Information Sciences 181 (2011) 2787–2796

Contents lists available at ScienceDirect

Information Sciences

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

Dynamic time warping constraint learning for large margin nearest neighbor classification

Daren Yu, Xiao Yu *, Qinghua Hu, Jinfu Liu, Anqi Wu

Harbin Institute of Technology, Harbin 150001, China

ARTICLE INFO

Article history:
Received 18 June 2010
Received in revised form 12 February 2011
Accepted 3 March 2011
Available online 14 March 2011

Keywords:
Time series classification
Dynamic time warping
Constraint learning
Large margin

ABSTRACT

Nearest neighbor (NN) classifier with dynamic time warping (DTW) is considered to be an effective method for time series classification. The performance of NN-DTW is dependent on the DTW constraints because the NN classifier is sensitive to the used distance function. For time series classification, the global path constraint of DTW is learned for optimization of the alignment of time series by maximizing the nearest neighbor hypothesis margin. In addition, a reduction technique is combined with a search process to condense the prototypes. The approach is implemented and tested on UCR datasets. Experimental results show the effectiveness of the proposed method.

1. Introduction

Time series classification is a challenging task in speech recognition, medical analysis, identification of moving objects, etc. [24,9,18,20,2]. Dynamic time warping (DTW) is considered to be the most commonly used method for similarity measurement [34,5] in time series classification. DTW was introduced into this domain by Berndt to overcome the weakness of Euclidean metric[1] in measuring the similarity between time series, where time phases of different series are distinct. Its superiority has been demonstrated by a large amount of work [13,15,35]. Although the performance of k-NN method is highly sensitive to the distance function [12], extensively empirical evaluations on more than 40 datasets have showed that 1NN-DTW classifier outperforms most of other techniques used in time series classification [5,33].

How to learn the Constraint of time warping is one of the most important issues in 1NN-DTW. For most classification problems, unconstrained DTW often leads to pathological warping and then a reduced performance of classification. These warps do not represent the proper mapping of a feature. An appropriate global path constraint of DTW can improve the locally out of phase phenomenon without pathological warping. Sakoe–Chiba band is the most commonly used global path constraint [27] and Itakura Parallelogram [11]. In [25], Ratanamahatana showed that narrow Sakoe–Chiba band (less than 10% of series length) performs better on many classification tasks. Moreover, DTW does not scale well to large databases because of its quadratic time complexity [28]. Global path constraints can be used to speed up the DTW algorithm. However, the appropriate width of a global constraint is always problem-dependent.

In this paper, we try to address the problem of similarity measurement of time series by adjusting the constraints of DTW. In the last decade, the large margin criterion has been widely discussed in feature evaluation, distance learning and classification modeling. According to the statistical learning theory, a classifier with large margin will produce good generalization performance. In this work, we introduce this criterion into the global constraint learning for dynamic time warping. Based on

* Corresponding author.
E-mail address: lostcrimson@gmail.com (X. Yu).

0020-0255/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.ins.2011.03.001

the learned constraint, we have designed a large margin nearest neighbor classifier for time series classification. The margin is used for evaluating the generalization ability of a classifier [32,19]. In addition, in order to reduce the computational cost, a technique is designed for prototype condensing. We present a set of experiments to show the effectiveness of the proposed techniques.

The rest of this paper is organized as follows. Section 2 describes the background of DTW and recent global constraint learning method. Section 3 shows large margin classification and the relationship between generalization bound and margin. Section 4 introduces the learning algorithm, including the constraint model of DTW and a speedup technique. Section 5 shows the experimental results. Finally conclusions are given in Section 6.

2. Background and related work

Euclidean metric is a popular method to define similarity and index time series, but it is very brittle in computing similarity between time series with different time phases [16,14]. DTW distance can overcome this problem by searching an optimal match between two given time series in spite of phase aberration[21]. DTW uses a dynamic programming technique to find the minimal distance between two time series, where sequences are warped by stretching or shrinking the time dimension.

We consider sequence C of length m and sequence Q of length n, where $C = c_1, c_2, \ldots, c_i, \ldots, c_m$, $Q = q_1, q_2, \ldots, q_j, \ldots, q_n$. A n-by-m matrix can be obtained where element (i,j) is computed by base distance $d_{base}(i,j) = (c_i - q_j)^{base}$. Generally, we use the square Euclidean distance as the base distance. An alignment between C and Q can be represented by warping path $W = w_1, w_2, \ldots, w_k, \ldots, w_L$, $max(m,n) \leqslant L \leqslant m + n - 1$, where $w_k = (i,j)_k$. We can find a path through the matrix which minimizes the cumulative distance. The DTW distance between two series is defined as:

$$DTW(C_i, Q_j) = d(C_i, Q_j) + min\begin{cases} DTW(C_i, Q_{j-1}) \\ DTW(C_{i-1}, Q_j) \\ DTW(C_{i-1}, Q_{j-1}) \end{cases} \tag{1}$$

Warping path W should satisfy several local constraints [27]:

- Boundary constraint: $w_1 = (1,1)$, $w_L = (m,n)$
- Monotonicity constraint: $w_k = (a,b)$, $w_{k+1} = (a',b')$, then $a' \geqslant a$, $b' \geqslant b$
- Continuity constraint: $w_k = (a,b)$, $w_{k+1} = (a',b')$, then $a' \leqslant a + 1$ and$b' \leqslant b + 1$

In practice, we do not need to compute all possible warping paths, because most of them correspond to pathological warping. Therefore, an optimal match with a certain limitation should be bound for computing DTW distance. Global constraints of warping path can be used in the matching process to decrease the number of paths [33].

The warping path changes if the DTW global path constraints are adjusted. An appropriate DTW distance function can be formulated by different constraints for a specific dataset. Ratanamahatana created an arbitrary shape and size of the band, which is called R-K band. This technique is appropriate for various datasets to give a learning algorithm[26,22]. Gaudin introduced a weighted DTW which is named adaptable time warping, and presented a learning process using a genetic algorithm[8]. These two techniques either minimize the error rate on training set using a leaving-one-out scheme[26,8], or minimize all the pairwise distances between intra-class samples and maximize the distance between inter-class samples using the Silhouette index[22]. However, there are two potential problems for these techniques. First, the conventional empirical risk minimization (ERM) on training data does not imply good generalization ability on unseen testing data. Second, for a nearest neighbor classifier, it is not necessary to compute all the pairwise distances between samples because only the near samples are useful for the classification for NN classifier.

3. Large margin classification

The principle of structural risk minimization (SRM) allows a tradeoff between training errors and model complexity [29]. According to this theory, a classifier with large margin would produce good generalization performance. Recently, some new techniques related to maximizing the uncertainty [30] or to combine multiple reducts of rough sets [31] have been proposed to improve the generalization of decision rules extracted from fuzzy decision trees. Several approaches are attempted to learn the distance function in various domains [4,7,36]. Bartlett discussed the distance between samples and the decision boundary and uses the sample margin to derive generalization bounds.

Theorem 1. *Let $\delta > 0$ and T be a set of size m. With probability $1 - \delta$ over the random choice of T, for any $\theta \in (0,1]$*

$$ER(h) \leqslant ER_T^{\theta}(h) + \sqrt{\frac{2}{m}\left(d \ln\left(\frac{34em}{d}\right)\log_2(578m) + \ln\left(\frac{8}{\theta^{\delta}}\right)\right)} \tag{2}$$

where $d = (64R/\theta)$ and h is a real valued function, i.e classifier. R is the ball radius of a feature space. The item $ER_T(h)$ means the training error in training set T and the latter item is the complexity of a classifier.

Two types of margin are discussed: sample margin and hypothesis margin [3,10]. Support vector machines use sample margin to measure the distance between the support vector and classification hyper-plane of a classifier while the hypothesis margin is the distance between the hypothesis and the closest hypothesis with different label to the given sample. The research on the large margin for a nearest neighbor classifier provides a criterion for the evaluation of the distance function and imbue the classifier with a good generalization ability [6,23]. It is of great importance for a k-NN classifier to select a good distance function.

In NN, a distance function is required to compute the similarity between objects. A good distance function should make the intra-class samples closer and the inter-class samples farther, so a large sample-margin can obtain. In this paper, a distance measurement is learned by adjusting the DTW constraints according to the large margin criterion. An example of the nearest neighbor hypothesis-margin is shown in Fig. 1.

For sample x, margin θ can be given to the nearest neighbor (1NN), and hypothesis margin can be defined as:

$$\theta(x) = (\|x - nearmiss(x)\| - \|x - nearhit(x)\|) \tag{3}$$

where $nearhit(x)$ and $nearmiss(x)$ are the nearest neighbor of x with the same label and the nearest neighbor with the different label, respectively. It was shown that hypothesis margin lower bounds the sample margin (SVM margin).

4. Constraint learning for large margin classification

In this section, the framework of constraint learning is introduced according to the large margin criterion and a condensing technique is proposed for highly effective speedup methods.

4.1. Global path constraint model

Intuitively, a good alignment path should not warp much from the diagonal. A window of width k is designed to limit warping constraints.

Sakoe–Chiba band is originally used for speech recognition [seeing Fig. 2 (a)] and it is one of the most popular global path constraints although the unified band width is not effective for some applications. l can be specified as a percentage of the length of time series. It should be noted that Euclidean distance between two sequences can be seen as a special case when $l = 0\%$. With width k of Sakoe–Chiba, element C_i can be aligned only to one of the k nearest elements of Q_i. Its definition is the same as the classical *DTW* except for the d_{base}:

$$d_{base}(i,j) = \begin{cases} (C_i - Q_j)^{base}, & \text{if } \|i-j\| \leqslant k \\ \infty, & \text{if } \|i-j\| > k \end{cases} \tag{4}$$

In [25], Ratanamahatana computed the classification accuracies using DTW for all the range of a global path constraint band. It can be seen from her results that the width of a global constraint band does have its effect on the accuracy of classification. We use DTW_S to represent DTW with constraint S. It can be concluded that the value of $DTW_S(C,Q)$ is monotonically non-decreasing with width k, so for any C and Q,

$$DTW(C,Q) \leqslant DTW_S(C,Q) \leqslant D_{Euclidean}(C,Q) \tag{5}$$

The R-K band represented by a one-dimensional array can be used to define an arbitrary constraint model. $\mathbf{r} = [r_1, r_2, \ldots, r_l]$ is a term used to define the allowed range of warping for given series. $r_k (1 \leqslant i \leqslant k)$ is the height above the diagonal in y axis and the width to the right of the diagonal in x axis. Warping path element $w_k(i,j)$ is constrained by r_k for $\| i - j\| \leqslant r_k$. It can be degenerated to Sakoe–Chiba band if r_k is independent of k.

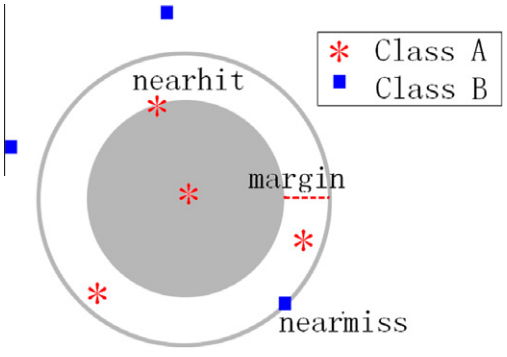

Fig. 1. Nearest neighbor margin.

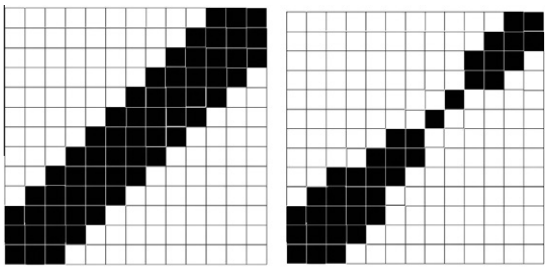

Fig. 2. Global path constraint of DTW distance matrix.

It is computationally intensive to learn the best **r** because no constraint is given. We should compute all the candidate cases. For simplicity, we take up with the reconciliation of Sakoe–Chiba band and R-K band by dividing the Sakoe–Chiba band into k equal-width intervals so that we can use $S = [s_1, s_2, \ldots, s_k,]$ to present the constraint. If $k = 1$, then S equals the Sakoe–Chiba band. We can also assign different parameters for S and k to achieve complex constraint. Fig. 2 (b) shows an example when $k = 4$ and $S = [2, 1, 0, 1]$. The width of a warping band also has its great impact on the efficiency for the computation of DTW. If the width is small, a large area of the matrix will not be examined, and the search for optimal warping path will be faster. We formulate the constraint space with $S = [s_1, s_2, \ldots, s_k,]$; here $s_i (1 \leqslant i \leqslant k)$ is the undetermined coefficient, and $0 \leqslant s_i \leqslant max\_width$. The *max_width* of a warping band can be set as a proportion of time series length or is decided by a specific task.

4.2. Evaluation function

Margin is used in guiding distance learning. It plays a crucial role in measuring the confidence of a classification. It ensures that the classifier will bound the generalization error on test data if the NN classifier selects a proper distance function.

Our approach is inspired by the work on learning Mahanalobis distance metric for k-nearest neighbor[32]. In [10], the authors have proved the generalization bound of a NN classifier with hypothesis margin. The theorem is presented as Theorem 2:

Theorem 2. *Let $\delta > 0$ and let T be the training set with k prototypes from each class. With probability $1 - \delta$ over the random choice of T, for any margin $\theta \in (0, 1]$*

$$ER(h) \leqslant ER_T(h) + \sqrt{\frac{8}{m}\left(d\log^2\frac{32m}{\theta^2} + \ln\left(\frac{4}{\delta}\right)\right)}, \tag{6}$$

where h is a NN classifier and VC dimension $d = min\left(n+1, \frac{64R^2}{\theta^2}\right)2k\log k^2$.

The goal of constraint learning based on large margin nearest neighbor is to find a global path constraint which can maximize the average margin of samples. The specific S in the constraint space is evaluated by the accumulative margin between *nearmiss* and *nearhit* using the DTW distance with global path constraint S. So the evaluation function can be defined as:

$$Eval(S) = \sum_{x \in T} \frac{\|x - nearmiss(x)\|_{DTW_S} - \|x - nearmiss(x)\|_{DTW_S}}{R_{DTW_S}}, \tag{7}$$

where x is a sample in training set T, and R means the radius of data. $\|x - y\|_{DTW_S}$ refers to the function where distance function DTW_S is calculated with DTW with the global path constraint S as defined in (3). Here we calculate the radius by $max(DTW_S(x, y))$, $x, y \in T$. The numerator of fractional expression is hypothesis margin θ_S. According to SRM theory, we expect the radius of data space can be as small as possible, while the margin between different class label samples are as large as possible.

Researches showed that the 1NN-DTW performs well on most datasets in UCR datasets (error rate $\delta_{1NN-DTW} \leqslant \delta_{1NN-Euclidean}$) [26,17,16]. We computed the average margin of UCR repository by 1NN-Euclidean and 1NN-DTW in order to bind its cause. The correlation of error rate and margin is shown in Fig. 3. where $\times$ and $\bigcirc$ correspond to a dataset of UCR repository.

$\times$ denotes the dataset where 1NN-DTW performs better than 1NN-Ed (error rate $\delta_{1NN-DTW} < \delta_{1NN-Euclidean}$), and $\bigcirc$ denotes the contrary situation. The x-axis and y-axis show the margin of two compared classifiers. The straight line has a slope of 1.0. The dataset labels below or above the straight line indicates that 1NN-Ed has a larger or smaller margin than 1NN-DTW. It can be seen from Fig. 3 that 1NN-DTW outperforms the 1NN-Euclidean in most of the datasets (the number of $\times$ is larger than $\bigcirc$'s). As shown in the graph, if the margin of classifier A is larger than the margin of classifier B, then the accuracy of A is probably higher than that B of. This example motivated the need of a distance function to produce a larger margin for the nearest neighbor query. The results enlightened us to engage in the distance learning based on large margin.

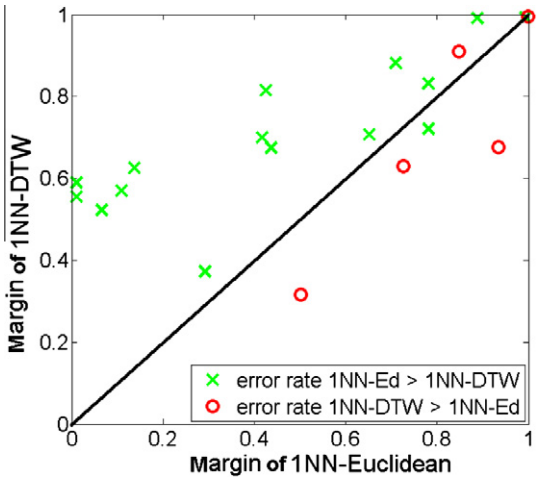

Fig. 3. The relationship between margin and error rate for 1NN-DTW and 1NN-Euclidean calculated on UCR dataset.

We used a brute-force solution to choose an appropriate global path constraint to maximize the evaluation function. Our goal is to adapt a global path constraint for DTW distance in an optimal way for the dataset. So we need to find the best constraint, i.e. a global path constraint to maximize the margin with DTW and k-NN. The brute force solution is used to perform the classification with all possible values, and keep the one with the best result only. Parameter *max_width*, the max value of each s_i, is restricted to 10% of the time series length, because a 10% constraint on warping is actually too large for real world data mining[25]. The number of segments of a constraint is chosen in terms of the size of a constraint space. The algorithm is shown in Table. 1.

4.3. Data condensing for evaluation function

It is a time-consuming process to search for *nearhit* and *nearmiss* under different constraint condition. Condensing methods is developed to pick out a consistent subset of prototypes for a problem. Given training set T and constraint S of DTW, the aim of condensing is to find subset $U \subseteq T$ to satisfy $f(\bullet)_U = f(\bullet)_T$, where $f(\bullet)_U$ is a function developed by calculating on special sample set U. The rival set U_x means the condensed set for sample x. *nearhit*(x) and *nearmiss*(x) can be found from the U_x but not training set T. Obviously, the neighborhood relationship between samples is dependent on the global path constraint used. In order to find *nearhit* and *nearmiss* for each sample x, we must call on all the other samples and select the nearest neighbor with same label and different label. So it is time-consuming to learn the distance function from a large number of training set.

Given training set T, and time series $x \in T$, we need *nearhit* and *nearmiss* of x from $T - \{x\}$ to evaluate constraint S. As we know, DTW distance is less than or equal to Euclidean distance because the accumulated distance on the distance matrix is minimized by warping path. For each DTW function with specific global path constraint, Euclidean distance and DTW distance are its upper and lower bounds, respectively. So element $y \in T$ can be ignored if $\exists z \in T$, $D_{Euclidean}(z,x) \leqslant DTW(y,x)$. As a result, we need not compute all pairwise distance between x and other samples to determine *nearhit*(x) and *nearmiss*(x) for a specific global path constraint. We can condense the redundant samples before learning as follows:

- Step 1: For $x \in T$, compute the distance between x and all the elements of rival set $T - \{x\}$ by Euclidean metric and get *nearhit*(x) and *nearmiss*(x).

Table 1
Learning algorithm.

	Learning algorithm.
1	Given the max_width and the number of segmentation;
2	max_margin = 0;
3	for each constraint S
4	Evaluate S using the samples in training set
5	if Evaluation (S) > max_margin
6	max_margin = Evaluation (S)
7	endif
8	endfor
9	Use the best S for classification on testing set

- Step 2: Compute the distance between x and all the elements of rival set $T - \{x\}$ by DTW. Delete element y from the rival set of x if $DTW(x,y) > max(D_{Euclidean}(x,nearhit(x)), D_{Euclidean}(x,nearmiss(x))$.
- Step 3: Evaluate all the global path constraints on the renewed rival set for each sample.

The algorithm is formulated as shown in Table. 2

As Eq. (5) showed, $DTW(x,y) \leqslant DTW_S(x,y) \leqslant D_{Euclidean}(x,y)$, for any $x, y \in T$. So, if $DTW(x,y) > D_{Euclidean}(x,nearhit(x))$, then $DTW_S(x,y) > D_{Euclidean}(x,nearhit(x))$, then $DTW_S(x,y) > DTW_S(x,nearhit(x))$. Therefore, y can't be *nearhit* of x when constraint S is used. Similarly, if $max(\|x - nearhit(x)\|,\|x - nearmiss(x)\|)$ is used y can be neither neighbor nearhit (x) nor $nearmiss(x)$. So sample y can be ignored. For example, *CBF* dataset can be used to explain the use of the algorithm. We select the first sample from 30 samples in the training set, and match it with other 29 samples using Euclidean distance and DTW, respectively. To evaluate each constraint S, we need traverse each $x \in T$ for its *nearhit* and *nearmiss*. Theoretically, each sample in training set T should be searched for *nearhit* and *nearmiss* and their distances should be calculated by DTW with global path constraint S. It is a computation-intensive process. In fact, we just need retain samples with DTW distances smaller than the $max(D_{Euclidean}(x,nearhit(x)), D_{Euclidean}(x,nearmiss(x))$ as shown in Fig. 4. The radius (distance from query point to ▲) is equal to $max(D_{Euclidean}(x,nearhit(x)), D_{Euclidean}(x,nearmiss(x))$. For *Coffee* dataset, only 10 samples satisfy the conditions for being retained.

5. Experimental results

5.1. Data sets

The proposed method is tested on several datasets from the UCR Time Series Data Mining Archive[17]. The datasets cover different domains (e.g., video trajectory recognition, computer vision, etc.). In some problems, there is no significant difference between classic 1NN-DTW and 1NN unconstrained warping DTW. That is to say, constraint is not the key factor deciding the classification results.

In UCR dataset, 1NN-best warping window DTW classifier in some datasets performs better than classic 1NN-DTW. We call these "constraint-sensitive datasets". Here we regard that these tasks are more sensitive to constraint schema and we use the datasets which meet this condition to test our method. They are *Gun Point*, *Swedish Leaf*, 50*Words*, *Face Four* and *ECG*. Meanwhile,we also listed dataset with no significant difference or worse than the latter. We randomly chose part of them and tested them with our method.

The number of classes varies between 2(Gun-Point) and 50(50Words). The *Gun Point* dataset comes from the video trajectory recognition domain. There are two classes, Gun and Point, each containing 100 examples. The major difference of the two classes lies in some little fluctuations which appear at the wave trough. *Face Four* dataset is collected to classify four peoples face image. The series data are converted from image edge. *Swedish Leaf* dataset contains 15 different Swedish tree species. The time series data are transformed from the edge map of leaf images. 50*words* is a dataset used for classification of handwritten documents. The dataset is tested for word image matching. The images of 50 common words such as "the", "and", etc. are taken from the George Washington collection. *ECG* dataset is from electrocardiogram diagnosis signal. On the other hand, some constraint-insensitive datasets are chosen. Table 3 summarizes the data sets.

5.2. Comparative study on error rate

Table 4 shows the comparative results obtained with four classifiers: the 1NN Euclidean, 1NN DTW, 1NN Best warping window DTW and 1NN DTW with learned global path constraint based on large margin criterion (1NN-DTWLM). The nearest neighbor works in terms of the minimal distance from the query instance to the training samples to determine the nearest neighbor. The first two classifiers with Euclidean metric and DTW are the most commonly used methods in time series classification. The approach of 1NN Best warping window DTW is to find the best width of the Sakoe–Chiba band[17] for DTW.

Table 2
Condensing algorithm.

	Evaluation algorithm by sample condensing
1	for each $x \in T$
2	U = T-{x};
3	Get $nearhit(x)$ and $nearmiss(x)$ from U by Euclidean metric
4	for each $y \in U$
5	if $DTW(x,y) > max(D_{Euclidean}(x,nearhit(x)), D_{Euclidean}(x,nearhit(x)))$
6	$U = U - \{y\}$;
7	endif
8	endfor;
9	endfor
10	Evaluate constraints on rival set U_x

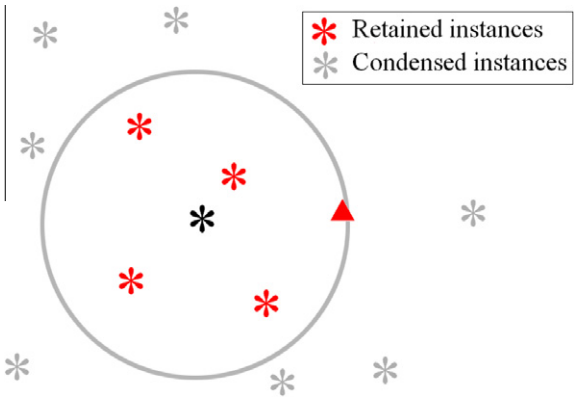

Fig. 4. Condensing redundant samples for learning algorithms.

Table 3
Description of datasets.

Dataset	Number of classes	Training set size	Testing set set size	Length
Constraint sensitive dataset				
Gun-Point	2	50	150	150
Swedish Leaf	15	500	625	128
50Words	50	450	455	270
Face Four	4	24	88	350
ECG	2	100	100	96
Constraint insensitive dataset				
CBF	3	30	900	128
Adiac	37	390	391	176
Fish	7	175	175	463

Table 4
Comparative error rate and margin.

Dataset	1NN Euclidean	1NN-DTW	1NN Best Warping Window DTW	1NN-DTWLM
Constraint sensitive dataset				
Gun-Point	0.087	0.093	0.087	**0.027**
Swedish Leaf	0.213	0.21	0.157	**0.152**
50Words	0.369	0.310	**0.242**	0.292
Face Four	0.216	0.170	0.114	**0.102**
ECG	0.12	0.23	0.12	**0.11**
Constraint insensitive dataset				
CBF	0.148	**0.003**	0.004	0.05
Adiac	0.389	0.396	**0.391**	0.396
Fish	0.217	0.167	**0.160**	**0.160**

The best results are presented in bold.

The number of segments $k = 4$ for predicting a class label for each unlabeled time series. According to the result of the best Sakoe–Chiba band, the *max_width* is the 8% of time series length. The classification error rates are presented in Table 4. The shapes of constraint are shown in Fig. 5. If the dataset is sensitive to DTW constraint (i.e., 1NN-DTW with best warping window has positive effects than 1NN-DTW), we can see that 1NN-DTWLM can further tap the potentialities of constraint than 1NN-DTW with best warping window. More accurate results can be obtained on most datasets. On the other hand, there is no significant difference among 1NN-DTW, 1NN-DTWLM and 1NN-DTW with best warping window on the rest 15 constraint-insensitive dataset. Here we only select three of them to show the results.

Dataset *Gun Point* [seeing Fig. 6 (a)] is chosen as an example to explain the relationship between margin and classification accuracy. We start off with Euclidean band ($\forall i, s_i = 0$). The section of the envelope is increased to compute the margin induced by this global path constraint, i.e., the margins are computed with warping windows gradually increasing. Accuracy curves on testing set calculated by DTW with each constraint S. As shown in Fig. 7, there are some interesting correlations between margin and classification accuracy. The great similarity between these two curves indicates that the effect of a

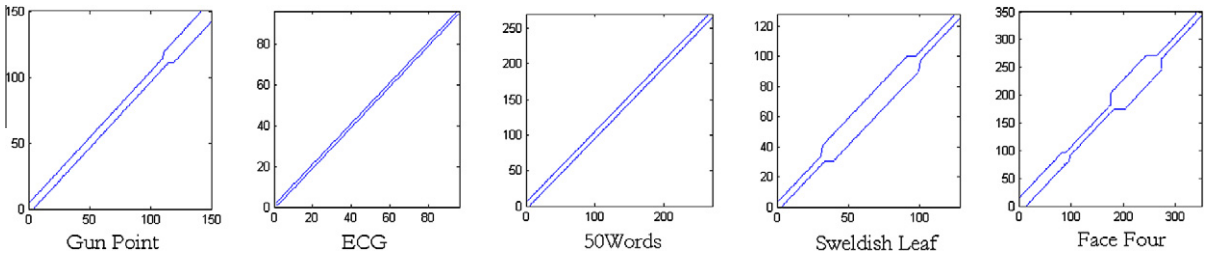

Fig. 5. Shapes of constraint.

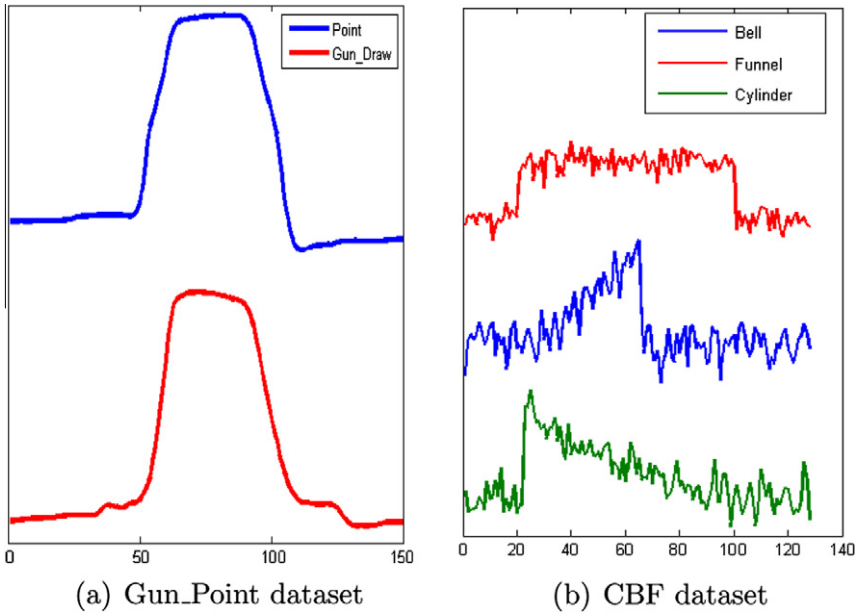

Fig. 6. Examples of dataset.

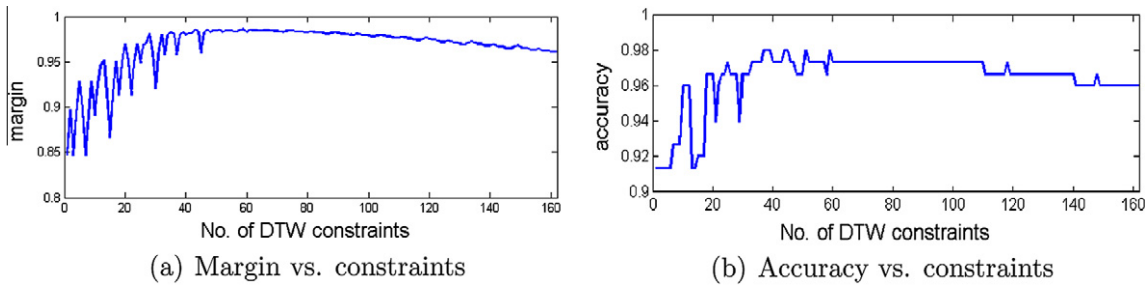

Fig. 7. Margin vs. classification accuracy.

classifier is highly dependent on the margin. For constraint-insensitive dataset, constraint is not the key point for the classification task, so we can simply use classic DTW method (e.g., CBF data, seeing Fig. 6 (b)). Large margin will probably lead to a good generalization performance for classification tasks, or vice verse.

A large margin classifier may not perform well enough on training samples but it can be used to obtain good generalization results on testing samples.

5.3. Condensing results

We have to compute $(n^2 - n)/2$ times DTW distance through a LOO nearest neighbor iteration. A condensing technique can be used to reduce the size of samples. As shown in Fig. 8, the rate of reduction ranges from 0.2 to 0.6.

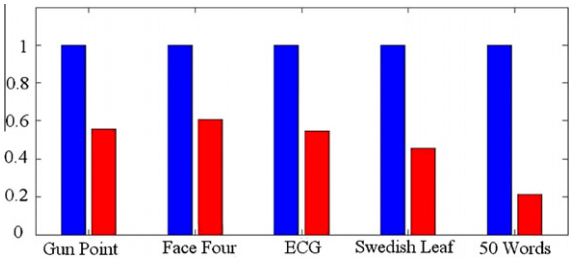

Fig. 8. Condensing rate for all data sets.

The condensing technique does not work if the number of training samples is very small. It is reasonable as there is no redundant samples in a very small dataset. The condensing technique becomes effective as the number of samples increases. This technique can also be used together with other DTW global path constraint learning algorithms to reduce the number of samples.

6. Conclusions

In some time series classification tasks, it is a common case that two time series are out of phase, even they share the same class label. An appropriate constraint of DTW can strongly improve the classification performance. In this paper, we introduce a framework of global path constraint so that the warping of varying degrees can be selected at different matching location. A learning algorithm is proposed for finding an appropriate global path constraint based on the large margin criterion. Good generalization ability can be guaranteed by the hypothesis margin for NN classifier. The learning algorithm is time-consuming but can meet the need of high accuracy of classification. In addition, a condensing technique has been proposed to speed up the evaluation process.

References

[1] D. Berndt, J. Clifford, Using dynamic time warping to find patterns in time series, in: AAAI-94 Workshop on Knowledge Discovery in Databases, pp. 229–248.
[2] Y. Chen, S. Wu, Y. Wang, Discovering multi-label temporal patterns in sequence databases, Information Sciences (2010).
[3] K. Crammer, R. Gilad-Bachrach, A. Navot, N. Tishby, Margin analysis of the LVQ algorithm, Advances in Neural Information Processing Systems (2003) 479–486.
[4] J. Davis, B. Kulis, P. Jain, S. Sra, I. Dhillon, Information-theoretic metric learning, in: Proceedings of the 24th International Conference on Machine Learning, ACM New York, NY, USA, pp. 209–216.
[5] H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, E. Keogh, Querying and mining of time series data: experimental comparison of representations and distance measures, in: Proceedings of the VLDB Endowment archive 1 (2008) pp.1542–1552.
[6] C. Domeniconi, D. Gunopulos, J. Peng, Large margin nearest neighbor classifiers, IEEE Transactions on Neural Networks 16 (2005) 899–909.
[7] Y. Freund, R. Schapire, Large margin classification using the perceptron algorithm, Machine learning 37 (1999) 277–296.
[8] R. Gaudin, N. Nicoloyannis, An Adaptable Time Warping Distance for Time Series Learning, in: Machine Learning and Applications, in: 5th International Conference on ICMLA'06, 2006, pp. 213–218.
[9] D. Gavrila, L. Davis, Towards 3-d model-based tracking and recognition of human movement: a multi-view approach, in: International workshop on automatic face-and gesture-recognition, pp. 272–277.
[10] R. Gilad-Bachrach, A. Navot, N. Tishby, Margin based feature selection-theory and algorithms, in: Proceedings of the Twenty-First International Conference on Machine Learning, ACM, p. 43.
[11] F. Itakura, Minimum prediction residual principle applied to speech recognition, IEEE Transactions on Acoustics, Speech and Signal Processing 23 (1975) 67–72.
[12] L. Jiang, Z. Cai, D. Wang, S. Jiang, Survey of improving k-nearest-neighbor for classification, in: Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on, volume 1.
[13] M. Kadous, Learning comprehensible descriptions of multivariate time series, in: Proceedings of the 16th International Machine Learning Conference, pp. 454–463.
[14] E. Keogh, S. Kasetty, On the need for time series data mining benchmarks: a survey and empirical demonstration, Data Mining and Knowledge Discovery 7 (2003) 349–371.
[15] E. Keogh, M. Pazzani, Scaling up dynamic time warping for datamining applications, in: Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM New York, NY, USA, pp. 285–289.
[16] E. Keogh, C. Ratanamahatana, Exact indexing of dynamic time warping, Knowledge and Information Systems 7 (2005) 358–386.
[17] E. Keogh, X. Xi, L. Wei, C. Ratanamahatana, The UCR Time Series Classification-Clustering Datasets, <http://www.cs.ucr.edu/~eamonn/time_series_data/> (2006).
[18] S. Kim, B. Jeong, Performance bottleneck of subsequence matching in time-series databases: observation, solution, and performance evaluation, Information Sciences 177 (2007) 4841–4858.
[19] P. Kumar, P. Torr, A. Zisserman, An invariant large margin nearest neighbour classifier, in: IEEE 11th International Conference on Computer Vision, pp. 1–8.
[20] A. Lee, Y. Chen, W. Ip, Mining frequent trajectory patterns in spatial-temporal databases, Information Sciences 179 (2009) 2218–2231.
[21] H. Lim, K. Whang, Y. Moon, Similar sequence matching supporting variable-length and variable-tolerance continuous queries on time-series data stream, Information Sciences 178 (2008) 1461–1478.
[22] V. Niennattrakul, C. Ratanamahatana, Z. Chen, D. Wen, C. Bessiere, E. Hebrard, B. Hnich, Z. Kiziltan, T. Walsh, M. Nekovee, et al., Learning DTW Global Constraint for Time Series Classification, Arxiv preprint arXiv:0903.0041 (2009).
[23] Y. Qian, J. Liang, W. Pedrycz, C. Dang, Positive approximation: an accelerator for attribute reduction in rough set theory, Artificial Intelligence (2010).

[24] L. Rabiner, B. Juang, Fundamentals of speech recognition, Prentice Hall, 1993.
[25] C. Ratanamahatana, E. Keogh, Everything you know about dynamic time warping is wrong, in: Third Workshop on Mining Temporal and Sequential Data.
[26] C. Ratanamahatana, E. Keogh, Making time-series classification more accurate using learned constraints, in: Proceedings of SIAM International Conference on Data Mining, Lake Buena Vista, Florida, pp. 11–22.
[27] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, Readings in speech recognition (1990) 159.
[28] S. Salvador, P. Chan, FastDTW: toward accurate dynamic time warping in linear lime and space, Intelligent Data Analysis 11 (2007) 561–580.
[29] V. Vapnik, The Nature of Statistical Learning Theory, Springer Verlag, 2000.
[30] X. Wang, C. Dong, Improving generalization of fuzzy IF–THEN rules by maximizing fuzzy entropy, IEEE Transactions on fuzzy systems 17 (2009) 556–567.
[31] X. Wang, J. Zhai, S. Lu, Induction of multiple fuzzy decision trees based on rough set technique, Information Sciences 178 (2008) 3188–3202.
[32] K. Weinberger, J. Blitzer, L. Saul, Distance metric learning for large margin nearest neighbor classification, Advances in Neural Information Processing Systems 18 (2006) 1473.
[33] X. Xi, E. Keogh, C. Shelton, L. Wei, C. Ratanamahatana, Fast time series classification using numerosity reduction, in: Proceedings of the 23rd International Conference on Machine Learning, ACM New York, NY, USA, pp. 1033–1040.
[34] L. Ye, E. Keogh, Time series shapelets: a new primitive for data mining, in: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM New York, NY, USA, pp. 947–956.
[35] B. Yi, H. Jagadish, C. Faloutsos, Efficient retrieval of similar time sequences under time warping, in: Proceedings of International Conference on Data Engineering, IEEE Computer Society, Los Alamitos, CA, USA, 1997, pp. 201.
[36] J. Yu, J. Amores, N. Sebe, P. Radeva, Q. Tian, Distance learning for limilarity estimation, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (2008) 451–462.