知能と複雑系　132－30
(2003. 3. 15)

多重スケールマッチングにより導出される類似度の性質

平野　章二†　　津本　周作†

† 島根医科大学医学部医学科医療情報学講座　〒693-8501 島根県出雲市塩冶町 89-1
E-mail: †hirano@ieee.org, tsumoto@computer.org

あらまし　本稿では，多重スケールマッチングにより導出される時系列の系列間相違度の性質について，いくつかの基本的波形において実験的に調べた結果を報告する．多重スケールマッチングは，局所的な部分系列の相違度を系列全体に渡り積算して得られる相違度の総和を最小化し，かつ対応漏れや重複のない最適対応を全ての平滑化スケールにわたって探索する方法である．これまでに我々は，同方法を時系列データの比較分類に用いるべく，回転角，経路長，位相，勾配の４成分からなる相違度基準を提案し，様々な時系列の類型化を試みてきた．しかし，最終的な最適対応関係は対応漏れの有無等様々な要素が複雑に絡み合い決定されるため，類型化の結果から各成分が実際の相違度にどのように寄与しているかを直感的に把握することは困難であった．そこで，正弦波を基本とし，振幅，位相などに変化を与えた評価用データを用意し，導出される相違度からそれぞれの要素の寄与の定量的評価を試みた．
キーワード　多重スケールマッチング，時系列解析，相違度

On Characteristics of Dissimilarity Measures for Multiscale Matching

Shoji HIRANO† and Shusaku TSUMOTO†

† Department of Medical Informatics, Shimane Medical University, School of Medicine
89-1 Enya-cho, Izumo, Shimane 693-8501, Japan
E-mail: †hirano@ieee.org, tsumoto@computer.org

Abstract　This paper presents some properties of the dissimilarity measures used in the multiscale matching. Multiscale matching is a method to compare planar curves by partly changing observation scales. For the multiscale comparison of one-dimensional temporal sequences, we proposed the dissimilarity measure of subsequences consists of four components, that are, rotation angle, length, phase and gradient. On the synthetic data, we empirically examined contribution of the four components to the resultant dissimilarity between subsequences.
Key words　multiscale matching, temporal data mining, proximity measure

1. 背　　景

時系列データマイニングは，データに潜む未知の時間的構成要素を発見する手段として，経営分析，医療をはじめ様々な領域で注目を集めている．時系列データの特徴の１つとして，内在するイベントが時間的連続性伴う点が挙げられる．例えば脳波のデータの場合では，発作時において特定の波形が繰り返し観察され，そのパターンに疾患の特徴が見られる場合も多い．したがって，時系列データの解析においては，データ中に頻回に観察される，共通の推移パターンをもった一連の部分系列を発見する作業が重要となる．このような目的で提案されたアプローチの１つが部分系列の類型化であり，これまでに様々な方法が提案されてきた [1].

部分系列の類型化において考慮を要する点として，部分系列の切り出し範囲の選択が挙げられる．これはイベントの観察期間に対応するものであり，獲得されるイベントの種類に直接的に影響するものであることからその選択には注意を要する．特にイベントに関する予備知識を持たない場合は特定の幅に固定すべきものではない．データマイニングにおいては，多くの場合，異なる期間で切り出した複数データセットを準備し，それぞれ類型化することで様々な長さをもつイベントに対応する．しかしながら，このアプローチには以下の問題がある．

（1）　データの特徴を無視して部分系列が切り出される

部分系列は単に指定した期間で原系列をマスキングしたものであり，波形としての特徴を加味して得られたものではない．したがって，あるパターンを持つ部分系列が類型化の結果得られたとしても，その部分系列が果たして本来のイベントの発生期間を適切に表現するか，例えば単純に，その開始位置と終了位置が本来のイベントの開始/終了位置と正しく対応するかどうかについても，保証されていない．

（2） 異なる期間をもつ連続したイベントを抽出できない

異なる期間で切り出された部分系列を連結する場合，それらの連続性が担保されない．すなわち，ある 2 つの異なる長さをもつ部分系列を連結しても，その端点の連続性が保証されないため，連結後の部分系列が原系列上の一連の区間と正確に対応するとは限らない．これは，前項に述べたとおり，原系列の構造情報を参照せずに部分系列を切り出していることによる．また，部分系列間の相違度の算出法に依存するが，一般的なユークリッド距離を用いる場合，比較対象の系列長が異なると距離の算出が困難となることも考慮すべきである．以上から，例えば 1 週間の増加のあと 2 週間の減少が生じ，その後再び 1 週間の増加が続くような，異なる期間をもつ連続したイベントの特徴を評価できない．

これらの点を克服するため，我々は多重スケールマッチング [2] に基づく系列の比較分類法を提案した [3]．多重スケールマッチングは元々2 次元図形の比較法として提案されたもので，図形輪郭を変曲点を基準に部分輪郭化し，それらの対応関係を調べることで図形構成要素の対応関係を認識する方法である．この方法では，(1) 原輪郭を様々なスケールパラメータを持つガウス関数と畳み込み平滑化し，様々なスケールで対象を観察する，(2) 変極点を基準に部分輪郭を切り出し，平滑化度の変化に伴う変極点構造の変化を追跡する，ことで，部分輪郭の連続性を損なうことなく，部分ごとに観察スケールを変化させながらマッチングを行なうことを可能としている．我々はこの特性を時系列解析に応用し，図 1. に示すように，(1) 変極点に基づく部分系列の切り出し，(2) 連続性を担保した異スケール間比較，を行なうことで上述した欠点の克服を試み，これまでに肝炎検査データ等においてその有効性を明らかにしてきた．

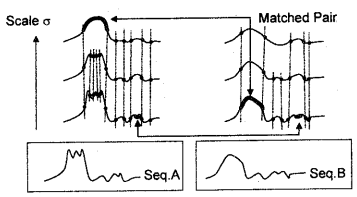

図 1 多重スケールマッチング

2 次元図形を対象とした多重スケールマッチングでは，異なるスケールで記述された部分系列間の類似性を評価するため，回転角，比系列長等の相似変換に不変な特性を用いて相違度を定義する．しかしながら，時系列データでは，振幅をはじめ系列値そのものが重要な要素となるため，異なる観点から相違度を再定義する必要が生じる．我々は，前出の 2 項に加え，位相，勾配に関する項を加えた新たな相違度基準を提案してきたが，実際の系列間相違度はこれらが複雑に絡み合い導出されるため，一般的なユークリッド距離が与える系列間相違度に比べ直感的に各項の寄与が理解しづらい側面があった．

本稿では，正弦波を基本とし，振幅，位相等に様々な変化を与えた人工データセットに多重スケールマッチングを適用し，系列間類似度の特性を解析した結果を報告する．

2. 多重スケールマッチングにおける相違度

多重スケールマッチングは，2 つの系列を A，B を入力とし，それぞれの部分系列の対応関係を異なるスケールに渡って探索し求めるものである．その手続きは，(1) 入力系列を様々な平滑化度で表現し，変極点を基準に部分系列セットを構築 (2) 系列 A,B それぞれの部分系列セットから，対応漏れがなく，かつ相違度を最小化する最適部分系列対応を求める，ことに帰着する．ここでは，詳細な手続きを割愛し，部分系列の比較に必要な相違度の導出について述べる．マッチングの詳細については文献 [4] を参照されたい．

まず，入力系列 A について，時刻 t をパラメータとする関数 $x(t)$ で表現する．このとき，スケール σ における系列は，$x(t)$ とスケールファクター σ をもつガウス関数 $g(t,\sigma)$ との畳み込みとして以下のように表現できる．

$$X(t,\sigma) = x(t) \otimes g(t,\sigma) = \int_{-\infty}^{+\infty} x(u) \frac{1}{\sigma\sqrt{2\pi}} e^{-(t-u)^2/2\sigma^2} du \qquad (1)$$

ここで，ガウス関数は畳み込みにおける加算重みの分布を規定する．その特性から明らかなように，σ が大きいほど遠方の加算重みが増し，より広範にわたる平滑化が行なわれる．すなわち，スケールの増加に従い近傍値との平滑化が進み，より変曲点の少ない滑らかな系列が得られる．系列上の各点における曲率は次式で与えられる．

$$K(t,\sigma) = \frac{X''}{(1+X'^2)^{3/2}} \qquad (2)$$

ここで，X'，X'' は $X(t,\sigma)$ の t による 1 次および 2 次微分である．$X(t,\sigma)$ の m 次微分 $X^{(m)}(t,\sigma)$ は，$x(t)$ と $g(t,\sigma)$ の m 次微分 $g^{(m)}(t,\sigma)$ の畳み込みとして次式により与えられる．

$$X^{(m)}(t,\sigma) = \frac{\partial^m X(t,\sigma)}{\partial t^m} = x(t) \otimes g^{(m)}(t,\sigma) \qquad (3)$$

次に，曲率の符号の変化から系列上の変曲点の位置を求め，隣接する変曲点を両端とする部分輪郭，すなわち凹凸セグメントを構築する．スケール $\sigma^{(k)}$ における検査値系列 $\mathbf{A}^{(k)}$ を N 個のセグメントの集合とすると，

$$\mathbf{A}^{(k)} = \left\{ a_i^{(k)} \mid i = 1, 2, \cdots, N^{(k)} \right\}$$

ここで，$a_i^{(k)}$ はスケール $\sigma^{(k)}$ における i 番目のセグメントを示す．ここまでの処理をもう 1 つの入力系列 B に適用し，以下のような系列組 $\mathbf{B}^{(h)}$ を得る．

$$\mathbf{B}^{(h)} = \left\{ b_j^{(h)} \mid j = 1, 2, \cdots, M^{(h)} \right\} \qquad (4)$$

ここで，$\sigma^{(h)}$ は系列 B の観察スケールである．多重スケールマッチングにおけるマッチング手続きは，図 1 に示す通り，A,B を構成する全てのセグメント組から，対応漏れの無い，かつこのセグメント組の相違度の総和を最小にする組を探索することに相当する．

上田ら [4] は，2 次元図形の部分輪郭間 $a_i^{(k)}$ と $b_i^{(h)}$ の相違度

$d(a_i^{(k)}, b_j^{(h)})$ を次式により定義した．

$$(a_i^{(k)}, b_j^{(h)}) = \frac{|\theta_{a_i}^{(k)} - \theta_{b_j}^{(h)}|}{\theta_{a_i}^{(k)} + \theta_{b_j}^{(h)}} \left| \frac{l_{a_i}^{(k)}}{L_A^{(k)}} - \frac{l_{b_j}^{(h)}}{L_B^{(h)}} \right| \tag{5}$$

ここで，$\theta_{a_i}^{(k)}$ および $\theta_{b_j}^{(h)}$ は各セグメントに沿った接ベクトルの回転角，$l_{a_i}^{(k)}$ および $l_{b_j}^{(h)}$ は各セグメントの長さ，$L_A^{(k)}$ および $L_B^{(h)}$ は対象系列 A，B のスケール $\sigma^{(k)}$，$\sigma^{(h)}$ における総セグメント長をそれぞれ示す．この定義による相違度は明らかに回転，拡大縮小などの相似変換に不変であり，図形の位置や傾きが一定でない条件下における図形認識や大きな変形を伴う物体の認識に適したものといえる．

一方，この方法を 1 次元の系列に適用する場合，以下のような問題点が生じる．

（1） 2 次元輪郭の回転は 1 次元系列の位相シフトに相当するため，系列間の位相差が吸収され，相違度に反映されない．すなわち，早期に発生したイベントと晩期に発生したイベントを時期に基づき区別することができない．

（2） 系列の勾配を評価できない．セグメント回転角はセグメントに沿った接ベクトルの回転角の総和として定義されるため，例えば縦軸（検査値軸）について対称関係にある 2 つのセグメント間では，勾配の符号が反転していても，すなわちセグメント両端点での系列値の差の正負が反転していても，相違度が 0 になる．このことは，その系列全体が上昇トレンドであるのか，下降トレンドであるのかを認識できないことを意味する．これらに対応するため，我々は相違度の定義を以下のとおり拡張した．

$$d(a_i^{(k)}, b_j^{(h)}) = \max(\theta, l, \phi, g), \tag{6}$$

ここで，θ, l, d and g はそれぞれ部分系列の回転角，長さ，位相，勾配の相違度を表したもので，以下のとおり定義する．

$$\theta(a_i^{(k)}, b_j^{(h)}) = \frac{|\theta_{a_i}^{(k)} - \theta_{b_j}^{(h)}|}{\theta_{a_i}^{(k)} + \theta_{b_j}^{(h)}} \tag{7}$$

$$l(a_i^{(k)}, b_j^{(h)}) = \left| \frac{l_{a_i}^{(k)}}{L_A^{(k)}} - \frac{l_{b_j}^{(h)}}{L_B^{(h)}} \right|, \tag{8}$$

$$\phi(a_i^{(k)}, b_j^{(h)}) = \left| \frac{\phi_{a_i}^{(k)}}{\Phi_A^{(k)}} - \frac{\phi_{b_j}^{(h)}}{\Phi_B^{(h)}} \right|, \tag{9}$$

$$g(a_i^{(k)}, b_j^{(h)}) = \begin{cases} 1, & \text{if } g_{a_i}^{(k)} \times g_{b_j}^{(h)} < 0 \\ \left| g_{a_i}^{(k)} - g_{b_j}^{(h)} \right|, & \text{otherwise.} \end{cases} \tag{10}$$

ここで，式 (6) において max をとる項のうち，$\theta(a_i^{(k)}, b_j^{(h)})$ 及び $l(a_i^{(k)}, b_j^{(h)})$ はそれぞれ式 (5) に示した上田らの定義における積項の第 1 項及び第 2 項と同一である．残る 2 つのうち，ϕ は位相差を特徴付けるもので，系列 A について，セグメント $a_i^{(k)}$ の 開始点位置 $\phi_{a_i}^{(k)}$ を系列 A の収集期間 $\Phi_A^{(k)}$ で除して正規化したものから系列 B のそれを引いたものである．最後の g は勾配差を特徴付けるもので，系列 A について，セグメント $a_i^{(k)}$ の両端の系列値をその標準偏差 σ で除したものから，系列 B のそれを引いたものである．図 2. にそれぞれの要素を図示する．この拡張により，(1) 系列値の上昇/下降の激しさ，(2) イベントの期間，(3) イベントの発生時期，(4) 系列値のトレンド，の 4 つの視点からイベントの類似性を評価している．また，これらの項を max により結合することで，最も違いの大きい要素を強調し，かつ系列間の分離度が向上するように配慮している．

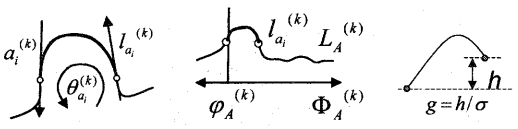

図 2 相違度の各要素

しかしながら，実際のマッチング過程では，すべてのセグメント組にわたる相違度の積算値を最小化し，かつ対応漏れの無いセグメント組が最終的に選択されるため，必ずしも局所的に最適な要素の積み重ねがそのまま結果となるわけではない．このことから，局所的な性質の変化が最終的にどのような形で大局的な相違度の変化として現れるのか，ひいては各要素が実際にどのように寄与しているのか，直感的に把握することが困難となっている．

3. 実験結果

前節に示した相違度の特性理解を目的として，人工データ上で多重スケールマッチングを適用し，得られる相違度の特性を調べる実験を行なった．実験に用いた系列は以下の 8 種類である．

$$w_1 : y = \sin(2.5t) \tag{11}$$

$$w_2 : y = 2\sin(2.5t) \tag{12}$$

$$w_3 : y = 0.5\sin(2.5t) \tag{13}$$

$$w_4 : y = \sin(2.5t + 0.5\pi) \tag{14}$$

$$w_5 : y = \sin(2.5t) + 0.2(t - 9.0) \tag{15}$$

$$w_6 : y = \sin(2.5t) - 0.2(t - 9.0) \tag{16}$$

$$w_7 : y = 0.5e^{0.1t}\sin(2.5t) \tag{17}$$

$$w_8 : y = 0.5e^{-0.1t+0.6\pi}\sin(2.5t) \tag{18}$$

t の値域 ($= \Phi$) は $0 \leqq t < 6\pi$ とし，これを $1/500T$ 間隔でサンプリングすることで各 500 点からなる時系列を作成した．肝炎データ等の実データにおけるマッチングと環境を同じとし，スケール σ は 1.0 から 1.0 刻みで 30 段階に変化させた．なお，これらの系列は何れも単一の正弦波を基本波としており，スケール変化が生じても基本的にセグメントの置換は生じない．実装上，端点近傍で例外処理が必要のためセグメント置換が生じるが，すべてのケースで同一であるため，その影響は相殺されるものとして取り扱う．

上記系列のうち，w_1 は他の系列の基礎となる正弦波系列である．w_2 及び w_3 は，それぞれ w_1 の振幅を倍増あるいは半減させたものである．これらの系列は，w_1 と振幅のみ異なり，変極点の位置をはじめ，各部分系列の長さの寄与，位相，勾配の

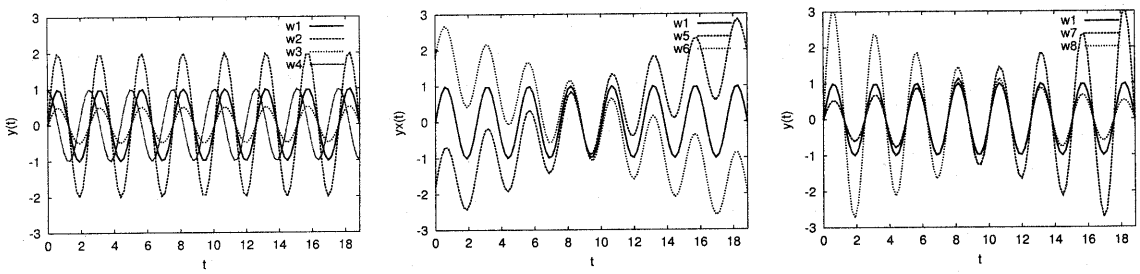

図 3 実験系列．左：w_1–w_4．中：w_5,w_6．右：w_7,w_8．

何れも同じになることから，回転角が上昇/下降の鋭さをどの程度表現するか評価するために用いている．次に，w_4 は，振幅を同じに保ちつつ w_1 の位相を 0.5π だけずらしたもので，位相項の評価に用いている．w_5 及び w_6 は，w_1 に対して t に比例する全体的な増加あるいは減少トレンドを付与したもので，勾配項を評価している．最後に w_7 及び w_8 は，t に対し非線形の振幅変化を与えた場合の相違度の変化を調べるために用いている．なお，長さの寄与のみが独立して変化する波形は作成が困難であるため，今回の評価では対象外とした．図 3. に，w_1 から w_8 までの各系列を示す．

表 1 に，各系列組に対し多重スケールマッチングを適用して得られた相違度行列を示す．ここでは相違度の基礎的性質をあわせて評価するため全ての行列要素を示しており，同表から全ての例で相違度は非負 ($d(w_i) \geq 0$) であり，反射性 ($d(w_i, w_i) = 0$) と対称性 ($d(w_i, w_j) = d(w_j, w_i)$) を満足していることが分かる．この 3 つの性質については実データにおいても成立するが，三角不等式については特に 3 系列の長さが著しく異なる場合やマッチングが成立しない場合に満足しないことが分かっている．

まず，w_1 と w_2,w_3 の比較から，$d(w_2, w_3) > d(w_1, w_2) > d(w_1, w_3)$ となり，振幅の差が大きいほど相違度が大きくなることが分かる．前述の通り，回転角以外の要素は全て同じになることから，これは回転角の項が振幅の差を形状の差として捉えていることを示している．次に，w_4 と w_1 の比較から，位相差が評価され相違度に現れていることが分かる．w_2,w_3 と合わせてみると，振幅の場合と異なり，この例では振幅差と相違度が必ずしも比例していない．これは，位相差によって端部で系列 w_4 の形状が変化しており，マッチングがより上位のスケールで行われ，その場合の置換コストが積算されているためと考えられる．続いて，w_5 及び w_6 を w_1 と比較すると，$d(w_5, w_6) >> d(w_1, w_6) > d(w_1, w_5)$ となり，平坦な w_1 に比べ，上昇，下降のトレンドが異なる w_5 と w_6 の間で大きな相違度が与えられていることが分かる．$d(w_1, w_5)$ と $d(w_1, w_6)$ は同一ではないが，これは作成した系列の性質上，開始直後に負から上方に振れる上昇系列の方がより小さい短い経路で最初に変極するためである．最後に w_7 及び w_8 に着目すると，$d(w_7, w_8) >> d(w_1, w_7) > d(w_1, w_8)$ となり，振幅の差に起因する形状の差が回転角の差として積算されて，その差が大きいほど大きな相違度が与えられていることが分かる．

表 1 相違度行列

	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8
w_1	0.000	6.317	5.008	7.087	5.949	6.725	6.499	7.281
w_2	6.317	0.000	8.519	8.244	8.095	7.421	7.052	6.327
w_3	5.008	8.519	0.000	9.178	7.926	7.327	6.448	8.099
w_4	7.087	8.244	9.178	0.000	8.980	7.683	8.792	6.989
w_5	5.949	8.095	7.926	8.980	0.000	10.09	7.952	6.868
w_6	6.725	7.421	7.327	7.683	10.09	0.000	6.870	7.982
w_7	6.499	7.052	6.448	8.792	7.952	6.870	0.000	11.00
w_8	7.281	6.327	8.099	6.989	6.868	7.982	11.00	0.000

4. むすび

本稿では，多重スケールマッチングで導出される系列間相違度の基本的性質について，正弦波に位相と振幅の変化を与えたテスト系列を用いて調べた結果を報告した．極めて単純化した条件下であるが，振幅の差が回転角の差として表現されていること，位相の差，トレンドの差も相違度に表現されていることが確認できた．しかし，それらの増減が必ずしも線形ではないことも示され，より直感にそぐう形での相違度の定義方法を考察する必要もあると考えられた．

5. 謝辞

本研究の一部は文部科学省科学研究費補助金特定領域研究 (領域番号 758)「情報洪水時代におけるアクティブマイニングの実現」の助成による．

文献

[1] E. Keogh (2001): Mining and Indexing Time Series Data. Tutorial at the 2001 IEEE International Conference on Data Mining.

[2] F. Mokhtarian and A. K. Mackworth (1986): Scale-based Description and Recognition of planar Curves and Two Dimensional Shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-8(1): 24-43

[3] S. Hirano and S. Tsumoto (2002): Mining Similar Temporal Patterns in Long Time-series Data and Its Application to Medicine. Proceedings of the IEEE 2002 International Conference on Data Mining: pp. 219–226.

[4] N. Ueda and S. Suzuki (1990): A Matching Algorithm of Deformed Planar Curves Using Multiscale Convex/Concave Structures. IEICE Transactions on Information and Systems, J73-D-II(7): 992–1000.