知 能 と 複 雑 系 128−12
（２００２．５．２３）

推移パターンの類似性に基づく時系列臨床検査データの解析

平野章二　　孫　暁光　　津本周作
島根医科大学医学部医療情報学講座

概要:　本稿では、位相情報を重視した多重スケールマッチングとラフクラスタリングによる時系列臨床検査データベースの解析法を提案する。多重スケールマッチングは、スケール変化に伴う部分輪郭の階層構造の変化を認識し、全体の連続性を崩さないように制限を加えつつスケール間マッチングを行う手法である。一方、ラフクラスタリングは、重心や分散を使用せず、相対的類似度と識別不能性に基づき対象を分類する方法である。本方法ではこれらを組み合わせ、多重スケールマッチングにより系列間の相対的類似度を求め、それをもとにラフクラスタリングを適用して系列を分類する。その際、多重スケールマッチングの類似度関数に位相シフトの抑制項を加え、過度の位相シフトを伴う誤対応を抑止する。実験では、肝炎データに本方法を適用し、類似系列が同一のクラスタに分類されるとともに、部分系列の対応関係が正しく獲得されることを示す。

Analysis of time-series medical data based on similarity of convex/concave structure of sequences

Shoji Hirano　　Xiaoguang Sun　　Shusaku Tsumoto
Department of Medical Informatics, Shimane Medical University

Abstract: This paper presents a method for analyzing time-series data on laboratory examinations based on the phase-constraint multiscale matching and rough clustering. Multiscale matching compares two subsequences throughout various scales of view. It has an advantage of preserving connectivity of subsequences even if the subsequences are represented at different scales. Rough clustering groups up objects according not to the topographic measures such as the center or deviance of objects in a cluster but to the relative similarity and indiscernibility of objects. We use multiscale matching to obtain similarity of sequences and rough clustering to cluster the sequences according to the obtained similarity. We slightly modified dissimilarity measure in multiscale matching so that it suppresses excessive shift of the phase that causes incorrect matching results. Experimental results on the hepatitis dataset show that the proposed method successfully clustered similar sequences into an independent cluster, and that correspondence of subsequences are also successfully captured.

1　はじめに

病院情報システム (Hospital Information System, HIS) は 1980 年代に導入が進み、10 余年の運用を経た現在では膨大な診療情報を蓄積するに至った。このような長期間にわたる継続的運用の成果の１つとして、数年から数十年の長期に渡り継続的に記録された同一患者のデータを用いることで、病態の時系列的変化を観察可能としたことが挙げられる。しかし、現状ではこれらの時系列情報は患者個人単位での参照、あるいは少数患者間での比較に利用されているのみであり、大規模データベースのもつスケールメリットを有効に活用できていない。これは、データの収集間隔が検査日間隔に対応して数日から数ヶ月の幅で不定期に変化すること、検査の有無により項目単位で欠損値が生じることなどから、データが不均質なものとなり、異なる患者を横断的に比較することが困難であることによる。

このようなデータにおいては、連続したデータが存在する場合にはより詳細な特徴をもとに、存在しない場合にはより大局的な特徴をもとにと、視野スケールを変化させつつ比較を行う必要がある。時系列データの特徴解析においては、波形が幾つかの基本波形の合成からなると考え、その主要な成分を比較する手法がしばしは用いられる。例えば Agrawal ら [1] および Chan ら [2] はぞれぞれ離散フーリエ変換 (Discrete Fourier Transformation,DFT) および離散ウエーブレット変換 (Discrete Wavelet Transformation, DWT) を適用し、周波数あるいは時間-周波数成分の類似性から系列間の類似性を求める方法を提案している。また、Korn ら [3] は特異値分解 (singular value decomposition, SVD) を用いて系列を幾つかの簡単な固有波に分解し、その類似性から系列間の類似性を求めている。一方、もう１つのアプローチとして、部分系列ごとに波形の類似性を比較する方法が挙げられる。Morinaka ら [4] は

線形近似した部分系列ごとに比較を行う L-index を提案している。また、Keogh ら [5] は個々の部分系列を固定長をもつ矩形関数で近似し、高速に比較する piecewise aggregate approximation (PAA) 法を提案している。

これらの方法では、周波数成分の数と範囲、あるいは部分系列に分解する際のウインドウ幅を適当に変化させることによって、様々な視野スケールで系列を比較できる。しかし、いずれにおいても、系列全体が同一スケールで記述される必要があり、部分系列ごとにスケールを変えて比較することは困難である。これは、スケール変化に伴う部分系列の階層構造の変化を把握していないため、部分ごとに異なるスケールで表現された系列を繋いだ場合の連続性を保証できないことによる。このことは、部分系列を連結する際に重なりや隙間が生じることを意味しており、系列の凹凸構造を比較する上で無視できない問題となる。

本稿では、部分系列の連続性を保存しつつ系列を様々な視野スケールにおいて比較する多重スケールマッチング [6] とラフクラスタリング [7] を組み合わせた時系列臨床検査データベースの解析法を提案する。多重スケールマッチングは、スケール変化に伴う部分輪郭の階層構造の変化を変曲点軌跡として保存し、全体の連続性を崩さないように制限を加えつつクロススケールマッチングを行う手法である。一方、ラフクラスタリングは、ラフ集合 [8] における識別不能性に基づき対象を分類する方法であり、対象間類似度が相対的類似度のみによって与えられる場合においても可読性の高いクラスタを生成することができる特徴をもつ。本方法では、これらを組み合わせ、多重スケールマッチングにより系列間類似度を求め、その類似度をもとにラフクラスタリングを適用して系列を分類する。分類された系列のもつ類似点は、多重スケールマッチングの結果から視覚的に容易に把握することができる。実験では、共通データである肝炎データの一部に本方法を適用し、類似系列が同一のクラスタに分類されるとともに、対応する部分系列が正しく獲得されることを示す。

2 時系列データのクラスタリング

2.1 位相情報を重視した多重スケールマッチング

Mokhtarian [6] らにより提案された多重スケールマッチングは、対象図形を様々な視野スケールで記述、比較する方法である。マッチングは部分輪郭ごとの類似性を基準にして行われ、対応する部分輪郭組は同一スケールのみならず、異なるスケールに渡って探索される。これにより、局所的な類似性だけでなく、より大局的な観点から観察した類似性に基づきマッチングを行うことが可能となる。この方法ではスケールを連続的に変化させる必要があり、計算量の問題が指摘されていたが、上田ら [9] が変曲点

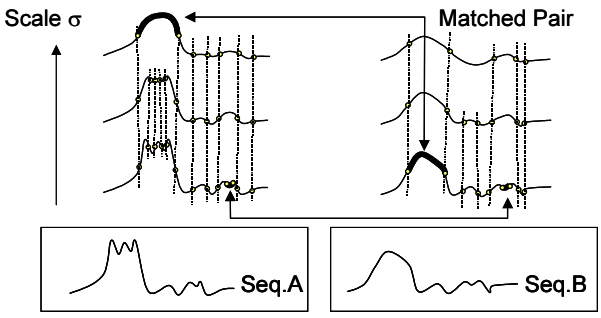

Figure 1: Multiscale matching.

間の凹凸セグメントをマッチング単位とすることで離散スケールの導入を可能とし、この問題を解決した。本方法では、上田らの方法を用いて患者間での検査値系列のマッチングを行う。ここでは、検査値の増減に起因して生じる系列の凹凸構造を部分輪郭の凹凸構造と対応させる。これにより、短期的な変化パターンの類似性のみならず、より長期的な変化パターンの類似性を評価する。

まず、時刻 t をパラメータとする関数 $x(t)$ で検査値の系列を表現する。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定義される。

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

図 1 に σ を変化させた場合の系列の変化を示す。同図および上式から明らかなように、スケールの増加とともに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。

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

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。

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

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。同様に、スケール $\sigma^{(h)}$ における比較

対象系列を $\mathbf{B}^{(h)}$ とすると、

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

と表現できる。このとき、セグメント $a_i^{(k)}$ と $b_i^{(h)}$ の相違度 $d(a_i^{(k)}, b_j^{(h)})$ を次式により定義する。

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

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。連続した奇数個のセグメントはより上位のスケールにおいて単一セグメントに置換され得るが、この場合のセグメント間相違度は、過度の置換を抑制する置換コスト関数を上式に加えたものとして定義される。

上式のセグメント間相違度は明らかに、部分系列の回転角および長さについての相似変換に不変であると同時に、系列間の位相差の変化に対しても不変であるという特徴をもつ。しかしながら、検査の系列に適用する場合、過度の位相シフトはイベント発生時期に関する情報を軽視する結果になるため、ここでは位相シフトを抑制する項を加え、相違度を以下のように拡張して再定義する。

$$d(a_i^{(k)}, b_j^{(h)}) = \frac{1}{3} \left(\left| \frac{d_{a_i}^{(k)}}{D_A^{(k)}} - \frac{d_{b_j}^{(h)}}{D_B^{(h)}} \right| + \frac{\mid \theta_{a_i}^{(k)} - \theta_{b_j}^{(h)} \mid}{\theta_{a_i}^{(k)} + \theta_{b_j}^{(h)}} + \left| \frac{l_{a_i}^{(k)}}{L_A^{(k)}} - \frac{l_{b_j}^{(h)}}{L_B^{(h)}} \right| \right)$$

ここで、$d_{a_i}^{(k)}$ および $d_{b_j}^{(h)}$ は初回検査日から該当セグメントの最初の検査日までの日数、$D_A^{(k)}$ および $D_B^{(h)}$ は初回検査日から最終検査日までの期間、すなわちデータの採取期間を示す。この拡張により、各項がそれぞれ示すイベントの (1) 発生日、(2) 上昇/下降の鋭さ、(3) 継続期間、の 3 つの特徴について、その全ての相違度を最小化させるセグメント組を捉えることが可能となる。

多重スケールマッチングにおけるマッチング手続きは、全てのセグメント組から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。図 1 上側に示すマッチング例では、系列 A の 5 つの連続したセグメントが上位スケールで 1 つのセグメントに置換され、これが系列 B の 1 セグメントと対応している。一方、同図下側に示すもう 1 つのマッチング例では、最下位スケールおいて対応するセグメントが見られる。このように、短期的に類似した傾向が見られる場合は下位スケールで、短期的には異なるが長期的には類似した傾向が見られる場合はより上位のスケールで対応がとられる。本方法では、マッチングの結果得られる残相違度を以後のラフクラスタリングにおける系列間相違度として用いる。なお、マッチングアルゴリズムの詳細については文献 [9] を参照されたい。

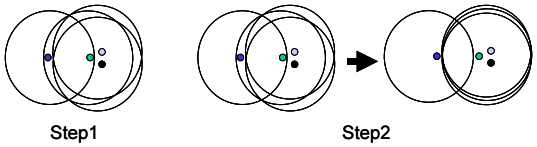

Figure 2: Rough clustering.

2.2 ラフクラスタリング

クラスタリングはある特徴量を基準に類似した対象を同一のクラスタにまとめる処理であり、これまでに k-means [10]、EM algorithm [11]、CLIQUE [12]、CURE [13]、BIRCH [14] など、様々な方法が提案されている。一般に、数値データを対象としたクラスタリング法では、クラスタ内分散を最小化しつつクラスタ間分散を最大化するよう、最適分割を決定していく。しかし、多重スケールマッチングで算出される系列間類似度は相対的類似度であり、三角不等式の成立が必ずしも保証されているものではないため、クラスタ中心、重心等の幾何学的特徴量を用いたこれらのクラスタリング法は単純には適用できない。古典的な階層的クラスタリング [15] は相対的類似度を取り扱うことができるが、クラスタリング結果が場合によっては処理手順に依存して変化することが知られている。

ラフクラスタリングは、ラフ集合論の識別不能性の概念に基づくクラスタリング法であり、対象のまとまり具合いを識別不能度として表現することで、相対的類似度で表現されたデータにおいても可読性の高いクラスタを生成することができる。多重スケールマッチングにより得られる類似度（相違度）は、任意の 2 検査系列間の類似性を表す相対的な尺度であるため、本方法ではラフクラスタリングを適用して系列を分類する。

まず、ラフクラスタリングに関連するラフ集合論の諸定義について述べる。$U \neq \phi$ を対象オブジェクトの集合とし、X を U の部分集合とする。U 上で定義される同値関係 R は、U を以下の条件を満たす部分集合 X の集合 $U/R = \{X_1, X_2, ...X_m\}$ に分割する。

$$\begin{array}{ll} (1) X_i \subseteq U, X_i \neq \phi & \text{for any } i, \\ (2) X_i \cap X_j = \phi & \text{for any } i, j, \\ (3) \cup_{i=1,2,...,m} X_i = U. & \end{array}$$

それぞれの部分集合 X_i はカテゴリと呼ばれ、U における R の同値類を示す。また、あるオブジェクト $x \in U$ を含む R のカテゴリを $[x]_R$ で表現する。さらに、与えられた同値関係の集合 $\mathbf{P} \subseteq \mathbf{R}$ に関して識別不能であるという識別不能関係を $IND(\mathbf{P})$ で

表し、次式により定義する。

$$IND(\mathbf{P}) = \bigcap_{R \in \mathbf{P}} IND(R)$$

ラフクラスタリングは、(1) 初期同値関係の構築、(2) 同値関係の再帰的更新、の 2 ステップから構成される。図 2 に各ステップの概略を示す。第 1 ステップでは、各対象に対して自らと類似したものと異なるものを分類する初期同値関係を与える。n 個の対象からなる全体集合を $U = \{x_1, x_2, ..., x_n\}$ としたとき、対象 x_i に対する同値関係 R_i は次式により定義される。

$$R_i = \{\{P_i\}, \{U - P_i\}\}$$

$$P_i = \{x_j |\ s(x_i, x_j) \geq S_i\}, \quad \forall x_j \in U$$

ここで、P_i は x_i と類似した対象の集合であり、類似度 s が閾値 S_i を越える対象の集合として定義される。ここでは、s として多重スケールマッチングの結果得られる 2 系列間の相違度の逆数を用いる。また、その閾値 S_i は類似度が著減する位置に自動的に定める。クラスタは、全ての同値関係を用いても識別不能な対象の集合、すなわち、$U/IND(\mathbf{R})$ のカテゴリ X_i として得られる。図 2 はこれらの概念をユークリッド空間上で表現したもので、各対象を中心とする円の半径は S_i に相当し、この中に存在する他の対象はすべて同値類とみなされる。2 つの対象間を交差する円がただ 1 つも存在しない場合、その対象は識別不能であるとみなされ、同一クラスタに分類される。この例では、4 つの対象が 3 つのクラスタに分類されている

第 2 ステップでは、第 1 ステップで個々に構築した初期同値関係を全体的な観点から修正し、より可読性の高いクラスタを生成する。まず、ある 2 つの対象 x_i と x_j が、他のどれだけ多くの対象から識別不能と見なされているかを示す識別不能度 γ を次式により定義する。

$$\gamma(x_i, x_j) = \frac{1}{|U|} \sum_{k=1}^{|U|} \delta_k(x_i, x_j)$$

$$\delta_k(x_i, x_j) = \begin{cases} 1, & \text{if } [x_k]_{R_k} \cap ([x_i]_{R_k} \cap [x_j]_{R_k}) \neq \phi \\ 0, & \text{otherwise} \end{cases}$$

ここで、$[x_i]_{R_i}$ は同値関係 R_i において x_i と同値類とみなされる対象の集合を示す。識別不能度 γ が高い対象は類似度が高く、同一のクラスタに分類されることが望ましい。逆に、識別不能度の高い対象を異なるクラスタに類別するような同値関係 R_i は詳細すぎる類別知識を与えているといえる。そこで、そのような同値関係を以下の手続きにより R'_i に修正し、詳細すぎる類別知識による細かなクラスタの生成を抑制する。

$$R'_i = \{\{P'_i\}, \{U - P'_i\}\}$$

$$P'_i = \{x_j | \gamma(x_i, x_j) \geq T_h\}, \quad \forall x_j \in U$$

ここで、T_h は対象を識別不能と見なす閾値であり、類別知識の粗さに対応づけられる。この T_h の値を徐々に低下させつつ再帰的に同値関係を更新することで、適度に粗い知識に基づくクラスタリング結果 $U/IND(\mathbf{R}')$ が得られる。図 2 の例では、2 つの同値関係を更新することでクラスタ数が 4 から 2 に減少している。なお、2 度目以降の同値関係の更新は、初期同値関係ではなく前回更新後の同値関係を用いる。

3 実験結果

実験として、肝炎データセットから抽出した時系列 GPT データに本方法を適用し、多重スケールマッチングおよびラフクラスタリングの時系列データ解析への適用可能性について調べた。ここでは、結果の解釈を容易とするため、ランダムに選択した 20 程度のシーケンスからなるサブセットを構築して提案方法を適用し、分類された各クラスタの特徴とマッチング結果の妥当性について視覚的に検討した。図 3 に前処理適用後の各シーケンスを示す。原データはそれぞれ数日から数週間までの異なる間隔で収集されていたが、その間隔は数週から数ヶ月まで 1 週間を単位として変化すること、すなわち、患者は曜日を決めて検査を受けていることが事前の基礎的解析から示唆されたため、リサンプリング間隔を 1 週間と設定した。

表 1 に、多重スケールマッチングにより導出された正規化後の類似度を示す。多重スケールマッチングでは、結果の対称性 ($s(A, B) = s(B, A)$) が成り立つため、類似度行列の左下半分は割愛した。同表と図 3 を見比べると、類似した系列に高い類似度が与えられていることが分かる。

この類似度組に対し、ラフクラスタリングは$U/IND(\mathbf{R})$ = {{1,2,9,11,17,19}, {4,3,8}, {7,14,15}, {10,12,13}, {5}, {6}, {16}, {18}, {20}} の 9 つのクラスタを生成した。クラスタリングのパラメータ Th は経験的に $Th = 0.6$ と定め、同値関係の更新は Th を $Th = 0.4$ まで減少させつつ 5 回行った。各クラスタに含まれる系列を図 3 と比較すると、類似したパターンをもち、類似度の高いものが同一クラスタに分類されていることが分かる。幾つかの系列、例えば#16 等は独立したクラスタに分類されている。これは、他の系列との類似度が極端に低いためで、多重スケールマッチングで対応する部分系列組が同定できなかったことに起因している。

図 4 に最も高い類似度を持つ系列#10 および#12 におけるマッチング結果を示す。今回の実験では、スケール σ を 1.0 から 13.5 まで 2.5 間隔で変化させている。同図最下部の 2 つの系列が $\sigma = 1.0$ における両系列と対応し、その上の 5 つの系列組が $\sigma =$ 3.5, 6.0, 8.5, 11.0、13.5 における両系列と対応する。異なる色で示される部分系列はそれぞれセグメントを示している。また、最上段はマッチング結果を示し、マッチすると判定されたセグメントは同一色で

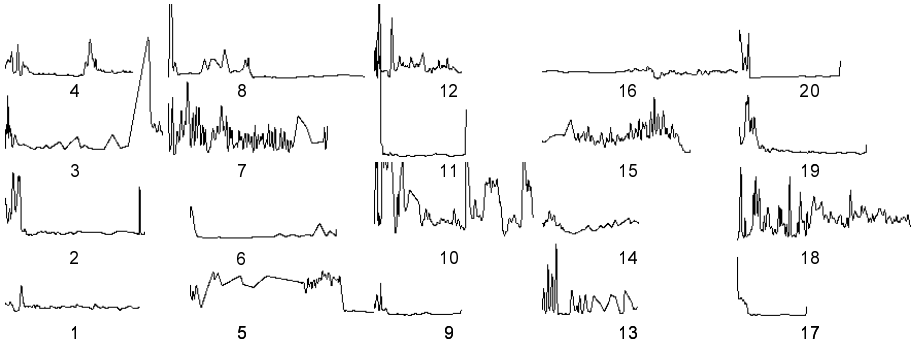

Figure 3: Test patterns.

Table 1: Similarity of the sequences

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1.00	0.70	0.68	0.78	0.00	0.63	0.48	0.71	0.72	0.61	0.73	0.66	0.64	0.72	0.50	0.00	0.53	0.00	0.74	0.45
2		1.00	0.61	0.73	0.00	0.68	0.22	0.46	0.68	0.67	0.72	0.73	0.72	0.68	0.54	0.00	0.68	0.00	0.77	0.41
3			1.00	0.75	0.45	0.51	0.68	0.47	0.71	0.70	0.69	0.73	0.71	0.81	0.68	0.00	0.62	0.00	0.72	0.55
4				1.00	0.00	0.60	0.52	0.47	0.75	0.71	0.64	0.79	0.75	0.82	0.47	0.00	0.60	0.00	0.75	0.48
5					1.00	0.23	0.62	0.49	0.33	0.53	0.44	0.45	0.50	0.44	0.56	0.01	0.00	0.26	0.53	0.30
6						1.00	0.00	0.00	0.59	0.00	0.58	0.39	0.61	0.65	0.00	0.00	0.47	0.00	0.47	0.48
7							1.00	0.49	0.54	0.80	0.57	0.73	0.73	0.59	0.76	0.00	0.00	0.44	0.62	0.39
8								1.00	0.53	0.47	0.57	0.56	0.51	0.49	0.54	0.00	0.00	0.00	0.66	0.51
9									1.00	0.00	0.68	0.00	0.00	0.00	0.00	0.00	0.82	0.00	0.00	0.00
10										1.00	0.59	0.83	0.76	0.75	0.81	0.00	0.47	0.11	0.59	0.37
11											1.00	0.76	0.54	0.68	0.00	0.00	0.74	0.00	0.76	0.00
12												1.00	0.81	0.78	0.67	0.00	0.70	0.00	0.63	0.40
13													1.00	0.75	0.00	0.00	0.64	0.00	0.67	0.35
14														1.00	0.00	0.00	0.66	0.00	0.71	0.00
15															1.00	0.00	0.43	0.20	0.55	0.39
16																1.00	0.00	0.00	0.43	0.19
17																	1.00	0.00	0.00	0.00
18																		1.00	0.39	0.03
19																			1.00	0.00
20																				1.00

示されている。例えば、セグメント A はセグメント A' と対応し、セグメント B はセグメント B' と対応する。同図から、大幅な上昇 (A および A')、突発的な上昇を含む小幅な減少 (B および B')、小幅な上昇 (C および C') のように、長期的な上昇/下降のパターンの類似性のみならず、セグメント B、B' に見られる突発的上昇のような短期的な特徴をも同時に捉えていることが分かる。セグメント $D-F$ と $D'-F'$ も同様に類似した上昇/下降パターンを有しており、それぞれが正しく対応付けられていることが確認される。また、両系列のように大きく異なる収集期間をもつ系列においても正しくマッチングが行われている。

4 むすび

本稿では、推移パターンの類似性に基づく時系列データの解析法を提案した。本方法では、多重スケールマッチングとラフクラスタリングを併用することで、部分系列の連続性を考慮しつつ異なるスケールにわたるマッチングを行い、その結果得られる類似度をもとに系列をクラスタリングした。これにより、単に系列を分類するのみではなく、分類結果における部分輪郭の対応をも同時に示すことができ、より理解しやすい結果の提示を可能とした。また、肝炎データに適用した基礎実験では、部分系列の類似性が長期的、短期的両方の観点から正しく評価され、得られた類似度から系列が直感的に正しく分類できていることが確認された。今後の課題として、計算量に関する検討と大規模データでの有効性の検討が挙げられる。

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 (B)(No.759))、「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。

References

[1] R Agrawal, C. Faloutsos, and A. N. Swami (1993): Efficient Similarity Search in Sequence Databases. Proceedings of the 4th International Conference on Foundations of Data Organization and Algorithms: 69–84.

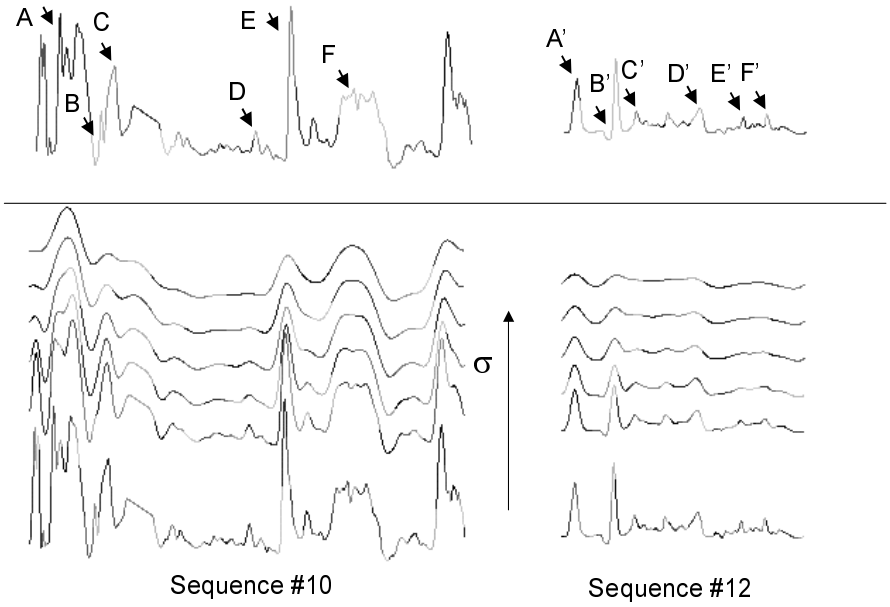

Figure 4: Matching result of sequences #10 and #12.

[2] K. P. Chan and A. W. Fu (1999): Efficient Time Series Matching by Wavelets. Proceedings of the 15th IEEE International Conference on Data Engineering: 126–133.

[3] F. Korn, H. V. Jagadish, and C. Faloutsos (1997): Efficiently Supporting Ad Hoc Queries in Large Datasets of Time Sequences. Proceedings of ACM SIGMOD International Conference on Management of Data: 289–300.

[4] Y. Morinaka, M. Yoshikawa, T. Amagasa and S.Uemura (2001): The L-index: An Indexing Structure for Efficient Subsequence Matching in Time Sequence Databases. Proceedings of International Workshop on Mining Spatial and Temporal Data, PAKDD-2001: 51-60.

[5] E. J. Keogh, K. Chakrabarti, M. J. Pazzani, and S. Mehrotra (2001): "Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases" Knowledge and Information Systems 3(3): 263-286.

[6] 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.

[7] S. Hirano and S. Tsumoto (2001): Indiscernibility Degrees of Objects for Evaluating Simplicity of Knowledge in the Clustering Procedure. Proceedings of the 2001 IEEE International Conference on Data Mining. 211–217.

[8] Z. Pawlak (1991): Rough Sets, Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Dordrecht.

[9] 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.cf

[10] S. Z. Selim and M. A. Ismail (1984): K-means-type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality. IEEE Transactions on Pattern Analysis and Machine Intelligence, 6(1): 81–87.

[11] A. P. Dempster, N. M. Laird, and D. B. Rubin (1977): Maximum likelihood from incomplete data via the EM algorithm. J. of Royal Statistical Society Series B, 39: 1–38.

[12] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan (1998): Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. Proceedings of ACM SIGMOD International Conference on Management of Data: 94–105.

[13] S. Guha, R. Rastogi, and K. Shim(1998): CURE: An Efficient Clustering Algorithm for Large Databases. Proceedings of ACM SIGMOD International Conference on Management of Data: 73–84.

[14] T. Zhang, R. Ramakrishnan, and M. Livny (1996): BIRCH: An Efficient Data Clustering Method for Very Large Databases. Proceedings of ACM SIGMOD International Conference on Management of Data: 103–114.

[15] M. R. Anderberg (1973): Cluster Analysis for Applications. Academic Press, New York.