Adaptive Feature Based Dynamic Time Warping

Ying Xie[†] and Bryan Wiltgen [††]

[†]Department of Computer Science and Information Systems
Kennesaw State University, Georgia, USA

[††] College of Computing
Georgia Institute of Technology, Georgia, USA

Summary

Dynamic time warping (DTW) has been widely used in various pattern recognition and time series data mining applications. However, as examples will illustrate, both the classic DTW and its later alternative, derivative DTW, may fail to align a pair of sequences on their common trends or patterns. Furthermore, the learning capability of any supervised learning algorithm based on classic/derivative DTW is very limited. In order to capture trends or patterns that a sequence presents during the alignment process, we first derive a global feature and a local feature for each point in a sequence. Then, a method called feature based dynamic time warping (FBDTW) is designed to align two sequences based on each point's local and global features instead of its value or derivative. Experimental study shows that FDBTW outperforms both classic DTW and derivative DTW on pairwise distance evaluation of time series sequences. In order to enhance the capacity of supervised learning based on DTW, we further design a method called adaptive feature based dynamic time warping (AFDBTW) by equipping the FDBTW with a novel feature selection algorithm. This feature selection algorithm is able to expand the learning capability of any DTW based supervised learning algorithm by a dual learning process. The first-fold learning process learns the significances of both the local feature and global feature towards classification; then the second-fold learning process learns a classification model based on the pairwise distances generated by the AFDBTW. A comprehensive experimental study shows that the AFDBTW is able to make further improvement over the FDBTW in time series classification.

Key words:

Dynamic time warping, DTW, Feature based DTW; Adaptive Feature based DTW; Time series classification; Pattern recognition; Data mining; Machine learning; Information retrieval

1. Introduction

As an algorithm for measuring similarity between time series sequences, Dynamic Time Warping (DTW) has been widely used in various pattern recognition applications, such as speech recognition [3, 9], handwriting recognition [1], gesture recognition [2], signature recognition [20], ECG pattern recognition[21], shape recognition [7] and others. Due to the huge amount of time series data that has been accumulated in different domains such as finance, manufacturing, process engineering, medicine, molecular biology, physics, and chemistry, recent years have also seen increasing interest in applying DTW to time series data mining. The involved data mining tasks include, but are not limited to, clustering [4, 22], classification [23, 24], association mining [25], and motif discovery [8, 26].

Unlike Linear Time Warping (LTW), which compares two sequences based on a linear match of the two temporal dimensions, DTW uses dynamic programming to search a space of mapping between the time axes of the two sequences in order to determine the minimum distance between them. Typically, certain constraints are imposed on DTW to optimize and expedite the search of the warping path. Major constraints outlined in [9] include monotonic condition, continuity condition, boundary condition, adjustment window condition, and slope constraint condition.

More formally, given two time series sequences R and Q as follows: $R = r_1 r_2 r_3 \ldots r_i \ldots r_M$, and $Q = q_1 q_2 q_3 \ldots q_j \ldots q_N$, DTW finds an optimal warping path between R and Q by using dynamic programming to calculate the minimal cumulative distance $\gamma(M,N)$, where $\gamma(i,j)$ is recursively defined as:

$$\gamma(i, j) = d(r_i, q_j) + \min(\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)) \quad (1)$$

As can be seen from formula 1, given a search space defined by two time series sequences and a set of constraints, DTW guarantees to find the warping path with the minimum cumulative distance among all possible

Manuscript received January 5, 2010
Manuscript revised January 20, 2010

warping paths that are valid in the search space. Furthermore, according to formula 1, the determinant factor for $\gamma(M,N)$ is all the $d(r_i, q_j)$'s, where $d(r_i, q_j)$ represents the distance between the data point r_i ($1 \leq i \leq M$) from the sequence R and the data point q_j ($1 \leq j \leq N$) from the sequence Q. In most situations, a data point in a time series sequence is a numerical value, so $d(r_i, q_j)$ is typically defined as either $| r_i - q_j |$ or $(r_i - q_j)^2$. In this paper, we refer to this type of classic DTW as *value based DTW*.

The fundamental problem of value based DTW is that the numerical value of a data point in a time series sequence is not the complete picture of the data point in relation to rest of the sequence. We will show in section 2 that, when a data point in a sequence is compared with another point in another sequence, its position in the sequence and relation to its neighbors should also be taken into consideration. In [5], a *derivative* DTW was proposed that replaces the value of each data point with its first derivation in the process of dynamic time warping. The derivation of a data point can be viewed as a local feature of the point that expresses its relationship with two adjacent neighbors. However, as will also be illustrated in section 2, only considering derivations in comparison may make derivative DTW lose sight of the overall shapes or significant features that occur in the involved sequences.

Based upon these observations of the essential problems of value based DTW and derivative DTW, we propose in this paper a novel approach called Feature Based Dynamic Time Warping (FBDTW) as a better technique for evaluating the similarity between two given time series sequences. When comparing two points coming from each of the two sequences in the process of dynamic time warping, FBDTW takes into consideration both the local and global features of the two points. By doing this, our algorithm gains a vision of not only the overall shapes of the sequence but also the local trend around the points. Experimental studies on the UCR time series classification/clustering test bed [6] with twenty different time series data sets show that FBDTW outperforms both value based (the classic) DTW and derivative DTW.

The second contribution presented in this paper is the enhancement of the supervised learning capacity of DTW through a learning algorithm. It is well known that time series classification has numerous important applications in different domains. Although a wide range of time series classification algorithms were proposed in the past decade, X. Xi, E. Keogh, C. Shelton, and L. Wei [10] claimed based on their experimental studies that the combination of one-nearest-neighbor (1-NN) with Dynamic Time Warping (DTW) distance "has proven exceptionally difficult to beat". Nevertheless, despite its superior performance over other alternatives, the combination of 1NN and DTW has limited learning capacity. In other words, in this combination, the pairwise distance evaluated by DTW is domain and application independent. In the study of the proposed FBDTW, we found that the pairwise distance between two time series sequences may be domain or application dependent. For instance, in some domains or applications, time series sequences are classified primarily based on the global trends of the sequences; while in others, the local trends of the sequences may carry more weights. Therefore, a learning capacity should be equipped to a time series classification approach to learn an optimized way to calculate pairwise distances from the training data. The proposed FBDTW, which aligns sequences based on both the local feature and global feature of each point, provides an excellent instrument for such an adaptive distance measure. The accumulative effect of the local features of all points in a sequence reflects local trends of that sequence; whereas the accumulative effect of the global features of points in a sequence reflects the global trend of the same sequence. Hence, we design the adaptive FBDTW (AFBDTW) where the contributions of global features and location features are leveraged by weighting factors. The weighting factors are learned from the training data by a newly designed feature selection algorithm. The AFBDTW, therefore, enhance the capacity of supervised learning for time series data, such that the combination of 1-NN and AFBDTW contains a dual learning process. The first-fold learning derives an optimized pairwise distance function for time series data; then, the second-fold learning is carried by 1-NN based on the learned distance. Our experimental study shows that the enhancement of learning capacity brought by AFBDTW makes further improvement on the classification accuracy for time series data.

The rest of the paper will be organized as follows. In section 2, we study the limitations of value based DTW and derivative DTW. Subsequently, our proposed FBDTW algorithm is presented in section 3. Next, in section 4, we describe the AFBDTW and the corresponding feature selection algorithm. In section 5 we conduct comprehensive experimental and comparative studies on AFBDTW, FBDTW, value based DTW, and derivative DTW. The time complexity of FBDTW and AFBDTW is given in section 6. Finally, we conclude our contributions and envision further development on FBDTW in section 7.

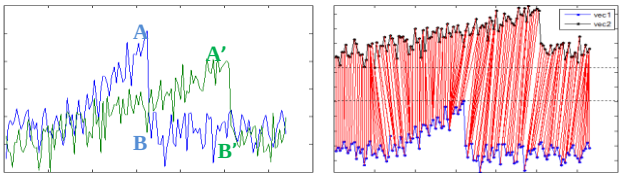
2. Limitation of Value Based DTW and Derivative DTW

In this section we show that both value based DTW and derivative DTW may lose sight of overall shapes of the involved sequences. First, Figure 1(a) presents two time series sequences that develop similar trends at almost the same pace. These two sequences are the first one third of the two sequences that belong to the same class of a data

set called Beef that is one of the UCR time series data sets [6]. Intuitively, little time warping is needed when aligning these two sequences. However, as shown in Figure 1(b), value based DTW maps almost the whole first sequence (the one with the highest peak) to one single point denoted as P in the second sequence. This alignment most certainly does not have a positive impact on the similarity evaluation of these two sequences. The reason why value based DTW generates this abnormal alignment is simply because P is the closest point of the second sequence towards any point in the first sequence in terms of value. In other words, this pure value-oriented comparison makes value based DTW ignore the context of points, such as their positions in local features and their relations to overall trends. One may ask if normalization of these two sequences could solve this problem. Figure 1(C) shows the alignment result after normalization of these two sequences. The problem is lessened a little but fundamentally still exists, i.e., the alignment is blind to the common trends developed by both sequences. Better alignments of these two sequences by methods proposed in this paper can be seen in figure 3.

Derivative DTW was proposed in [5] to remedy the weakness of value oriented mapping. However, the following example will illustrate that derivative oriented comparison may also neglect significant features of the involved sequences. The two time series sequences shown in Figure 2(a) belong to the same class of a data set called CBF that is one of UCR time series data sets [6]. These two sequences share a common feature, which is a significant drop of value from point A to B in the first sequence or from A' to B' in the second sequence. An ideal time warping would match the point A to A' and B to B'. However, this significant common feature is not detected by derivative DTW, which generates the alignment shown in Figure 2(b). Better alignments of these two sequences by methods proposed in this paper can be seen in figure 4.

These two examples suggest that in order to be able to identify and match common trends and patterns presented by a pair of sequences in the warping process, more features are needed to describe each point rather than just using pure value or only the first derivative.



a. Two time series sequences from the same class of CBF

b. Alignment generated by derivative DTW

Figure 2 visualizing the limitation of derivative DTW

3. Feature Based Dynamic Time Warping

Given two time series sequences R and Q as follows: $R = r_1r_2r_3...r_i...r_M$, and $Q = q_1q_2q_3...q_j...q_N$. A $N \times M$ Matrix is created to find an optimal warping path by using dynamic programming. The node (i,j) of the matrix is assigned with the distance between the data point r_i and q_j, which is denoted as $dist(r_i, q_j)$. By the FBDTW algorithm, $dist(r_i, q_j)$ is evaluated based upon both the local and global features of r_i and q_j.

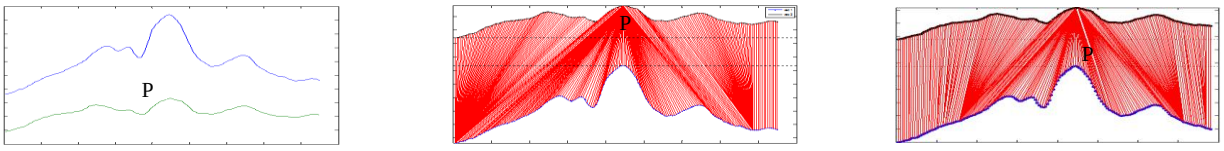
3.1 Local Feature of a Data Point

The local feature of the data point r_i, which is denoted as $f_{local}(r_i)$, is defined as a vector of two components: $f_{local}(r_i) = (r_i - r_{i-1}, r_i - r_{i+1})$. We feel that this definition can better reflect the local trend on the point r_i than the first derivation of r_i used in [5], which is defined as a single value $Dev(r_i) = ((r_i - r_{i-1}) + (r_{i+1} - r_{i-1})/2)/2$. For example, consider the following two groups of curves, where each curve has 3 points. All the middle points within each group have the same deviation despite the fact that the local trends on them are very different. By using our definition of local features, the different trends related to the middle points can be correctly expressed.

Group 1: (1, 5, 3) vs. (1, 3.5, 6) vs. (8/3, 6, 6)
Group 2: (3, 1, 7) vs. (1, 1, 1) vs. (1, 3, -3)

3.2 Global Feature of a Data Point

The global feature of a data point in the given sequence



a. Two time series sequences with similar trends

b. Alignment generated by value based DTW

c. Alignment generated by value based DTW after normalization

Figure 1 visualizing the limitation of value based DTW

should reflect the position of that point in the global shape of the sequence. As illustrated in section 2, the derivation of a data point contains no global information of the data point. The value of a data point can be viewed as a global feature; however it may not be in the same scale as the components of the local feature, so as to make it difficult to combine the global and local features. In this paper, we define the global feature of a data point r_i in a sequence $R = r_1r_2r_3...r_i...r_M$ as a vector of two components:

$f_{global}(r_i) = (r_i - \sum_{k=1}^{i-1} r_k/(i-1), r_i - \sum_{k=i+1}^{M} r_k/(M-i))$. That is, the first component of the vector is the difference between the value of r_i and the average value of the first *i-1* points in the sequence *R*; while the second component of the vector is the difference between the value of r_i and the average value of the last *M-i* points in *R*.

3.3 Evaluation of dist(r_i, q_j)

Based on the global feature and local feature we defined in section III. A and III. B, a point *p* is described by two vectors $f_{local}(p)$ and $f_{global}(p)$. Given two time series sequences *R* and *Q* as follows: $R = r_1r_2r_3...r_i...r_M$, and $Q = q_1q_2q_3...q_j...q_N$, we define the distance between the point r_i and q_j as follows:

$$dist(r_i, q_j) = dist_{local}(r_i, q_j) + dist_{global}(r_i, q_j), \quad (2)$$

where $dist(r_i, q_j)$ is the overall distance between r_i and q_j, $dist_{local}(r_i, q_j)$ is the distance between r_i and q_j based on their local features, and $dist_{global}(r_i, q_j)$ is the distance between r_i and q_j based on their local features. We further design two methods to evaluate both $dist_{local}(r_i, q_j)$ and $dist_{global}(r_i, q_j)$.

By method 1, we have the following:

- $dist_{local}(r_i, q_j) = |(f_{local}(r_i))_1 - (f_{local}(q_j))_1| + |(f_{local}(r_i))_2 - (f_{local}(q_j))_2|$ (3.1)
- $dist_{global}(r_i, q_j) = |(f_{global}(r_i))_1 - (f_{global}(q_j))_1| + |(f_{global}(r_i))_2 - (f_{global}(q_j))_2|$ (3.2)

where $\vec{v}_i$ represents the i^{th} component of vector $\vec{v}$.

Method 2 uses vector operations to calculate local and global distances, where

- $dist_{local}(r_i, q_j) = |f_{local}(r_i) - f_{local}(q_j)|$ (4.1)
- $dist_{global}(r_i, q_j) = |f_{global}(r_i) - f_{global}(q_j)|$ (4.2)

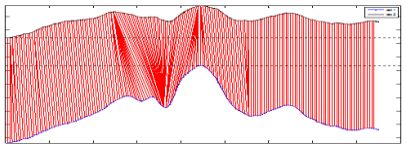
The DTW algorithm where the $dist(r_i, q_j)$ is evaluated based on method 1 is called *Feature Based DTW with Distance Function 1* (FBDTW1 for short), and the DTW algorithm where $dist(r_i, q_j)$ is evaluated based on method two is called the *Feature Based DTW with Distance Function 2* (FBDTW2 for short). As with value based DTW and derivative DTW, both FBDTW1 and FBDTW2 find an optimal warping path between *R* and *Q* by using dynamic programming to calculate the minimal cumulative distance $\gamma(M,N)$, where $\gamma(i,j)$ is recursively defined as is recursively defined in equation 1. Finally, the distance between sequence *R* and sequence *Q* is expressed as $\gamma(M,N)/(M+N)$, where *M* and *N* are sizes of *R* and *Q* respectively. Please note that the local feature and global feature have no definition for the first and last points in a sequence, therefore, both FBDTW1 and FBDTW2 calculate the optimal warping path starting with the second points of the two sequences and ending at their penultimate points. The time complexity of FBDTW is the same as value based DTW and derivative DTW, which is O(*MN*).

3.4 Visually Comparing FBDTW with Value Based DTW and Derivative DTW

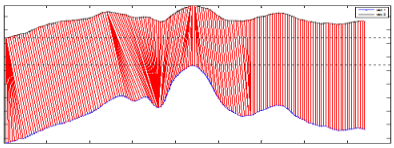
First we visually show that that FBDTW1 and FBDTW2 are able to remedy both the problem caused by value based DTW on the two sequences shown in Figure 1(a), and the problem caused by derivative DTW on the two sequences shown in Figure 2(a). As Figure 3 presents, both FBDTW1 and FBDTW2 align the two sequences shown in Figure 1(a) along their common track of feature development in general. Furthermore, as Figure 4 shows, both FBDTW1 and FBDTW2 are able to detect and match the common significant features embedded in the two sequences shown in Figure 2(a).

Next, we visually compare these four DTW methods on another pair of time series sequences from a data set called Wafer, which is also part of the UCR time series classification and clustering test bed. These two sequences are in different type than the sequences shown in Figure 1(a) & 2(a). As illustrated in Figure 5(b), the stable line parts of these two sequences are similar to each other, whereas the unstable parts of the two sequences are the major source of their dissimilarity. However, both value based DTW and especially derivative DTW generate two large singularities on the straight line part, as shown in Figure 5(c) & (d). That is, by these two methods, a large group of consecutive data points from one sequence match with one single point from the other sequence. This example shows again that value based or derivative DTW may have the tendency of overlooking overall shapes or global features of the involved sequences. On the contrary, both the proposed FBDTW1 and FBDTW2 generate more reasonable warping results by matching the stable line part

of the first sequence to the stable line part of the second sequence, as shown in Figure 5(e) and (f).

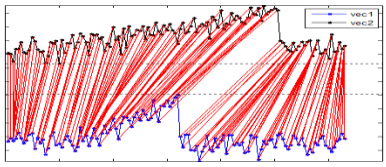


a. Alignment generated by FBDTW1

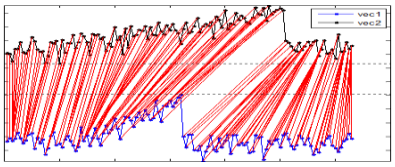


b. Alignment generated by FBDTW2

Figure 3 Alignments generated by FBDTW1&2 on the two sequences shown in Figure 1(a)



a. Alignment generated by FBDTW1



b. Alignment generated by FBDTW2

Figure 4 Alignments generated by FBDTW1&2 on the two sequences shown in Figure 2(a)

4. Feature Based Dynamic Time Warping

The experimental studies showed that the combination of one-nearest-neighbor (1-NN) with Dynamic Time Warping (DTW) distance "has proven exceptionally difficult to beat" [10]. Nevertheless, despite its superior performance over other alternatives, the combination of 1NN and DTW has limited learning capacity. In other words, in this combination, the pairwise distance evaluated by DTW is domain and application independent; the only learning ability comes from 1NN. In the study of the proposed FBDTW, we found that the pairwise distance between two time series sequences may be domain or application dependent. In other words, in some domains or applications, time series sequences may be classified primarily based on the global trends of the sequences; while in others, the local trends of the sequences may carry more weights. Taking one of the UCR time series data sets *Synthetic Control* as example, if we conduct dynamic wrapping solely based on local features of points, the classification accuracy is only around 50%; whereas the accuracy rate is above 90% if dynamic warping is based on only global features of points. This example implies that the classification labels were assigned to training sequences much more based on global trends of sequences than their local features. Conversely, for the UCR time series data set *Coffee*, the classification solely based on local features leads to accuracy rate close to 90%; whereas classification solely based on global features leads to accuracy rate only close to 80%. This implies that, for this data set, local feature of sequences are more important factors for classification.

Therefore, a learning capacity should be equipped to a time series classification approach to learn an optimized way to calculate pairwise distances from the training data. The proposed FBDTW, which aligns sequences based on both the local feature and global feature of each point, provides an excellent instrument for such an adaptive distance measure. The accumulative effect of the local features of points in a sequence reflects local trends of that sequence; whereas the accumulative effect of the global features of points in a sequence reflects the global trend of the same sequence. Hence, we design the adaptive FBDTW (AFBDTW) where the contributions of global features and local features are leveraged by weighting factors.

More specifically, given two time series sequences R and Q as follows: $R = r_1 r_2 r_3 \ldots r_i \ldots r_M$, and $Q = q_1 q_2 q_3 \ldots q_j \ldots q_N$, we define the adaptive distance between the point r_i and q_j as follows:

$$dist(r_i, q_j) = w_1 \cdot dist_{local}(r_i, q_j) + w_2 \cdot dist_{global}(r_i, q_j), \qquad (5)$$

where $dist(r_i, q_j)$ is the overall distance between r_i and q_j; $dist_{local}(r_i, q_j)$ is the distance between r_i and q_j based on their local features; $dist_{global}(r_i, q_j)$ is the distance between r_i and q_j based on their local features; and $w_1 + w_2 = 1$, $0 \leq w_1 \leq 1$, $0 \leq w_2 \leq 1$.

Then, the AFBDTW find an optimal warping path between R and Q by using dynamic programming to calculate the minimal cumulative distance $\gamma(M,N)$, where $\gamma(i,j)$ is recursively defined in equation 1. Finally, the distance between sequence R and sequence Q is expressed as $\gamma(M,N)/(M+N)$, where M and N are sizes of R and Q respectively.

Given that two methods were designed to evaluate both $dist_{local}(r_i, q_j)$ and $dist_{global}(r_i, q_j)$ in section 3.C, we denote the AFBDTW that uses the first method (equation 3.1 &

3.2) as AFBDTW1, and the AFBDTW that uses the second method (equation 4.1 & 4.2) as AFBDTW2.

Comparing equation 5 that is used by AFBDTW with equation 2 that is used by FBDTW, we can see that both AFBDTW and FBDTW take advantages of the local feature and the global feature of each point; however, AFBDTW further leverages the contributions of global features and local features by using weighting factors. The weighting factors used in equation 5 can be learned from the training data, which makes the evaluation of distances between time series sequences no longer domains and applications irrelevant. We design the following algorithms called *In-Class-Range Weighting Algorithm* to learn the weighting factors w_1 and w_2 from the training data. This algorithm evaluates the distinguishability of the local feature and the global feature one at a time by setting the corresponding weighting factor in equation 5 to be 1 and the other weighting factor to be 0. The algorithm defines an in-class range for each sequence in the training set as the distance between this sequence and the farthest sequence in the same class. Then, for each training sequence, it calculates the difference between the number of same-class sequences within the in-class range and the number of different-class sequences within the in-class range. Finally, the value of the normalized accumulated differences among all the training sequences is used as the value of the weighting factor for the corresponding feature. The algorithm is presented in details as follows.

Algorithm: *In-Class-Range Weighting Algorithm*
Input: training data set consisting of
- A set of time series sequences $\boldsymbol{S} = \{S_1, S_2, \ldots, S_n\}$
- A set of class labels $\boldsymbol{C} = \{c_1, c_2, \ldots, c_m\}$
- A mapping from $\boldsymbol{S}$ to $\boldsymbol{C}$.

Output: w_1 and w_2

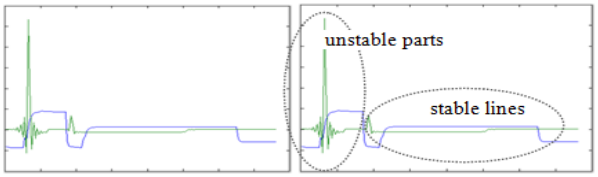
for each $w_{i\ (i = 1 \text{ or } 2)}$ in { w_1, w_2} //i.e., i=1 for the first iteration, and =2 for the second iteration.
set $w_i = 1$ and $w_{j\ (i \neq j,\ j=1 \text{ or } 2)} = 0$; //i.e., j=2 for the first iteration, and =1 for the second iteration.
for any two sequences S_x, S_y $(x \neq y)$ in $\boldsymbol{S}$
use AFBDTW to calculate the distance between S_x and S_y by using equation 5.
end for
for each sequence S_x in $\boldsymbol{S}$
let $inClass_x$ store all the sequences in $\boldsymbol{S}$ that have the same class label as S_x.
calculate $maxDistInClass_x$, which is the maximum of all the distances between S_x and a sequence in $inClass_x$
calculate $numSameClass_x$, which is the total number of sequences with the same class label as Si
calculate $numDiffClassInRange_x$, which is the total number of sequences with different class label than Si and with distance to S_x smaller than or equal to $maxDistInClass_x$.

$$w_i = \sum_{x=1}^{n} (numSameClass_x - numDiffClassInRange_x)$$

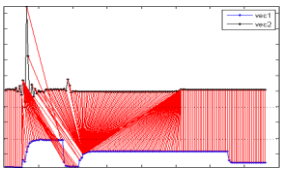
end for
end for
normalize(w_1, w_2).

Procedure: *normalize* (w_1, w_2)

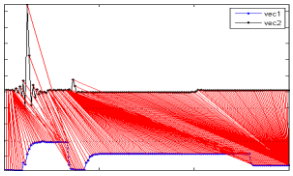
if (w_1>0 && w_2 >0) $w_1 = w_1 / (w_1 + w_2)$;
$w_2 = w_2 / (w_1 + w_2)$
else if (w_1>0 && w_2 <=0) $w_1 = 1$; $w_2 = 0$;
else if (w_1 <=0 && w_2 >0) $w_1 = 0$; $w_2 = 1$;
else if (w_1 < 0 && w_2 < 0) $w_1 = -w_2 / -(w_1 + w_2)$,
$w_2 = -w_1 / -(w_1 + w_2)$
else $w_1 = w_2 = 0.5$;



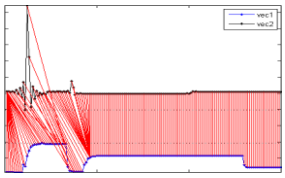
a & b. Two sequences from Wafer



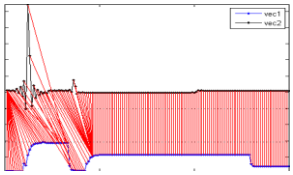
c. Alignment by value based DTW



d. Alignment by derivative DTW



e. Alignment by FBDTW1



f. Alignment by FBDTW2

Figure 5 Visually compare value based DTW, derivative DTW, FBDTW1&2 based on two curves from Wafer.

5. Experimental Results

In order to test the effectiveness of applying FBDTW to evaluate the similarity between time series sequences as well as the capacity of AFBDTW in time series supervised learning, we used all the 20 data sets published on UCR Time Series Classification/Clustering Page (as of

12/28/2009) as our test bed [6]. These 20 data sets contain time series data in different domains, such as electrocardiogram, control chart, microelectronics fabrication, video surveillance, and various contour data [6, 27]. Each data set is divided into training set and test set. Some characteristics of the 20 data sets, which are copied from the UCR webpage are listed in table 1 for easy reference.

Table 1. Characteristics of the 20 Data Sets (directly from [6])

Data Set	# of Classes	Training size	Testing size	Time Series Length
50words	50	450	455	270
Adiac	37	390	391	176
Beef	5	30	30	470
CBF	3	30	900	128
Coffee	2	28	28	286
ECG200	2	100	100	96
FaceAll	14	560	1690	131
FaceFour	24	24	88	350
Fish	7	175	175	463
Gun_Point	2	50	150	150
Lighting2	2	60	61	637
Lighting7	7	70	73	319
OliveOil	4	30	30	570
OSULeaf	6	200	242	427
SwedishLeaf	15	500	6.25	128
Synthetic Control	6	300	300	60
Trace	4	100	100	275
Two_Patterns	4	1000	4000	128
wafer	2	1000	6174	152
yoga	2	300	3000	426

1-NN Classification algorithms are implemented on the following DTW algorithms: value based DTW, FBDTW1 & FBDTW2, and AFBDTW1 & AFBDTW2 (Since 1NN + derivative DTW has much worse performance than 1NN + value based DTW on most of the 20 data sets based on our experimental results, we don't include it in our further comparison). We use the accuracy rate of the classification results as the performance measure. The accuracy rate for 1NN + value based DTW is calculated by 1- *error_rate*, where *error_rate* for 1NN + value based DTW is directly obtained from UCR Time Series Classification/Clustering web page.

The experimental results on each data set are recorded in table 1 (in this table, we denote value based DTW as DTW for simplicity). We have the following observations on the experimental results:

1) All of the proposed methods including FBDTW 1&2 and AFBDTW 1&2 get better results on majority of the 20 data sets than the value based DTW.

2) FBDTW1 gains better results on 14 out of 20 data sets over the value based DTW; ties with value based DTW on 3 data sets; and gets worse results on 3 data sets. Among the 14 data sets where FBDTW1 gains improvements, there are 9 data sets with accuracy improvement great than 5 percent; 5 data sets with accuracy improvement greater than 10 percent; 2 data sets with accuracy improvement greater than 20 percent; and 1 data set with accuracy improvement greater than 30 percent.

3) AFBDTW1 makes further improvement over FBDTW1 on 8 data sets, ties with FDBTW1 on 10 data sets, and gets worse results than FBDTW1 on 2 data sets.

4) FBDTW2 gains better results on 12 out of 20 data sets over the value based DTW; ties with value based DTW on 2 data sets; and gets worse results on 6 data sets.

	Accuracy Rate of Classification								
Data Set	1NN + DTW	1NN + FBDTW1	Accuracy Improved	1NN + FBDTW2	Accuracy Improved	1NN + AFBDTW1	Accuracy Improved	1NN + AFBDTW2	Accuracy Improved
50words	0.69	0.787	14.06%	0.802	16.23%	0.787	14.06%	0.807	16.96%
Adiac	0.604	0.657	8.77%	0.683	13.08%	0.66	9.27%	0.683	13.08%
Beef	0.5	0.667	33.40%	0.633	26.60%	0.667	33.40%	0.633	26.60%
CBF	0.997	0.9	-9.73%	0.919	-7.82%	0.996	-0.10%	0.979	-1.81%
Coffee	0.821	0.857	4.38%	0.857	4.38%	0.821	0	0.864	5.24%
ECG200	0.77	0.87	12.99%	0.88	14.29%	0.88	14.29%	0.88	14.29%
FaceAll	0.808	0.81	0.25%	0.803	-0.62%	0.811	0.37%	0.802	-0.74%
FaceFour	0.83	0.875	5.42%	0.875	5.42%	0.875	5.42%	0.875	5.42%
Fish	0.833	0.903	8.40%	0.943	13.21%	0.903	8.40%	0.949	13.93%
Gun_Point	0.907	0.973	7.28%	0.98	8.05%	0.98	8.05%	0.98	8.05%
Lighting2	0.869	0.885	1.84%	0.869	0	0.885	1.84%	0.885	1.84%
Lighting7	0.726	0.726	0	0.699	-3.72%	0.712	-1.93%	0.699	-3.72%
OliveOil	0.867	0.833	-3.92%	0.833	-3.92%	0.833	-3.92%	0.8	-7.73%
OSULeaf	0.591	0.719	21.66%	0.711	20.30%	0.731	23.69%	0.756	27.92%
SwedishLeaf	0.79	0.883	11.77%	0.883	11.77%	0.891	12.78%	0.886	12.15%
Synthetic Control	0.993	0.89	-10.37%	0.827	-16.72%	0.977	-1.61%	0.947	-4.63%
Trace	1	1	**0**	0.99	-1%	1	0	1	0
Two_Patterns	1	1	**0**	1	**0**	1	0	1	0
wafer	0.98	0.993	1.33%	0.993	1.33%	0.993	1.33%	0.994	1.43%
yoga	0.836	0.868	3.83%	0.865	2.90%	0.868	3.83%	0.866	3.59%

Table 2. Experimental results on the UCR time series classification/clustering test bed

Among the 12 data sets where FBDTW2 gains improvement, there are 9 data sets with accuracy improvement greater than 5 percent; 7 data sets with accuracy improvement greater than 10 percent; and 2 data sets with accuracy improvement greater than 20 percent.

5) AFBDTW2 make further improvement over FBDTW2 on 11 data sets, ties with FDBTW2 on 7 data sets, and gets worse results than FBDTW2 on 2 data sets.

We further compare the proposed AFBDTW with value based DTW by plotting all the data sets in Figure 6 with x-axis representing accuracy rates obtained by value based DTW and y-axis representing accuracy rates obtained by AFBDTW (1&2). From these two figures, it is clearly show that majority of data sets favor AFBDTW. The few data sets where value based DTW gains better results are actually close to the diagonal line, which means that the performance differences on those few data sets between AFBDTW and value based DTW are actually very minor. Therefore, the experimental results suggest that AFBDTW is a better alternative to valued based DTW in time series classification in terms of classification accuracy.

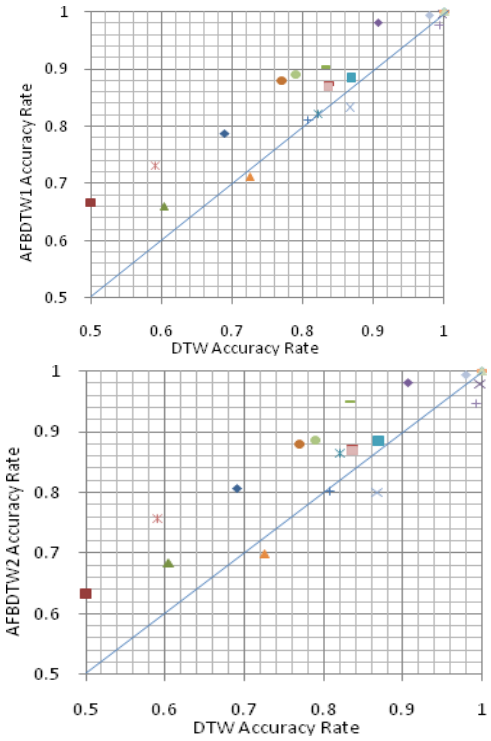


Figure 6 Comparison of AFBDTW with value based DTW on all 20 data sets

6. Time Complexity of FBDTW & AFBDTW

Assume the size of the training set is N_1, the size of the testing set is N_2, and the length of each sequence is M. Then the time complexity of 1NN+FBDTW is the same as 1NN + DTW, which is $O(N_1 N_2 M^2)$. Since 1NN + AFBDTW adopts a dual learning strategy, its time complexity is $O(N_1^2 M^2) + O(N_1 N_2 M^2)$, where $O(N_1^2 M^2)$ is the time complexity of the first learning process that is to calculate the weighting factors for global features and local features; and $O(N_1 N_2 M^2)$ is the time complexity of 1NN classification. Given that the training size is typically much smaller than the testing size (i.e., $N_1 << N_2$) in real situations, the time complexity of 1NN+AFBDTW is reduced to $O(N_1 N_2 M^2)$, which is theoretically the same as 1NN+DTW.

Quite a few techniques have been proposed to reduce the quadratic time complexity of DTW in sequence length from different aspects, such as imposing constraints on warping windows [9, 13], reducing sequence dimension by data abstraction or transformation [14, 15, 19], indexing sequences with lower bonds [16, 17, 18], as well as methods that combined two or more above strategies [10, 12]. It is not difficult to see that most of these techniques can be easily adapted to the proposed FBDTW and AFBDTW. In our future work, we will study the effectiveness of different speeding techniques on FBDTW and AFBDTW, based on which come up with linear or near-linear versions of FBDTW and AFBDTW without sacrificing their performance on accuracy rate.

7. Conclusions

In this paper, we first analyzed some major limitation of value based DTW and derivative DTW. Since the value or the deviation of a point may not reflect the position of this point in global or local trends of the sequence, both value based DTW and derivate DTW may fail to align a pair of sequences along their common trends or patterns. In order to solve this issue, we first define a global feature and a local feature for each point in a time series sequence, then proposed the FBDTW algorithm that dynamically aligns two time series sequences based on both the global features and local features of each points in the sequences. Experiments show that FBDTW generates better classification results on majority of the UCR time series data sets.

The proposed FBDTW make it possible to enhance the learning capacity of DTW based classification algorithms. Through our study, we first found out that the significance of global features and local features in classification may vary from one domain/application to another. Then we further propose an adaptive version of FBDTW that is called AFBDTW to learn the weighting factors for global features and local features from the training data. Experiments show that AFBDTW is able to make further improvement on classification accuracy over FBDTW.

Our future focus will be put on studying algorithms that are able to improve the speed of AFBDTW without sacrificing its classification accuracy.

References

[1] C. Bahlmann, B. Haasdonk & H. Burkhardt (2002) 'On-Line handwriting recognition with support vector machine: a kernel approach', *Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition*, 2002, pp. 490-495.

[2] L. Campbell, D. Becker, A. Azarbayejani, A. Bobick & A. Pentland (1996) 'Invariant features for 3d gesture recognition', *Proceedings of international workshop on automatic face and gesture recognition*, Vermont, USA, pp.157-162.

[3] C. Godin & P.Lockwood(1989) 'DTW schemes for continuous speech recognition, a unified view', *Computer Speech and Language*, vol. 3, No. 2, pp. 169-198.

[4] J. Hu, B. Ray & L. Han (2006) 'An Interweaved HMM/DTW Approach to Robust Time Series Clustering', *Proceedings of the 18th International Conf. on Pattern Recognition*, Washington, DC, pp. 145-148.

[5] E. Keogh, M. Pazzani (2001) 'Derivative dynamic time warping', *Proceedings of the First SIAM International Conference on Data Mining*, Chicago, USA, 2001.

[6] E. Keogh, X. Xi, L. Wei & C. Ratanamahatana, The UCR Time Series Classification/Clustering Hompage: www.cs.ucr.edu/~eamonn/time_series_data/ (as of 12/28/2009)

[7] A. Marzal & V. Palazón (2005) 'Dynamic time warping of cyclic strings for shape matching', *Pattern Recognition and Image Analysis*, Springer Berlin/Heidelberg, pp.644-652.

[8] D. Minnen, T. Starner, I. Essa, and C. Isbell (2007) 'Improving activity discovery with automatic neighborhood estimation', *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, Hyderabad, India.

[9] H. Sakoe & S. Chiba (1978) 'Dynamic programming algorithm optimization for spoken word recognition', *IEEE Transanctions on Acoustics, Speech, and Signal Process*, Vol. 26, pp.43-49.

[10] X. Xi, E. Keogh, C. Shelton, L. Wei & C. Ratanamahatana (2006) 'Fast time series classification using numerosity reduction', *Proceedings of the 23rd international conference on Machine learning*, Pittsburgh, USA, pp.1033-1040.

[11] K. T. Islam, K. Hasan, Y. Lee, & S. Lee (2008) 'Enhanced 1-NN Time Series Classification Using Badness of Records', *Proceedings of the 2nd international conference on Ubiquitous information management and communication*, New York, USA, pp. 108-113.

[12] S. Salvador & P. Chan (2004) 'FastDTW: toward accurate dynamic time warping in linear time and space', *Proceedings of 3rd KDD workshop on mining temporal and sequential data*, pp.70-80.

[13] F. M. Itakura (1975) 'Prediction residual principle applied to speech recognition', *IEEE Trans. Acoustics, Speech,and Signal Proc*. vol. ASSP-23, pp 52-72.

[14] S. Chu, E. Keogh, D. Hart & M. Pazzani (2002) 'Iterative deepening dynamic time warping for time series' *Proceedings of the Second SIAM International Conference on Data Mining*, Arlington, Virginia.

[15] E. Keogh & M. Pazzani (2000) 'Scaling up dynamic time warping for data mining applications', *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Boston, USA.

[16] S. Kim, S. Park & W. Chu (2001) 'An index-based approach for similarity search supporting time warping in large sequence databases', *Proceedings of 17th International Conference on Data Engineering*, Heidelberg, Germany, pp. 607-614.

[17] E. Keogh (2005) 'Exact indexing of dynamic time warping', *Knowledge and Information Systems*, Vol. 7, Issue 3, pp. 358-386.

[18] B. Yi, K. Jagadish & H. Faloutsos(1998) 'Efficient retrieval of similar time sequences under time warping', *Proceedings of 14th International Conference on Data Engineering*, pp 23–27

[19] F. Chan, A. Fu & C. Yu (2003) ' Haar wavelets for efficient similarity search of time-series: with and without time warping', IEEE Transations on Knowledge and Data Engineering, Vol 15, No. 3, pp. 686–705.

[20] M. Faundez-Zanuy (2006) 'On-line signature recognition based on VQ-DTW', *Pattern Recognition*, Vol. 40, Issue 3, pp. 981–992.

[21] Huang B, Kinsner W (2002) 'ECG frame classification using Dynamic time warping' *Proceedings of the 2002 Canadian Conference on Electrical and Computer Engineering*, Los Alamitos, USA, pp. 1105–1110

[22] T. Oates, M. Schmill & P. Cohen (2000) 'A method for clustering the experiences of a mobile robot that accords with human judgments', *Proceedings of the 17th National Conference on Artificial Intelligence*, pp 846-851.

[23] R. Muscillo, S. Conforto, M. Schmid, P.Caselli & T. D'Alessio (2007) 'Classification of motor activities through derivative dynamic time warping applied on accelerometer data', *Proceedings of 29th IEEE International Conference on Engineering in Medicine and Biology Society*, Lyon, pp. 4930-4933.

[24] B. Legrand, C. Chang, S. Ong, S. Neo & N. Palanisamy (2008) 'Chromosome classification using dynamic time warping', *Pattern Recognition Letters*, Vol.29, Issue 3, pp. 215-222.

[25] B. Sarker & K Uehara (2006) 'Efficient parallelism for mining sequential rules in time sereis data: a lattice based approach', *International Journal of Computer Science and Network Security*, Vol. 6, No. 7A, pp. 137-143.

[26] K. Makio, Y Tanaka & K Uehara (2007) 'Discovery of skills from motion data', *New Frontiers in Artificial Intelligence*, Spinger, ISBN 978-3-540-71008-0.

[27] A. Ratanamahatana & E. Keogh (2004) 'Everything you know about dynamic time warping is wrong', *Proceedings of 3rd KDD Workshop on Mining Temporal and Sequential Data*, Seattle, USA, 2004.

Dr. Ying Xie is an Assistant Professor of Computer Science at the Kennesaw State University. His research interests include information retrieval, data mining, bioinformatics and granular computing. His research was sponsored by a couple of US companies. He holds several pending US/International patents. He has been involved in organizing several international conferences and workshops in data mining and granular computing. He was an Invited Speaker at the 2008 IEEE international conference on granular computing.

Bryan Wiltgen holds a BS in Computer Science from Kennesaw State University. He is currently a graduate student at the Georgia Institute of Technology working towards his Master's degree in Computer Science, specializing in artificial intelligence and cognitive science. As a member of the Design & Intelligence Lab at Georgia Tech, he conducts research into the role of knowledge representation in understanding, analogical reasoning, and design.