Adaptive Time Warping

Adaptive Time Warping

大桃 諭♥ 陳 漢雄♦
古瀬 一隆♣ 大保 信夫♠

Satoshi OOMOMO Hanxiong CHEN
Kazutaka FURUSE Nobuo OHBO

大規模な時系列データベースに対する高速な類似検索が重要な課題となっている．これまでにユークリッド距離に基づく類似検索技法が多く提案されているが，最近の研究によって，タイムワーピングによって得られた距離が時系列データの類似検索において有効であることが実証されてきた．しかし，タイムワーピングの計算時間は非常に長く，しかも距離公理の三角不等式が成り立たないため，伝統的な索引技法の適用は困難である．この問題を解決するために，時系列を固定長に区切って圧縮する手法が提案されているが，本論文では，時系列を可変長に区切って圧縮する手法を提案する．そして，本論文で提案する手法が有効であることを実験によって実証する．

Recently, time series produced and treated in many fields, and fast similarity search in large time series databases becomes important. For this subject, similarity search techniques based on Euclidean distance are straightforward. However, recent studies have shown that time warping distance is more robust distance in similarity search for time series. Difficulty is that time warping distance is computationally expensive. Traditional indexing techniques are powerless because it violates the triangle inequality. To overcome this problem, compression is considered to be effective and some techniques which divide time series into fix length for compressing have been proposed. In this paper we propose an adaptive compressing technique. Prefer to fix length division, our approach divides a time series according to its characters hence obtains higher effect while keeping lower information lost in compression. The effectiveness has been confirmed in out experimental results.

1. 序論

近年，時系列データは科学，医療，経済，工学などの様々な分野で扱われるようになり，それぞれの分野でデータマイニングが行われるようになっている．そして，時系列データに対してデータマイニングを行うには，時系列データの類似度を計算することが必要となる．

現在，ユークリッド距離が類似度として最もよく使用されている．しかし，ユークリッド距離には時間軸における小さな歪みにも影響を受けやすいという欠点がある．そして，この欠点を解決する技法に，タイムワーピング（Time Warping: TW）がある．TWでは，一方の時系列の1つの点と他方の時系列の連続する複数の点を対応させることによって，時間軸に伸縮性を持たせて歪みの影響を抑えることができる（図1）．

図 1 ユークリッド距離とタイムワーピング距離

Fig.1 Euclidean distance and Time Warping distance

これまでの研究によってTWの有効性は実証されているが，計算時間が非常に長いという欠点がある．そこで，計算時間を短くするための様々な手法が提案されている．その中の1つに，時系列を圧縮する手法がある．そして，圧縮する手法の1つに，時系列を固定長に区切って圧縮するPiecewise Dynamic Time Warping（PDTW）という手法がある．

PDTWを行うことによって，正確性の低下を抑えつつ高速化できることが実証されているが，全ての時系列を固定長で区切って圧縮するよりも，時系列の変動の度合に応じて可変長で区切って圧縮した方が，圧縮によって生じる誤差を抑えつつ高速化できると考えられる．そこで，本論文では，時系列を変動の度合に応じて可変長に区切って圧縮するAdaptive Time Warping（ATW）の手法を提案する．

2. タイムワーピングと固定長圧縮

2.1 タイムワーピング

この章では，タイムワーピング（TW）のアルゴリズムを簡単に述べる[1]．

長さn,mの2つの時系列$X=x_1,\cdots,x_i,\cdots,x_n$，$Y=y_1,\cdots,y_j,\cdots,y_m$に対して，まずは，以下の式で定義される2つの点$x_i,y_j$間の距離$d(x_i,y_j)$を(i,j)要素の値とするn×m行列を作成する．

$$d(x_i,y_j)=(x_i-y_j)^2 \quad (1)$$

次に，ワーピングパス$W=w_1,w_2,\cdots,w_k,\cdots,w_K$を求める．ワーピングパスは，以下の3つの条件を満たす行列の要素の順列で表現される．ワーピングパスの例を図2に示す。

- **[境界条件]** $w_1=(1,1)$,$w_K=(n,m)$とする．
- **[連続性]** $w_k=(a,b)$,$w_{k-1}=(a',b')$とすると，$a-a'\leqq 1$かつ$b-b'\leqq 1$となる．
- **[単調性]** $w_k=(a,b)$,$w_{k-1}=(a',b')$とすると，$a-a'\geqq 0$かつ$b-b'\geqq 0$となる．

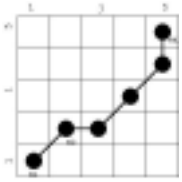

図 2 ワーピングパスの例

Fig.2 An example of warping path

♥ 学生会員 筑波大学大学院システム情報工学研究科
oomomo@dblab.is.tsukuba.ac.jp
♦ 正会員 筑波大学電子・情報工学系 {chx, furuse}@dblab.is.tsukuba.ac.jp
♠ 筑波大学電子・情報工学系 ohbo@dblab.is.tsukuba.ac.jp

上記の条件を満たすワーピングパスの中からワーピングコストが最小のパスを見つけて，距離を以下の式から求める．

$$TW(X,Y)=\min\left\{\sqrt{\sum_{k=1}^{K}w_k}\right\}\Big/K \quad (3)$$

コストが最小のパスを求めるには以下の再帰式を利用する．

$$\gamma(i,j)=d(x_i,y_j)+\min\begin{Bmatrix}\gamma(i-1,j-1)\\ \gamma(i-1,j)\\ \gamma(i,j-1)\end{Bmatrix} \quad (4)$$

この再帰式を利用すると，式(3)は以下のように表される．

$$TW(X,Y)=\sqrt{\gamma(n,m)}\Big/K \quad (5)$$

2.2 固定長圧縮を利用したタイムワーピング

この章では，時系列を固定長に区切って圧縮した後にタイムワーピングを適用する手法（PDTW）の概要を述べる[2]．

まずは，長さがnの時系列X=x1,…,xnを，長さNに圧縮する．圧縮後の時系列を，X=x1,…,xNとすると，Xのi番目の点xiの値は以下のように定義される．

$$\underline{x}_i=\frac{N}{n}\sum_{j=\frac{n}{N}(i-1)+1}^{\frac{n}{N}i}x_j \quad (6)$$

実際に時系列を圧縮した様子を図3に示す．

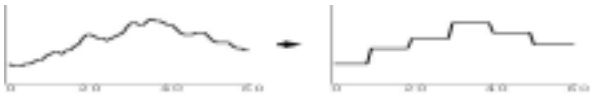

図 3 時系列の固定長圧縮

Fig.3 Fixed length compression of time series

圧縮の程度を指定するパラメータとして，以下のように定義される圧縮率cを用いる．

$$c=n/N \quad (7)$$

時系列を圧縮した後は，TWをそのまま適用する．

3. 可変長圧縮を利用したタイムワーピング

第2.2章で述べたPDTWは，時系列を固定長に区切って圧縮しているが，変動の度合に応じて可変長に区切って圧縮した方が，圧縮による誤差を小さくできると考えられる．そこで，時系列を可変長に区切って圧縮する方法を第3.1章で，圧縮後にTWを適用する方法を第3.2章で述べる．

3.1 時系列の可変長圧縮

まず初めに，可変長圧縮の流れを示す．

1. 時系列の区切り箇所を決定して複数の部分時系列に分ける．この部分時系列のことをフレームと呼ぶ．
2. 1 つのフレームを 1 つの点に圧縮することによって時系列全体を圧縮する．フレームの値は，そのフレーム内の点の値の平均値とする．

可変長圧縮を行う上で重要なことは，短い計算時間で，圧縮によって生じる誤差を一定以下に抑えつつ，フレームの数をなるべく少なくすることである．圧縮による誤差が小さいほど，TW を適用して得られる距離が圧縮前の距離に近似し，フレームの数が少ないほど，TW を適用したときの計算時間が短くなる．時系列を可変長圧縮した例を図 4 に示す．

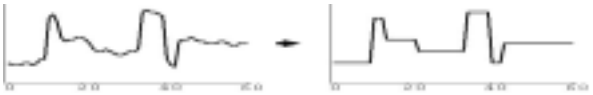

図 4 時系列の可変長圧縮

Fig.4 Variable length compression of time series

計算時間がO(n)で時系列X=x1,…,xnをX=x1,…,xNに可変長圧縮を行うアルゴリズムを以下に示す．このアルゴリズムに現れる τは閾値であり，これについては後で説明する．

1. s=1, N=1
2. find minimum e(s≦e<n) such that
   $\max\{x_s,\cdots,x_e\}-\min\{x_s,\cdots,x_e\}\le\tau$
   $\max\{x_s,\cdots,x_{e+1}\}-\min\{x_s,\cdots,x_{e+1}\}>\tau$
   if e dose not exist, go to 5.
3. calculate frame value $\underline{x}_N.v$ and length $\underline{x}_N.l$.
   $\underline{x}_N.v=\sum_{i=s}^{e}x_i\Big/(e-s+1)$
   $\underline{x}_N.l=e-s+1$
4. s=e+1, N=N+1, and go to 2.
5. calculate last frame value $\underline{x}_N.v$ and length $\underline{x}_N.l$.
   $\underline{x}_N.v=\sum_{i=s}^{n}x_i\Big/(n-s+1)$
   $\underline{x}_N.l=n-s+1$

このアルゴリズムに従って，τ=4 として時系列をフレーム分けする様子を図 5 に示す．点 A から点 B までの部分時系列の最大値と最小値の差は 3 であり，τ以下の値となっているが，点 B の次の点である点 C まで進むと差は 6 になり，閾値より大きくなる．よって，点 A から点 B までが 1 つのフレームとなり，点 C が次のフレームの開始点になる．これを時系列の最後まで繰り返すと，フレーム分けは終了する．

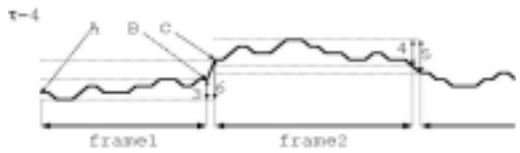

図 5 時系列のフレーム分け

Fig.5 Division of time series into frames

次に，閾値 τ の決定方法について説明する．まずは，圧縮の程度をユーザが直接指定できるパラメータとして，許容振幅率 ρというものを定義する．この ρから，閾値 τは以下の式で決定される．

$$\tau=\rho\cdot\left(\max\{x_1,\cdots,x_n\}-\min\{x_1,\cdots,x_n\}\right) \quad (8)$$

これによって，時系列の取り得る値の範囲に関係なく，0 から 1 の値で圧縮の程度を指定できるようになる．

3.2 可変長圧縮された時系列に対する TW の適用

時系列を可変長圧縮した後は，TW を適用して距離を計算することになる．しかし，フレームの長さが一定でないために，そのまま TW を適用することはできない．そこで，フレームの長さを考慮に入れて TW を適用する方法を述べる．

長さ N,M の 2 つの圧縮時系列 X=x1,…,xi,…,xN, Y=y1,…,yj,…,yMに対して，まずは，第 2.1 章で述べた方法で，コストが最小となるワーピングパスを求める．次に，こ

のワーピングパスから距離を求めるのだが，その方法について述べる前にワーピングパスについて詳しく考えてみる．

第 2.1 章で述べた方法でワーピングパスを求めると，図 6 の P_1 に示すような‘ 1 ’対‘ 1 ’の関係と，P_2 や P_3 に示すような‘ 1 ’対‘ 多 ’の関係は存在するが，P_4 に示すような‘ 多 ’対‘ 多 ’の関係は存在しない．この証明に関しては省略する．

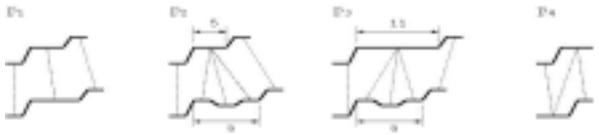

図 6　ＡＴＷにおける時系列の対応関係

Fig.6　Relations of time series in ATW

そして，可変長圧縮された時系列におけるワーピングパスでは，‘ 1 ’対‘ 多 ’の関係をさらに 2 つのパターンに分けることができる．図 6 の P_2 のような‘ 1 ’の方のフレームの長さが‘ 多 ’の方のフレームの長さの和よりも短いパターンと，P_3 のような‘ 1 ’の方が‘ 多 ’の方の和よりも長いパターンである．可変長圧縮された時系列におけるワーピングパスは P_1，P_2，P_3 の 3 つのパターンで構成されているので，図 7 に示すようにワーピングパスを $g_1,\cdots,g_n$ にグループ分けして，パターン別のグループの値 $g_h.v$ と大きさ $g_h.l$ を求める．

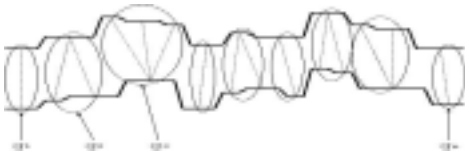

図 7　ワーピングパスのグループ分け

Fig.7　Grouping of warping path

- **[パターン P_1]** グループ g_h における関係が$\{\underline{x}_i\}$と$\{\underline{y}_j\}$の‘ 1 ’対‘ 1 ’であるとすると，

$$g_h.v = d(\underline{x}_i, \underline{y}_j) \cdot \max\{\underline{x}_i.l, \underline{y}_j.l\} \quad (9)$$

$$g_h.l = \max\{\underline{x}_i.l, \underline{y}_j.l\} \quad (10)$$

- **[パターン P_2]** グループ g_h における関係が$\{\underline{x}_i\}$と$\{\underline{y}_j,\cdots,\underline{y}_k\}$の‘ 1 ’対‘ 多 ’であり，フレームの長さが $\underline{x}_i.l \le \sum_{s=j}^{k} \underline{y}_s.l$ であるとすると，

$$g_h.v = \sum_{s=j}^{k} \left(d(\underline{x}_i, \underline{y}_s) \cdot \underline{y}_s.l \right) \quad (11)$$

$$g_h.l = \sum_{s=j}^{k} \underline{y}_s.l \quad (12)$$

- **[パターン P_3]** グループ g_h における関係が$\{\underline{x}_i\}$と$\{\underline{y}_j,\cdots,\underline{y}_k\}$の‘ 1 ’対‘ 多 ’であり，フレームの長さが $\underline{x}_i.l > \sum_{s=j}^{k} \underline{y}_s.l$ であるとすると，

$$g_h.v = \frac{\underline{x}_i.l}{\sum_{s=j}^{k} \underline{y}_s.l} \sum_{s=j}^{k} \left(d(\underline{x}_i, \underline{y}_s) \cdot \underline{y}_s.l \right) \quad (13)$$

$$g_h.l = \underline{x}_i.l \quad (14)$$

これらの式は，圧縮前の時系列に対して TW を適用すると図 8 のようになるという考えから得られたものである．

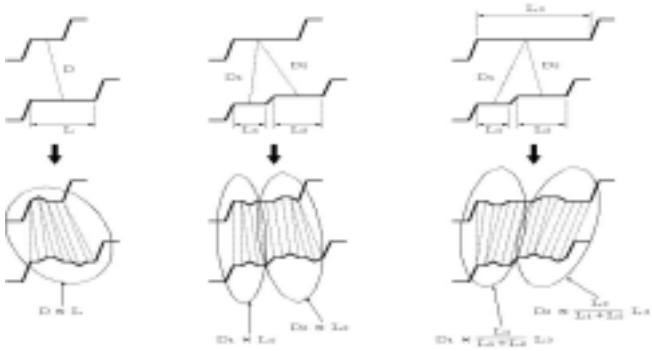

図 8　圧縮前の時系列に対するＴＷの適用

Fig.8　Application of TW to original time series

パターン毎の計算式から得られた n 個のグループ $g_1,\cdots,g_n$ の値と大きさから，ATW 距離を以下のように定義する．

$$ATW(X,Y) = \sqrt{\sum_{h=1}^{n} g_h.v} \Big/ \sum_{h=1}^{n} g_h.l \quad (15)$$

4. 実験結果

この章では，以下の 4 種類の手法の比較実験を行った結果について述べる．

- **[Euclidean]** ユークリッド距離
- **[TW]** 圧縮を行わずに TW を適用
- **[PDTW]** 固定長で圧縮してから TW を適用
- **[ATW]** 可変長で圧縮してから TW を適用

実験には，合成データセットと実データセットを使用した．合成データセットには，論文[2]で使用されている CBF データセットを用いた．CBF データセットには，Cylinder, Bell, Funnel の 3 種類のクラスが存在する（図 9）．実データセットには，Keogh,E.&Folias,T.(2002).The UCR Time Series Data Mining Archive (http://www.cs.ucr.edu/~eamonn/TSDMA/index.html)の時系列を使用した．これらの時系列の中から，異なる特徴を持つ 7 種類の時系列をクラスとして選択した．この 7 種類の時系列を平均値が 0，最小値と最大値の差が 1 となるように正規化した後，一定の大きさに分割して，それらを集めてデータセットとした．このデータセットに含まれる 7 種類のクラスのデータを図 10 に示す．

図 9　ＣＢＦデータ

Fig.9　CBF data

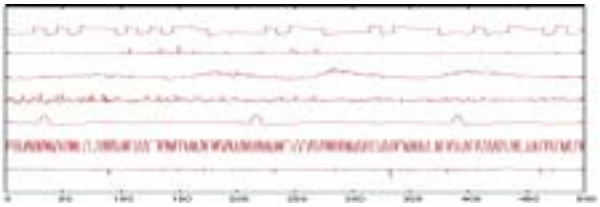

図 10　実データ

Fig.10　Examples of real data

この 2 つのデータセットを用いた分類実験の内容は以下の通りである．まず，それぞれのクラスからランダムに 1 つデータを選んで基準データとする．次に，残りの全てのデータを最も類似度の高い基準データのクラスに分類する．分類が終わったら，全てのデータの中で正しく分類されたデータの割合と要した時間を求める．これを繰り返して平均の結果を求める．

このように実験を行った結果を表 1, 2 と図 11, 12 に示す．図 11，12 のグラフは，PDTW と ATW に関してパラメータを変えて得られた結果を，横軸を実行時間，縦軸を正解率として線で結んだものである．このグラフから，今回提案した手法の ATW が, PDTW よりも優れた性能を持っているということができる．

	parameter	time(ms)	correct(%)
Euclidean		3.4	64.97
TW		1086.3	83.30
PDTW	2	270.1	77.28
	4	67.8	74.56
	8	18.2	68.14
ATW	0.1	366.9	80.26
	0.2	106.3	75.95
	0.4	8.5	80.54

表 1　ＣＢＦデータセットを使用した実験の結果

Table 1　Result of experiments with CBF dataset

	parameter	time(ms)	correct(%)
Euclidean		70	26.19
TW		100020	80.95
PDTW	2	26090	79.05
	4	6140	66.19
	8	1600	59.05
ATW	0.1	8320	82.86
	0.2	3810	81.90
	0.4	1740	67.62

表 2　実データセットを使用した実験の結果

Table 2　Result of experiments with real dataset

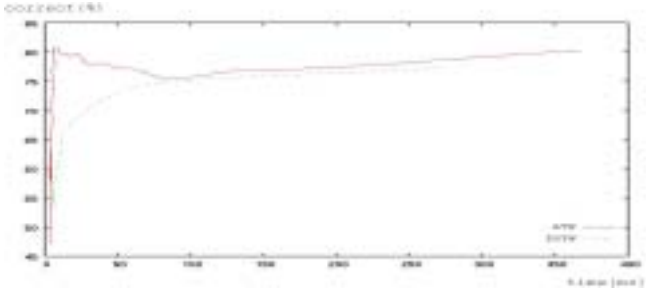

図 11　ＣＢＦデータセットを用いた実験結果

Fig.11　Result of experiments using CBF dataset

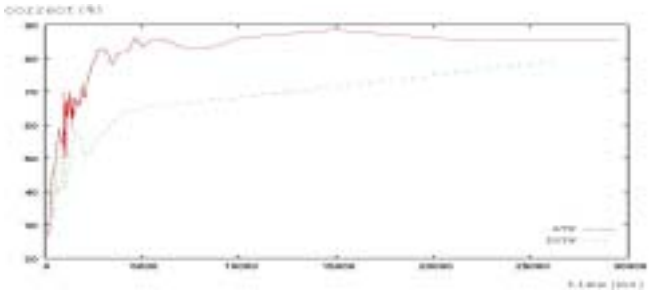

図 12　実データセットを用いた実験結果

Fig.12　Result of experiments using real dataset

5. 結論

本研究では, 時系列の類似度を定義する上でユークリッド距離よりも優れているタイムワーピング（TW）に対して，可変長圧縮の手法を用いることによって，欠点であった計算の遅さを改善することを試みた．実験によって，今回提案した手法（ATW）は，すでに提案されている固定長圧縮の手法（PDTW）よりも優れた性能を持っていることが実証された．

今後の課題としては, パラメータと実行時間と正解率の関係について考察して，適切なパラメータの自動設定を行う手法の提案ができれば良いと考えている．この他にも，今回提案した手法を応用して，時系列の類似検索においてフィルタリングを行う手法についても考案している．

[文献]

[1] Joseph B. Kruskall, and Mark Liberman, ``The Symmetric Time-Warping Problem: From Continuous to Discrete,'' Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparison, pp.125--161, Addison-Wesley, 1983.

[2] Keogh,E. and Pazzani,M. ``Scaling up Dynamic Time Warping for Datamining Applications,'' Proc. of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.285-289, Boston, MA, USA, August.2000

大桃 諭　Satoshi OOMOMO

2003 年筑波大学情報学類卒業．現在，同大大学院博士課程システム情報工学研究科在学中．データベースシステムに関する研究に従事．日本データベース学会学生会員．

陳 漢雄　Hanxiong CHEN

1993 年筑波大学大学院博士課程工学研究科了．現在，同大電子・情報工学系講師．データベースシステムに関する研究に従事．工博．日本データベース学会正会員．

古瀬 一隆　Kazutaka FURUSE

1993 年筑波大学大学院博士課程工学研究科了．現在，同大電子・情報工学系講師．データベースシステムに関する研究に従事．工博．日本データベース学会正会員．

大保 信夫　Nobuo OHBO

1968 年東京大学理学部卒．1970 年同大大学院修士課程了．同年同大理学部助手. 1980 年筑波大学電子･情報工学系講師. 1995 年同大電子・情報工学系教授．現在に至る．データベースシステムに関する研究に従事．理博．