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.