構造的類似性に着目した多変量時系列医療データのクラスタ分析

平野　章二†　　津本　周作†
†島根大学医学部医学科医療情報学講座

1　はじめに

近年のセンサ技術・情報通信技術の発展により，医療，気象，社会安全など様々な分野で多変量時系列データを自動的に収集し蓄積することが可能となった．膨大なデータの横断的解析により，これまで未知であった時間変量間の関連性，時間変化の全体的，部分的共通性あるいは例外事例の存在など有益な知識の獲得が期待される．多変量時系列の比較分類においては，(1) 変量間の共変化関係をどのように捉えるか，(2) 時間変化をどの粒度で捉えるか，等が問題となるが，有効な対応手法の確立には至っていない。本研究では，従来の周波数分析，抽象化等に基づく時系列解析とは異なり，時空間軌跡の有する方向性と幾何的特徴に着目し，その類似性を多重スケール比較する新たな自動分類法の開発を試みる．

2　軌跡の多重スケール比較・分類

提案法では最初に「軌跡のどの部分を対応させて比較すべきか」を構造的な類似性に基づき決定する．次に，対応づけられた各部分軌跡について原系列値の差異を積算し，これを軌跡間相違度として軌跡のクラスタリングを行う．構造的類似性の評価には多重スケール比較法 [1, 2] を応用し，観察粒度の問題に対応する．

2.1　時系列の軌跡表現

属性数 M の時系列データを考える．時間を t とし，$m \in M$ 番目の属性に対応する時系列を $ex_m(t)$ とする．このとき，M 次元の時系列軌跡を $c(t) = \{ex_1(t), ex_2(t), \ldots, ex_M(t)\}$ により表す．次に，観察スケール σ における m 番目の時系列を，もとの時系列 $ex_m(t)$ と n 次の修正ベッセル関数 $I_n(\sigma)$ との畳み込み演算により $EX_m(t,\sigma) = \sum_{n=-\infty}^{\infty} e^{-\sigma} I_n(\sigma) ex_m(t-n)$ と得る [3]．この演算を時系列 $ex_m(t), \forall m \in M$ に適用し，スケール σ における軌跡 $C(t,\sigma) = \{EX_1(t,\sigma), EX_2(t,\sigma), \ldots, EX_M(t,\sigma)\}$ を得る．図 1 に軌跡の多重スケール表現例を示す．スケール σ を変化させることで，様々な視野から軌跡を表現する．

Cluster Analysis of Multivariate Time-series Medical Data Based on the Strucutral Comparison Technique.
†Shoji HIRANO　Shusaku TSUMOTO
†Department of Medical Informatics, Shimane University, School of Medicine

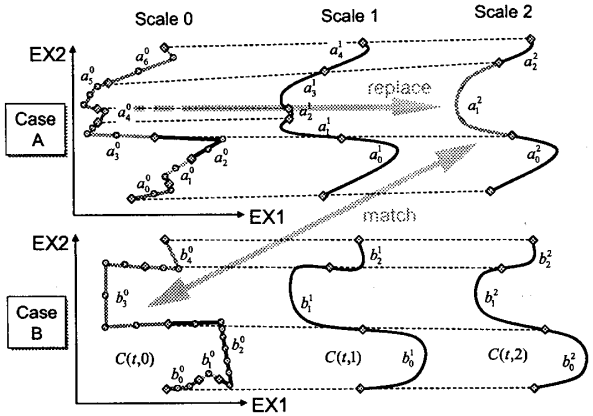

図 1: 軌跡の多重スケール比較.

2.2　セグメント置換関係の追跡とマッチング

多重スケール表現された各軌跡 $C(t,\sigma)$ を，変曲点を両端とする部分軌跡（セグメント）に分割する．軌跡 A がスケール k において $N_A^{(k)}$ 個の部分軌跡から構成されるとき，これをセグメント集合 $\mathbf{A}^{(k)} = \{a_i^{(k)} \mid i = 1, 2, \cdots, N_A^{(k)}\}$ として表現する．ここで，$a_i^{(k)}$ はスケール k における i 番目のセグメントを示す．同様に，スケール h における比較対象軌跡 B はセグメント集合 $\mathbf{B}^{(h)} = \{b_j^{(h)} \mid j = 1, 2, \cdots, N_B^{(h)}\}$ により表現される．

続いて，最上位スケールから下位スケールへ順に，隣接スケール間で変曲点の対応付けを行い，スケール変化に対するセグメントの置換関係を明確化する．その後，入力軌跡 A,B の全てのセグメント対から以下の条件を満たす最適対応組を探索する．(1) 完全対応: セグメントの結合により原軌跡が空隙や重複無く構成される．(2) 相違度の最小化: 各セグメント組の局所相違度を積算した総相違度が最小化される．探索は全スケールを横断して行われ，局所的に類似した傾向がみられる場合は下位スケールで，局所的には異なるが大局的には類似した傾向がみられる場合はより上位のスケールで対応がとられる．その可能な組合せの中から上記 2 条件を満足する最適対応組が決定される．アルゴリズムの詳細については文献 [1] を参照されたい．

最適対応の決定に用いるセグメント間の局所相違度 $d(a_m^{(k)}, b_n^{(h)})$ は次式により定義する．

$$d(a_m^{(k)}, b_n^{(h)}) = \sqrt{(g(a) - g(b))^2 + (\theta(a) - \theta(b))^2} + |v(a) - v(b)| + \gamma \{cost(a) + cost(b)\}$$

ここで，$a_m^{(k)}$ と $b_n^{(h)}$ は比較対象のセグメントである．$g(a)$，$\theta(a)$，$v(a)$ はセグメント $a_m^{(k)}$ の開始点勾配，回転角及び速度を示す形状パラメータであり，それぞれ，セグメント開始時における軌跡の方向性，終端までの方向的変化量，その変化の早さを表現する．$cost()$ は過度の置換を抑制するためのコスト関数 [1]，γ は置換コストの重み係数である．

2.3 軌跡間相違度

構造的類似性に基づく比較で部分系列の最適対応関係を獲得した後，対応づけられたセグメントの各組について原時系列値の差異を求めて積算し，その平均値を軌跡間相違度 $D_{val}(A,B)$ とする．セグメント組の系列値の差は各属性 ($m \in M$) ごとにセグメントの頂点及び両端点において求め，それを加重結合する．

3 実験結果

人工データ及び医療データを対象にクラスタリング実験を行い，提案法の有効性を検証した．

3.1 人工データ

1 から 9 まで 9 種類の数字について，Arial フォントの骨格線から 2 次元筆跡データを作成した．各点の座標に $N(0,1)$ に従うガウスノイズを印可し，各数値について 10 個，計 90 個の 2 次元軌跡データを生成した．

まず，提案法の出力する軌跡間相違度が軌跡の分類問題に適用可能か否かを調べるため，以下の手続きによる擬似的な再近隣分類実験を行った．(1) 90 個の軌跡の対比較を行い，90×90 要素からなる相違度行列を作成する．(2) 1 つの軌跡を選択し，相違度が最小である軌跡の属するクラスに分類する．(3) 選択/ 分類の手続きを全ての軌跡に適用し，分類誤差を評価する．以上の手続きを 100 個のデータセットに対して適用した結果，分類誤差は 0.016 ± 0.014 (平均 ± SD) であり，本データにおいては概ね 98% の軌跡が正しい数値として分類される結果となった．

次に，提案法が出力する軌跡間相違度のクラスタリングへの適用可否を調べるため，以下の手続きによるクラスタ構築実験を行った．(1) 前出の 90×90 の相違度行列を用い，群平均法による階層型クラスタリング [4] を適用，数字種と同じ 9 個のクラスタを構築する．(2) 各クラスタの代表ラベルを voting により決め，そのラベルに対する各事例の分類誤差を評価する．以上の手続きを 100 個のデータセットに対して適用した結果，分類誤差は 0.118 ± 0.057 (平均 ± SD) であり，本データにおいては概ね 88% の軌跡が正しい数値として分類される結果となった．

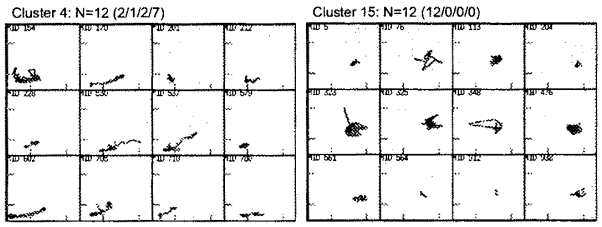

図 2: 分類された軌跡の例．左：クラスタ 4, 右：クラスタ 15．横軸は CHE，縦軸は PLT を示す．

3.2 医療データ

慢性ウイルス性肝炎症例の時系列検査データ（コリンエステラーゼ (CHE) と血小板数 (PLT) の 2 次元軌跡）を対象としてクラスタリングを行った．結果の一部を図 2 に示す．左側は肝線維化進行例を多く含むクラスタに分類された軌跡の例であり，CHE(横軸) と PLT(縦軸) が時間経過と共に基準範囲外へ減少する強い左下がりの軌跡を示している．また，PLT の低下が CHE の低下に先行する傾向が見られた．一方，同図右側は肝線維化度の低い例を多く含むクラスタに分類された軌跡の例であり，CHE，PLT 共にほぼ基準範囲内で推移しており，また時間的に明瞭な方向性が見られなかった．このように，特徴的な推移傾向を呈する事例のクラスタ化が可能であったほか，クラスタ構成と肝線維化度の間に関連が見られた等，興味深い知識が獲得された．

4 まとめ

構造的類似性に基づく時空間軌跡の自動分類法について開発及び基礎実験を行った．今後，高次元軌跡への拡張，計算量の削減，クラスタからのルール生成システムの開発等を進めていく予定である．

参考文献

[1] N. Ueda and S. Suzuki (1990): A Matching Algorithm of Deformed Planar Curves Using Multiscale Convex/Concave Structures. IEICE Trans. Inf. & Syst., J73-D-II(7): 992–1000.

[2] F. Mokhtarian and A. K. Mackworth (1986): Scale-based Description and Recognition of planar Curves and Two Dimensional Shapes. IEEE Trans. PAMI, PAMI-8(1): 34-43.

[3] T. Lindeberg (1990): Scale-Space for Discrete Signals. IEEE Trans. PAMI, 12(3):234–254.

[4] B. S. Everitt, S. Landau, and M. Leese (2001): Cluster Analysis Fourth Edition. Arnold Publishers.