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２８日現在

研究種目：基盤研究(C)
研究期間：2007～2009
課題番号：１９５００１３０
研究課題名（和文）　多次元データに内在する高次共変性検出の研究

研究課題名（英文）　Detection of higher order covariate relations embedded in multi-dimensional data

研究代表者
市野　学（ICHINO MANABU）
東京電機大学・理工学部・教授
研究者番号：40057245

研究成果の概要（和文）：本研究は、 量的・質的記述の混在を許した多次元データ（シンボリック・データとよぶ）に内在する、単調な構造や、 部分的に単調な高次の共変的な関係を検出するための、一般的な仕組みの開発を目的としている。主な成果は、以下の３項目である。
（１）「区分的に単調な、一般的な鎖状構造の検出法の実現」
多次元空間において、シンボリック・オブジェクト群が鎖状（ループ状）に連なる構造を想定したとき、その形状が全体的に単調な構造ばかりでなく、局所的に単調であれば、高次多項式や正弦波のような構造をしていても、検出可能な方法を開発した。
（２）「高次の共変関係を評価可能とする、一般化相関係数の開発」
良く知られた相関係数を、与えられた各データサンプルの局所領域に適用し、累積することによって、高次の共変関係検出を可能とする一般化された相関係数を開発した。
（３）「入れ子構造による単調性の特性化と、そのシンボリック・データ解析への応用」
膨大なデータの要約にヒストグラムが良く用いられるが、ヒストグラムを値とする多次元データの主成分分析法を、m 分位数による方法として提案した。
また分位数法は、単に主成分分析法の範囲に留まらず、シンボリック・データのより広範囲の分析に適用可能であることが判明してきた。

研究成果の概要（英文）：This research aims to develop new methods applicable to detect monotone and locally monotone higher order covariate relations embedded in multi-dimensional symbolic data. We obtained the following three major results.
(1) Detection of locally monotonic chain structures embedded in multidimensional symbolic data: We developed a method that is able to detect higher order covariate relations. By this method we can detect higher order polynomial structures, sinusoidal structures, and others in multidimensional symbolic data table without functional identification process.
(2) A generalized correlation coefficient: By applying a well known correlation coefficient to local regions associated with each data sample and by aggregating the local correlations, we have a generalized correlation coefficient that is able to evaluate higher order covariate relations between two feature variables.
(3) The characterization of monotone structures by the nesting property and its application to symbolic data analysis: We frequently use histogram representations in order to reduce given huge data tables. By the virtue of monotone property of the cumulative distribution functions, we developed the quantile method to the principal component analysis for histogram valued data tables. The quantile method may also be able to treat other research problems in symbolic data analysis.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2007 年度	1,000,000	300,000	1,300,000
2008 年度	500,000	150,000	650,000
2009 年度	500,000	150,000	650,000
年度			
年度			
総　計	2,000,000	600,000	2,600,000

研究分野：総合領域
科研費の分科・細目：情報学・知識発見とデータマイニング
キーワード：データマイニング、シンボリック・データ解析、一般化相関係数、単調構造の検出、単調性の評価、ヒストグラム、主成分分析、分位数

１．研究開始当初の背景

パターン認識やデータ解析の一般化の試みは、広範な学問分野で、独立に、あるいは相互連携をしながら試みられている。シンボリック・データ・アナリシスは、ヨーロッパを中心に立ち上がってきたデータ解析の一般化に対する一つの流れである。ここでは、シンボリック・オブジェクト（数量や区間、記号や集合などが混在する形式で記述される、複雑なデータ）を扱うための、一般的なデータ解析法（知識獲得法）を確立することが目的とされており、最近ではデータマイニングと連携する会議やワークショップが多数開催されている。

多次元データに埋もれた因果関係を見いだす相関分析は、シンボリック・データ・アナリシスやデータマイニングおける主要なテーマの一つである。もし、多次元データの中に関数構造のような因果関係が存在するとすれば、関数構造を与える特徴組に関して、データは幾何学的に薄い構造を有するはずである。したがって、一般的なシンボリック・データに対して、「幾何学的に薄い構造を評価する仕組みの開発」が望まれる。

２．研究の目的

一般的なシンボリック・データ・テーブルにおいて、各個体(シンボリック・オブジェクト)は数量的特徴ばかりでなく、名義的特徴などが混在する形で記述される。報告者の前課題（課題番号 16500089）の研究成果として、２つの個体の相対的な近隣性に着目して、鎖状接続（chain connection）の概念を定義した。直感的な説明として、２つの個体が他の個体に関して相対的に近隣関係にあれば、これら２つの個体が鎖の構成要素　(セグメント)を成すと考える。近隣関係の個体同士をつなぎ合わせると、一般的には短い鎖が複雑につながった構造になる。ここで興味があるのは、鎖の形状が長い紐状になった構造である。多次元空間の中で、このような長い紐状の構造を検出することは、最終的には複数の特徴が示す複雑な共変関係の検出につながる。このような長い紐状の構造の内、特に始点を固定したとき、各部分鎖がより大きな部分鎖の入れ子構造になっている場合を、単調な構造と呼んでいる。古典的な主成分分析の主な目的が、多次元空間内の直線的な構造の検出にあることに対比すれば、多次元シンボリック・データに内在する単調な構造の検出は、古典的な主成分分析の概念を特別な場合に含む、一般的な共変関係の評価、発見の方法を提供することになる。研究目標を達成するために、以下の項目を立てた。

（１）問題の定式化：報告者の提案する、シンボリック・データを扱うための数学モデル（カルテシアン・システム・モデル（ＣＳＭ））に基づいて行う。

（２）単調な構造の定義：ＣＳＭのカルテシアン・ジョインと呼ぶ演算に基づいて入れ子構造を定める。

（３）特徴間の類似性の評価：ＣＳＭに基づいて、各個体の各特徴組に関する近隣集合を求め、良く知られた Jaccard index を利用して定める。

（４）（３）の類似性の評価尺度を利用して、特徴のクラスタリング（融和に基づく方法）を行う。

（５）単調性の評価：クラスタリングによって得られた特徴組において、与えられたデータの近隣関係が理想的な単調構造に比べてどの程度近いかで評価する。

（６）各種の実際的シンボリック・データ・テーブルに適用して、提案の方法の有用性を

確かめる。
（７）さらに、各個体の近隣性の一般化を行うことで、単調な構造ばかりでなく、複雑な構造（各種の非線形な関数構造や、より一般的にデータの散布状況が幾何学的に薄い構造）への拡張を試みる。

３．研究の方法

本研究の目標である「多次元データに内在する高次共変性検出の研究」の基礎になる概念は、多次元データに内在する幾何学的に薄い構造の発見である。与えられた有限個の個体が、多次元空間内で、関数関係のような共変的関係に従うとすれば、これら有限個の個体の存在する領域は、ある狭い範囲に限定されると考えられる。すなわち、個体の記述空間における散布構造は、幾何学的に薄くなるはずである。したがって、多次元データに内在する幾何学的に薄い構造を検出できれば、関数関係を含むより広い共変的な関係の検出ができることになる。このような考えの下で、報告者等は、多次元空間に内在する、幾何学的に薄く見える特徴組の選択法を開発した（Y. Ono and M. Ichino, Feature selection to extract functional structures based on geometrical thickness, IEICE Trans. Inform. And Sys., Vol. E81-D, (6),1998;　Y. Ono and M. Ichino, A new feature selection method based on geometrical thickness, Research in Official Statistics, Vol.1, No.2, 1998）。また、ピアソンの積率相関係数の一般化として、多項式関数や、X型のような線形の重ね合わせ構造を評価・検出できるカルホーン相関係数を提案した（電子情報通信学会論文誌2002）。しかしながら、これらの成果は、主に数量的特徴（区間を値とする場合も含む）を対象としており、シンボリック・データに対する一般的な方法には至っていなかった。

本研究においては、各個体の近隣集合に関する考察から、鎖状接続(chain connection)や鎖状接続被覆(chain connected covering)の概念に到達し、量的質的記述の混在する多次元データに対して，単調な鎖(monotonic chain)の概念や単調性の程度の評価尺度の構成が可能であることが分かってきた。これらの成果の一部は，国際会議録（ICCR2005, Karnataka, INDIA）に収録されている。以上を背景として、研究計画は以下の通りとした。
（１）シンボリック･データ解析の分野でよく用いられる幾つかのデータについて、特徴のクラスタリングと、単調性評価の仕組みの妥当性を検証する。
（２）　特徴毎に入れ子構造に基づくランキングを行うことで、古典的な主成分分析をシンボリック・データに拡張可能する。
（３）　他にも古典的主成分分析の一般化の方法は存在しており、これら各種の方法の利害得失の検証を行う。
（４）スピアマンの順位相関係数を局所的に適用し、その結果を総合することで、非線形構造を含む広範な共変性を評価する仕組みを創る。
（５）シンボリック・データに対する主成分分析法の一般化を行う。

４．研究成果

（１）「区分的に単調な、一般的な鎖状構造の検出法の実現」

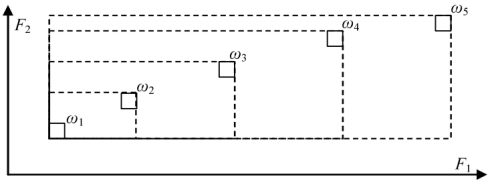

図１

図１は、ω_1からω_5までの個体の関係を示している。ω_1とω_2を頂点とする矩形領域（CSM ではカルテシアン・ジョイン領域とよぶ）は、ω_1とω_3を頂点とする、領域に含まれ、以下同様な入れ子の構造がω_1とω_5による領域まで続く。このような入れ子の性質を満たす個体の系列ω_1～ω_5を、単調な鎖とよぶ。このような単調な構造においては、個体同士の相対的な隣接関係が拘束される。実際、ω_1とω_2、ω_2とω_3、…、　ω_4とω_5の各対の領域には、他の個体が含まれないという性質があり、そのような隣接関係をもつ個体対を辺で結べば、上図においてω_1からω_5までの個体の系列が鎖状につながった姿に見える。多次元空間における単調な鎖においては、各個体の隣接関係がどの部分空間においても保存される性質がある。つまり、各軸（特徴）における個体の隣接関係をもとに、類似した特徴の組を評価･検出可能である。その結果、多次元空間に埋もれた単調な構造の発見が可能となる（分担著書参照）。
以上の考え方を、区分的にのみ単調なより一般的な鎖構造まで一般化可能である。実際、発表論文２）においては、２次関数、円、三角関数などを、乱数による無関係な特徴を加えた多次元の空間から、評価検出可能であることを報告している。
（２）「高次の共変関係を評価可能とする、一般化相関係数の開発」
２つの特徴間の共変的関係を評価する尺度として、ピアソンの相関係数、スピアマンの順位相関係数、ケンドールの順位相関係数などが知られている。これらの相関係数は、２つの特徴間の直線的な関係あるいは単調な関係を評価可能とする。しかし、図２のよう

な、２特徴間の大局的な共変的関係は評価が困難である。このとき、例えば図２(a)(b)のように幾つかの局所領域を想定すると、各局所領域においては、単調な構造が見える。発

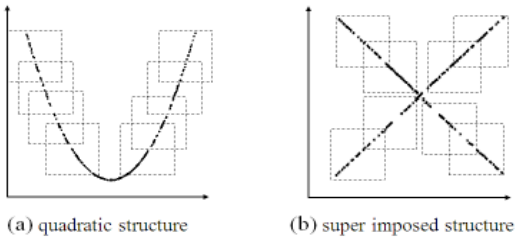

図２

表論文１）においては、与えられた個体群の各々を中心として、局所領域を設定し、その局所領域における特徴間の共変性の程度をスピアマンの順位相関係数を評価し、それらを積算することで、全体的共変性を評価する方法を報告している。

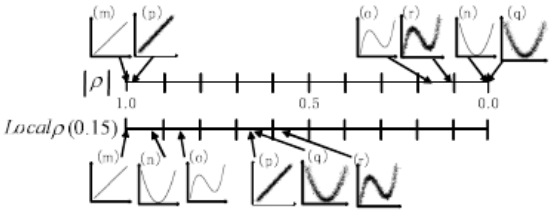

図３

この方法においては、局所性の程度を制御するパラメターを選択する任意性は残るが、各種の複雑な共変関係が評価可能である。図３の上部は、各種分布のスピアマン順位相関係数による評価結果と、局所領域にスピアマン順位相関係数を適用する提案法の結果をまとめている。提案の方法では、非単調な構造にノイズを加えた場合においても、適正な評価値が得られている。

（３）「入れ子構造による単調性の特性化と、そのシンボリック・データ解析への応用」

単調性の入れ子構造による特性化は、広範なシンボリック・データに適用可能である。表１は、シンボリック・データの典型例である。８つの油脂の種類があり、各個体は５つの特徴に関して記述されている。このうち、はじめの４つの特徴が区間を値としており、また５番目の特徴は脂肪酸の種類を値としいる。国際会議発表①において、発表③を一般化した、ヒストグラムを値とするシンボリック・データの主成分分析法を提案した。一般に、ヒストグラム・データにおいては、ヒストグラムを構成するビンの数が、個体毎に、また特徴毎に異なる。そこで、各ヒストグラムに累積分関数を想定することで、定まった個数の分位数による表現に還元する。これによって、与えられたシンボリック・データは、通常の数値データに還元され、伝統的な主成分分析の方法が適用可能となる。

表 1 Fats and oils data.

Name	Specific gravity (F1)	Freezing point (F2)	Iodine value (F3)	Saponification (F4)	Major acids (F5)
Linseed	0.930～0.935	-27～-18	170～204	118～196	L,Ln,O,P,M
Perilla	0.930～0.937	-5～-4	192～208	188～197	L,Ln,O,P,S
Cotton	0.916～0.918	-6～-1	99～113	189～198	L,O,P,M,S
Sesame	0.920～0.926	-6～-4	104～116	187～193	L,O,P,S,A
Camellia	0.916～0.917	-21～-15	80～82	189～193	L,O
Olive	0.914～0.919	0～6	79～90	187～196	L,O,P,S
Beef	0.860～0.870	30～38	40～48	190～199	O,P,M,S,C
Hog	0.858～0.864	22～32	53～77	190～202	L,O,P,M,S,Lu

具体的に、表１のデータについて説明する。今 Linseed に注目すると、この個体は、F１から F４までの特徴に関して、４次元の矩形によって記述されており、またこの矩形領域は、最小値の組による頂点（最小個体）(0.930, -27, 170, 118)と最大値の組による頂点（最大個体）(0.935, -18, 204, 196)のカルテシアン・ジョインによって生成される領域として再現される。したがって、もし F5 に関しても適当な方法によって各個体に対する最小値 a と最大値 b が定められれば、Linseed を F1 から F5 の５次元空間において、最小個体である Linseed(min) = (0.930, -27, 170, 118, a)と、最大個体である Linseed(max) = (0.935, -18, 204, 196, b)に分解して表現することができる。国際会議論文③において提案した主成分分析のための個体分解法（object splitting method）は、このような考えかたから、８行×５列のシンボリック・データを（８×２）行×５列の数値データに還元して主成分分析法を実現している。

さて、区間はビンの数が１という、特別な形のヒストグラムである。もし、例えば４分位数を想定すると、一様分布に対応する累積分布関数を想定することで、最小値と各分位数を含む５つの値を定めることができる。また、会議論文①において提案したように、F5 に関しても、名義的な場合のヒストグラムと対応する分布関数を通じて、同様に５つの値を定めることができる。一方、既に述べたように、各個体は最小個体と最大個体によって記述される事実と、累積分布関数の単調性から、各個体のそれぞれの分位数に対応する５次元の頂点も、最小個体と最大個体による矩形領域の中に、入れ子構造を保証する形で埋め込まれている。したがって、表１のシンボリ

ック・データは、８つの油脂の各個体が、５次元の特徴空間において、５つの部分個体に分解して記述されることになり、結局（８×５）行×５列の数値データに還元できることになる。この数値データに通常のピアソンの相関行列に基づく主成分分析を適用した。第１主成分と第２主成分に関する４０個の部分個体の主成分得点をプロットし、各個体に対する５つの部分個体を矢印で結んだ結果が図４となる。一つの矢印は４分位に相当する。寄与率は、第１主成分が 56.21%、第２主成分が 26.75%となっており、同じ問題に対する他に提案されている主成分分析法に比較しても、高い累積寄与率が得られている。また、因子平面におけるシンボリック・オブジェクト（個体）の新たな表現を与えている事も強調したい。すなわち、各個体の相対的な位置関係の情報ばかりでなく、矢印の向きや長さといった、各個体に関する新たな情報も表現されている。
このような情報は、個体のクラスタリングなどの、他のシンボリック・データ解析手法と組み合わせることで、様々な利点を発揮すると期待している。

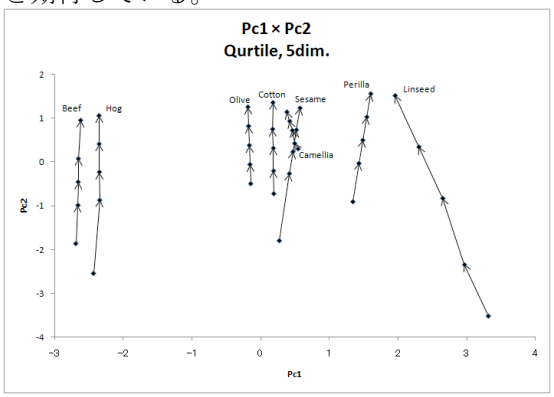

図４

あとがき
本研究の目的である、「多次元データに内在する高次共変性の検出」に関して、ほぼ目標を達成する具体的方法論が実現できたと考えている。特に、ヒストグラムは、膨大なデータをまとめるときに常套的に用いられる方法であり、この報告で一部述べた分位数によるシンボリック・データの数量化の方法は、今後、各種の解析法の実現に寄与すると期待している。
最後に、本課題のご支援に感謝するとともに、引き続いて内定をいただいている課題「分位数法に基づくシンボリック・データ・アナリシスの提案」においても新たな気持ちで取り組みたい。

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２件）
① 石川慎也、市野学、多次元データに内在する共変関係の評価について、電子情報通信学会論文誌、査読有り、A Vol. J-92-A No.11 pp.1-8, 2009.
② A. Nagoya, Y. Ono, M. Ichino, Detection of chain structures embedded in multi-dimensional symbolic data, Pattern Recognition Letters, refereed, Vol. 30, pp. 951-959, 2009.

〔学会発表〕（計３件）
① M. Ichino, Symbolic PCA for histogram -valued data, IASC2008, Yokohama, 5-8 December 2008.
② Bapu B. Kiranagi, D.S. Guru, M. Ichino, Exploitation of multivalued type proximity for symbolic feature selection, ICCTA'07, Kolkata, 5-7 March 2007.
③ M. Ichino, Symbolic principal component analysis based on the nested covering, ISI2007, Lisbon, 22-29 August 2007.

〔図書〕（計１件）
①Book Chapter:
M. Ichino, Feature clustering method to detect monotonic chain structures in symbolic data, p.95-102 in (P. Brito, et.al Eds.) Selected Contribution in Data Analysis (634pages), Springer, 2007.

〔研究会〕（計１件）
①小野、名児耶、市野、鎖状構造抽出に有効な特徴選択法の高速化、電子情報通信学会研究会、2010 年 2 月 18 日、東京農工大、PRMU 2009-210, pp. 19-24.

〔ホームページ〕
http://www.csm.ia.dendai.ac.jp

６．研究組織
(1)研究代表者
市野　学（ICHINO MANABU）
東京電機大学・理工学部・教授
研究者番号：40057245
(2)研究分担者　なし

(3)連携研究者　なし