様式 C-19

科学研究費補助金研究成果報告書

平成２２年　５月１３日現在

研究種目：若手研究（B）
研究期間：2008～2009
課題番号：20700140
研究課題名（和文）　情報粒度に着目した多変量時系列の自動分類に関する研究

研究課題名（英文）　Research on automatic grouping of multivariate time series based on the information granularity

研究代表者
平野　章二（HIRANO SHOJI）
島根大学・医学部・准教授
研究者番号：60333506

研究成果の概要（和文）：
本研究では，多変量時系列を様々な粒度で観察し比較分類する方法の開発に取り組んだ。提案法では，まず多変量時系列から軌跡を構築して多重スケール表現し，曲率極大点の位置に基づき基本構造単位である granule へ区分する。次に細粒度から粗粒度にわたる granule の構造変化を追跡し，粒度を変化させながら軌跡間の最適対応を求め比較を行う。医療データへの適用実験では，類似した推移傾向を呈する症例クラスタを獲得できたほか，各クラスタの線維化度の分布に特徴が見られるなど，興味深い知見を得ることができた。

研究成果の概要（英文）：
In this research we have developed a multiscale comparison method for multivariate time series. Our method firstly constructs multidimensional trajectories from the time series, and represent them using multiscale representation. Next, it splits the trajectories into data granules according to the positions of curvature maxima. Then it traces the hierarchical structure of data granules and performs granule-by-granule matching across the scales to find the best correspondences between the trajectories. Experimental results on a medical dataset showed that our method could generate groups of trajectories that exhibited similar temporal courses, and some of the clusters showed interesting characteristics about the distribution of fibrotic stages.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００８年度	1,800,000	540,000	2,340,000
２００９年度	1,300,000	390,000	1,690,000
年度			
年度			
年度			
総　計	3,100,000	930,000	4,030,000

研究分野：総合領域
科研費の分科・細目：情報学・知能情報学
キーワード：データマイニング，知識工学，時系列，情報システム

１．研究開始当初の背景

計測技術と情報通信技術の進展により，医療，気象，社会安全など様々な分野で多属性かつ大量の時系列データを収集・蓄積することが可能となった。例えば医療分野では数十～数百項目からなる臨床検査の記録が日々蓄積されており，その数は島根大学医学部附属病院の場合で年間 15 万件以上に及ぶ。多変量時系列のデータベースは，個々の観測対象が内的状態の変化あるいは外部との相互作用を受けて複雑に遷移する様を多面的に計測した膨大な記録の集合体であり，その横断的比較と分類を通じて多くの対象に共通して観察される特徴的な遷移パターン，異なる属性間の全体的・部分的な共変化関係，あるいは例外事例の存在など，個々の事例の観察からでは容易に気づき得ない様々な未知の知識の発見が期待される。国際的にも多変量時系列データからの知識発見に関する研究は注目を集めており，例えばデータマイニングに関する代表的な国際会議である ICDM では，2007 年に regular paper として採択された time series セッションで発表された２件の論文のうち，１件は統計量に基づく多変量時系列の自動分類に関するものであった他，時空間データマイニングに関する Workshop が併設されるなど，ホットトピックとして急速な成長を見せつつある。しかしながら，(1)時間変化の観察粒度の設定，(2)異なる変量間の共変化関係の考慮，は多変量時系列の分類問題が内包する困難な課題であり，未だ有効な解決法は確立されていない。

２．研究の目的

本研究では，前項の課題に対処すべく，変量間の共変化関係を反映した多変量時系列の情報粒度表現に関する基礎理論の構築とその応用による自動分類法の開発を目指す。

３．研究の方法

視野スケールを変化させて様々な粒度で対象を観察するアプローチは 1980 年代から 90 年代を中心にパターン認識分野で精力的に研究されてきた。特に Witkin による曲率スペースフィルタリングは，平滑化に用いるガウスカーネルが causality (スケールが増加したとき新たな極点を生成しない)を満足する理想的な性質を有するため，様々な多重スケール表現の基礎となっている。Mokhtarian らは物体輪郭の多重スケール表現への応用を示し，上田ら輪郭をセグメントに区分し多重スケール比較する手法を考案している。これら手法の多くは，物体輪郭の凹凸構造を曲率の符号変化(curvature zero-crossings; 変曲点)により特徴付けている。しかしながら，3 次元以上の空間曲線においては曲率が符号を持たないため従来の変曲点を基準としたマッチング方法は適用できない。一方，Mokhtarian らは曲率と共に空間曲線の主要なパラメータである捩率(torsion)を用いた捩率スケールスペースを提案しているが，極点の単調性が担保できないことが問題となる。

これらに対し提案法では，多変量時系列を軌跡として多重スケール表現し，その極大点を基準として階層構造を表現する。

（１）軌跡の多重スケール表現

時系列 x(t)，y(t)，z(t)からなる 3 次元軌跡を c(t)={x(t),y(t),z(t)}とし，軌跡の観察スケールを σ とする。このとき，スケール σ における時系列 X(t, σ)は，もとの時系列 x(t)と平滑化カーネル In(σ)との離散畳み込み演算により以下のごとく得られる。

$$X(t,\sigma) = x(t) \otimes g(t,\sigma) = \sum_{n=-\infty}^{\infty} e^{-\sigma} I_n(\sigma) x(t-n)$$

ここで，In(σ)は n 次の修正ベッセル関数である。同演算を y(t)及び z(t)についても適用し，スケール σ における軌跡 C(t, σ) = {X(t, σ),Y(t, σ),Z(t, σ)}を得る。図１に軌跡の多重スケール表現例を示す。なお，図では簡単のため 2 次元で表記している。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できる。

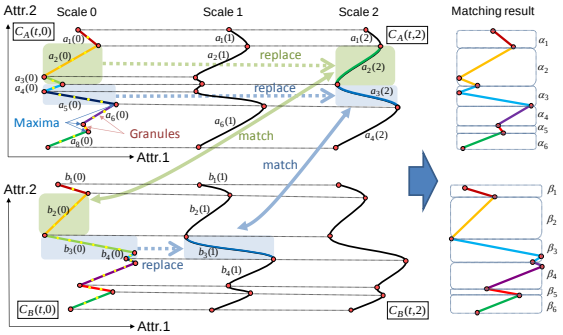


図１：軌跡の多重スケール表現/比較

（２）Granule の構築と階層化

多重スケール表現された軌跡上の各点について，曲率の極大点を求め，導出した極大点を t-σ を軸とする曲率スケールスペースに布置する。図２に例を示す。横軸が時間 t，縦軸がスケール σ，「+」が極大点に対応する。極大点の数はスケール増加に伴い単調的に減少する。

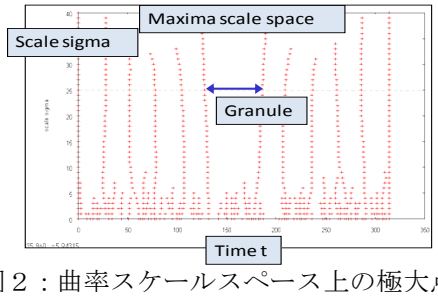


図２：曲率スケールスペース上の極大点

ここで，各スケールにおいて極大点で区分される部分軌跡（セグメント）を granule と考える。曲率の極大点は軌跡が周囲に比して大きく方向を変える点であり，推移の傾向が類似している部分区間が一つの granule となる。低スケールでは軌跡の微細な構造が反映されるため，粒度の細かな granule が構成され，高スケールでは平滑化により抽象化された構造が反映されるため粒度の粗い granule が形成される。なお，本方法で取り扱う時系列は離散的であり，構造上最も細かな granule は図 1 に示すように隣接データ点間を結ぶ直線セグメントとなる。

続いて上位スケールから下位スケールに向けて隣接するスケール間での極大点の対応付けを行う。これにより granule の「階層」を構築し，異なる粒度で軌跡を観察した場合に各部がどのように変化するかを表現することが可能になる。

（３） 可変粒度比較

いま，比較を行う二つの３次元軌跡を $c_A(t)$, $c_B(t)$ とする。図１に示すように，$c_A(t)$, $c_B(t)$ の多重スケール表現 $c_A(t,\sigma)$, $c_B(t,\tau)$ から派生する全ての granule のペア（$a_i(\sigma)$, $b_j(\tau)$）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: granule 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 granule 組の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粒度の細かな下位スケールで，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観察粒度を粗くした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。

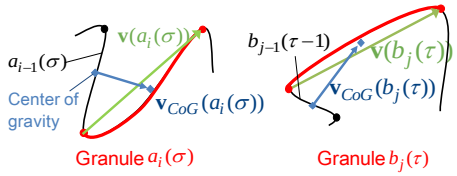


図３：Granule の形状パラメータ

相違度の導出に用いられる Granule の形状パラメータを図 3 に示す。これらのパラメータに基づき，Granule 間の相違度 $d(a_i(\sigma), b_j(\tau))$ を次式により定義する。ここで，$v(a_i(\sigma))$ 及び $v(b_j(\tau))$ はそれぞれ

$$d(a_i(\sigma),b_j(\tau)) = \mathrm{vdiff}(a_i(\sigma),b_j(\tau)) \cdot \mathrm{cost}(a_i(\sigma),b_j(\tau))$$

$$\mathrm{vdiff}(a_i(\sigma),b_j(\tau)) = \left\| \mathbf{v}(a_i(\sigma)) - \mathbf{v}(b_j(\tau)) \right\| + \left\| \mathbf{v}_{CoG}(a_i(\sigma)) - \mathbf{v}_{CoG}(b_j(\tau)) \right\|$$

$$\mathrm{cost}(a_i(\sigma),b_j(\tau)) = \left(\frac{n_A(0)}{n_A(\sigma)} \cdot \frac{n_B(0)}{n_B(\tau)} \right)^{\lambda}$$

granule $a_i(\sigma)$, $b_j(\tau)$ の両端を結ぶ３次元ベクトルであり，その差は granule の方向性の違いに相当する。また，$vCoG(a_i(\sigma))$ は一つ前の granule $a_{i-1}(\sigma)$ の重心と $a_i(\sigma)$ の重心を結ぶベクトルであり，$vCoG(b_j(\tau))$ との差は granule の位置に関する差異を示す。$n_A(0)$ と $n_B(0)$ はそれぞれ基底となる $C_A(t,0)$, $C_B(t,0)$ における granule の数を示す。Cost() は過度の置換を抑制するコスト項で，granule の結合に比例する形で増加していく。λ はコストの重み付けパラメータである。

図 1 の例では，C_A のスケール 0 の granule $a_2(0)$ と $a_3(0)$ が粗粒度のスケール 2 において $a_2(2)$ として表現され，その形状が C_B の $b_2(0)$ と類似しているため対応がとられる。また，$a_4(0)$ と $a_5(0)$ が $a_3(2)$ に置換され，これが $b_3(0)$ と $b_4(0)$ の粗粒度表現である $b_3(1)$ と類似するため対応づけられる。結果として，同図右に示すような（α_1, β_1）から（α_6, β_6）までの 6 組の対応が得られる。

最後に，対応づけられた部分軌跡の各組（α_m, β_m）について原時系列値の差異を求めて積算し，これを軌跡間相違度として以降のクラスタリング等で用いる。

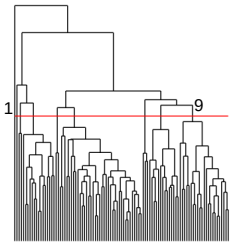
４．研究成果

提案法を慢性ウイルス性肝炎の検査データに対して適用し，有効性を検証した。本データは ECML/PKDD discovery challenge 2002-2004 の共通データセットとして使用されたもので，B 型と C 型の計 771 症例に関する時系列検査データ等が含まれている。本実験では，そのうち C 型でインターフェロン非適用例を対象とし，肝機能状態と関連する血小板数(PLT)，アルブミン(ALB)，コリンエステラーゼ(ChE)の３項目に関する軌跡を構築し症例間の類似性を調べた。目的は，(1)共通した推移傾向をもつ群の生成，(2)得られた群と肝線維化度との関連性の分析である。検査項目に不備のある事例を除いた 99 例を対象とした。

分析は，99 例の軌跡の全ペアについて前述の可変粒度比較を行い相違度行列を作成した後，階層的クラスタ分析を行う方法により実施した。マッチングに用いたパラメータは以下のとおりである：開始スケール=1.0，最大スケール=800，最小スケール間隔=0.2，置換コストウエイト=1.0。クラスタの結合基準は群間平均を用いた。スケール間隔は極大点数の変化に応じて動的に決定している。

図４左に生成された樹状図を示す。比較的上位において結合時の相違度が大きく変化する箇所で分割して 9 クラスタを構築し，各クラスタの例数を線維化度別に集計したものが図４右側の表である。F0 から F4 は線維化の進行度を示し，F0 が線維化の無い状態，F4 が肝硬変の状態を示す。同表から，クラスタ 3，4 のように進行例の割合が多い群，ク

ラスタ 8，9 のように非進行例の割合が多い群が構成されており，軌跡の類似性に基づき構成された群と，線維化度の分布の間に関連性が見いだされる。一方，クラスタ 6 のように多様な線維化度の例が混在している群も存在する。



Cluster	No of cases/Fibrotic stage				Total
	F0+F1	F2	F3	F4	
1	1	0	0	0	1
2	0	1	0	0	1
3	0	0	1	1	2
4	1	2	2	10	15
5	0	0	1	1	2
6	22	5	7	4	38
7	2	1	0	0	3
8	13	1	0	0	14
9	19	2	2	0	23
Total	58	12	13	16	99

図４：(左)生成された樹状図 (右)9 クラスタ分割時の線維化度別例数構成

図５から図７にクラスタ 4，6，9 に分類された軌跡の例を示す。クラスタ 4 の事例においては，PLT が既に基準下限を下回った状態で ChE と ALB が減少しており，PLT が先に下がる傾向を呈している。クラスタ 6 の事例においては PLT と ChE の両方が減少していく傾向を呈しており，クラスタ 4 とは異なるパターンが観察される。一方，クラスタ 9 ではいずれも基準範囲を中心に推移している。

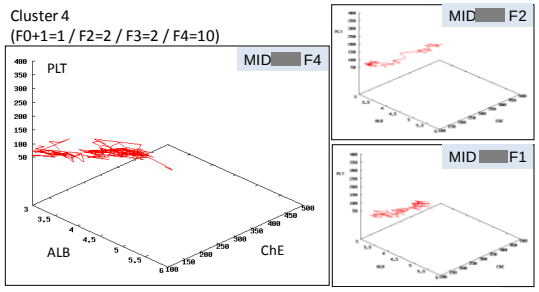


図５：クラスタ 4 に分類された軌跡

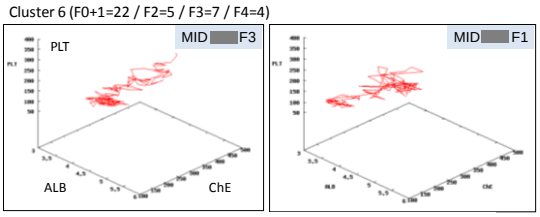


図６：クラスタ 6 に分類された軌跡

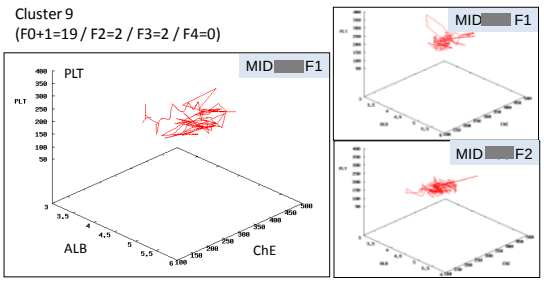


図７：クラスタ 9 に分類された軌跡

図８にクラスタ 4 に分類された軌跡に関する階層粒度表現とマッチングの例を示す。両者は細部において異なるが，大局的な推移の傾向は類似しており，また局所的に類似した変化を呈する部分もある。両者には基底スケールでそれぞれ 68 個，58 個の granule が存在しているが，同図中央の階層構造に基づいて可変粒度での比較が行われ，最終的に同図右に示す 41 個の granule による対応付けが行われている。同図に矢印で示す A-B，A'-B' のように，類似したパターンで変化する部分が適切に対応づけられていることが分かる。

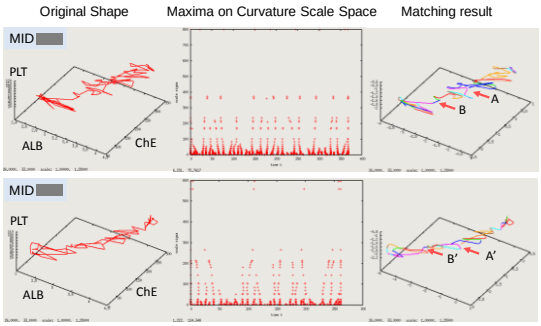


図８：ALB-PLT-ChE 軌跡の階層粒度表現．左から，原軌跡，曲率スケールスペース上の極大点変化，マッチング結果

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）

① Shoji Hirano, Shusaku Tsumoto, Multiscale Comparison of Three-dimensional Trajectories Based on the Curvature Maxima and Its Application to Medicine, Lecture Notes in Computer Sciences, 6007, 128-137, 2010, 査読有

〔学会発表〕（計１件）

① Shoji Hirano, Shusaku Tsumoto, Multiscale Comparison of Three-dimensional Trajectories: A Preliminary Step, 12th Int'l Conf. on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing 2009, Delhi, India, 2009.12.16.

６．研究組織
(1)研究代表者
平野　章二（HIRANO SHOJI）
島根大学・医学部・准教授
研究者番号：60333506