経済データに対する値と形状に基づく時系列類似尺度の比較

白浜, 公章

(Citation)
国民経済雑誌,204(5):71-79

(Issue Date)
2011-11

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/81008373

(URL)
https://hdl.handle.net/20.500.14094/81008373

経済データに対する値と形状に基づく時系列類似尺度の比較

白　浜　公　章

国民経済雑誌　第 204 巻　第 5 号　抜刷
平 成 23 年 11 月

経済データに対する値と形状に基づく時系列類似尺度の比較

白　浜　公　章

日々の株価変動のような時系列データから，有用な傾向，規則性，異常性といったパターンを抽出するためには，時系列間の類似度を適切に測ることが重要になってくる。本論文では，２つの時系列に対して時間軸を伸縮させながら最適な対応付けを行う“Dynamic Time Warping”（DTW）をベースとして，各時点での値に基づく従来の DTW と，各時点での差分（すなわち，形状）に基づく“Derivative DTW”（DDTW）を比較する。実験では，2,940社の企業の株価変動を分類する問題を対象として，DTW と DDTW との比較を行い，株価変動の解析には，値と形状のどちらが重要か検証する。

キーワード　　データマイニング，時系列，類似尺度，DTW，DDTW

１　は　じ　め　に

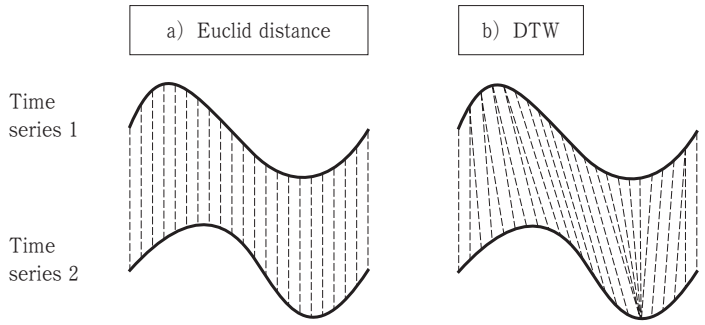
“データマイニング”とは，大量のデータの中から，過去には知られていなかった興味深いパターンを抽出する技術である（Han and Kamber 2006）。例えば，マーケティングの分野では，大量の購買データから，「パンを買う人の70％は牛乳も買う」といった併せ買いのパターンを抽出して，商品陳列，在庫管理，顧客の志向分析などに応用されている。また，医療の分野では，大量の電子カルテから，特定の病気と関連する要因をパターンとして抽出して，治療法の改善や新薬の開発などに応用されている。本論文では，経済データ，特に株価データを対象として，株価変動のパターンを抽出することを目的としている。抽出されたパターンは，株価予測や経営戦略の決定などに有用な知識となる。

株価変動のような時系列データからのパターン抽出では，時系列間の類似度を適切に測ることが重要になってくる。理由は，パターン抽出では，類似した時系列が何個存在するか，１つの時系列の中に類似した部分系列が何回出現するか，というように頻度をカウントすることが基本となるからである。株価変動の解析では，企業間での株価変動の期間は，全く同一ではなく，時間的なばらつきがあるという点を考慮しなければならない。言い換えると，２つの企業の株価は，全く同一の期間に上昇（もしくは，下降）する場合もあれば，片方が

他方より先に上昇（もしくは，下降）する場合も多々ある。そこで，時間的なばらつきに柔軟に類似度を測るために，２つの時系列に対して時間軸を伸縮させながら最適な対応付けを行う“Dynamic Time Warping”（DTW）（Keogh 2005）を用いる。

図１を用いて，DTW のアイデアを説明する。まず，図４（a）は，最も単純な類似尺度である“ユークリッド距離”を表している。ユークリッド距離では，２つの時系列における同一時点の値を対応付けているだけである。すなわち，各時点間を独立に扱っている。これに対して，図４（b）の DTW では，一方の時系列のある時点の値が，もう一方の時系列の複数の時点の値と対応付けられる。これにより，値の推移を考慮できるようになり，時系列の値の変動に時間的なばらつきがあっても柔軟に類似度を測ることが可能になる。

図１　ユークリッド距離（a）と DTW（b）による時系列の対応付けの比較



しかしながら，従来の DTW では，各時点での「値」を用いて２つの時系列を対応付けている点に問題がある。例えば，片方の時系列では値が上昇し，もう片方では値が下降していても，値自体に大差がなければ，これらの時系列を誤って類似していると判定してしまうことがある。そこで，“Derivative DTW”（DDTW）（Keogh and Pazzani 2001）という，各時点での「差分（微分値）」を用いて，２つの時系列を対応付ける DTW の拡張型を導入する。すなわち，DDTW では，形状に基づいて時系列を対応付けしていることになる。実験では，DTW と DDTW のどちらが，株価データに対する類似度計測に有用か比較検討する。また，DTW の枠組みで，値と差分の２つを同時に考慮して時系列を対応付ける手法についても検討する。

２　DTW と DDTW

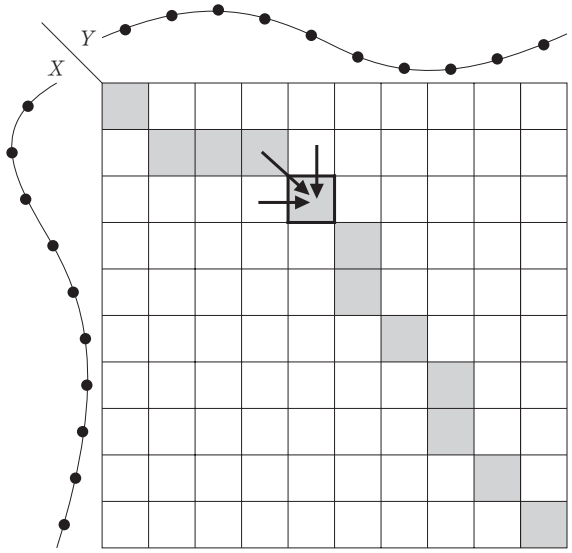
本節では，まず，DTW を用いた値に基づく時系列間の類似度算出法について説明する。次に，DTW を DDTW に拡張して，形状に基づく類似度算出法について述べる。さらに，DTW と DDTW を組み合わせて，値と形状の両方を考慮した類似尺度について説明する。最後に，得られた時系列間の類似度を用いたクラスタリング手法について概説する。

2.1　DTW：値に基づく時系列類似尺度

今，2つの時系列 $X=x_1, x_2, \cdots, x_M$（$x_i \in R$, $1 \leqq i \leqq M$），$Y=y_1, y_2, \cdots, y_N$（$y_i \in R$, $1 \leqq j \leqq N$）が与えられているとする。時間軸を伸縮させながら，X と Y の値を対応付けるために，図2のような行列を作成する。図2では，行（縦）方向に長さ $M=10$ の X，列（横）方向に長さ $N=10$ の Y が配置されている（$M \neq N$ の場合も同じである）。行列の要素 (i, j) には，x_i と y_j が対応付けられており，その値は，x_1 から x_i までの部分系列と y_1 から y_j までの部分系列の類似度を表す。例えば，図2の太線の要素 (3, 5) は，x_1 から x_3 までと y_1 から y_5 までの部分系列の類似度を表している。DTW では，このような行列を用いて，X と Y の最適な対応付け，すなわち類似度が最大となる対応付けを探索する。最終的に，図2の網掛けされた要素のような，どの時点の X の値とどの時点の Y の値が対応付けられたかを表すパスが得られる。

DTW では，以下の式を用いて，時系列間の類似度を算出する。

図2　DTW による時系列間の値の対応付けと非類似度の算出方法



$$D(i, j)=\min\begin{cases} D(i, j-1)+d(i, j) \\ D(i-1, j-1)+2\times d(i, j) \\ D(i-1, j)+d(i, j) \end{cases} \qquad (1)$$

$D(i, j)$ は，x_1 から x_i までの部分系列と y_1 から y_j までの部分系列の非類似度，$d(i, j)$ は x_i と y_i の非類似度を表す。つまり，一番上の式は，$D(i, j)$ を，x_1 から x_i と y_1 から y_{j-1} ま

での部分系列の非類似度 $D(i, j-1)$ と，x_i と y_j の非類似度 $d(i, j)$ を足し合わせて算出している。同様に，真ん中と一番下の式では，それぞれ $D(i-1, j-1)$ に $d(i, j)$，$D(i-1, j)$ に $d(i, j)$ を足し合わせて $D(i, j)$ を算出している。すなわち，式(１)では，直前までの部分系列の非類似度（$D(i, j-1)$，$D(i-1, j-1)$，$D(i-1, j)$）に x_i と y_j の非類似度 $d(i, j)$ を足し合わせて，部分系列の非類似度 $D(i, j)$ を反復的に算出している。特に，$d(i, j)$ が常に０以上であるとき，上記の反復によって，最適な部分系列の対応付け（$D(i, j)$ が最小）が求まることが保証されている。また，対応付けの基本要素となる $d(i, j)$ には x_i と y_j の差（$|x_i-y_j|$）が採用されることが多く，２つの時系列は，値に基づいて対応付けされていると言える。

図２を用いて，$D(i, j)$ の算出方法を図示する。今，太線の（3, 5）要素に対応する $D(3, 5)$ を計算しているとする。このとき，式(１)の $D(i, j)$ の最小化は，(3, 5) 要素に向かう３つのパスのいずれを選ぶかという問題に帰着できる。すなわち，式(１)の一番上の式が (3, 4) 要素，真ん中の式が (2, 4) 要素，一番下の式が (2, 5) 要素からのパスを表している。そして，図２の網掛けされた要素の配置から分かるように，結果的に (2, 4) 要素からのパス，すなわち式(１)の真ん中の式によって $D(3, 5)$ が算出されている。このような $D(i, j)$ の計算を，行列の (1, 1) 要素から始めて，１行もしくは１列ごとに $D(i, j)$ を算出し，$D(M, N)$（図２では $D(10, 10)$）が求まれば，X と Y の最適な対応付けと非類似度が求まったことになる。正確には，$D(M, N)$ が X，Y の長さ M，N に独立になるように，$D(M, N)/(M+N)$ が最終的な非類似度となる。

2.2　DDTW：形状に基づく時系列類似尺度

2.1 節の DTW が値に基づいて２つの時系列 X，Y を対応付けしている理由は，式(１)で，x_i と y_j の差として $d(i, j)$ を定義しているためである。DDTW では，形状に基づいて時系列を対応付けるために，以下のような微分係数の差として $d(i, j)$ を定義する。

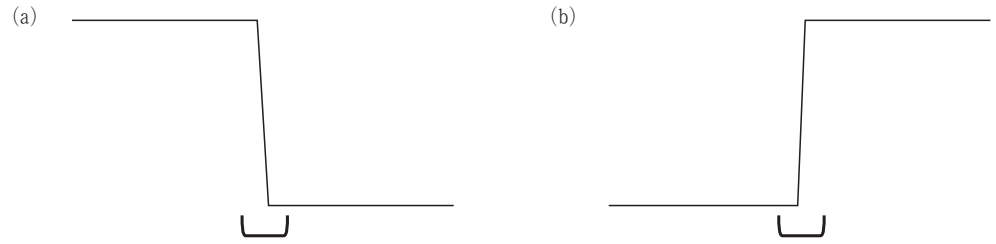
$$d(i, j)=|D(x_i)-D(y_j)| \quad \text{ここで} \quad D(x_i)=\frac{(x_i-x_{i-1})+((x_{i+1}-x_{i-1}/2))}{2} \qquad (2)$$

上式では，３つの時点 $i-1$，i，$i+1$ から，X の時点 i における微分係数 $D(x_i)$ を算出している。$D(y_j)$ についても，Y の時点 j における微分係数 $D(y_j)$ を同様に算出できる。最終的に，式(１)の枠組みで，微分係数の差としての $d(i, j)$ の和が最小になるように X と Y を対応付ければ，形状に基づいて対応付けしたことになる。

しかしながら，形状だけではうまく時系列を対応付けできないときがある。具体的には，図３のように，短時間に値が急激に変化する時系列に対して，DDTW はうまく機能しない。つまり，図３(a) と (b) の時系列は全く特徴の異なる時系列であるが，括弧で示した時間以

外では形状は同一であるため，誤って類似していると判定されてしまう。すなわち，形状だけでは，全く異なる値をとっている時系列を類似していると誤判定してしまう。

図 3　DDTW がうまく機能しない時系列の例



そこで，値と形状の両方を考慮した時系列の対応付けを行うために，以下のように，x_i と y_i の差とそれらの微分係数の差を足し合わせ，もしくは x_i と y_i の差と微分係数を掛け合わせて $d(i, j)$ を定義するアプローチについても検証する。すなわち，以下のように，式（1）の $d(i, j)$ を定義する。

$$d(i, j)=\begin{cases} d_1+d_2 \\ \text{もしくは，ここで，} d_1=|x_i-y_j|,\ d_2=|D(x_i)-D(y_j)| \\ d_1 \cdot d_2 \end{cases} \qquad (3)$$

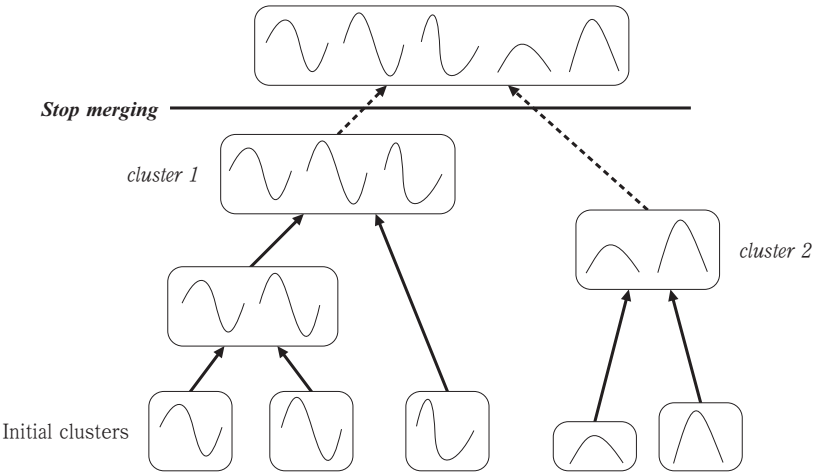
以下では，簡単のため，式（3）上部の $d(i, j)$ を用いた類似尺度を“DTW＋DDTW”，下部を用いた類似尺度を“DTW*DDTW”と呼ぶ。実験では，DTW，DDTW，DTW＋DDTW，DTW*DDTW を用いて，株価データに対して類似度を算出する。そして，2.3 節で述べるクラスタリング手法を用いて分類し，どの類似尺度が株価データの分類に最も適切であるか検証する。

2.3　クラスタリング

時系列間の類似度に基づいて，時系列を類似した時系列からなるクラスター（グループ）に分割する。ここで，N 個の時系列を K 個のクラスターに分割する場合，最大 K^N 通りの分割方法が存在する。しかしながら，全ての考えうる分割方法を検証して，最適な K 個のクラスターを決定することは，計算時間の観点から不可能である。そこで，何らかの指針に基づいて，最適な K 個のクラスターの近似解を求める必要がある。特に，本論文では，計算時間の高速性から，“階層的クラスタリング”を用いる（Jain et al. 1999）。図 4 に示すように，階層的クラスタリングは，1 つの時系列だけからなるクラスターから始めて，最も類似した 2 つのクラスターをボトムアップにマージしていくアプローチである。ただし，あま

りに類似していないクラスターをマージしないように，クラスター間の類似度がしきい値以下になったとき，もしくはクラスター数が特定の数以下になったとき，クラスタリングを終了する。このような終了条件を用いて，図４では，最終的に，３つの時系列からなる *cluster 1* と２つの時系列からなる *cluster 2* という２つのクラスターが抽出されている。

図４　ボトムアップ型階層的クラスタリングの概念図



階層的クラスタリングにおいて，DTW（もしくは，DDTW，DTW＋DDTW，DTW*DDTW）を用いて算出された時系列間の非類似度を用いて，下記のようにクラスター間の非類似度を定義する。

$$DC(c_i,\ c_j)=1/n_i n_j \sum_{x_i \in c_i} \sum_{x_j \in c_j} D(x_i,\ x_j) \qquad (4)$$

ここで，クラスター c_i，c_j にはそれぞれ n_i，n_j 個の時系列が含まれているとし，c_i，c_j に含まれる時系列を x_i，x_j と表現する。すなわち，c_i と c_j の非類似度 $DC(c_i,\ c_j)$ は，クラスターに含まれる時系列の非類似度の平均である。最終的に，上記のクラスター間の非類似度を用いて，類似した時系列を含むクラスターを反復的にマージしていく。

３　実　験　結　果

実験データとして，“トムソン・ロイター　データストリーム（http://thomsonreuters.com/products_services/financial/financial_products/a-z/datastream/）”を用いて，2002/3/1 から2009/8/24までの間，日本の株式市場に上場されていた2,940社の株価変動を表す時系列を収集した。計算時間の問題から，各社の１日ごとの株価変動を表す長さ1,955の時系列を，平日５日間の株価の平均をとって長さ391の時系列に簡単化している。また，ある企業の株価

は1,000円台，別の企業の株価は100円台というような，値域の異なる時系列を比較するために，各時系列を平均 0 ，分散 1 に正規化している。上記の前処理を施した後で，DTW，DDTW，DTW＋DDTW，DTW*DDTW を用いて，時系列間の類似度を算出しクラスタリングを行った。ここで，類似尺度の違いだけを検証するために，いずれの類似尺度を用いた場合でも500個のクラスターに分割するようにしている。

表 1 に，クラスタリング結果の概要を示す。2 行目は，500個のクラスターの中で，最大のクラスターのサイズ（クラスターが含んでいる時系列の数）を表している。3 行目は，最小のクラスターのサイズを表しているが，どのクラスタリングにおいても，1 つの時系列しか含まないクラスターが多数抽出されていたため，括弧内に，1 つの時系列しか含まないクラスターの数を示している。表 1 の 2 行目から，DDTW，DTW*DDTW を用いたクラスタリングでは，全2,923個の時系列のうちの大半が 1 つのクラスターにまとめられてしまっていることが分かる。この結果から，形状（DDTW）だけでは，株価データを区別するには不十分であると言える。実際，DDTW によって算出される非類似度は，どの時系列のペアに対しても小さな値になっており違いが少ない。DTW*DDTW に関しては，DTW と DDTW により算出された非類似度の積をとると，DDTW により算出された小さな非類似度の値の影響を大きく受けてしまい，時系列間の区別がつかなくなってしまっていると考えられる。

表 1　DTW，DDTW，DTW＋DDTW，DTW*DDTW を用いたクラスタリング結果の概要

	DTW	DDTW	DTW＋DDTW	DTW*DDTW
クラスターの最大サイズ	551	2,193	642	1,519
クラスターの最小サイズ	1（290）	1（429）	1（337）	1（331）

残る DTW と DTW＋DDTW を比較したところ，定性的[1)]な評価ではあるが，DTW＋DDTW が株価データを分類するのに最も適切であると考えられる。具体的には，DTW では，図 5 に示す 2 つの時系列のように，「形状は大きく異なっているが，値が平均的に類似している時系列」が誤って同一のクラスターに分類されることがしばしばある。一方，DTW＋DDTW では，図 6 の時系列のように，「値は多少違っているが，形状が大まかに類似した時系列」が誤って同一のクラスターに入ることがある。ただし，得られたクラスターを全て確認したところ，DTW＋DDTW が失敗する割合よりも，DTW が失敗する割合の方が大幅に高かった。ゆえに，DTW＋DDTW が，株価データを分類・解析するのに最適な類似尺度であると考えられる。

図５　DTW による類似度計測がうまく機能しない時系列の例

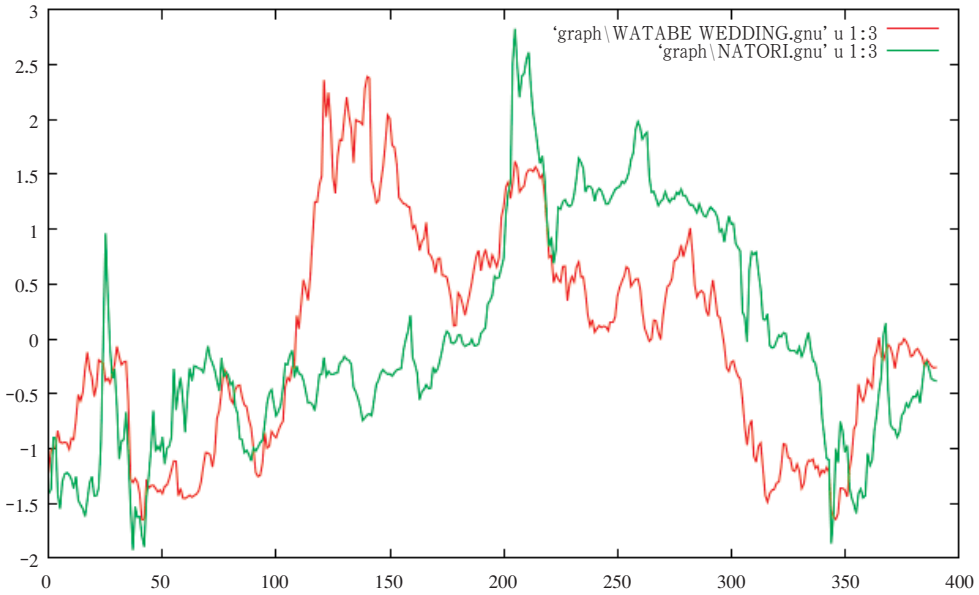
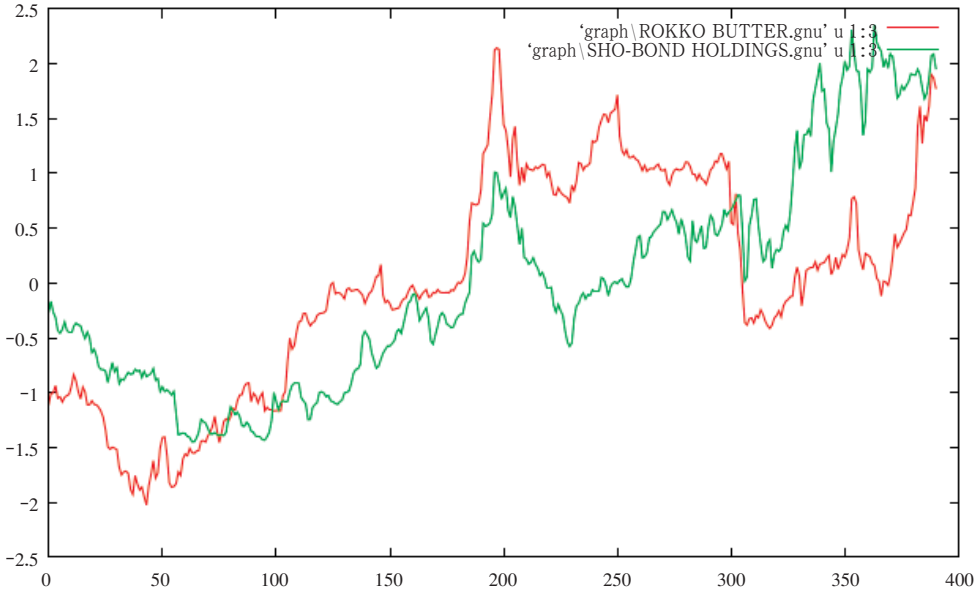


図６　DTW＋DDTW による類似度計測がうまく機能しない時系列の例



4　お　わ　り　に

本論文では，株価データを対象として，4種類の時系列類似尺度の性能比較を行った。具体的には，値に基づく時系列類似尺度DTW，形状に基づく類似尺度DDTW，値と形状の両方を考慮したDTW＋DDTWとDTW*DDTWを比較した。実験結果から，DTW＋DDTWが株価データを分類・解析するのに最適な類似尺度であることが分かった。今後は，DTW＋DDTWを以下のように改良する予定である。まず，同じ重み（すなわち，1）でDTWとDDTWの値を単純に足し合わせるのではなく，例えば交差検定[2)]などを行って，DTWとDDTWの値の重みを決定するアプローチについて検討する予定である。また，常に同一の重みを用いるよりも，比較する時系列によって重みを動的に変えるべきであると考えられる。そこで，例えば時系列を直線近似して，時系列に含まれる値の分布や形状を大まかに推定する。そして，2つの時系列の値の分布が大きく異なっている，もしくは形状が大きく異なっていると推定された場合は，DTW，もしくはDDTWの重みを大きくするアプローチについて検討する予定である。

注

本論文を作成するにあたり，研究背景や解析手法等に関して，幅広く助言いただきました神戸大学大学院工学研究科　上原邦昭教授に感謝いたします。また，実験に関して多大な助力をいただいた，神戸大学大学院システム情報学研究科　辻本貴昭氏に感謝いたします。

1）どの時系列とどの時系列が同一のクラスターに属すべきかを表す正解データが存在しないため，著者が各クラスターに含まれる時系列を確認して，類似した時系列であるかどうか目視で検証している。

2）人手により，類似していると判定されるべき時系列のペアをあらかじめ注釈付けしておく必要がある。

参　考　文　献

Han J. and Kamber M. (2006), *Data Mining: Concepts and Techniques* (*Second Edition*), Morgan Kaufmann Publishers.

Jain A., Murty M. and Flynn P. (1999), "Data Clustering: A Review," *ACM Computing Surveys,* Vol. 31, No. 3, pp. 264-323.

Keogh E. (2005), "Exact Indexing of Dynamic Time Warping," *Proc. of the 28-th International Conference on Very Large Data Bases,* pp. 406-417.

Keogh E. and Pazzani M. (2001), "Derivative Dynamic Time Warping," *Proc. of the 2001 SIAM International Conference on Data Mining,* pp. 1-11.