From Dynamic Time Warping (DTW) to Hidden Markov Model (HMM)

Final project report for ECE742 Stochastic Decision

Chunsheng Fang
University of Cincinnati, 2009/3/19

Abstract

Dynamic Time Warping (DTW) and Hidden Markov Model (HMM) are two well-studied non-linear sequence alignment (or, pattern matching) algorithm. The research trend transited from DTW to HMM in approximately 1988-1990, since DTW is deterministic and lack of the power to model stochastic signals. In this report, I make a comprehensive literature study into this transition, and show that DTW and stochastic DTW, HMM are actually sharing the same idea of DP (dynamic programming). Some experiments are also performed to address this problem.

Introduction

Non-linear sequence alignment (or, pattern matching) has vast range of applications in DNA matching, string matching, speech recognition, etc. It seeks an optimal mapping from the test signal to the template signal, meanwhile allowing a non-linear, monotonic distortion (warping) in the test signal. However, the brute force algorithm to handle this has exponential complexity. Therefore, dynamic programming show it's power to cut the complexity to O(nm), where n, m is the length of signals. This algorithm is proposed in 1978 [1], which is called Dynamic Time Warping (DTW). DTW has been applied to mostly in speech recognition, since it's obvious that the speech tends to have different temporal rate, and alignment is very important for a good performance.

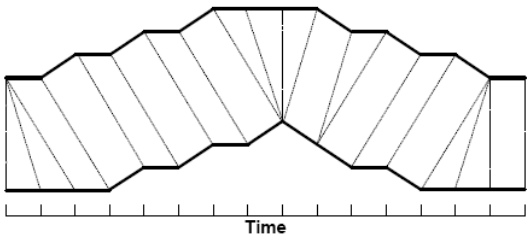

Figure 1, a warping between two temporal signals

Figure 2 and 3 shows the result of matching 2 speech signal "University of Cincinnati", pronounced by the author. The result is generated by [5].

The first signal in Figure 1 is slowly pronounced which lasts 1 second, and second one is quickly pronounced lasting 2 seconds. We can see that these 2 signals are mostly identical except that there is a temporal difference.

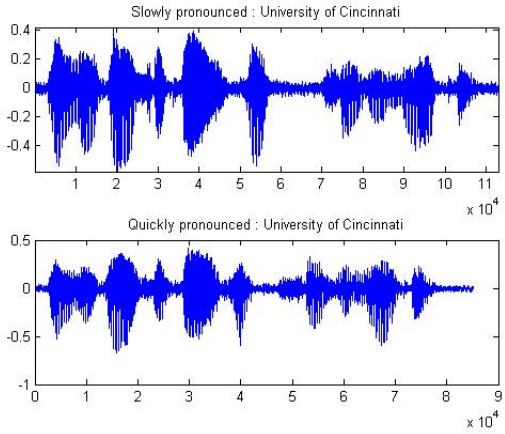

Figure 2. Two different speech signals for “University of Cincinnati”

Figure 3 illustrates the result of standard DTW for signals above. The red line is an optimal cost path from the beginning to the end of both signals. The left figure shows DTW path in similarity matrix, which denotes the correlation (similarity) of two signals. Hence the path will tend to pick darker blocks since it’ll maximize the matching performance. **Note that minimizing the distance is identical to maximizing the similarity.**
The right figure shows the “minimum cost to arrive” matrix with the same DTW path. We can see that as the path stretching from top left to bottom right, it becomes darker since the cost is monotonically increasing; meanwhile optimal DTW path is taking as small cost as possible.

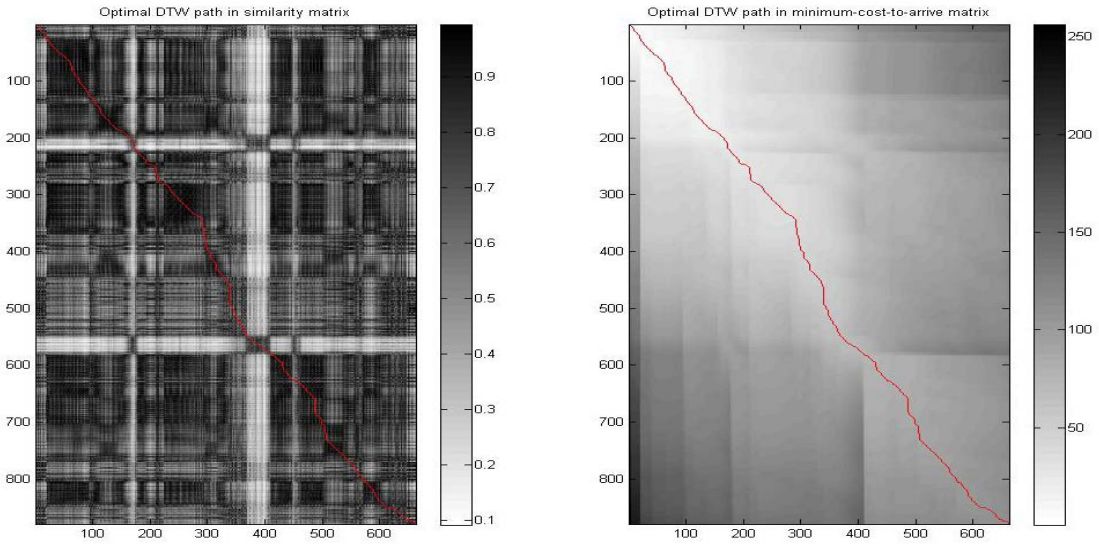

Figure 3. Result of standard DTW.

The standard DTW is basically using the idea of deterministic DP. However, a lot of real signals are stochastic processes, such as speech signal, video signal, etc. Therefore, in 1988, a new algorithm called “stochastic DTW” is proposed [2]. In this method, conditional probabilities are used instead of local distance in standard DTW, and transition probabilities instead of path costs. This actually is related to HMM. Its author also show the equivalence between stochastic DTW and HMM. The result of stochastic DTW improved the recognition rate from 89.3% to 92.9% in word recognition experiments.
In HMM, Viterbi algorithm is used for searching the optimal state transition sequence, for a given observation sequence. It turns out to be another application of DP to cut down its computations.

Problem formulation [3]

The dynamic time warping problem is stated as follows: Given two time series X, and Y, of lengths |X| and |Y|,

$$X = x_1, x_2, \ldots, x_i, \ldots, x_{|X|}$$
$$Y = y_1, y_2, \ldots, y_j, \ldots, y_{|Y|}$$

construct a warp path W

$$W = w_1, w_2, \ldots, w_K \qquad \max(|X|, |Y|) \le K < |X| + |Y|$$

where K is the length of the warp path and the kth element of the warp path is

$$w_k = (i, j)$$

where i is an index from time series X, and j is an index from time series Y. The warp path must start at the beginning of each time series at w1 = (1, 1) and finish at the end of both time series at wK = (|X|, |Y|). This ensures that every index of both time series is used in the warp path. There is also a constraint on the warp path that forces i and j to be monotonically increasing in the warp path, which is why the lines representing the warp path in Figure 1 do not overlap. Every index of each time series must be used. Stated more formally:

$$w_k = (i, j), w_{k+1} = (i', j') \qquad i \le i' \le i+1, \;\; j \le j' \le j+1$$

The optimal warp path is the warp path is the minimum-distance warp path, where the distance of a warp path W is

$$Dist(W) = \sum_{k=1}^{k=K} Dist(w_{ki}, w_{kj})$$

Dist(W) is the distance (typically Euclidean distance) of warp path W, and Dist(wki, wkj) is the distance between the two data point indexes (one from X and one from Y) in the kth element of the warp path.

Therefore, we can see that the function that we are going to minimize, Dist(W), is the same as Forward DP deterministic shortest path algorithm [4]. Therefore it can be converted from (ordinary) backward to forward, and can be also modeled as a finite state system.

Standard DTW

Here is the standard DTW algorithm:

1. Initial condition:

$$D(1,1) = 0;$$

2. Recurrence:

$$D(i, j) = Dist(i, j) + \min[D(i-1, j), D(i, j-1), D(i-1, j-1)]$$

Finally, the optimal cost will sit in D(|X|, |Y|), and we can also easily trace back the whole warping path.
Note that in this case, the possible monotonic warping can be illustrated as Figure 4, which means only 3 direction will be searched from previous step to construct the current step.

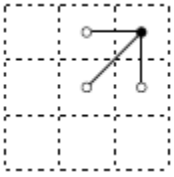

Figure 4, possible search grid in DP

In practice, there are many possible search grids in DP, such as follows:

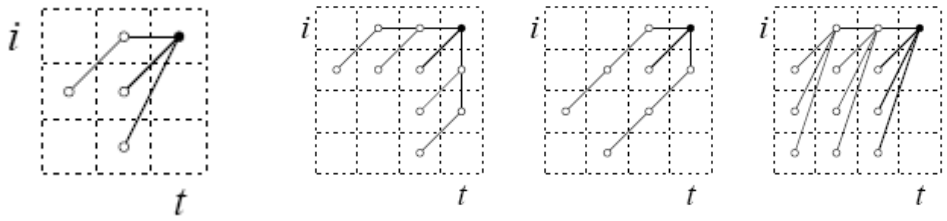

Figure 5, other examples of search grid in DP

Hidden Markov Model (HMM)

According to Wikipedia:

“A hidden Markov model (HMM) is a statistical model in which the system being modeled is assumed to be a Markov process with unknown parameters; the challenge is to determine the hidden parameters from the observable data. The extracted model parameters can then be used to perform further analysis, for example for pattern recognition applications. An HMM can be considered as the simplest dynamic Bayesian network.

In a regular Markov model, the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible, but variables influenced by the state are visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states.”

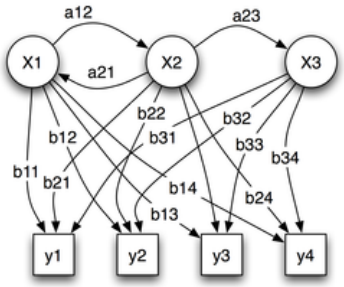

Figure 6, example of HMM, X is hidden states, Y is observations.

There are 3 canonical problems in HMM, the one we care most is finding the most likely hidden state sequence given an observation sequence:

Given the parameters of the model, find the most likely sequence of hidden states that could have generated a given output sequence. This problem is solved by the Viterbi algorithm, which turns out to be another DP algorithm.

Viterbi algorithm for HMM

The Viterbi algorithm was conceived by Andrew Viterbi in 1967 as an error-correction scheme for noisy digital communication links.

The Viterbi algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden states – called the Viterbi path – that results in a sequence of observed events, especially in the context of Markov information sources, and more generally, hidden Markov models. The forward algorithm is a closely related algorithm for computing the probability of a sequence of observed events. These algorithms belong to the realm of information theory.

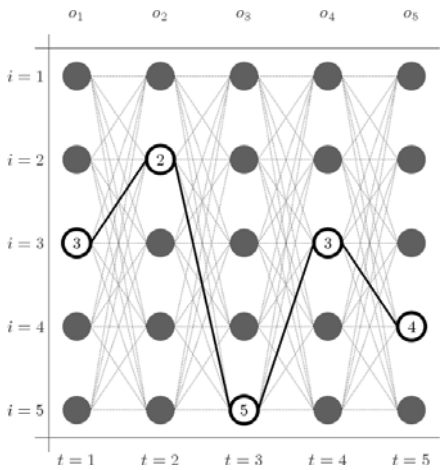

Figure 7, example of Viterbi algorithm for HMM

Stochastic DTW

In [2], the author mention that some research shows that if the underlying structure in DTW and Markov modeling is equ-probable, the standard DTW with linear predictive modeling and distortion measurements is equivalent to the probabilistic modeling except that it searches for the best transition path to minimize the accumulative distortion, while the probabilistic technique sums the density along every possible path.

In [2], stochastic DTW algorithm is first proposed, which start from replacing the deterministic cost with probabilities:

$$
\begin{aligned}
&P\,(a_1\,a_2 \ldots\, a_i;\, b_j) = \\
&\quad P\,(a_1\,a_2 \ldots a_{i-2};\, b_{j-1}) * q\,(b_{j-1} \rightarrow b_j) * P\,(a_{i-1},\, a_i \mid b_{j-1} \rightarrow b_j) \\
&\quad + P\,(a_1\,a_2 \ldots a_{i-1};\, b_{j-1}) * q\,(b_{j-1} \rightarrow b_j) * P\,(a_i \mid b_{j-1} \rightarrow b_j) \\
&\quad + P\,(a_1\,a_2 \ldots a_{i-1};\, b_{j-2}) * q\,(b_{j-2} \rightarrow b_{j-1}) * P\,(a_i \mid b_{j-2} \rightarrow b_{j-1}) \\
&\qquad\qquad * q\,(b_{j-1} \rightarrow b_j) * P\,(\phi \mid b_{j-1} \rightarrow b_j) \\
&\quad = P\,(a_1\,a_2 \ldots\, a_{i-2};\, b_{j-1}) * q\,(b_{j-1} \rightarrow b_j)' \\
&\qquad * P\,(a_{i-1} \mid b_{j-1} \rightarrow b_j') * q(b_j, \rightarrow b_j) * P(a_i \mid b_j' \rightarrow b_j) \\
&\quad + P\,(a_1\,a_2 \ldots a_{i-1};\, b_{j-1}) * q\,(b_{j-1} \rightarrow b_j) * P\,(a_i \mid b_{j-1} \rightarrow b_j) \\
&\quad + P\,(a_1\,a_2 \ldots a_{i-1};\, b_{j-2}) * q\,(b_{j-2} \rightarrow b_{j-1}) * \; P(a_i \mid b_{j-2} \rightarrow b_{j-1}) \\
&\qquad\qquad * q\,(b_{j-1} \rightarrow b_j) * P\,(\phi \mid b_{j-1} \rightarrow b_j)
\end{aligned}
$$

Where A is the observations, and B is the hidden states, P(a1, a2, … ai; bj) denotes the observation sequence {a1 to ai} until sitting in bj hidden state. It can be expressed by 3 previous probabilities, based on the asymmetric DP path which proposed by the author, in Figure 8.

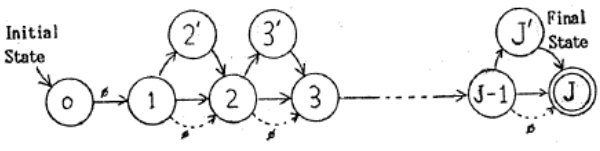

Figure 8. HMM corresponding to asymmetric DP path

Then, replace the right hand side with maximum probability, and taking the logarithm for P, we have the following approximate equation for Viterbi:

$$
\begin{aligned}
&Q(i,j) = \\
&\max\begin{cases}
Q(i-2,j-1)+\log P(a_{i-1} \mid b_{j-1}\rightarrow b_j')+\log P(a_i \mid b_j' \rightarrow b_j)+\log P_a(b_{j-1}) \\
Q(i-1,j-1)+\log P(a_i \mid b_{j-1}\rightarrow b_j)+\log\{1-P_a(b_{j-1})\} \\
Q(i-1,j-2)+\log P(a_i \mid b_{j-2}\rightarrow b_{j-1}) \\
\qquad +\log P(\phi \mid b_{j-1}\rightarrow b_j) \\
\qquad +\log q(b_{j-2}\rightarrow b_{j-1})\cdot\log q(b_{j-1}\rightarrow b_j)
\end{cases} \\
&\text{where } Q(i,j) = \log P(a_1 a_2 \ldots a_i; b_j)
\end{aligned}
$$

The author of [2] points out that this equation is actually similar to standard DTW equation, from this analogy, they propose the following general equation of stochastic DTW method:

$$
\begin{aligned}
&Q(i,j) = \\
&\max\begin{cases}
Q(i-2,j-1)+\log P(a_i \mid b_j)+\log P_{DP1}(j) \\
Q(i-1,j-1)+\log P(a_i \mid b_j)+\log P_{DP2}(j) \\
Q(i-1,j-2)+\log P(a_i \mid b_{j-1})+\log P(a_i \mid b_j) \\
\qquad +\log P_{DP3}(j),
\end{cases}
\end{aligned}
$$

Where DP1, DP2, DP3 are 3 predefined asymmetric DP path as Figure 9.

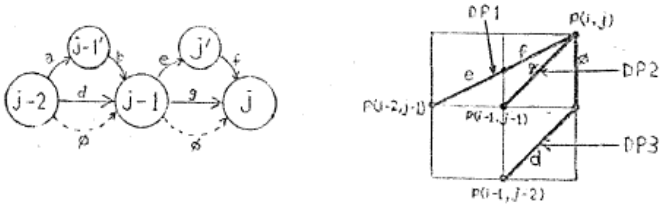

Figure 9, asymmetric DP path

Therefore, the author successfully show the equivalence between DTW and stochastic DTW, where the latter one is actually neatly modeled by HMM.

Recent development for DTW

Since the standard DTW has become a baseline algorithm in most of state-of-the-art speech recognition research, there are some effort in making the deterministic standard DTW faster, and applying to other applications.

FastDTW [3] show a linear time complexity algorithm, which uses the idea of coarse-to-fine to scale the problem down.

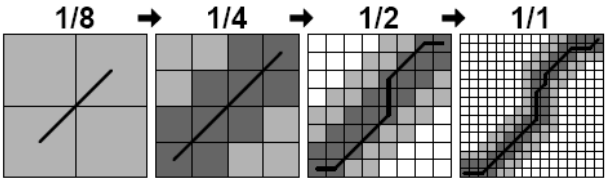

Figure 10. The four different resolutions evaluated during a complete run of the FastDTW algorithm.

Human motion sequence is also a temporal signal, therefore in [9] the authors employed DTW to classify motions from video.

DTW for 1D signal can be done in polynomial time, however, surprisingly, DTW for 2D image alignment is NP

Complete! A recent paper in Pattern Recognition Letter 2003 show the reduction from DTW 2D to 3SAT problem which is a classical NP complete problem.

References:

[1] Sakoe, H. and Chiba, S., *Dynamic programming algorithm optimization for spoken word recognition*, IEEE Transactions on Acoustics, Speech and Signal Processing, **26**(1) pp. 43- 49, 1978, ISSN: 0096-3518

[2] Seiichi Nakagawa, etc, Speaker-Independent English consonant and Japanese word recognition by a Stochastic Dynamic Time Warping method, Journal of Institution of Electronics and Telecommunication Engineers, 1988.

[3] Stan Salvador and Philip Chan, FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space, Intelligent Data Analysis, 2007.

[4] Dimitri P. Bertsekas, Dyanmic Programming and Optimal control, volume 1, 2nd Ed, Chapter 2, 2000.

[5] D. Ellis (2003). Dynamic Time Warp (DTW) in Matlab
Web resource, available: http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/.

[6] http://en.wikipedia.org/wiki/Hidden_Markov_model

[7] http://en.wikipedia.org/wiki/Viterbi_algorithm

[8] Daniel Keysers, Elastic Image Matching is NP-Complete, Pattern Recognition Letter 2003

[9] Kevin Adistambha et'al, Motion Classification Using Dynamic Time Warping, International Workshop on Multimedia Signal Processing 2008.