様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成２５年６月７日現在

機関番号：３２６５７
研究種目：基盤研究（C）
研究期間：2010～2012
課題番号：２２５００１３８

研究課題名（和文）　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

研究課題名（英文）　An approach to symbolic data analysis based on the quantile method.

研究代表者
市野　学（MANABU ICHINO）
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）：
シンボリック・データ・アナリシスは、膨大なデータの統合・要約で生ずる、シンボリック・データ（ヒストグラム、区間、有限集合などの記述による複雑なデータ）の解析を目的としている。ヒストグラム、区間、有限集合などによる記述から、適当な分布関数を介して分位数に還元する方法は、シンボリック・データを数値データに帰着させる、統一的な数量化の方法を提供する。本課題の成果として、シンボリック・データに対して、分位数の単調性に基づく、主成分分析の方法を開発した。分位数法による主成分分析においては、d 個の特徴で記述される各シンボリック・オブジェクト（事例）が、予め選択された分位数 m に対して、(m+1)個の d 次元（数値）ベクトル（サブオブジェクト）の組として表現される。従って、与えられた N(オブジェクト)×(d 特徴)のシンボリック・データは、(N×(m+1)サブオブジェクト)×(d 特徴)の数値データに変換される。変換後の数値データに対して、Spearman もしくは Kendall の順位相関行列に基づく主成分分析を実行する。各シンボリック・オブジェクトは、因子平面上で、(m+1)個のサブオブジェクトの連鎖として再現される。本方法の有用性は、Journal of Statistical Analysis and Data Mining に報告した。

研究成果の概要（英文）：
Symbolic data analysis aims to analyze complex data table described by the mixture of histograms, intervals, finite sets, and others. We usually obtain symbolic data tables by the process of aggregation and summarization of vastly many data sets. We assume proper cumulative distribution functions for feature values of histograms, intervals, finite sets, and others. Then. We can obtain the respective (m+1) vectors of quantile values. A main contribution to this study is the realization of the quantile method of principal component analysis for symbolic data tables. The quantile method transforms the given (N objects) ×(d features) symbolic data table to a standard numerical data table of the size (N×(m+1) sub-objects)×(d features) for a preselected integer number m which controls the representation quality for each symbolic object. We apply the standard principal component analysis to the transformed data table. In the obtained factor planes, each symbolic object is reproduced as a series of m connected arrow lines that combine (m+1) sub-objects. We reported the usefulness of the proposed method to the Journal of Statistical Analysis and Data Mining.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2010 年度	900,000	270,000	1,170,000
2011 年度	500,000	150,000	650,000
2012 年度	500,000	150,000	650,000
年度			
年度			
総　計	1,900,000	570,000	2,470,000

研究分野：総合領域
科研費の分科・細目：知能情報学
キーワード：シンボリック・データ・アナリシス、データマイニング、分位数、数量化、多変量解析、主成分分析、単調性、分布関数

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、あるいは相互連携しながら試みられてきた。シンボリック・データ・アナリシスは、ヨーロッパを中心に３０年程を掛けて立ち上がってきた、データ解析法の一般化に対する一つの流れである。ここでは、膨大なデータの統合・要約で生ずるシンボリック・データ（ヒストグラムや区間、記号の集合等の混在した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)の確立を目的としており、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。

このようなシンボリック・データを対象として、主成分分析、クラスター分析、判別分析、重回帰分析など、伝統的な多変量解析の方法を拡張・一般化することが、主要なテーマの一つとなっている。

２．研究の目的

ヒストグラムや区間、有限集合などの混在する形式で記述されるシンボリック・データを対象とするとき、どのようにすれば、既存の多変量解析の一般化法を実現可能かが、本研究の主題である。ここでは、ヒストグラムや区間、有限集合などの記述に対して、適当な累積分布関数を想定し、予め定められた個数 m の分位数を求める。すると、d 種類の特徴で記述される対象事例（オブジェクト）は、d 個の(m+1)次元数値ベクトルの組によって表現可能である。従って、(N オブジェクト)×(d 特徴)のシンボリック・データは、（N×(m+1) ベクトル）×(d 特徴) の数値データ・テーブルに変換可能である。このように、分位数に注目すると、異なる記述が混在するデータを、統一的に数値データに変換する「数量化の方法」を提供することになる。このような数量化の下で、各種の一般化された多変量解析法を実現可能であるが、原理的な方法の確立した、シンボリック・データに対する主成分分析の方法を以下に報告する。

３．研究の方法

主成分分析法をシンボリック・データに拡張する試みに関して多くの報告があるが、そのほとんどが、K. Pearson の定式化の一般化を目指している。一方、本報告の方法は、分位数に基づく数量化と、単調性の性質である入れ子構造の特性に基づく定式化である。以下、各オブジェクトが d 次元区間で表現される場合の単調性に基づく主成分分析の方法から、分位数法による一般化された方法に至る過程の概要を述べる。

（１）d 次元区間の単調性と主成分分析

N 個のオブジェクトの各々が、d 個の特徴それぞれに対して、区間を値とする場合を想定する。つまり、各オブジェクトは d 次元区間を形成している。今 N 個のオブジェクト $W_1, W_2, \cdots, W_N$ のそれぞれが d 次元区間であり、これら N 個の区間が W_1 を始点とし W_N を終点とする単調構造を有すると仮定する。このとき、W_1 は W_1 と W_2 の張る d 次元区間（W_1 と W_2 を同時に含む最小の d 次元区間、以下同様）に包含され、その区間はまた W_1 と W_3 が張る区間に含まれると言うように、入れ子の構造が成立している。このような入れ子の性質は、各座標軸に遺伝的に継承される。一方、任意の d 次元区間は、d 個の最小値の組で表現される最小頂点ベクトルと、d 個の最大値の組で表現される最大頂点ベクトルの２つのベクトルの張る領域として表現される。これら 2N 個の頂点ベクトルは、元の N 個のオブジェクトの入れ子構造の制約を受けており、N 個の最小頂点ベクトルも N 個の最大頂点ベクトルも同様の入れ子構造を有することになる。このような入れ子構造の性質は、2N 個の頂点ベ

クトルの各座標における並び順に正確に反映されることから、任意の座標対における頂点ベクトルの並び順の類似性を、Spearman もしくは Kendall の順位相関係数で評価可能となる。つまり、N 個の d 次元区間で記述されるオブジェクトの単調性を、2N 個の頂点ベクトルに関する順位相関係数による評価に還元することが可能である。従って、2N 個の頂点ベクトルに関する d 次の順位相関行列に基づく固有値問題を通じて、主成分分析が可能となる。各因子平面上で、与えられたそれぞれのオブジェクトは、最小頂点と最大頂点を結ぶ矢印として表現可能となる。

（２） 分位数法による一般化

オブジェクト W を記述する(m+1)個の d 次元ベクトル（以下分位ベクトルとよぶ）を $Q_0, Q_1, \cdots, Q_m$ によって表す。Q_0 と Q_m は、それぞれ最小頂点ベクトルと最大頂点ベクトルに対応しており、また他の分位ベクトルは、各成分である分位数の単調性から、ベクトルとしての順位 $Q_0 \le Q_1 \le \cdots \le Q_m$ が自動的に保証される。従って、与えられた N 個のオブジェクト $W_1, W_2, \cdots, W_N$ が単調性を有すれば、その性質は、N×(m+1)個の分位ベクトルの単調性として受け継がれることになる。よって、(N×(m+1)分位ベクトル)×(d 特徴)のデータ・テーブルに対して、Spearman もしくは Kendall の順位相関行列を基にして主成分分析が可能となる。因子平面上で、各オブジェクトは、最小分位ベクトルから最大分位ベクトルを結ぶ m 個の矢印の連鎖として表現可能となる。

４．研究成果

（１）ヒストグラム・データの分析

最初に、Histogram data by the U.S. Geological Survey, Climate-Vegetation Atlas of North America, http:// pubs.usge.gov /pp/ p1650-b/から引用したデータの解析を通じて、提案法の有用性を示す。

表１ 年間平均気温のヒストグラム・データ

	Annual Temperature (°C)							
Taxon name	N	0%	10%	25%	50%	75%	90%	100%
ACER EAST	6865	-2.3	0.6	3.8	9.2	14.4	17.9	23.8
ACER WEST	1954	-3.9	0.2	1.9	4.2	7.5	10.3	20.6
ALNUS EAST	10144	-10.2	-4.4	-2.3	0.6	6.1	15.0	20.9
ALNUS WEST	4761	-12.2	-4.6	-3.0	0.3	3.2	7.6	18.7
BETULA	16815	-13.4	-8.4	-5.1	-1.0	3.9	12.6	20.3
CARYA	4638	3.6	7.5	10.0	13.6	17.2	19.4	23.5
CASTANEA	2216	4.4	8.6	11.3	14.9	17.5	19.2	21.5
FRAXINUS EAST	8565	-2.3	1.4	4.3	8.6	14.1	17.9	23.2
FRAXINUS WEST	1095	2.6	9.4	11.5	17.2	21.2	22.7	24.4
JUGLANS EAST	4138	1.3	6.9	9.1	12.4	15.5	17.6	21.4
JUGLANS WEST	526	7.3	12.6	14.1	16.3	19.4	22.7	26.6
OSTRYA/CARPINUS	5348	1.2	4.4	7.0	11.4	16.0	19.2	28.0
QUERCUS EAST	7360	-1.5	3.4	6.3	11.2	16.4	19.1	24.2
QUERCUS WEST	1942	-1.5	6.0	9.5	14.6	17.9	19.9	27.2
TILIA	3792	1.1	3.8	5.8	8.8	12.0	14.4	19.9
ULMUS	8028	-2.3	1.7	4.9	9.7	15.3	18.6	23.8

使用したデータは、１６種類の広葉樹が８種類のヒストグラム・データで記述されている。表１は、１６種類の広葉樹が、年間平均気温に関して、予め設定された領域（領域数は N で示されている）での生育の可能性に関し、どのような分位数をとるかを示している。使用した特徴は、以下の８種類である。

F_1:年間平均気温(ANNT) (ºC)
F_2: １月平均気温(JANT) (ºC)
F_3: ７月平均気温(JULT) (ºC)
F_4:年間雨量(ANNP) (mm)
F_5: １月雨量(JANP) (mm)
F_6: ７月雨量(JULP) (mm)
F_7:成長率(GDC5)
F_8:湿度(MITM)

表２ 分位ベクトル・データの一部

Taxon name	F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8
ACER EAST 1	-2.3	-24.6	11.5	415	10	56	0.5	0.62
2	0.6	-18.3	16.6	720	23	77	1.2	0.89
3	3.8	-12.3	18.2	835	40	89	1.5	0.94
4	9.2	-5.1	22.2	1010	69	100	2.5	0.97
5	14.4	2.3	25.8	1200	96	113	3.6	0.99
6	17.9	7.9	27.3	1355	127	135	4.8	0.99
7	23.8	18.9	28.8	1630	166	222	6.8	1.00
ACER WEST 1	-3.9	-23.8	7.1	105	5	0	0.1	0.14
2	0.2	-11.8	11.3	380	28	8	0.5	0.49
3	1.9	-10.1	12.8	505	54	23	0.7	0.61
4	4.2	-6.9	14.9	750	92	38	1.1	0.75
5	7.5	-1.3	17.6	1175	176	52	1.6	0.91
6	10.3	3.3	19.9	1860	267	71	2.2	0.98
7	20.6	11.0	29.2	4370	616	160	5.6	1.00
ALNUS EAST 1	-10.2	-30.9	7.1	220	9	28	0.1	0.22
2	-4.4	-26.5	13.2	380	19	58	0.6	0.53
3	-2.3	-22.7	14.8	475	23	74	0.8	0.69
4	0.6	-18.1	16.5	770	46	91	1.1	0.93
5	6.1	-8.0	19.8	1060	80	108	1.9	0.99
6	15.0	3.7	25.7	1235	106	126	3.7	0.99
7	20.9	14.1	29.1	1650	166	212	5.9	1.00
ALNUS WEST 1	-12.2	-30.5	7.1	170	4	0	0.1	0.22
2	-4.6	-25.7	11.5	335	18	21	0.5	0.49
3	-3.0	-21.6	12.8	410	23	41	0.7	0.59
4	0.3	-15.1	14.4	510	37	57	0.9	0.72
5	3.2	-7.6	15.6	790	93	74	1.1	0.87
6	7.6	-0.8	17.5	1385	199	87	1.6	0.97
7	18.7	10.8	28.3	4685	667	452	4.8	1.00

各特徴の７分位数自身を新たな特徴に見立てると（16 オブジェクト)×(8×7 特徴)のデータ・テーブルを想定可能であるが、通常の意味での主成分分析は適用できない。一方、提案の分位数法においては、各オブジェクトが７個の分位ベクトルの組として表されることから、(16×7 分位ベクトル)×(8 特徴)のデータ・テーブルに対する問題に帰着される。構成されたデータ・テーブルの一部を、表２に示している。

Spearman 相関行列に基づいて主成分を求めると、第１主成分が大きさの因子を示し、その寄与率は 87.41%である。一方、第２主成分の寄与率は 8.38%であり、ANNP、JANP、JULP と MITH が正の重みを示し、他の４特徴は負の重みを示している。特に MITH が、大きな

正の重みを示す結果となっている。図１は、最初の因子平面において、１６個のオブジェクトを６個の矢印の連鎖として再現した、分位数法による主成分分析の結果である。

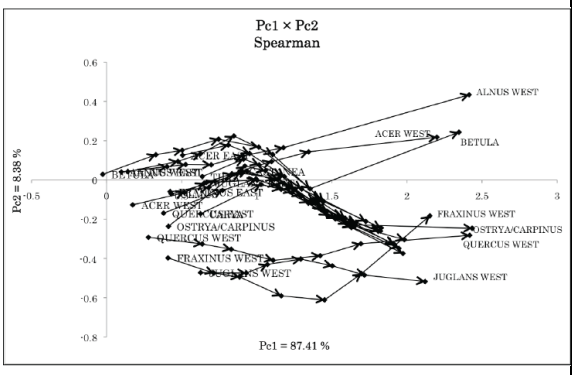

図１　分位数法による主成分分析の結果

東部に生育するほとんどの広葉樹は、一旦右上方に上ってから右下に合流する変化を示している。一方、西部に生育する広葉樹は、樹種それぞれが個性的な変化を示す。特に、Acer West（楓）、Alnus West(ハンノキ)、Betula（樺）、および Fraxinus West(トネリコ)が、最後の分位において右上に跳ね上がる特徴的な変化を示している。この変化は主に、雨量と湿度に依存していると解釈される。

以上のように、８つのヒストグラム・データを、分位数法によって一括して主成分分析に掛けられるという、提案法の利点の一端を示した。

（２）異種特徴の混在したデータの場合

上に示した例においては、元々の８種類のデータが４分位と１０分位を組み合わせた「分位数のデータ・テーブル」として表現されている。一方、より一般的な問題として、N 個のオブジェクトが、種類の異なる d 種類のデータ・テーブルの組として提示される場合が考えられる。この場合も、予め選択された分位数 m を定め、d 種類のデータ・テーブルのそれぞれを、適当な累積分布関数に基づいて(N オブジェクト)×((m+1)分位数)) に変換(数量化)することが可能である。その後、各オブジェクト対して、(m+1)個の d 次元分位ベクトルを構成することで、最終的に(N×(m+1)分位ベクトル)×(d 特徴)のデータ・テーブルにまとめることが可能である。後は、上に示した例のように、Spearman もしくは Kendall の順位相関係数に基づく主成分分析に帰着可能である。また多くの例において、Pearson の相関行列による解析も、類似の解析結果が得られることを確認している。

（３）他の多変量解析法への応用

分位数法によるシンボリック・データの数量化は、多くの多変量解析法を適用可能にするが、今後「分位数法に基づくクラスタリング」や、「分位数法に基づく判別分析」など独自の方法に取り組んでおり、順次その成果を報告する予定である。

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３件）

① A. Nagoya, Y. Ono, M. Ichino: A generalized measure of covariant relations based on relative neighborhood relations, Far East Journal of Theoretical Statistics,(査読有), 37, 2, pp. 125-143, 2011.

② 石川，市野：局所決定係数を用いた多次元データにおける共変性の評価について，電子情報通信学会論文誌 A,(査読有), J94-A, 5, pp. 372－382,2011.

③ Manabu Ichino: The quantile method for symbolic principal component, analysis, Statistical Analysis and Data Mining,(査読有), 4, 2, pp.184-198, 2011.

〔学会発表〕（計４件）

① M. Ichino, P. Brito: The data accumulation PCA to analyze periodically summarized multiple data tables, COMSTAT-2012, (査読無), August 27-31, Limassol, Cyprus, 2012.

② M. Ichino, P. Brito: The data accumulation method for symbolic principal component analysis, ISI 2011, (査読無), August 21-26, Dublin, Ireland, 2011.

③ P. Brito, M. Ichino: Symbolic clustering based on quantile representation, COMSTAT 2010 (19th International Conference on Computational Statistics), (査読無), August 22-27, Paris, France, 2010.

④ M. Ichino, P. Brito: The quantile method for symbolic hierarchical clustering, GfKl 2010 Symposium Karsruhe, (査読無), July 21-23, Karlsruhe, Germany, 2010.

〔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