} = d_{i,j}^{d} = d_{i,j}^{ud} \\ d_{i,j}^{ud} &= (|q_i - o_j|^p + \sum_{l=1}^{L} |q_{\max(1,i-l*s*n_c)} - o_{\max(1,j-l*s*n_c)}|^p \\ &\quad + \sum_{l=1}^{L} |q_{\min(n,i+l*s*n_c)} - o_{\min(m,j+l*s*n_c)}|^p)/(2L+1) \end{aligned} \tag{11}$$

Here L, s and n_c have the same meaning as Equation (10), and p means to use L_p norm distance. For the classification task, s and L can be determined through cross-validation on the training set. For our experiments in this paper, L is taken as 1, and s can be taken as 0.05 or determined through cross-validation. Since L is constant, the complexity of the CDTW is $O(nm)$.

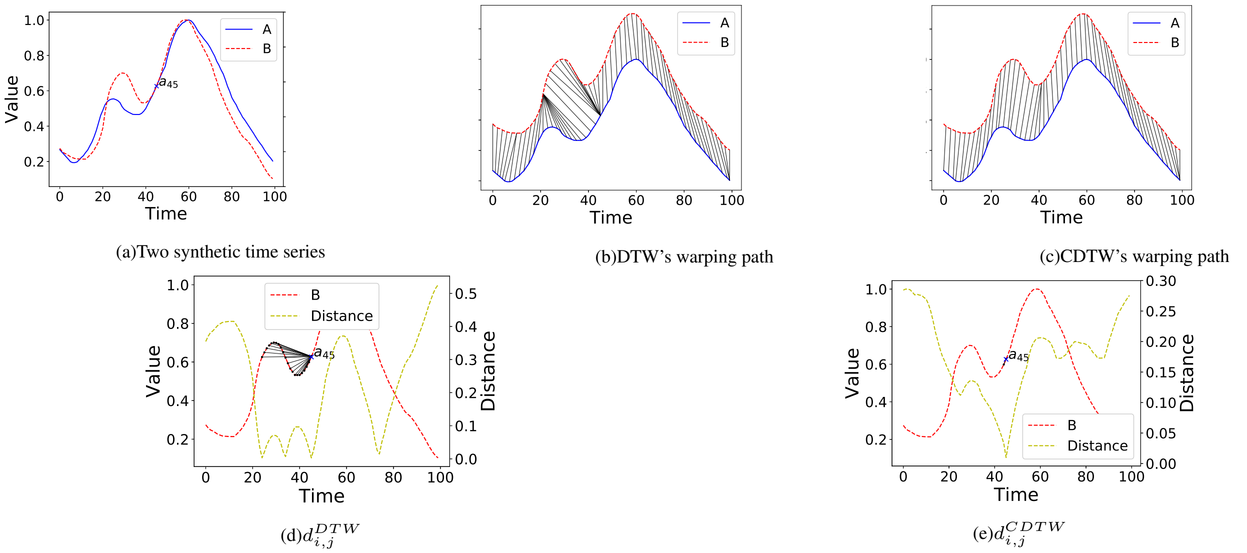
Mitigating Singularities

Before discussing singularities, we first introduce a few concepts.

Definition 1 (Warping Point)**.** *In the warping path W of time series Q and O, we consider $P(q_i)$ as a set of point pairs that contains q_i, $P(o_j)$ as a set of point pairs that contains o_j. If $size(P(q_i))$ or $size(P(o_j))$ is greater than 1, we define q_i or o_j as a warping point, $size(P(q_i))$ or $size(P(o_j))$ is recorded as $degree(q_i)$ or $degree(o_j)$, which indicates the warping degree of q_i or o_j.*

When $degree(q_i)$ or $degree(o_j)$ is much greater than 1 [Keogh and Pazzani, 2001], we call q_i or o_j a singularity. We can get the relationship between the warping path length K, $ELD(Q,O)$ and $degree$ as the follows.

$$K = \sum_{i=1}^{n} degree(q_i) = \sum_{j=1}^{m} degree(o_j) \tag{12}$$



(a)Two synthetic time series

(b)DTW's warping path

(c)CDTW's warping path

(d)$d_{i,j}^{DTW}$

(e)$d_{i,j}^{CDTW}$

Figure 3: CDTW mitigates the problem of singularities.

$$ELD(Q,O) = \sum_{i=1}^{n} \sum_{(q_i,o_{j'})\in P(q_i)} \tilde{d}_{i,j'} = \sum_{j=1}^{m} \sum_{(q_{i'},o_j)\in P(o_j)} \tilde{d}_{i',j} \tag{13}$$

Here q_i is the ith point of time series Q, o_j is jth point of O, n and m are the lengths of time series Q and O, respectively. The value of p means that $\vec{d_{i,j}}$ uses L_p norm distance, and $degree(q_i)$ or $degree(o_j)$ is the degree of warping at q_i or o_j, respectively. $P(q_i)$, $P(o_j)$ are the point pairs containing q_i, o_j in the warping path W. The value of $\tilde{d}_{i^k,j^k}$ is one element of $[d_{i,j}^h, d_{i,j}^v, d_{i,j}^d, d_{i,j}^{ud}]$ according to the direction of step (w_{k-1}, w_k).

Before studying the principle of mitigating singularities, we should dig into the properties of singularities.

Property 1. *According to Equation (12), the occurrence of singularities will make the warping path longer.*

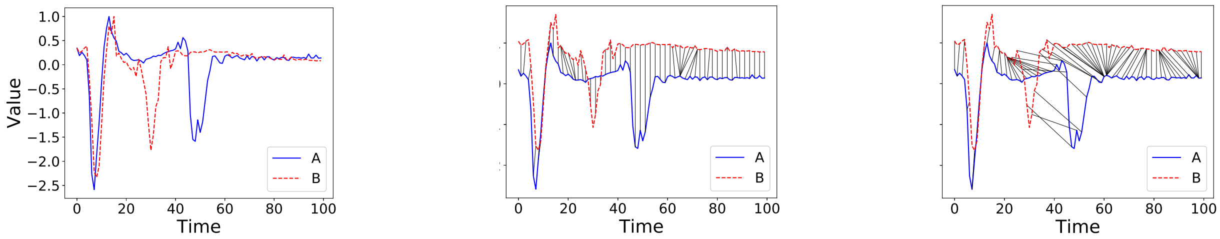
Property 2. *According to Equation (13), the occurrence of singularities will increase the number of $\tilde{d}_{i,j'}$ terms.*

As the number of $\tilde{d}_{i,j'}$ terms increases due to the singularities, we get an intuitive observation that the distances $\tilde{d}_{i,j'}$ of all point pairs of the singularities should be relatively small, otherwise $ELD(Q,O)$ is less possible to be the minimum cumulative distance. Figure 3 shows an example of singularities suppressing by CDTW. For two synthetic sequences A and B, when DTW is used to calculate the distance, two large singularities appear, which can be eliminated using CDTW. For the singularity a_{45}, the yellow line in Figure 3(d) shows the distance from a_{45} to all the points of the sequence B by the DTW. We find that the distances between a_{45} and its connected points in the sequence B have smaller distances, so a_{45} can warp with these points without causing a substantial increase of the cumulative distance $ELD(A,B)$. The yellow line in Figure 3(e) shows the distances from a_{45} to all the points of the sequence B by the CDTW. It is found that a_{45} only has a small distance with few points in the sequence B, so it is difficult to warp at a_{45}.

Our CDTW calculates the point pair distance $\vec{d_{i,j}}$ by using $\hat{q}_i$ and $\hat{o}_j$, which makes the distance between point pairs with different contexts larger, while the distance between point pairs with similar contexts is still smaller. Due to the diversity of multi-dimensional features, the points tend to be scattered. However, the points with similar contexts are still adjacent. Therefore, the singularities are suppressed while retaining the warping ability in our CDTW.

Keeping the Warping Ability

As we discuss in section 1, one kind of approaches for reducing singularities is to mechanically limit the warping path [Chen *et al.*, 2012; Keogh and Pazzani, 2001; Giorgino and others, 2009; Keogh and Ratanamahatana, 2005; Giorgino and others, 2009; WDT, 2011], but these approaches fail to consider the values of time series. Therefore, while limiting singularities, they are easy to miss a good warping path. For example, the penalty term of WDTW is only related to phase difference not to the values of time series. But our CDTW uses the context of the current point to limit the singularities. The point pair distance is only related to the context and has nothing to do with the phase difference. For example, in Figure 4, WDTW has a phase difference penalty term, so the distances between the points on the trough of the sequence A and the points on the trough of the sequence B are large, which leads to a large accumulative distance and the troughs cannot be aligned correctly with each other. For our CDTW, since their context is similar, the points on the trough of the sequence A and those on the trough of the sequence B have a smaller distance and the troughs can be aligned correctly. It can be seen that our CDTW can not only limit the singularity but also retain the warping ability.



(a)Two case of CricketY [Dau *et al.*, 2018a]

(b)WDTW's warping path

(c)CDTW's warping path

Figure 4: CDTW keeps the warping ability.

Noise tolerance

Another kind of approach for limiting singularities is to use derivatives, e.g. DDTW. The DDTW uses derivatives to calculate $\vec{d_{i,j}}$, and then corresponds a sequence segment with a positive derivative to a segment with a positive derivative in another sequence, or a segment with a negative derivative to a negative derivative in another sequence. However, the derivative is easily affected by noise, and the small perturbation of the value of time series will cause the derivative value to change drastically. As shown in Figure 5, For our DDTW, a small change of the time series can cause a drastic change in the derivative value, which has a great impact on the folding path, causing the two noise peaks to align with the middle signal peak as shown in Figure 5(b) and preventing the long common subsegment defined below from appearing.

Definition 2 (Common Subsegment)**.** *On the warping path W of the time series Q and O, if the subsection $Q[i_1 : i_2]$ and $O[j_1 : j_2]$ correspond to each other one point by one point, then we call $Q[i_1 : i_2]$ or $O[j_1 : j_2]$ as common subsegment.*

For the CDTW, due to the addition of context information, it is less possible to align the peaks of the noise with the peaks of the signal, and often the peaks of the signals can be aligned correctly. As shown in Figure 5(c), a long common subsegment can be observed. The anti-noise ability can be measured by the number of long common subsegments.

4.2 Extending the Concept of Context to other Elastic Distances

Does the context-aware idea only apply to DTW? We find that the difference between elastic distances lies in $\vec{d_{i,j}}$. What we need to do to apply the context-aware idea to other elastic distances is to apply the multi-dimensional features $\hat{q}_i$ and $\hat{o}_j$ to the calculation of $\vec{d_{i,j}}$. Next, we take MSM [Stefan *et al.*, 2013] as an example. We change the condition $if\ r \leq z \leq t\ or\ r \geq z \geq t$ in Equation (9) to $if\ d(r,t) \geq \max(d(z,r), d(z,t))$ to adapt to multi-dimensional features and then get the following equation.

$$
\begin{aligned}
&C(z,r,t) = \begin{cases} c, if\ d(r,t) \geq \max(d(z,r), d(z,t)); \\ c + \min(d(z,r), d(z,t)), \quad otherwise. \end{cases} \\
&d(a_x, b_y) = (|a_x - b_y| + \sum_{l=1}^{L} |a_{\max(1,x-l*s*n_c)} - \\
&\qquad b_{\max(1,y-l*s*n_c)}| + \sum_{l=1}^{L} |a_{\min(u,x+l*s*n_c)} \\
&\qquad - b_{\min(v,y+l*s*n_c)}|)/(2L+1) \\
&\qquad A = a_1, a_2, a_3, ..., a_x, ..., a_u \\
&\qquad B = b_1, b_2, b_3, ..., b_y, ..., b_v
\end{aligned}
\tag{14}
$$

Here both A and B can be Q(or O). u and v are the lengths of A and B, respectively. The notation a_x and b_y represent the values passed to the function d, and x and y is subscripts of a_x and b_y, respectively. L and s is the same as those in Equation (10). Since L is constant, the complexity of the CMSM is also $O(nm)$.

5 Experiments

5.1 Experimental Setting

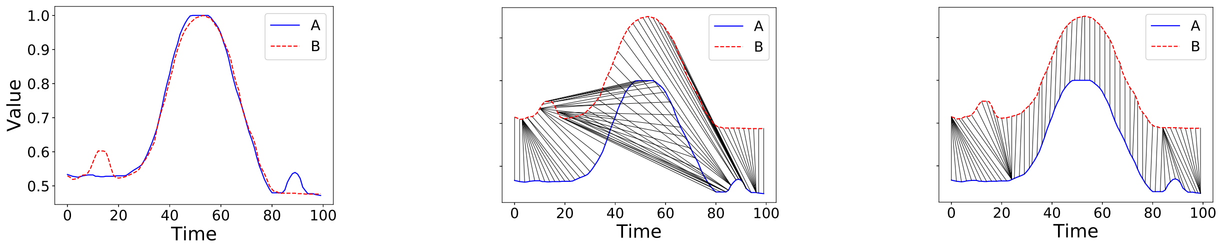
In this paper, the UCR [Dau *et al.*, 2018a] dataset is used as the experimental dataset. UCR is a collection of datasets for time series classification task. It is the most common dataset for time series classification tasks [Bagnall *et al.*, 2017].

In the experiments, our CDTW and CMSM are compared with DTW, DDTW, WDTW, ERP, TWE, and MSM for verifying the validity of the context-aware idea. The kNN classifier is used for all comparisons, since kNN requires the calculation of distance for time series. For convenience, we set k to 1 in the experiments.

For CDTW, the L of Equation (11) is taken as 1, and s takes 0.05. For CDTW$_{CV}$ (which is the cross-validation version of CDTW), L takes 1, and s is determined from $\{0, 0.01, 0.02, ..., 0.09\}$ through cross-validation on the train set. For convenience, the value c of CMSM is chosen from $\{0.01, 0.085, 0.676, 4.96, 31.6\}$ through cross-validation, and L of Equation (14) takes 1 and s takes 0.05.

5.2 Validating Warping Path Properties

In order to verify the advantages of CDTW discussed in Section 4.1, we analyzed all the warping paths of each test case



(a) Two synthetic time series

(b) DDTW's warping path

(c) CDTW's warping path

Figure 5: CDTW is tolerant to noise.

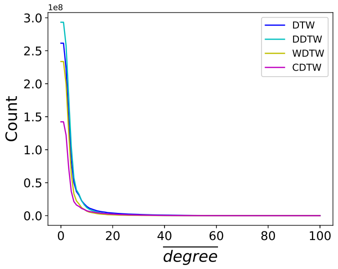


Figure 6: Mitigating singularities

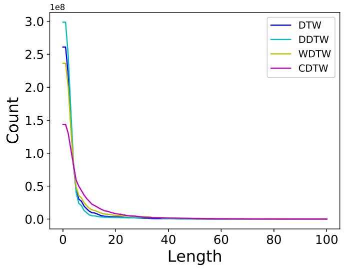


Figure 7: Common subsegments

with each training case on multiple datasets using CDTW, DTW, DDTW, and WDTW respectively. For these paths, we do statistics of singularities, common subsegments, phase differences and the path lengths.

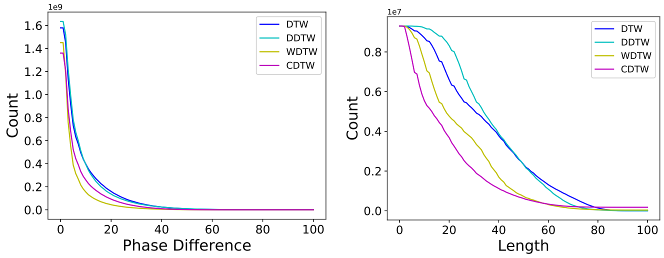
Mitigating Singularities

As we discuss the principle of CDTW suppressing singularities in Section 4.1, our experimental results also verify this conclusion. As in Figure 6, the x axis is the relative degree $\overline{degree}$ of the warping point, where $\overline{degree} = \lceil degree * 100/len(\text{time series})\rceil$. The vertical axis $Count$ is the total number of all warp points whose relative degree greater than or equal to the value of $\overline{degree}$. As introduced in [Keogh and Pazzani, 2001], a singularity is a warping point whose degree is much greater than 1. Therefore, the relative degree of a singularity is also much greater than 1. In Figure 6, we can see that the magenta line representing CDTW is below others, which means that CDTW generally leads to less singularities. It can be concluded that CDTW can alleviate the singularities problem of DTW, and the effect is better than DDTW and WDTW.

Tolerant to Noise

We have discussed the ability of CDTW to resist noise in Section 4.1. Notice that in the generation of the warping path between two time series by the DDTW, the noise peak is likely to misalign with the signal peak, and thus the alignment of the signal peak is destroyed. However, our CDTW is insensitive to noise and can ensure the alignment of signal peaks, which is beneficial to the formation of long common subsegment (Common subsegment is defined in Definition 2). In Figure 7, the horizontal axis represents the relative length of common subsegment. For a common subsegment $Q[i_1 : i_2]$, its relative length is $Length = \lceil(|i_1 - i_2| * 100)/len(\text{time series Q})\rceil$. The vertical axis $Count$ is the total number of all common subsegments whose relative lengths are no less than the value of $Length$. In Figure 7, we can see that our CDTW has more long common subsegments (whose $Length$ are over 5) than DTW, DDTW, and WDTW. It indicates that CDTW has stronger noise immunity and better capability to align the signal peak correctly.

Keeping the Warping Ability



(a) Phase Difference

(b) Path Length

Figure 8: CDTW keep the warping ability.

Our CDTW's warping capabilities have been discussed in Section 4.1. On the warping path W of Q and O, each point w_k on the path W corresponds to the point pair (q_{i_k}, o_{j_k}), and $|i_k - j_k|$ is called phase difference. The larger the phase difference, the greater the degree of warping of the point pair (q_{i_k}, o_{j_k}). In Figure 8(a), the horizontal axis is phase difference (phase difference $= |i_k - j_k|$), and the vertical axis $Count$ is the number of the point pairs whose phase differences are greater than or equal to the corresponding values of Phase Difference axis. Since CDTW suppresses the singularities, its warping ability is weaker than DTW, and its

Table 1: Performance on UCR datasets

Approaches	DTW	DDTW	WDTW	ERP	TWE	MSM	CDTW	CDTW$_{CV}$	CMSM
Mean accuracy	0.7575	0.7251	0.7802	0.775	0.7786	0.784	0.7824	0.7871	**0.7903**

phase difference is also smaller than DTW. Therefore, it can be seen that the magenta line representing CDTW is below DTW in Figure 8(a). At the same time, since CDTW finds the warping path based on the values of time series, its warping ability is stronger than WDTW and has a greater phase difference as shown in Figure 8(a). Besides, our experiments also show that CDTW can generate long paths. As shown in Figure 8(b), the horizontal axis is the relative path length, $Length = \lceil (K - len(\text{time series})) * 100/len(\text{time series}) \rceil$, CDTW has more long paths with $Length$ greater than 85 than DTW, DDTW, and WDTW. It also shows that CDTW can generate ultra-long folding paths.

5.3 Performance for Time Series Classification

Since the time series of the same class on the UCR dataset have more similar values than the time series of different class, a reasonable distance measurement approach should give time series with similar values smaller distances. That is, distances between time series of the same class are smaller, so better classification results. We use classification accuracy to evaluate the performance of other approaches and our proposed approach on an independent dataset. We first get the accuracy of each approach on a single dataset by Equation (15), and then calculate the mean accuracy for all the tested datasets by Equation (16).

$$Accuracy = \frac{\text{Number of test cases classified correctly}}{\text{Total number of test cases}} \tag{15}$$

$$Mean\ accuracy = \frac{\text{Sum of accuracy on each dataset}}{\text{Total number of datasets}} \tag{16}$$

The mean accuracies of various approaches are shown in Table 1. It can be seen that our CDTW$_{CV}$ and CMSM have achieved higher performance. Furthermore, it can be concluded that the classification accuracy of CDTW is better than DTW, DDTW, and WDTW. CDTW$_{CV}$, which sets the value of parameter s through cross-validation, has better classification performance than CDTW. Similarly, the classification accuracy of CMSM is better than MSM. It can be seen that except for DTW, our context-aware idea is also applicable to other elastic distances, such as MSM which is based on edit distance.

6 Conclusion

In this paper, we propose a context-aware distance analysis approach for time series. We show that it is reasonable that when computing the point pair distances of time series, context-aware point pairs of time series often have smaller distances. We present the context-aware DTW (called CDTW), then analyze the properties of CDTW, and show that our CDTW can not only suppress singularities but also retain both warping ability and anti-noise ability. Our context-aware distance analysis approach is not only applicable to DTW, but also to other elastic distances. Correpondingly, we also present a context-aware MSM, called CMSM. We have verified the effects of CDTW and CMSM on time series classification by using UCR datasets. The experimental results show that CDTW and CMSM have higher classification accuracy than the original DTW and MSM respectively, and demonstrate the effectiveness of our context-aware distance analysis approach for time series.

Acknowledgements

This work was supported in part by the Scientific & Technological Innovation 2030 - "New Generation AI" Key Project (No. 2021ZD0114001; No. 2021ZD0114000), and the Science and Technology Commission of Shanghai Municipality (No. 21511102200).

References

[Bagnall *et al.*, 2015] A. Bagnall, J. Lines, J. Hills, and A. Bostrom. Time-series classification with cote: The collective of transformation-based ensembles. *IEEE Transactions on Knowledge and Data Engineering*, 27(9):2522–2535, Sep. 2015.

[Bagnall *et al.*, 2017] Anthony Bagnall, Jason Lines, Aaron Bostrom, James Large, and Eamonn Keogh. The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, 31(3):606–660, May 2017.

[Chen and Ng, 2004] Lei Chen and Raymond Ng. On the marriage of lp-norms and edit distance. In *Proceedings of the Thirtieth International Conference on Very Large Data Bases - Volume 30*, VLDB '04, pages 792–803. VLDB Endowment, 2004.

[Chen *et al.*, 2005] Lei Chen, M. Tamer Özsu, and Vincent Oria. Robust and fast similarity search for moving object trajectories. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, pages 491–502, 2005.

[Chen *et al.*, 2012] Qian Chen, Guyu Hu, Fanglin Gu, and Peng Xiang. Learning optimal warping window size of dtw for time series classification. In *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, pages 1272–1277. IEEE, 2012.

[Dau *et al.*, 2018a] Hoang Anh Dau, Eamonn Keogh, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, Yanping, Bing Hu, Nurjahan Begum, Anthony Bagnall, Abdullah Mueen, Gustavo Batista, and Hexagon-ML. The ucr time series classification archive, October 2018. https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.

[Dau *et al.*, 2018b] Hoang Anh Dau, Diego Furtado Silva, François Petitjean, Germain Forestier, Anthony Bagnall, Abdullah Mueen, and Eamonn Keogh. Optimizing dynamic time warping's window width for time series data mining applications. *Data mining and knowledge discovery*, 32(4):1074–1120, 2018.

[Ding and Chang, 2016] Jr Ding and Che-Wei Chang. Feature design scheme for kinect-based dtw human gesture recognition. *Multimedia Tools and Applications*, 75(16):9669–9684, 2016.

[Giorgino and others, 2009] Toni Giorgino et al. Computing and visualizing dynamic time warping alignments in r: the dtw package. *Journal of statistical Software*, 31(7):1–24, 2009.

[Górecki and Łuczak, 2013] Tomasz Górecki and Maciej Łuczak. Using derivatives in time series classification. *Data Mining and Knowledge Discovery*, 26(2):310–331, Mar 2013.

[HIRSCHBERG, 1977] D. S. HIRSCHBERG. Algorithms for the longest common subsequence problem. *J. ACM*, 24(4):664–675, 1977.

[Holder *et al.*, 2024] Christopher Holder, Matthew Middlehurst, and Anthony J. Bagnall. A review and evaluation of elastic distance functions for time series clustering. *Knowl. Inf. Syst.*, 66(2):765–809, 2024.

[Holznigenkemper *et al.*, 2023] Jana Holznigenkemper, Christian Komusiewicz, and Bernhard Seeger. Exact and heuristic approaches to speeding up the MSM time series distance computation. In *Proceedings of the 2023 SIAM International Conference on Data Mining, SDM 2023, Minneapolis-St. Paul Twin Cities, MN, USA, April 27-29, 2023*, pages 451–459. SIAM, 2023.

[Hsu *et al.*, 2011] Hui-Huang Hsu, Andy C Yang, and Ming-Da Lu. Knn-dtw based missing value imputation for microarray time series data. *Journal of computers*, 6(3):418–425, 2011.

[Kate, 2016] Rohit J Kate. Using dynamic time warping distances as features for improved time series classification. *Data Mining and Knowledge Discovery*, 30(2):283–312, 2016.

[Keogh and Pazzani, 2001] Eamonn J. Keogh and Michael J. Pazzani. Derivative dynamic time warping. In *Proceedings of the 2001 SIAM International Conference on Data Mining*, 2001.

[Keogh and Ratanamahatana, 2005] Eamonn Keogh and Chotirat Ann Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and Information Systems*, 7(3):358–386, Mar 2005.

[Latecki *et al.*, 2005] L. J. Latecki, V. Megalooikonomou, Q. Wang, R. Lakaemper, C. A. Ratanamahatana, and E. Keogh. Elastic partial matching of time series. In Alípio Mário Jorge, Luís Torgo, Pavel Brazdil, Rui Camacho, and João Gama, editors, *Knowledge Discovery in Databases: PKDD 2005*, pages 577–584, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[Lines and Bagnall, 2015] Jason Lines and Anthony Bagnall. Time series classification with ensembles of elastic distance measures. *Data Mining and Knowledge Discovery*, 29(3):565–592, May 2015.

[Lines *et al.*, 2016] J. Lines, S. Taylor, and A. Bagnall. Hivecote: The hierarchical vote collective of transformation-based ensembles for time series classification. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 1041–1046, 2016.

[Marteau, 2009] P. Marteau. Time warp edit distance with stiffness adjustment for time series matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):306–318, 2009.

[Sakoe and Chiba, 1978] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.

[Stefan *et al.*, 2013] A. Stefan, V. Athitsos, and G. Das. The move-split-merge metric for time series. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1425–1438, 2013.

[Tim, 2015] Time-series clustering – a decade review. *Information Systems*, 53:16 – 38, 2015.

[WDT, 2011] Weighted dynamic time warping for time series classification. *Pattern Recognition*, 44(9):2231 – 2240, 2011.

[Yadav and Alam, 2018] Munshi Yadav and Afshar Alam. Dynamic time warping (dtw) algorithm in speech: a review. *Int. J. Res. Electron. Comput. Eng*, 6, 2018.

[Yurtman *et al.*, 2023] Aras Yurtman, Jonas Soenen, Wannes Meert, and Hendrik Blockeel. Estimating dynamic time warping distance between time series with missing data. In *Machine Learning and Knowledge Discovery in Databases: Research Track - European Conference, ECML PKDD 2023, Turin, Italy, September 18-22, 2023, Proceedings, Part V*, volume 14173 of *Lecture Notes in Computer Science*, pages 221–237. Springer, 2023.

[Zhang *et al.*, 2015] Zheng Zhang, Ping Tang, and Rubing Duan. Dynamic time warping under pointwise shape context. *Information sciences*, 315:88–101, 2015.