様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　２９　年　　４　月　３０　日現在

機関番号：　３２６５７
研究種目：　基盤研究(C)（一般）
研究期間：　2013～2016
課題番号：　２５３３０２６８

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）The quantile method for symbolic data analysis.

研究代表者
市野　学（Ichino, Manabu）

東京電機大学・理工学部・名誉教授

研究者番号：４００５７２４５

交付決定額（研究期間全体）：（直接経費）　　2,400,000円

研究成果の概要（和文）：　個々の事例が数値による記述であっても、データの規模が大きくなると区間やヒストグラムによる要約が行われる。シンボリック・データは、このような要約されたデータの総称であり、シンボリック・データ・アナリシスは、要約されたデータを対象とする解析法である。本研究の分位数法は、シンボリック・データを統一的に数値データに再変換し、主成分分析やクラスタリングなどの伝統的な方法を適用可能とする方法である。本研究では、階層的概念クラスタリング法、シンボリック・データの可視化の方法、およびルックアップ・テーブル型回帰モデルを開発した。

研究成果の概要（英文）：　We often use intervals and histograms to summarize the given large numbers of numerical data sets. We use the term symbolic data to call such summarized data and data by the aggregation of different data tables.The quantile method transforms the given symbolic data table to a different sized numerical data table by a unified way. Then, we realize various data analysis methods on the transformed data. This research report includes three quantile methods for symbolic data: (1) A hierarchical method of conceptual clustering; (2) Visualization of multidimensional symbolic data; and (3) The lookup table regression model.

研究分野：　知能情報学

キーワード：　symbolic data analysis　data mining　Cartesian system model　quantile method　PCA　hierarchical clustering　visualization　regression model

１．研究開始当初の背景

情報処理技術の進展に伴い、大量のデータの収集・蓄積・処理が容易になり、データマイニングやビッグ・データなどの呼称も定着してきた。データベース技術の延長線で考えられていたデータマイニングの方法も、統計的手法と合流し、したがって 1980 年代後半にヨーロッパを中心に立ち上がってきた、シンボリック・データ・アナリシスの分野とも合流することとなった。シンボリック・データ・アナリシスは、区間やヒストグラム、有限集合といった、より一般的な記述を許すシンボリック・オブジェクトの集合（シンボリック・データとよぶ）の解析を目標としており、伝統的な統計的手法の一般化が一つの大きな柱となっている。

２．研究の目的

報告者は、オブジェクトが量的記述や質的記述の混在する形式で与えられたとき、オブジェクト集合の概念記述を行う数学モデルとして、カルテシアン・ジョイン・システムを提案した（引用文献④）。以来、数学モデルもカルテシアン・システム・モデルに発展し、科学研究費の継続的ご支援を得て、「分位数に基づくシンボリック・データ・アナリシスの提案」（引用文献③）に至り、シンボリック・データを対象とした、一般的な主成分分析法を開発した（引用文献①、②）。

本研究は、引き続き分位数法に基づくシンボリック・データ・アナリシスの提案として、

（１）　分位数による、シンボリック・データの階層的概念クラスタリング法の開発。

（２）　累積概念グラフによるシンボリック・データの可視化。

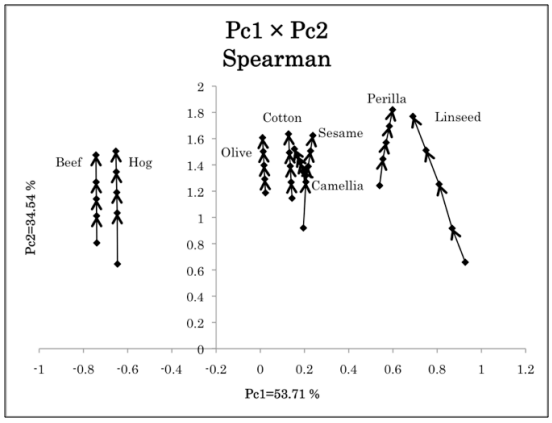
（３）　シンボリック・データに対するルックアップ・テーブル型回帰モデルの開発

などを目標とした。

３．研究の方法

各オブジェクトが、区間やヒストグラム、さらには有限集合の混在した形式で記述されるとき、分位数法においては、各異なる種類の記述に適当な分布関数を想定し、各々を分位数の組として表現する。その上で、与えられたＮ個のオブジェクトがｄ種類の異なる特徴で記述された、Ｎ×ｄのサイズのシンボリック・データを、{(N×(m+1)サブオブジェクト}×(d 特徴) の数値データに再変換する。ここで、ｍは予め定められた１以上の整数で、４分位を選択するとすれば、m=4 とする。つまり、各シンボリック・オブジェクトは、m+1 個のサブオブジェクトの組として表現されており、各サブオブジェクトは、d 次元の数値ベクトル（分位ベクトルと呼ぶ）によって記述されている。ここで、m+1 個の分位ベクトルは、最小分位ベクトルから最大分位ベクトルまで単調性を保証している。特別な場合として、m=1 場合に限定すると、各シンボリック・オブジェクトは、d 次元の最小、最大の２つの分位ベクトルで記述されている。このようなシンボリック・データの例として、選択されたＮ都市を１月から 12 月まで、各月の最低平均気温と最高平均気温で記述した気温データが挙げられる。ここでは各都市が 12 個の区間データで記述されている。つまり、各都市は、12 次元の超区間として記述されている。一方、分位数法では、各都市を記述する 12 次元の超区間を、それを張る最小分位ベクトルと最大分位ベクトルに分けて表現していることになる。一般の場合として、例えばｍ=4 の時は、最小分位ベクトルと最大分位ベクトルの張る 12 次元超区間の中に、他の３つの分位ベクトルは各次元に沿って単調性を満たすように配置されている。

このような分位数法の考え方の基で、シンボリック・データに対する主成分分析（引用文献②）においては、各シンボリック・オブジェクトを、因子平面上で(m+1)個のサブオブジェクトを繋ぐｍ個の矢印の連鎖として表現する（下図は、４次元の区間データ（比重、凝固点、イオン価、鹸化価）と、１個の有限集合（主な脂肪酸）を値とする油脂データに４分位を適用した結果)。また、この方法は、単調性を担保するようにデータを融和させることで、３相データの主成分分析に対する新たなツールを提供する事を示した（引用文献①、学会発表③)。



主成分分析以外の分位数法に基づく個々の方法については、次節の研究成果においてまとめる。

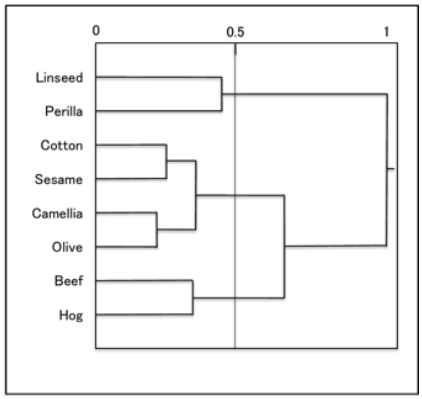
４．研究成果

分位数法によるシンボリック・データ・アナリシスの包括的な報告は、学会発表④で行ったが、個々の方法に関する、成果の概要を以下にまとめる。

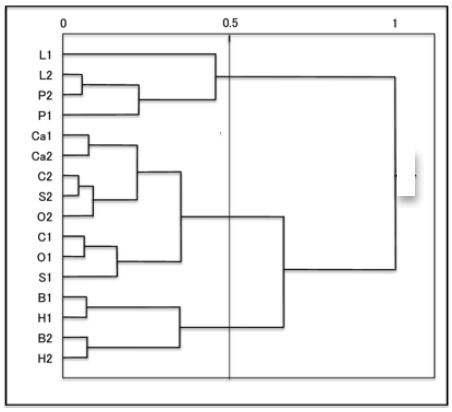
（１）階層的概念クラスタリングの方法

ここでは、各概念クラスターをｄ個の特徴に関する超区間で記述する。各オブジェクトを記述するｄ次元ユークリッド空間において、オブジェクト対、あるいはクラスター対が生成する概念の大きさ（超区間の大きさ）を、

コンパクトネスとよぶ尺度により定める。したがって、本階層的概念クラスタリングにおいては、各ステップにおいてクラスターのコンパクトネスを最小となるようにクラスターの融和を繰り返す。下図は、油脂データの各オブジェクトが、５個の区間値で定められるとした時の結果であり、コンパクトネスの大きさを 0.5 で切った時に、明確に３つのクラスターが得られる（主成分分析の結果参照）。

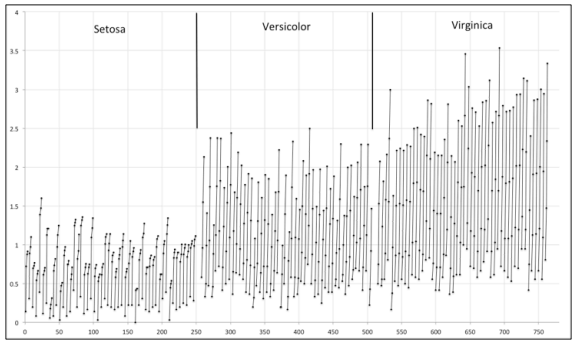


各オブジェクトが(m+1)個の分位ベクトルで記述される時、オブジェクトの主要部分を表現する２つの分位ベクトルを選択し、マクロな視点からクラスタリングを行う方法と、各分位ベクトルを独立したオブジェクトとみなす、サブオブジェクトに基づくクラスタリングの方法が考えられる（学会発表⑧、⑨、①）。後者を併用することで、マクロなクラスター間の構造に加えて、クラスター間の局所的な関係をサブオブジェクトの繋がりから把握可能である。（下図参照。ここで例えばＬ１、Ｌ２は、それぞれ Linseed の最小分位ベクトルと最大分位ベクトルを意味する）



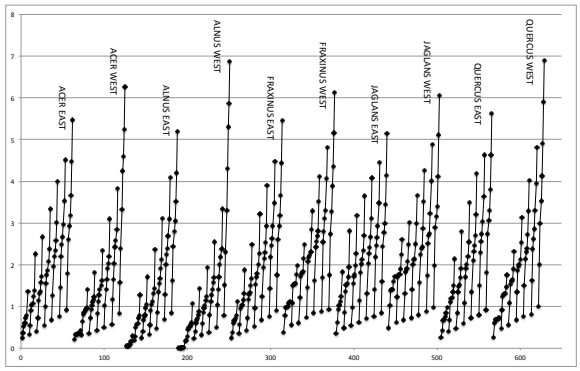
本階層的概念クラスタリングにおいて、オブジェクト対あるいはクラスター対のコンパクトネスを最小化することは、最も高い類似性を有する対を選択し融和する事を意味している。一方、その事は、各対の生成する概念が、全概念に対して最も非類似性が高い事を意味している。したがって、コンパクトネスは、融和すべき対象の類似性尺度の役割を果たすと同時に、生成されるクラスターの良さを評価する尺度の役割も果たしていることを明らかにした（学会発表②）。

（２）累積概念グラフによるシンボリック・データの可視化の方法



多次元データの可視化のために、各種の方法が開発されている。本研究では、概念の大きさに基づく、累積概念グラフを提案した。各オブジェクトを、特徴毎に 0-1 区間で正規化した値の累積値を、折れ線グラフとして表現する。例えば、上図は良く利用されるアヤメのデータである。一つのアヤメを、正規化された４つの特徴値であるガクの長さ、幅、花びらの長さ、幅と逐次累積して３本の線分から成る単調な折れ線グラフを構成し、オブジェクトの数だけ並べて表示している。

累積概念グラフの利点は、マクロな大きさが一見して比較可能であり、特異なオブジェクトの検出に役立つ．累積概念グラフの形は積算する特徴の順序に依存するが、累積値を示すマーカーの位置と長さの比較により、オブジェクト間の違いが見える。



上図は、シンボリック・データの累積概念グラフの例である。北米に生育する広葉樹合わせて 10 種が、年間平均気温、年間雨量等８種類の特徴で記述されており、各々がヒストグラムデータである。各ヒストグラムデータは、0、10、25、50、75、90、100%の７分位数で構成されており、したがって、8 種類の特徴毎に、(10 種類の広葉樹）×（7 分位数）のデータ表による、３相データを形成している。提案の分位数法においては、各広葉樹は７個の８次元分位ベクトルで記述される。したがって、各分位ベクトルを８種類の 0-1 区間で正規化された分位数を累積する事で、オブジェクト毎に７本の単調な折れ線グラフの組として表現可能となる。分位ベクトル

（サブオブジェクト）のマクロな性質と同時に、サブオブジェクト間の局所的性質の違いが、２次元図形の微妙な違いとして捉えられる（学会発表⑥、⑦）。

（３）ルックアップ・テーブル型回帰モデル

関数の構造を仮定しない回帰モデルとして、ルックアップ・テーブル型のモデルを提案した（学会発表⑤）。

油脂データを例に、手続きを述べる。８種類の油脂が、５種類の区間データにより記述されている。８種類の油脂を５次元の最小分位ベクトル、最大分位ベクトルに分割する。イオン価を目的変数として、イオン価の最小値から最大値に向けて、各分位ベクトルを並び替える。

	Iodine Value	Specific Gravity	Freezing Point
B1	40	0.86	30
B2	48	0.87	38
H1	53	0.858	22
H2	77	0.864	32
O1	79	0.914	0
Ca1	80	0.916	−21
Ca2	82	0.917	−15
O2	90	0.919	6
C1	99	0.916	−6
S1	104	0.92	−6
C2	113	0.918	−1
S2	116	0.926	−4
L1	170	0.93	−27
P1	192	0.93	−5
L2	204	0.935	−18
P2	208	0.937	−4

Iodine value	Specific gravity	Freezing point
[40, 77]	[0.858, 0.870]	[30, 38]
[79, 113]	[0.914, 0.920]	
[79, 208]		[−27, 6]
[116, 208]	[0.926, 0.937]	

次に、各説明変数について、ブロック間に単調性が保証されるように区切る（Block segmentation）。一つのブロックしか作れない特徴は削除する。このようにして得られた結果が上表左である。鹸化価と主な脂肪酸は削除され、また比重は３つのブロックに、また凝固点は２つのブロックにそれぞれ区分けされた。結果を右の表にまとめている。

目標値の推定は、与えられた説明変数の値（数値、もしくは区間）と右表の説明変数の各ブロックを表す区間への帰属度を計算し、最大の帰属度を与える区間に対応する目的変数の区間を指定する事で行う。例えば、凝固点が区間[30,38]への帰属度が最大であれば、イオン価の推定値は[40,77]である。

帰属度には、非対称の親近性尺度（学会発表②）の有用性を確認している。また、推定の精度を上げるのには、クラスタリングを行い、複数のルックアップ・テーブルを利用することが可能である。詳細は、今後の残された課題としたい。

本報告に関連して、シンボリック・データに対する、教師付き階層的概念クラスタリングとパターン認識への応用や、高次元シンボリック・データに埋もれた高次共変性の検出、さらに、高次元シンボリック・データにおける欠損値の推定法なども引き続き検討課題としたい。

＜謝辞＞

最後に、本研究に至るまでに賜った、継続的なご支援に感謝する。

＜引用文献＞

① M. Ichino and P. Brito, The data accumulation PCA to analyze periodically summarized multiple data tables, COMSTAT2012, August 27-31, Limassol, Cyprus.
② M. Ichino, The quantile method for symbolic principal component analysis, Statistical Analysis and Data Mining, Vol.4, No.2, pp184-198, 2011.
③ 市野、分位数に基づくシンボリック・データ。アナリシスの提案、 基盤研究(C)、２０１０～２０１２年度。
④ 市野、カルテジアン・ジョイン・システムに基づく新しいパターン認識法の研究、科学研究費補助金一般研究（C)、１９８６年度。

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 大用、市野、高橋、緩い対称性を持つ因果的価値観数の認知的妥当性とＮ本木バンディット問題におけるその有効性、人工知能学会誌、30(2-A)403-416、 2015.
② 名児耶、小野、市野、鎖状構造に基づく主成分分析の一般化、電子情報通信学会論文誌 D,J98-D(3) 501-511、2015.

〔学会発表〕（計９件）

① K. Umbleja and M. Ichino, Predicting students' behavior during an E-learning course using data mining, International Conference on Interactive Collaborative Learning, ICL2016, September 21-23, Belfast, UK.
② M. Ichino and K. Umbleja, Similarity and dissimilarity measures for mixed feature-type symbolic data, 48th Scientific Meeting of the Italian Statistical Society, SIS2016, June 8-10, Salerno, Italy.
③ M. Ichino, The data accumulation method: Dimensionality reduction, PCA, and PCA like visualization, Symbolic Data Analysis Workshop, SDA2015, Nov.17-19, Orleans, France.
④ M. Ichino, The quantile method for symbolic data analysis, Tutorial of Symbolic Data Analysis Workshop, SDA2015, Nov.17-19, Orleans, France.
⑤ M. Ichino, The lookup table regression

model for symbolic data, Data Science Workshop, Nov.12-13, Paris-Dauphin, France.
⑥ P. Brito and M. Ichino, The data accumulation graph (DAG): Visualization of high dimensional complex data, COMSTA2014, August19-22, Geneva, Swiss.
⑦ M. Ichino and P. Brito: The data accumulation graph (DAG) to visualize multi dimensional symbolic data, Workshop in Symbolic Data Analysis, SDA2014, Junen13-16, Taipei, Taiwan.
⑧ M. Ichino and P. Brito, A hierarchical conceptual clustering based on the quantile method for mixed data, JOCLAD2014, April 10-12, Lisbon, Portugal.
⑨ M. Ichino and P. Brito, A hierarchical conceptual clustering based on the quantile method for mixed feature-type data, 59th World Statistics Congress of the International Statistical Institute, August 25-30, Hong Kong.

〔その他〕
ホームページ等
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　名誉教授
研究者番号：40057245

(2)研究協力者
BRITO, Paula
Faculty of Economics of the University of Porto, Associate Professor

UMBLEJA, Kadri
Department of Computer Control, Tallinn University of Technology, Ph.D