Pattern Recognition 44 (2011) 2231–2240

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Weighted dynamic time warping for time series classification

Young-Seon Jeong [a], Myong K. Jeong [a,b,c,*], Olufemi A. Omitaomu [d]

[a] Department of Industrial and Systems Engineering, Rutgers University, Piscataway, NJ, USA
[b] Rutgers Center for Operations Research, Rutgers University, Piscataway, NJ, USA
[c] Department of Industrial and Systems Engineering, KAIST, Daejon, Korea
[d] Geographic Information Science & Technology Group, Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

ARTICLE INFO

Article history:
Received 20 November 2009
Received in revised form
29 August 2010
Accepted 30 September 2010
Available online 14 October 2010

Keywords:
Dynamic time warping
Adaptive weights
Weighted dynamic time warping
Modified logistic weight function
Time series classification
Time series clustering

ABSTRACT

Dynamic time warping (DTW), which finds the minimum path by providing non-linear alignments between two time series, has been widely used as a distance measure for time series classification and clustering. However, DTW does not account for the relative importance regarding the phase difference between a reference point and a testing point. This may lead to misclassification especially in applications where the shape similarity between two sequences is a major consideration for an accurate recognition. Therefore, we propose a novel distance measure, called a weighted DTW (WDTW), which is a penalty-based DTW. Our approach penalizes points with higher phase difference between a reference point and a testing point in order to prevent minimum distance distortion caused by outliers. The rationale underlying the proposed distance measure is demonstrated with some illustrative examples. A new weight function, called the modified logistic weight function (MLWF), is also proposed to systematically assign weights as a function of the phase difference between a reference point and a testing point. By applying different weights to adjacent points, the proposed algorithm can enhance the detection of similarity between two time series. We show that some popular distance measures such as DTW and Euclidean distance are special cases of our proposed WDTW measure. We extend the proposed idea to other variants of DTW such as derivative dynamic time warping (DDTW) and propose the weighted version of DDTW. We have compared the performances of our proposed procedures with other popular approaches using public data sets available through the UCR Time Series Data Mining Archive for both time series classification and clustering problems. The experimental results indicate that the proposed approaches can achieve improved accuracy for time series classification and clustering problems.

1. Introduction

There has been a long-standing interest for time series classification and clustering in diverse applications such as pattern recognition, signal processing, biology, aerospace, finance, medicine, and meteorology [1,2,8,12,14,18,23,25,26], and thus some notable techniques have been developed including nearest neighbor classifier with a given distance measure, support vector machines, and neural networks [2,4,20]. The nearest neighbor classifiers with dynamic time warping (DTW) has shown to be effective for time series classification and clustering because of its non-linear mappings capability [7,18,25]. The DTW technique finds an optimal match between two sequences by allowing a non-linear mapping of one sequence to another, and minimizing the distance between two sequences [8,7,12,22]. The sequences are ''warped'' non-linearly to determine their similarity independent of any non-linear variations in the time dimension. The technique was originally developed for speech recognition, but several researchers have evaluated its application in other domains and have developed several variants such as derivative DTW (DDTW) [11,21,22]. Fig. 1 shows the example of process of aligning two out of phase sequences by DTW.

The methodology for DTW is as follows. Assume a sequence A of length m, $A=a_1, a_2, \ldots, a_i, \ldots, a_m$ and a sequence B of length n, $B=b_1, b_2, \ldots, b_j, \ldots, b_n$. We create an m-by-n path matrix where the (ith, jth) element of matrix contains the distance between the two points a_i and b_j such that $d(a_i,b_j)=||(a_i-b_j)||_p$, where $||\cdot||_p$ represents the l_p norm. The warping path is typically subject to several constraints such as [22]

Endpoint constraint: the starting and ending points of warping path have to be the first and the last points of the path matrix, that is, $u_1=(a_1, b_1)$ and $u_k=(a_m, b_n)$.
Continuity constraint: the path can advance one step at a time. That is, when $u_k=(a_i, b_j)$, $u_{k+1}=(a_{i+1}, b_{j+1})$ where $a_i-a_{i+1}\le 1$ and $b_i-b_{i+1}\le 1$.
Monotonicity: the path does not decrease, i.e., $u_k=(a_i, b_j)$, $u_{k+1}=(a_{i+1}, b_{j+1})$ where $a_i \ge a_{i+1}$ and $b_i \ge b_{i+1}$.

[*] Corresponding author at: Department of Industrial and Systems Engineering, Rutgers University, 640 Bartholomew Road-Room 115, Piscataway, NJ 08854, USA. Tel.: +1 732 445 4858; fax: +1 732 445 5472.
E-mail address: mjeong@rci.rutgers.edu (M.K. Jeong).

0031-3203/$ - see front matter © 2011 Published by Elsevier Ltd.
doi:10.1016/j.patcog.2010.09.022

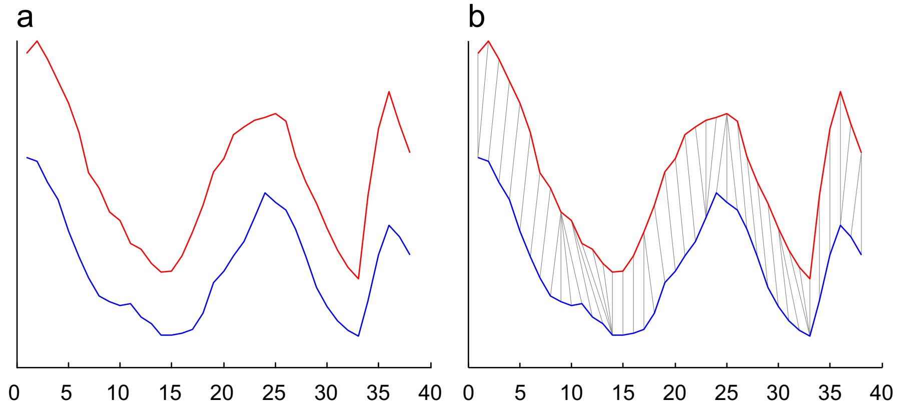

Fig. 1. Alignment of sequences based on DTW: (a) two similar sequences, but out of phase and (b) alignment by DTW.

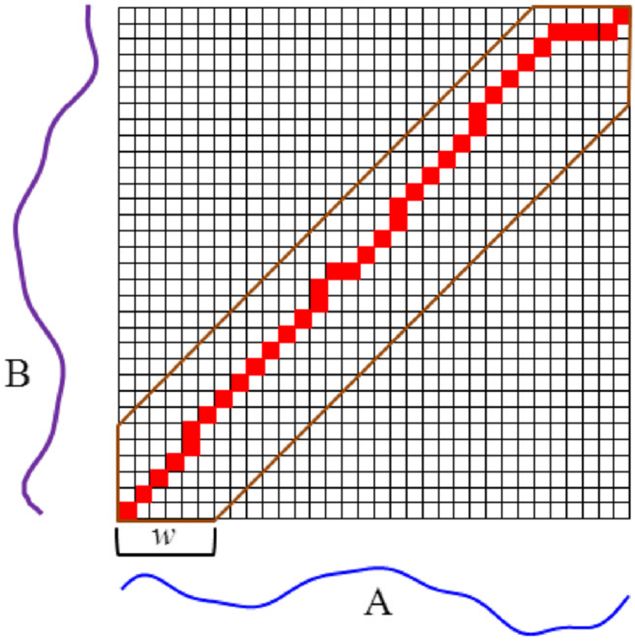

Fig. 2. Warping matrix and optimal warping path by DTW.

The best match between two sequences is the one with the lowest distance path after aligning one sequence to the other. Therefore, the optimal warping path can be found by using recursive formula given by

$$DTW_p(A,B) = \sqrt[p]{\gamma(i,j)}$$

where $\gamma(i,j)$ is the cumulative distance described by

$$\gamma(i,j) = |a_i - b_j|^p + \min\{\gamma(i-1,j-1), \gamma(i-1,j), \gamma(i,j-1)\} \quad (1)$$

As seen from Eq. (1), given a search space defined by two time series DTW_p guarantees to find the warping path with the minimum cumulative distance among all possible warping paths that are valid in the search space. Thus, DTW_p can be seen as the minimization of warped l_p distance with time complexity of $O(mn)$. By restraining a search space using constraint techniques such as Sakoe–Chuba Band [22] and Itakura Parallelogram [7], the time complexity of DTW can be reduced. Fig. 2 shows the warping matrix and optimal warping path between two sequences by DTW. In Fig. 2, a band with width w is used to constrain the warping.

However, the conventional DTW calculates the distance of all points between two series with equal weight of each point regardless of the phase difference between a reference point and a testing point. This may lead to misclassification especially in applications such as image retrieval where the shape similarity between two sequences is a major consideration for an accurate recognition, thus neighboring points between two sequences are more important than others. In other words, relative significance depending on the phase difference between points should be considered.

Therefore, this paper proposes a novel distance measure, called the weighted dynamic time warping (WDTW), which weights nearer neighbors more heavily depending on the phase difference between a reference point and a testing point. Because WDTW takes into consideration the relative importance of the phase difference between two points, this approach can prevent a point in a sequence from mapping the further points in another one and reduce unexpected singularities, which are alignments between a point of a series with multiple points of the other series. Some practical examples will be presented to graphically illustrate possible situations where WDTW clearly is a better approach.

In addition, a new weight function, called the modified logistic weight function (MLWF), is proposed to assign weights as a function of the phase difference between a reference point and a testing point. The proposed weight function extends the properties of logistic function to enhance the flexibility of setting bounds on weights. By applying different weights to adjacent points, the proposed algorithm can enhance the detection of similarity between series.

Finally, we extend the proposed idea to other variants of DTW such as derivative dynamic time warping (DDTW) and propose the weighted version of DDTW (WDDTW). We compare the performances of our proposed procedures with other popular approaches using public data sets available through UCR Time Series Data Mining Archive [13] for both time series classification and clustering problems. The experimental results show that the proposed procedures achieve improved accuracy for time series classification and clustering problems.

This remainder of the paper is organized as follows. In Section 2, we review some related literatures on times series classification and its methodologies. Section 3 explains the rationale of the advantage of the proposed idea. In Section 4, we describe the proposed WDTW and the modified logistic weight function for automatic time series classifications. The experimental results are presented and discussed in Section 5. The paper ends with concluding remarks and future works in Section 6.

2. Related works

As a result of the increasing importance of time series classification in diverse fields, lots of algorithms have been proposed for different applications. Husken and Stagge [6] utilized recurrent neural networks for time series classification and Guler and Ubeyli [4] presented the wavelet-based adaptive neuro-fuzzy inference system model for classification of ectroencephalogram (EEG) signals. Rath and Manmatha [21] used DTW for word image matching and compared the performance of DTW with other popular techniques, including affine-corrected Euclidean distance mapping, the shape context algorithm, and correlation using sum of squared differences. Gullo et al. [5] developed a time series representation model, called Derivative time series Segment Approximation (DSA), which combines the notions of derivative estimation, segmentation and segment approximation, for supporting accurate and fast similarity detection in time series data. Eads et al. [2] introduced a hybrid classification algorithm that employs evolutionary computation for feature extraction, and a support vector machine for classification with the selected features. They tested their algorithm on a lightning classification task using data acquired from the Fast On-orbit Recording of Transient Events (FORTE) satellite.

In the area of new distance measures for time series classification and clustering, Keogh and Pazzani [11] proposed a modification of DTW, called Derivative Dynamic Time Warping (DDTW), which transforms an original sequence into a higher level feature of shape by estimating derivatives. By preventing the production of unexpected singularities, DDTW has showed promising results for several special cases such as (1) two sequences differ in the *Y*-axis as well as *X*-axis, (2) cases in which there are local differences in the *Y*-axis, for instance, a peak in one sequence may be higher that the corresponding peak in the other sequences.

However, DDTW retains the assumption that all points in the sequence are weighted equally; that is, it is possible that a point of a series may be matched with further neighboring points of the other series, generating a similar problem as DTW. With a similar concept to DDTW, Xie and Wiltgen [27] recently proposed an adaptive feature based DTW, which was designed to align two sequences with local and global features of each point in a sequence instead of its value or derivative.

3. Rationale for the performance advantages of WDTW

In this section, we will present the rationale underlying the proposed WDTW with practical examples to graphically illustrate situations where WDTW shows better performance than conventional DTW. The first example deals with automatic classification of defect patterns on semiconductor wafer maps. Fig. 3(a)–(d) shows four common classes of defect patterns on wafer maps. Jeong et al. [9] presented the effectiveness of using spatial correlograms (i.e., time series data) as new features for the classification of wafer maps instead of original binary input variables for each pixel where 1 represents the defective chip (black color) and 0 indicates the good chip (white color). Fig. 3(e)–(h) shows the corresponding spatial correlograms of Fig. 3(a)–(d), respectively. In correlograms, *X*-axis represents the spatial lags and *Y*-axis indicates their corresponding statistic value.

The correlogram plots the standardized value of $T(d)$ over the spatial lag d where $T(d)$ is given as follows for a given defective rate (p) [9]:

$$T(d) = pc_{00}(d)+(1-p)c_{11}(d),$$

where $c_{00}(d)$ and $c_{11}(d)$ represents the total number of normal (0)-to-normal (0) chip and defective (1)-to-defective (1) chip joins at a lag d for a given wafer map, respectively (for more details, see [9]). Higher value of $T(d)$ means that defective chips or good chips exist together at lag d. Fig. 4 shows the definition of neighbors (or joins) at lag d under a Rook-move neighborhood (RMN) construction rule. In Fig. 4, the black square represents a reference chip and red lines indicate neighboring chips (i.e. neighbors of a reference chip) with spatial lag $d=1$. Similarly, blue lines present neighboring chips with spatial lag $d=2$.

If $T(d)$ is large, the neighbors at distance d from a reference defective chip (normal chip) include more defective chips (normal chips) than expected. If $T(d)$ is small, a reference defective chip (normal chip) tends to have normal chips (defective chips) as its neighbor at distance d. For example, in case of a cluster defect pattern, correlogram in Fig. 3(f), shows larger value of $T(d)$ for the 1st–5th lag, meaning that at those distances, defective chips are clustered at certain areas. From 20th to 30th lags, statistic value is a large negative, indicating that at that distance, defective chips (normal chips) are joined with normal chips (defective chips).

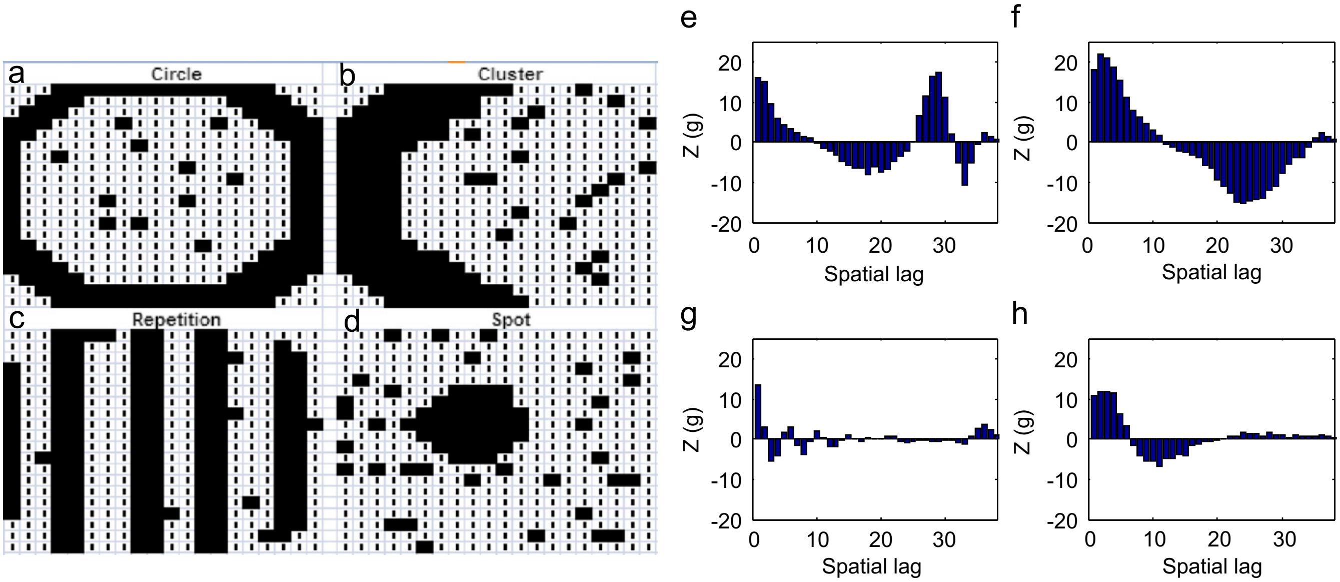

Fig. 3. Typical defect patterns on wafer map and their corresponding correlograms.

Thus, the comparison of statistic value at the *same* lag (or *neighboring* lags) between two correlograms (or sequences) is more meaningful when they are compared for defect pattern classification and WDTW may choose higher value of g where g is the control parameter for the penalization level in weighting function. The higher g value, the more penalizing to points with higher phase difference to determine the optimal weights (see Section 4.2 for the detailed introduction of a weight function).

Figs. 5 and 6 show the classification results of a new observation in testing data using DTW and WDTW, respectively. The red line represents a new time series data that should be classified into one of the classes, and blue and pink lines represent the training data set. Fig. 5(a) shows the result of alignment using DTW, showing the nearest distance among training data set. The distance is 41.31. Fig. 5(b) shows the result of alignment using DTW, showing the second nearest distance among training data set. The distance is 41.82. In case of DTW, some points in circle sequence (testing data, red line) are matched with further points in cluster sequence, distorting a minimum distance. Thus, a new testing sequence, which should be classified into a circle class, is misclassified into a clustering class. However, as shown in Fig. 6, our proposed distance measure accurately classifies testing circle pattern into a same class because it penalizes more a point with higher phase difference between points, in other words, by preventing a point in a sequence from matching further points in another one. Note that for this case study, the optimal parameter g value for WDTW, which was optimized using the validation data set, was found to be 0.4, implicating much more penalizing for further points to increases the classification accuracy because the matching between points with same or neighboring lags is more meaningful for the classification of defect patterns.

The second motivating example considers time series from “UCR Time Series Data Mining Archive.” The data consists of six classes (Normal, Cycle, Increasing trend, Decreasing trend, Upward shift, and Downward shift) [19]. Figs. 7 and 8 represent the alignments generated by DTW and WDTW, respectively. The red line indicates a new observation (in the test data) which is a

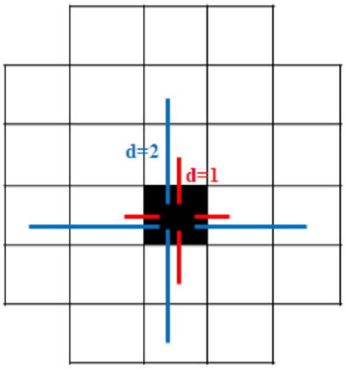

Fig. 4. RMN neighborhood construction rules.

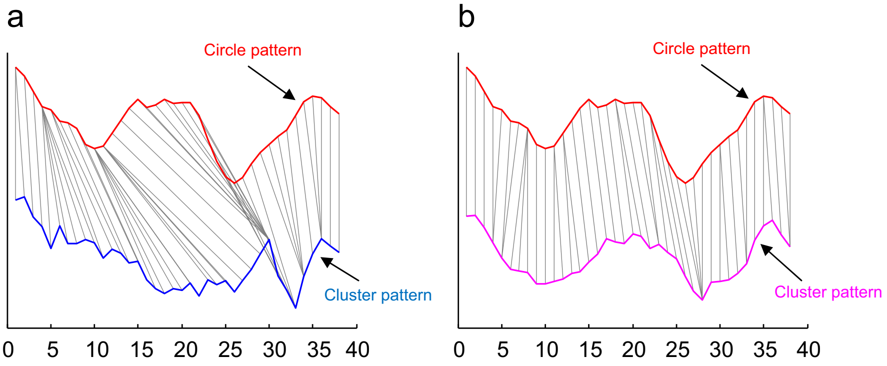

Fig. 5. Alignment results generated by DTW. (a) Circle pattern (a new observation in testing data, red line) vs. cluster pattern (an observation with the minimum distance using DTW in training data, blue line); DTW distance=41.31. (b) Circle pattern (a new observation in testing data, red line) vs. circle pattern (an observation with the second minimum distance using DTW in training data, pink line); DTW distance=41.82. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

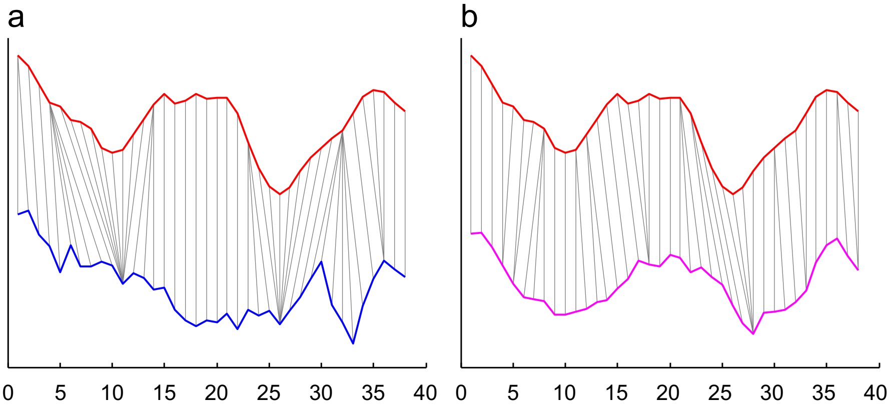

Fig. 6. Alignment results generated by WDTW (g=0.4). (a) Circle pattern (a new observation in testing data, red line) vs. cluster pattern (an observation that showed the minimum distance using DTW in training data, blue line); WDTW distance=0.16. (b) Circle pattern (a new observation in testing data, red line) vs. circle pattern (an observation with the minimum distance using WDTW in training data, pink line); WDTW distance=0.03. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

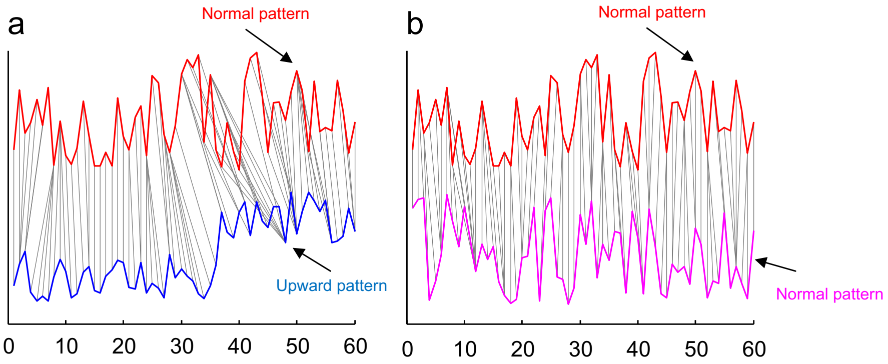

Fig. 7. Control chart pattern alignments generated by DTW (a) normal (a new observation in testing data, red line) vs. upward shift (an observation with the minimum distance using DTW in training data, blue line); DTW distance = 17.4. (b) Normal (a new observation in testing data, red line) vs. normal (an observation with the second minimum distance using DTW in training data, pink line); DTW distance = 18.6. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

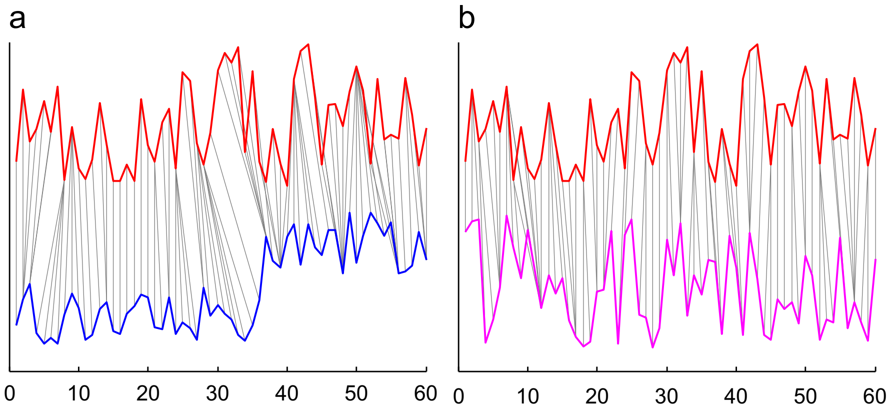

Fig. 8. Control chart pattern alignments generated by WDTW ($g=0.3$). (a) Normal (a new observation in testing data, red line) vs. upward shift (an observation that showed the minimum distance using DTW in training data, blue line); WDTW distance = 0.134. (b) Normal (a new observation in testing data, red line) vs. normal (an observation with the minimum distance using WDTW in training data, pink line); WDTW distance = 0.123. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

"Normal" pattern, and blue and pink line represents "Upward shift" and "Normal" pattern in the training data, respectively. In order to correctly classify a given sequence, a point in the series should be matched with nearer neighbors of the other series because all sequences in the same class have similar shape. As shown in Fig. 7, which shows the alignment by DTW, DTW maps a point in the red sequence to the points with further distance in the blue sequence. This alignment certainly does not have a positive impact on the similarity evaluation of these two sequences even though they have a minimum DTW distance between them. For example, Fig. 7(a) presents the alignments by DTW between Normal (a new observation in the testing data, red line) and Upward shift (training data, blue line) with 17.4 of DTW distance while Fig. 7(b) shows the alignments by DTW between Normal (a new observation in the testing data, red line) and Normal (training data, pink line) with 18.6 of DTW distance. Thus, DTW selects Upward shift sequence as the best match for a new sequence of Normal class, causing a misclassification. Meanwhile, Fig. 8(a) presents the alignment by WDTW between Normal (a new observation in the testing data, red line) and Upward shift (training data, blue line) with 0.134 of WDTW distance while Fig. 8(b) shows the alignment by WDTW between Normal (a new observation in the testing data, red line) and Normal (training data, pink line) with 0.123 of WDTW distance, correctly classifying Normal sequence. For WDTW, parameter g value was optimized using validation data set and was set to 0.3 in this case.

4. Proposed algorithm for time series classification

This section presents the proposed WDTW measure and a new weighting function, so called modified logistic weight function (MLWF) for time series data.

4.1. Weighted dynamic time warping

As mentioned earlier, the standard DTW calculates the distance of all points with equal penalization of each point regardless of the phase difference. The proposed WDTW penalizes the points according to the phase difference between a test point and a reference point to prevent minimum distance distortion by outliers. The key idea is that if the phase difference is low, smaller weight is imposed (i.e., less penalty is imposed) because neighboring points are important, otherwise larger weight is imposed.

In the WDTW algorithm, when creating an m-by-n path matrix, the distance between the two points a_i and b_j is calculated as $d_w(a_i,b_j) = ||w_{|i-j|}(a_i-b_j)||_p$ where $w_{|i-j|}$ is a positive weight value between the two points a_i and b_j. The proposed algorithm implies that when we calculate the distance between a_i in a sequence A and b_j in a sequence B, the weight value will be determined based on the phase difference $|i-j|$. In other words, if the two points a_i and b_j are near, smaller weights can be imposed. Thus, the optimal distance between the two sequences is defined as the minimum path over all possible paths as follows:

$$\mathrm{WDTW}_p(A,B)=\sqrt[p]{\gamma^*(i,j)} \tag{2}$$

where $\gamma^*(i,j)=|w_{|i-j|}(a_i-b_j)|^p+\min\{\gamma^*(i-1,j-1),\gamma^*(i-1,j),\gamma^*(i,j-1)\}$.

Based on the classical analysis of l_p spaces, we present the following propositions that show some mathematical properties of WDTW such as WDTW_p distance decreases monotonically as p increases and the opposite can be obtained under the specific condition on the measured space.

Proposition 1. *For* $0<p<q\le\infty$, $\mathrm{WDTW}_p(a_i,b_j)\ge\mathrm{WDTW}_q(a_i,b_j)$.

Proposition 2. *For* $0<p<q\le\infty$, $\mathrm{WDTW}_p(s_i,r_j)\le(2n-2)^{(1/p)-(1/q)}$ $\mathrm{WDTW}_q(s_i,r_j)$, *where n is the length of the two sequences.*

Given the lengths of two sequences are m and n, respectively, the time complexity of WDTW is the same as DTW, which is $O(mn)$. There are weight factors to a distance calculation in WDTW, but each cell in an m-by-n path matrix should be filled in with the same time. Also, the best distance measure is related to the selection of p because WDTW_p can be seen as the minimization of the warped l_p weighed distance. Even though optimal p depends on applications, l_1 and l_2 are usually good choices to classify time series data set [15,17].

4.2. Modified logistic weight function

The next issue is how to systematically assign weights as a function of the phase difference between two points. In this section, we present our proposed modified logistic weight function (MLWF). One of the most popular classical symmetric functions that use only one equation is the logistic function. However, the standard form of logistic function is not flexible in setting bounds on weights. Therefore, in this paper, we propose modified logistic weight function (MLWF), which extends the properties of logistic function.

The weight value $w_{(i)}$ is defined as

$$w_{(i)}=\left[\frac{w_{max}}{1+\exp(-g(i-m_c))}\right] \tag{3}$$

where $i=1,\ldots,m$, m is the length of a sequence and m_c is the midpoint of a sequence. w_{max} is the desired upper bound for the weight parameter, and g is an empirical constant that controls the curvature (slope) of the function; that is, g controls the level of penalization for the points with larger phase difference. The value of g could range from zero to infinity, but we investigate the characteristics of MLWF for four special cases. The characteristics of these four cases are summarized as follows: (1) *Constant weight*: This is the case in which all points are given the same weight. This can be achieved when $g=0$. (2) *Linear weight*: This is applicable to cases in which the weight is linearly proportional to the extent of the distance. This is the case when $g=0.05$, then the value of $w_{(i)}$ is nearly a linearly increasing relationship. (3) *Sigmoid weight*: Different sigmoid pattern can be achieved using different values of g. For example, the weight function follows a sigmoid pattern when $g=0.25$. (4) *Two distinct weights*: In this case, the first one-half is given one weight and the second one-half is given another weight. This is possible when $g=3$. The pictorial representations of the different weights for these g values are shown in Fig. 9. Fig. 9 also shows that the profile for MLWF is symmetric around the midpoint (m_c) of the total length of a sequence. For Fig. 9, the m and w_{max} are set to 100 and 1, respectively. It has been shown that a linear weighting profile and a sigmoidal pattern of weighting profile can be obtained by setting $g=0.05$ and $g=0.25$, respectively. Setting $g=3$ results in two distinct weights.

Remark 1. Conventional DTW and Euclidean distance measures are special cases of the proposed WDTW. For example, when $w_{|i-j|}$ is constant, i.e., $g=0$ in MLWF, with regard to phase $|i-j|$, WDTW is equivalent to DTW. However, as $w_{|i-j|}$ becomes smaller, i.e., g becomes larger, for the points in nearer phase $|i-j|$, WDTW will be closer to Euclidean distance because it does not allow non-linear alignments of one point to another. By choosing the appropriate g value, WDTW can achieve improved performance in diverse situations.

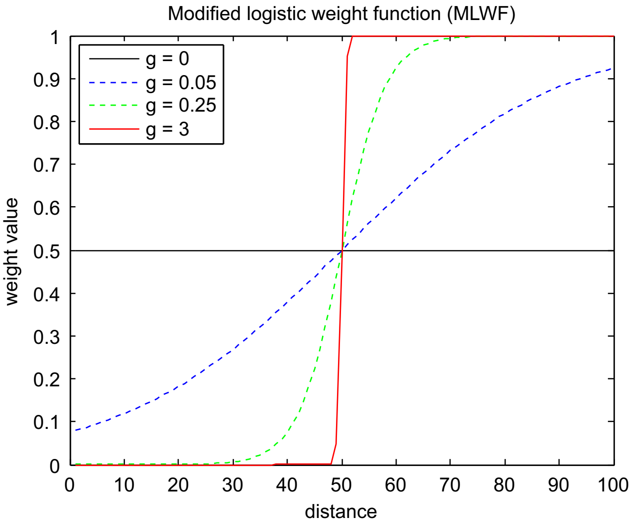

Fig. 9. The pictorial representations of MLWF with different values of g.

Remark 2. Based on our empirical study, the range of optimal g is distributed from 0.01 to 0.6. Smaller g means the less penalty for further points in the sequence, thus WDTW performance is similar to DTW. For example, in case of the signals with common initial phase shift, smaller penalty (or g) will be selected. For larger g, WDTW considers higher penalty for further points, leading to a similar performance of Euclidean distance.

4.3. Weighted derivative dynamic time warping (WDDTW)

The proposed weighted concept can be extended to variants of DTW. In this subsection, we extend the proposed idea to derivative dynamic time warping (DDTW) [11], which is one popular variant of DTW, and propose the weighted version of DDTW (WDDTW). Because DTW may try to explain variability in the Y-axis by warping the X-axis, this may lead to the unexpected singularities, which are alignments between a point of a series with multiple points of the other series, and unintuitive alignments. In order to overcome those weaknesses of DTW, DDTW transforms the original points into the higher level features, which contain the shape information of a sequence. The estimate equation for transforming data point a_i in the sequence A is given by [11]

$$D_A(d_i^a)=\frac{(a_i-a_{i-1})+((a_{i+1}-a_{i-1})/2)}{2},\quad 1<i<m$$

where m is the length of sequence A. Because the first and last estimates are not defined, it is considered that $d_1^a=d_2^a$ and $d_m^a=d_{m-1}^a$.

The weighted version of DDTW is given as follows:

$$\mathrm{WDDTW}_p(D_A,D_B)=\sqrt[p]{\xi^*(i,j)} \tag{4}$$

where $\xi^*(i,j)=|w_{|i-j|}(d_i^a-d_j^b)|^p+\min\{\xi^*(i-1,j-1),\ \xi^*(i-1,j),\xi^*\ (i,j-1)\}$, and D_A and D_B are the transformed sequences from sequence A and B, respectively.

5. Experimental results

5.1. Performance comparison for time series classification

In this section, we perform extensive experiments to verify the effectiveness of the proposed algorithm for time series classification and clustering. All data sets, which include real-life time series, synthetic time series, and generic time series, come from different application domains and are obtained from "UCR Time Series Data Mining Archive" [13]. For the detailed descriptions of the data sets, please see Ratanamahatana and Keogh [20].

Euclidean distance, conventional DTW, and DDTW techniques are selected for comparison with the proposed algorithm. In addition, for comparison with state-of-art for time series similarity search, we implement the Longest Common Subsequence (LCSS), which is one of the popular methods for time series similarity because of its robustness to noise [24]. LCSS measure has two parameters, δ and ε, which should be optimized using validating data set. The constant δ, which is usually set to less than 20% of the sequence length, controls the window size in order to match a given point from one sequence to a point in another sequence. The constant ε, where $0<\varepsilon<1$, is the matching threshold (please refer to [24] in details). In this paper, we use 1-nearest neighbor classifier because the 1-nearest neighbor classifier with DTW showed very competitive performance and has been widely used for time series classification [26].

For WDTW, two parameters should be fixed prior to the evaluation of testing performance. Different w_{max} does not affect its performance, thus, we set w_{max} to 1 in this work. In addition, because an optimal g value is different depending on the application domains, we choose the optimal g value using the validation data set after we divide the given data set into training, validating, and testing sets.

Table 1 shows the classification accuracy of the four different procedures for each data set. In this work, the error rate is calculated as follows:

$$\text{Error rate} = \frac{(\text{total number of testing data})-(\text{total number of correctly classified data})}{(\text{total number of testing data})}$$

As seen in Table 1, our proposed distance measures, WDTW and WDDTW, clearly outperform standard DTW, DDTW, and LCSS measures. In most of cases, the accuracies of WDTW and WDDTW is better (or equal in a few cases) than those of DTW and DDTW. In addition, we can see that depending on the application domains, DDTW results in better accuracy than DTW. The experimental results indicate that our proposed procedures are quite promising for automatic time series classifications in diverse applications. Note that when g becomes smaller, the error rate for WDTW becomes similar to that of DTW.

5.2. Effect of parameter values in WDTW

For WDTW, two parameters should be considered prior to the evaluation of testing performance. The w_{max}, which is used to set the maximum of weight values, does not influence on the accuracy of experimental results in this study because weight is positive and w_{max} represents the full scale of weights in MLWF. For example, Fig. 10 presents the MLWF with different w_{max} values. Regardless of w_{max} value, MLWF retains its shape, implying that MLWF assigns weights with constant ratios to points in a sequence.

In addition, WDTW should choose the optimal g value depending on the application domains. Fig. 11 shows the effect of g to the error rates of the validation data for the "Swedish Leaf" data set. "Swedish Leaf" data set was split into a training set of 500 samples, a validation set of 313 samples, and a test set of 312 samples. As shown in Fig. 11, at the beginning, as g value increases, error rate decreases because nearer points are heavily weighed so that it is highly possible that sequence with a similar shape is chosen with minimum distance. However, as g value increases continuously, error rate increases after reaching the minimum error rate (0.115) because too large g value does not allow non-linear alignments of one point to another. In order words, WDTW with large g value will achieve similar performance to Euclidean distance measure as

Table 1
Summary of classification performance.

Data name	Number of classes	Size of training set	Size of validating set	Size of testing set	Time series length	Error rates					
						ED[a]	DTW	WDTW (g)	DDTW	WDDTW (g)	LCSS (δ^*, ε)
Synthetic control	6	300	150	150	60	0.153	0.007	**0.002** (0.3)	0.433	0.433 (0.01)	0.033 (5, 0.6)
Gun-point	2	50	75	75	150	0.093	0.080	0.040 (0.2)	**0**	**0** (0.1)	0.027 (6, 0.1)
CBF	3	30	450	450	128	0.136	**0.002**	**0.002** (0.08)	0.418	0.418 (0.01)	0.004 (6, 0.3)
Face (all)	14	560	845	845	131	0.319	0.258	0.257 (0.01)	0.144	**0.131** (0.1)	0.300 (2, 0.1)
OSU leaf	6	200	121	121	427	0.438	0.388	0.372 (0.6)	0.116	**0.091** (0.01)	0.231 (11, 0.2)
Swedish leaf	15	500	313	312	128	0.218	0.210	0.138 (0.03)	0.115	**0.096** (0.6)	0.122 (5, 0.2)
50 words	50	450	228	227	270	0.352	0.317	**0.194** (0.1)	0.330	0.216 (0.1)	0.255 (6, 0.1)
Trace	4	100	50	50	275	0.240	**0**	**0** (0.01)	**0**	**0** (0.01)	0.100 (2, 0.2)
Two patterns	4	1000	1000	3000	128	0.09	**0**	**0** (0.01)	0.002	0.003 (0.1)	0.002 (14, 0.1)
Wafer	2	1000	1000	5164	152	0.005	0.004	**0.002** (0.3)	0.023	0.006 (0.1)	0.004 (3, 0.5)
Face (four)	4	24	44	44	350	0.182	0.136	0.136 (0.1)	0.273	0.250 (0.1)	**0.023** (2, 0.1)
Lightning-2	2	60	31	30	637	0.200	**0.100**	**0.100** (0.1)	0.367	0.133 (0.03)	0.167 (4, 0.1)
Lightning-7	7	70	37	36	319	0.472	0.222	**0.200** (0.1)	0.278	0.228 (0.1)	0.277 (5, 0.3)
ECG	2	100	50	50	96	0.180	0.180	**0.140** (0.5)	0.220	0.160 (0.6)	0.16 (2, 0.2)
Adiac	37	390	196	195	176	0.390	0.390	0.364 (0.1)	0.426	**0.333** (0.4)	0.569 (3, 0.1)
Yoga	2	300	1000	2000	426	0.174	0.165	0.165 (0.1)	0.176	0.175 (0.1)	**0.141** (4, 0.1)
Fish	7	75	88	87	463	0.184	0.1379	0.126 (0.01)	0.126	**0.023** (0.1)	0.057 (6, 0.1)
Beef	5	30	15	15	470	0.600	0.600	0.600 (0.2)	0.400	**0.333** (0.1)	0.800 (1, 0.1)
Coffee	2	28	14	14	286	0.200	0.133	0.133 (0.01)	0.071	**0** (0.4)	0.2667 (1, 0.4)
Olive oil	4	30	15	15	570	0.188	**0.188**	**0.188** (0.01)	0.313	0.313 (0.01)	0.857 (1, 0.3)

[a] ED: Euclidean distance, δ: % of sequence length.

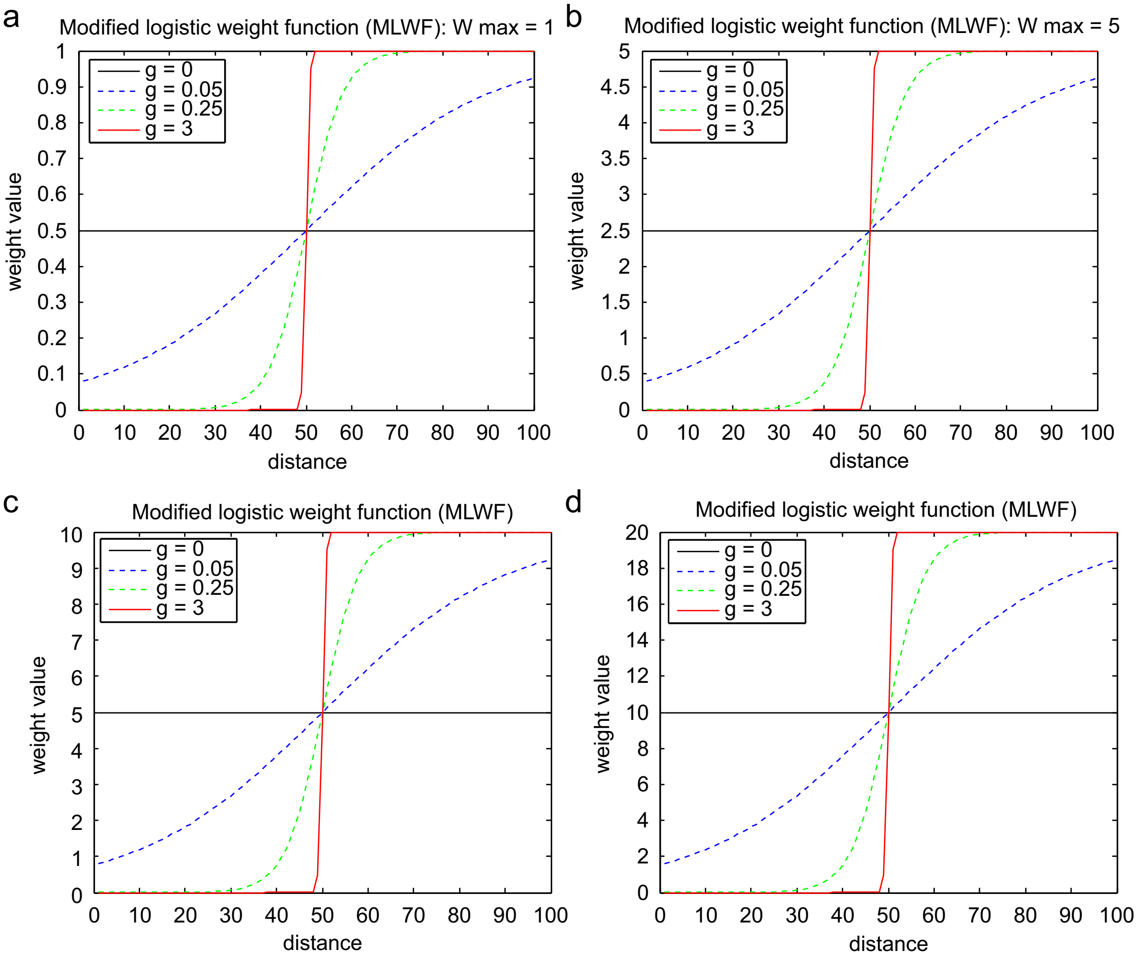

Fig. 10. MLWF with different value w_{max}: (a) $w_{max}=1$, (b) $w_{max}=5$, (c) $w_{max}=10$ and (d) $w_{max}=20$,

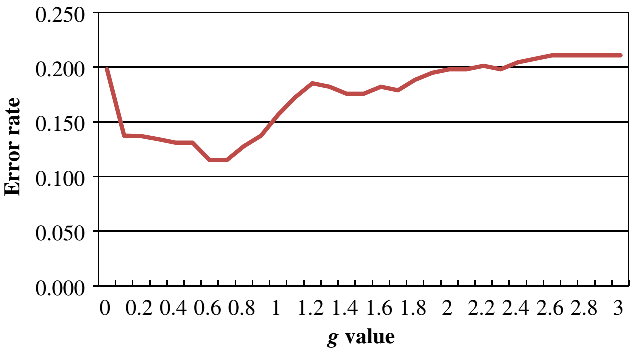

Fig. 11. Effect of g to the error rates of validation data for the "Swedish Leaf" data set.

shown in Table 1. This example indicates that WDTW can adjust the level of penalization of the phase difference on each point by using different g value depending on applications.

5.3. Performance comparison for time series clustering

Since WDTW is essentially a distance measure that can be generally used with different data mining tasks that consider the distance between two observations, we can extend the applications of WDTW to different tasks such as a clustering problem. Following the procedures of several literatures [10,18,25], which presented DTW-based K-means method for time series clustering, we compare the performance of WDTW with that of DTW. As evaluation measures for validating a clustering quality, we used entropy and F-measure for external cluster validity and average within-cluster-distance (the intra-cluster compactness) and average between-cluster-distance (the inter-cluster separation) for internal cluster validity [16,28].

Given data set belonging to I classes and partitioning them into J clusters using clustering algorithms, let n be the size of data set, n_i be the size of class i, n_j be the size of cluster j, and n_{ij} be the number of data belonging to both class i and cluster j. Then, Entropy and F-measure can be calculated as follows [16]:

$$\text{Entropy} = \sum_{j=1}^{J} \frac{n_j}{n}\left(-\sum_{i=1}^{I} P(i,j)\log_2 P(i,j)\right)$$

$$F\text{-measure} = \sum_{i=1}^{I} \frac{n_i}{n} \max_{0<j<J}\left[\frac{2\times R(i,j)\times P(i,j)}{R(i,j)+P(i,j)}\right]$$

where $R(i,j)=n_{ij}/n_i$ and $P(i,j)=n_{ij}/n_j$. The lower the value of entropy, the higher the clustering quality, on the contrary, the higher the value of F-measure, the better the clustering quality. For internal cluster criteria, average within-cluster-distance (d_{ave_within}) and average between-cluster-distance (d_{ave_bet}) are calculated by [10]

$$d_{ave\_within} = \frac{1}{KN_i}\sum_{i=1}^{K}\sum_{j=1}^{N_i} d(C_i,X_j)$$

$$d_{ave\_bet} = \frac{1}{M}\sum_{i=1}^{K}\sum_{j>i}^{K} d(C_i,C_j)$$

Table 2
Summary of clustering performance.

Data name	Number of classes	Data size	Length	External cluster validity						Internal cluster validity					
				Entropy			*F*-measure			Average within-cluster-distance			Average between-cluster-distance		
				ED[a]	DTW	WDTW	ED[a]	DTW	WDTW	ED[a]	DTW	WDTW	ED[a]	DTW	WDTW
Gun-point	2	200	150	1.012	0.999	**0.336**	0.5	0.505	**0.886**	3.989	3.865	**3.797**	7.223	7.384	**7.549**
Trace	4	200	275	1.807	**1.621**	**1.621**	0.482	**0.588**	**0.588**	4.399	**4.391**	4.806	15.969	**18.080**	17.901
Face (four)	4	112	350	0.925	**0.877**	0.916	0.758	**0.797**	0.778	13.566	13.653	**12.108**	11.957	12.021	**16.274**
Lighting 2	2	121	637	0.953	0.943	**0.868**	0.579	0.595	**0.612**	20.112	**18.112**	18.693	8.297	14.335	**16.566**
ECG	2	200	96	0.807	0.807	**0.752**	0.737	0.737	**0.769**	5.809	4.909	**4.461**	2.533	7.523	**8.079**
Beef	5	60	470	1.916	1.917	**1.906**	0.503	0.504	**0.542**	0.394	0.384	**0.354**	1.667	1.878	**2.069**
Coffee	2	56	286	0.891	**0.719**	**0.719**	0.631	**0.773**	**0.773**	35.769	34.817	**32.722**	82.319	79.539	**83.561**
Olive oil	4	60	570	1.319	1.235	**1.214**	0.636	0.669	**0.685**	0.079	0.079	**0.053**	0.126	0.125	**0.183**

[a] ED: Euclidean distance.

where $M=\sum_{m=1}^{K-1} m$ is the number of pairs of cluster centers, $d(C_i,X_j)$ is the distance between time series j in the cluster i and the cluster center of cluster i, and $d(C_i,C_j)$ is the distance between cluster centers of cluster i and cluster j. In addition, K and N_i the number of clusters and the number of items in cluster i, respectively. The smaller the value of average within-cluster-distance, the more compact each cluster, and the bigger the value of average between-cluster-distance, the more separate the clusters.

Table 2 shows the clustering results of 8 data sets out of 20 data sets. The cluster validity measures in Table 2 present the average values of 5 runs with the same data set. As for the value of g for WDTW, we used the selected value in Table 1 instead of optimizing it for a clustering purpose. As shown in Table 2, in most cases, WDTW outperforms both Euclidean distance and DTW even though we did not optimize the value of g for WDTW in terms of both external and internal cluster validity measures. Even though we used only data sets that have either small number of observations or low dimension of an input vector due to the limitation of computational time, similar conclusion can be made for the remaining data sets.

6. Conclusion

A new distance measures for time series data, WDTW and WDDTW, are proposed to classify or cluster time series data set in diverse applications. Compared with the conventional DTW and DDTW, the proposed algorithm weighs each point according to the phase difference between a test point and a reference point. The proposed method is the generalized distance measure of Euclidean distance, DTW, and DDTW, and maximizes its effectiveness with optimal g value depending on different applications. A new weighting function, called modified logistic weight function, is developed to systematically assign weights depending on the distance between time series points.

The extensive experimental results using public data sets from diverse applications indicate that WDTW and WDDTW with optimal weights have great potential for improving the accuracy for time series classification and clustering. As a part of future research, our proposed algorithm could be combined with some of the pruning techniques such as LB_Keogh and warping-window-DTW to reduce computational time for more massive time series data sets.

Acknowledgements

The authors acknowledge the support of Dr. Eamonn Keogh in providing us the experimental data set. Also, the authors would like to thank the anonymous reviewers for their valuable comments that improved our paper dramatically. The part of this work was supported by the National Science Foundation (NSF) Grant no. CMMI-0853894. Dr. Olufemi A. Omitaomu acts in his own independent capacity and not on behalf of UT-Battelle, LLC, or its affiliates or successors.

Appendix

Proof of Proposition 1

By classical analysis of l_p spaces [3, pp. 181–186], for $0<p<q\le\infty$, we obtain that $||\mathbf{x}||_p \ge ||\mathbf{x}||_q$ where $\mathbf{x}$ is a sequence. Let $\mathbf{a}$ and $\mathbf{b}$ denote two sequences with the same length, respectively. Given the two aligned sequences $\mathbf{a}^*$ and $\mathbf{b}^*$, it is true $||\mathbf{a}^*-\mathbf{b}^*||_p \ge ||\mathbf{a}^*-\mathbf{b}^*||_q$, so $||\mathbf{w}(\mathbf{a}^*-\mathbf{b}^*)||_p \ge ||\mathbf{w}(\mathbf{a}^*-\mathbf{b}^*)||_q$ due to $\mathbf{w}>0$. Therefore, $\mathrm{WDTW}_p(\mathbf{a}^*,\mathbf{b}^*) \ge \mathrm{WDTW}_q(\mathbf{a}^*,\mathbf{b}^*)$.

Proof of Proposition 2

By classical analysis of l_p spaces [3, pp. 181–186], given $\mathbf{x}$ sequence with n length, $||\mathbf{x}||_p \le |(n)^{(1/p)-(1/q)}||\mathbf{x}||_q$ for $0<p<q\le\infty$. In addition, the length of a minimal warping path in DTW is at most $2n-2$ when $n>1$ [15]. Given the two aligned sequences $\mathbf{a}^*$ and $\mathbf{b}^*$, it is true that $||\mathbf{a}^*-\mathbf{b}^*||_p \le (2n-2)^{(1/p)-(1/q)}||\ \mathbf{a}^*-\mathbf{b}^*||_q \le .$ Thus, $||\mathbf{w}(\mathbf{a}^*-\mathbf{b}^*)||_p \le (2n-2)^{(1/p)-(1/q)}||\mathbf{w}(\mathbf{a}^*-\mathbf{b}^*)||_q$ due to $\mathbf{w}>0$. Therefore, $\mathrm{WDTW}_p(\mathbf{a}^*,\mathbf{b}^*) \le (2n-2)^{(1/p)-(1/q)}\ \mathrm{WDTW}_q(\mathbf{a}^*,\mathbf{b}^*)$.

References

[1] C.D. Dietrich, G. Palm, K. Riede, F. Schwenker, Classification of bioacoustic time series based on the combination of global and local decision, Pattern Recognition 37 (2004) 2293–2305.

[2] D. Eads, D. Hill, S. Davis, S. Perkins, J. Ma, R. Porter, J. Theiler, Genetic algorithms and support vector machines for time series classification, Proceeding SPIE 4787 (2002) 74–85.

[3] G.B. Folland, Real Analysis. Modern Techniques and their Applications, Wiley, New York, 1999.

[4] I. Guler, E.D. Ubeyli, Adaptive neuro-fuzzy inference system for classification of EEG signals using wavelet coefficient, Journal of Neuroscience Methods 148 (2005) 113–121.

[5] F. Gullo, G. Ponti, A. Tagarelli, S. Greco, A time series representation model for accurate and fast similarity detection, Pattern Recognition 42 (2009) 2998–3014.

[6] M. Husken, P. Stagge, Recurrent neural networks for time series classification, Neurocomputing 50 (2003) 223–235.

[7] F. Itakura, Minimum prediction residual principle applied to speech recognition, in: Proceedings of the IEEE Transactions on Acoustics, Speech, and Signal, 1975, pp. 52–72.

[8] A.C. Jalba, M.H.F. Wilkinson, J.B.T.M. Roerdink, M.M. Bayer, S. Juggins, Automatic diatom identification using contour analysis by morphological curvature scale spaces, Machine Vision and Applications 16 (4) (2005) 217–228.

[9] Y.S. Jeong, S.J. Kim, M.K. Jeong, Automatic identification of defect patterns in semiconductor wafer maps using spatial correlogram and dynamic time warping, IEEE Transactions on Semiconductor Manufacturing 21 (2008) 625–637.
[10] E. Keogh, J. Lin, Clustering of time series subsequences is meaningless: implications for previous and future research, Knowledge and Information Systems 8 (2005) 154–177.
[11] E. Keogh, M. Pazzani, Derivative dynamic time warping, in: Proceedings of the SIAM International Conference on Data Mining, Chicago, 2001.
[12] E. Keogh, C.A. Ratanamahatana, Exact indexing of dynamic time warping, Knowledge and Information Systems 3 (2005) 358–386.
[13] E. Keogh, X. Xi, L. Wei, C.A. Ratanamahatana, The UCR Time Series Data Mining Archive. Available at: 〈http://www.cs.ucr.edu/~eamonn/time_series_data〉, 2006.
[14] D.J. Lee, R. Schoenberger, D. Shiozawa, X. Xu, P. Zhan, Contour matching for a fish recognition and migration monitoring system, in; Proceedings of the SPIE Optics East, Two and Three-Dimensional Vision Systems for Inspection, Control, and Metrology II, 5606-05, Philadelphia, PA, 2004, pp. 37–48.
[15] D. Lemire, Faster retrieval with a two-pass dynamic-time-warping lower bound, Pattern Recognition 42 (2009) 2169–2180.
[16] Y. Lu, Y. Ouyang, H. Sheng, Z. Xiong, An incremental algorithm for clustering search results, in: Proceedings of the IEEE International Conference on Signal Image Technology and Internet Based Systems, 2008.
[17] M.D. Morse, J.M. Patel, An efficient and accurate method for evaluating time series similarity, in: Proceedings of the ACM SIGMOD International on Information and Knowledge Management, 2006, pp. 14–23.
[18] V. Nieeattrakul, C. Ratanamahatana, On clustering multimedia time series data using *K*-means and dynamic time warping, in: Proceedings of the IEEE International Conference on Multimedia and Ubiquitous Engineering, 2007.
[19] D.T. Pham, A.B. Chen, Control chart pattern recognition using a new type of self-organizing neural network, Journal of Systems and Control Engineering 112 (1998) 115–127.
[20] C.A. Ratanamahatana, E. Keogh, Making time-series classification more accurate using learned constraints, in: Proceeding of the Fourth SLAM International Conference on Data Mining, 2004.
[21] T.M. Rath, R. Manmatha, Word image matching using dynamic time warping, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003.
[22] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, IEEE Transactions on Acoustics, Speech, and Signal Process (1978) 43–49.
[23] E.D. Ubeyli, Wavelet/mixture of experts network structure of ECG signals classification, Expert Systems with Applications 34 (2008) 1954–1962.
[24] M. Vlachos, G. Kollios, D. Gunopulos, Discovering similar multidimensional trajectories, in: Proceeding of the International Conference Data Engineering, 2002.
[25] F. Yu, K. Dong, F. Chen, Y. Jiang, W. Zeng, Clustering time series with granular dynamic time warping method, in; Proceedings of the IEEE International Conference on Granular Computing, 2007.
[26] X. Xi, E. Keogh, L. Wei, C.A. Ratanamahatana, Fast time series classification using numerosity reduction, in: Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA, 2006.
[27] Y. Xie, B. Wiltgen, Adaptive feature based dynamic time warping, International Journal of Computer Science and Network Security 10 (2010) 264–273.
[28] W. Zhao, E. Serpedin, E.R. Dougherty, Spectral preprocessing for clustering time-series gene expressions, EURASIP Journal on Bioinformatics and Systems Biology (2009) 1–10.

Young-Seon Jeong is now working toward his Ph.D. degree in the Department of Industrial and Systems Engineering, Rutgers University, New Brunswick, NJ. His research interests include spatial modeling of wafer map data, wavelet application for functional data analysis, and statistical modeling for intelligent transportation system

Myong K. Jeong is an Assistant Professor in the Department of Industrial and Systems Engineering and the Center for Operation Research, Rutgers University, New Brunswick, NJ. His research interests include statistical data mining, recommendation systems, machine health monitoring, and sensor data analysis. He is currently an Associate Editor of *IEEE Transactions on Automation Science and Engineering and International Journal of Quality, Statistics and Reliability.*

Olufemi A. Omitaomu is a Research Scientist at Geographic Information Science & Technology Group, Computational Sciences and Engineering Division in Oak Ridge National Laboratory Oak Ridge, TN. He is also an Adjunct Assistant Professor at Department of Industrial and Information Engineering in University of Tennessee, Knoxville, TN. His research areas include streaming and real-time data mining, signal processing, optimization techniques in data mining, infrastructure modeling and analysis, and disaster risk analysis in space and time.