様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月　 ５ 日現在

機関番号：３４３１０
研究種目：基盤研究(C)
研究期間：2012～2014
課題番号：２４５００３５２

研究課題名（和文）大規模複雑関連性データの解析法に関する総合的研究

研究課題名（英文）Proximity data analysis for large and complex data

研究代表者
宿久　洋（Yadohisa, Hiroshi）

同志社大学・文化情報学部・教授

研究者番号：５０２４４２２３

交付決定額（研究期間全体）：（直接経費）　　4,100,000円

研究成果の概要（和文）：関連性データとは，２者間の類似関係についてのデータである．関連性データの代表的な解析法としては，多次元尺度構成法やクラスタリング法があるが，近年のデータの大規模化，複雑化により，既存のデータ解析法を適用しても，結果の解釈が不可能であったり，計算量のために解析が不可能である場合が多い．そこで本研究では，大規模な関連性データに対して，シンボリックデータ解析を用いる方法，部分空間を用いる方法，次元縮約法と既存手法の同時分析を用いる方法という３つのアプローチによる新たな解析手法を提案した．

研究成果の概要（英文）：The proximity data is made of similarity or dissimilarity between two objects. The typical statistical methods to analyze the proximity data include Multidimensional Scaling (MDS) and Clustering methods. However, since the data becomes larger and more complicated recently, sometimes the existing method does not provide the interpretable result and/or does not work because of the amount of computation.
Therefore, in this study, for large and complicated proximity data, we propose new statistical methods via an approach by symbolic data analysis, by using subspace, by simultaneous analysis with existing method and dimensional reduction method.

研究分野：多変量データ解析

キーワード：ビッグデータ　シンボリックデータ　多次元尺度構成法　クラスタリング　行列分解型多変量解析

１．研究開始当初の背景

類似性データは，対象間の２者関係について記述したデータであり，その値が高ければ２者間の関係は近いとみなし，そうでなければ遠いとみなす．実際に類似性データ心理学，社会学等の様々な分野で観測される．例えば，心理学分野では実験課題を通して２つの刺激の関連性が観測され，社会学分野では世代間の職業移動や地域間の移住の件数を集計することで観測される．このような類似性データ行列が得られた際，対象間の近さ遠さの関係から，データの構造を明らかにすることを目的として様々な分析手法が提案されている．具体的な分析手法としては対象間の近さ遠さを視覚的に把握することを目的とした多次元尺度構成法，対象のグループ（以下クラスター）を検知するクラスタリング法が挙げられる．また，類似性データについて，２つの数学的特徴が挙げられ，それぞれの数学的特徴を考慮した分析手法が提案されている．具体的に類似性データが持つ特徴の１つめは対象間の関係について対称な関係と非対称な関係による特徴である．２つめの特徴は２者関係のみではなく，２者関係を拡張した多者関係によって観測される多相多元類似性データによって特徴づけられることである．

しかし，近年の情報技術の発達に伴い，データが大規模複雑化し，そのような大規模複雑データを扱う必要性が高まっている．類似性データも同様にそのような大規模複雑化しており，１次データとして，もしくは大規模複雑データを加工することで大規模非類似性データが得られ，分析する必要性が高まっている．具体的に，情報科学の分野では各 Web ページを対象とみなし，Web ページ間の移動履歴より非対称類似性データが観測され，各 Web ページ間の関係を把握することが重要となる．マーケティング分野では商品のブランドを対象とし，ブランドの購買履歴より同様の非対称類似性データを扱う必要性が高まっている．

このような大規模複雑データを分析する上で３つの問題点が挙げられる．１つめは，ゼロの値，もしくは NA の値が高い割合でデータに含まれる場合が多く，既存の分析手法では意図した結果を得ることが困難となる点である．これはゼロの値が多いことから，対象間を識別する情報が不足しているため生じる問題である（スパース性の問題）．２つめは，対象数が膨大になるため，各対象間の識別性を解釈することを目的とする既存の分析手法ではデータの特徴を把握することが困難となる．３つめは，対象数が膨大となるため，計算時間が膨大になってしまうという問題点となる．上記の問題を克服する方法として，Diday(1988)によって提案された大規模データを集約するための方法である Symbolic Data 解析や次元縮約と既存分析法の同時分析法の開発，計算時間を少なくするために既存手法等に関する Majorizing 関数の導出，等が考えられる．しかし，データが大規模複雑化することによって，分析目的が多種多様化していることから，これらの解決方法に対応する分析方法を整備し，応用上有用であるような分析手法の開発が望まれる．

２．研究の目的

先にも述べた問題を解決するため，以下の４点を達成することを目的とする．

（１）大規模複雑類似性データに対する分析法の総合的調査

（２）大規模複雑類似性データの新たな分析手法の提案

（３）提案手法の実装及び高速化

（４）実データに対して提案手法を適用し，新たな知見の獲得

（１）で述べた大規模類似性データとは先に述べた非対称および対称類似性データ，

多相多元類似性データに加え，Symbolic Data と呼ばれる「区間値」や「多値」，「分布値」といったデータ値を取りうる類似性データのことを指す．また，クラスタリング法では入力データとして類似性データではなく，多変量データ型のデータが入力となる場合もあるが，例えば k-means 等のアルゴリズムは対象とクラスター重心間の非類似性データ行列の系列を生成してクラスタリングを行うことから，広義の意味での類似性データの分析とみなすこととする．

(1)での目的は手法の性質に基づいて各手法を性質ごとに分類し，数理的に特徴付けることである．

（２）では先に述べた３つの問題点を克服する新たな分析方法を提案することを目的とする．特に類似性データの主な分析法であるクラスタリング法および多次元尺度構成法はいずれもデータの特徴把握を目的とした方法であることから，分析結果を容易に解釈可能な手法の開発が望まれる．そこで，「次元縮約」，「視覚化」に注目して新たな分析の開発を行うこととした．

（３）について，たとえ大規模複雑データの構造を捉え，解釈できるような分析手法を開発したとしても，計算量等の観点より適用困難では目的を達成することは困難となる．そこで，既存の分析方法および新たな分析方法について実装可能であるようなアルゴリズムの開発を行うことを目的とする．

（４）については，（１）（２）（３）を考慮して，実際に実データを分析することにより，提案手法の実用性について分析結果から検討し，より実用性の高い手法の開発を目指すことを目的とする．

３．研究の方法

研究の方法についても先に述べた（１）（２）（３）（４）に沿って説明することとする．

（１）で述べた大規模複雑データに関する分析法の総合的調査の目的である各手法の数理的特徴付けを行う為には，各分析法を比較可能な観点，定式化が必要となる．任意の類似性データ分析，およびクラスタリング法で扱うデータ分析法は「観測されたデータ（の関数）」と「分析法のモデル」の誤差を最小にする形式で記述可能であることから，統一的基準に基づく行列表記に基づく各手法の定式化を行い，分析手法の特徴について数理的特徴付けを行う．

（２）で述べた分析手法については，「スパース性の問題」や「解釈性の問題」を克服する方法として，各対象をその上位概念等によって１つの値ではなく，「分布値」，「区間値」，「多値」に集約して，各問題を克服しその「バラつき」の情報を保存し考慮する多次元尺度構成法等の視覚化手法を提案する．また，k-means 法ベースの手法を大規模複雑データに適用しても変量数が多いことから解釈が困難となる．そこで様々な形式のデータに適用可能な「次元縮約クラスタリング法」を提案する．次元縮約クラスタリング法とは対象のクラスタリング結果と各クラスターの特徴が識別可能な低次元空間を同時に検知する方法である．

（３）を達成する方法としては大きく分けて２つの方法が考えられる．一方は対象数を減らす方法であり，他方はパラメータを推定するためのアルゴリズムを改良する方法である．前者については，先にも述べた通り各対象を対象の上位概念であるコンセプトに集約する代わりに，上位概念のバラつきを保存した Symbolic Data で表現し，分析を行う方法である．後者については「特異値分解」に基づくアルゴリズムと「Majorizing 関数」に基づくアルゴリズムの導出である．古典的な多変量データ解析の多くでは分析手法のパラメータを特異値分解により陽に求めることが可能である．しかし，多くの提案手法

ではパラメータを逐次推定によって行い，かつ制約条件として直交制約を充たす必要性がある．その際，逐次推定を行う度に制約条件を充たしつつ繰り返しの数値計算法を要求することは大規模複雑データへの適用を困難にする．そこで交互最小二乗法の枠組みである一定の条件を充たせば逐次推定でのあるパラメータを特異値分解によって陽に導出できることを用いて，その枠組みでパラメータ推定を行えるように提案手法を開発する．特に次元縮約クラスタリング法で各パラメータを同時最適化する際に活用することを考える．また，Majorizing 関数についても同様に逐次計算において各パラメータの推定を陽に導出できるようになるが，この方法は非負値制約と相性が良いという性質を持つ．

（4）についてはＰＯＳデータやシングルソースデータ，アクセスログデータと呼ばれるマーケティング分野や情報通信技術のデータを主に想定して，提案手法を適用し，実際に応用上問題がないか否かを検討する．
する．必要であれば，提案手法の改良に取り組む．

４．研究成果

本研究では，大規模な関連性データに対して，シンボリックデータ解析を用いる方法，部分空間を用いる方法，次元縮約法と既存手法の同時分析を用いる方法という３つのアプローチによる新たな解析手法を提案した

（1）での研究成果は行列表記に基づく統一的記法による大規模複雑データの分析方法の定式化である．具体的には提案手法を与えられたデータ行列（の関数）をモデルによって近似する定式化に基づいて開発し，既存手法との関係を容易に把握できるようにしたことである．

（2）（3）主な成果としてここでは２つ挙げる．一つめは，分布値非類似性データが与えられた際に，既存手法では同心に基づく分布の形状のみしか表現できなかった分布値非類似性データに対する多次元尺度構成法を同心ではない分布の形状も表現できるような多次元尺度構成法を提案したことである．さらに，Majorizing 関数を導出し，それに基づくアルゴリズムを提案したことによって大規模複雑データに適用可能な方法となった．

２つめは質的多変量データや区間値データに対して，既存手法では解釈不可能であった解釈が行える新たな「次元縮約クラスタリング法」を提案したことである．この方法についても特異値分解に基づくアルゴリズムで推定可能なものであるため，大規模複雑データに適用可能な方法となった．

（4）これらの方法は５で挙げた論文や学会発表で実データに適用し，その有用性を示した．また，データ解析コンペティションなどで実際に適用した際，応用上の観点より一定の評価を得ることができた．

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９件）（すべて査読付き）

[1] Mitsuhiro, M. and Yadohisa, H. (2014): Reduced k-means clustering with MCA in a low-dimensional space, to appear in Computational Statistics. (10.1007/s00180-014-0544-8)
[2] Abe, H., Tanioka, K. and Yadohisa, H. (2014): Clusterwise Linear Regression Model for Modal Multi-Valued Data, Proceedings of the International Conference on Mathmatics, Statistics, and Financial Mathematics 2014, p.35-p.40.
[3] Tamura, K., Hatano, K. and Yadohisa, H. (2012): A Retrieval method based on language model considering neighboring contents, Journal of Digital Information Management, 10(1), p1-p9.
[4] Tanioka, K. and Yadohisa, H. (2012): Subspace hierarchical clustering for three-way three-mode data using quadratic regularization, Procedia Computer Science, 12, p248-p253.
[5] Tanioka, K. and Yadohisa, H. (2012): Effect of data standardization on the result of k-means clustering, Challenges

at the interface of Data Analysis, ComputerScience, and Optimization (W, Gaul et al. eds), Springer, Heidelbarg, p59-p67.
[6] Saito, Y. and Yadohisa, H. (2012): Visualization of asymmetric clustering result with digraph and dendrogram, Challenges at the interface of Data Analysis, Computer Science, and Optimization (W, Gaul et al. eds), Springer, Heidelbarg, p151-p159.
[7] Tanioka, K. and Yadohisa, H. (2012): Three-way asymmetric hierarchical clustering based on regularized similarity model, COMPSTAT 2012: Proceedings in Computational Statistics, p789-p800
[8] Mitsuhiro, M., Ota, Y. and Yadohisa, H. (2012): New mathematical approach for an inverse problem in financial markets, COMPSTAT 2012: Proceedings in Computational Statistics, p585-p594.
[9] Kitano, M. and Yadohisa, H. (2012): An overlapping clustering method for signed graphs, COMPSTAT 2012: Proceedings in Computational Statistics, p391-p402, (Best Paper Award: ERS IASC Young Researchers Award).

〔学会発表〕(計３５件)
[1] 谷岡健資, 宿久洋 (2015): 非負値制約に基づくカテゴリカルデータのクラスタリング法について, 応用統計学会 2015 年度年会 (於 京都大学, 期日：3月１４日).
[2] 土田潤, 宿久洋 (2015): 対数線形モデルを用いたベイジアン Unfolding について, 第 9 回日本統計学会春季集会 (於 明治大学, 期日：3月８日)
[3] Tsuchida, J. and Yadohisa, H. (2014): Two-mode three-way asymmetric MDS using the log linear model, 7th International Conference of the European Research Consortium for Informatics and Mathematics Working Group on Computational and Methodological Statistics 2014, University of Pisa, Italy, December 6.
[4] Abe, H., Tanioka, K. and Yadohisa, H. (2014): Reduced k-means for multivalued quantitative symbolic variables, 7th International Conference of the European Research Consortium for Informatics and Mathematics Working Group on Computational and Methodological Statistics 2014, University of Pisa, Italy, December 6.
[5] 土田潤, 宿久洋(2014): n 相 m 元データに対する対数線形モデルを用いた非対称 MDS について, 計算機統計学会第 28 回シンポジウム(於 沖縄科学技術大学院大学, 期日：１１月１４日).
[6] 長谷川公宏, 谷岡健資, 宿久洋(2014): コンセプト間非類似度における外部展開法について, 日本計算機統計学会第 28 回シンポジウム (於 沖縄科学技術大学院大学, １１月１４日).
[7] 谷岡健資, 宿久洋 (2014): 非対称 (非) 類似度データに対する k-medoids 法と制約付き MDS の同時分析法について, 行動計量学会第 42 回大会 (於 東北大学,期日:９月２日).
[8] 土田潤, 宿久洋 (2014): 対数線形モデルを用いた 2 相 3 元非対称 MDS について, 行動計量学会第 42 回大会 (於 東北大学, 期日：９月２日).
[9] 有重文平, 宿久洋(2014): 2 変量群の関係を考慮した主成分分析と解のスパース推定, 行動計量学会第 42 回大会 (於 東北大学, 期日：９月２日).
[10] 阿部寛康, 谷岡健資, 宿久洋 (2014): 連続値の多値シンボリック変数に対する RKM 法について, 行動計量学会第 42 回大会(於 東北大学, 期日：９月２日).
[11] Tsuchida, J. and Yadohisa, H. (2014): Partial least squares logistic regression using F-measure, COMPSTAT 2014, Geneva, August 19.
[12] Tanioka, K. and Yadohisa, H. (2014): K-mode clustering with dimensional reduction for categorical data, European Conference on Data Analysis 2014, Bremen, Germany, July 2.
[13] Takagi, I., Tanioka, K. and Yadohisa, H. (2014): Constrained VPCA for interval-valued data, 2014 Workshop in Symbolic Data Analysis, Taipei, Taiwan, June 14.
[14] 谷岡健資, 宿久洋 (2014): 正則化に基づく次元縮約を伴う k-mode 法について, 日本計算機統計学会第 28 回大会, (於 中央大学, 期日：５月１７日).
[15] 高岸茉莉子, 谷岡健資, 宿久洋 (2014): 雑音を考慮した独立成分分析混合モデルについて, 日本計算機統計学会第 28 回大会, (於 中央大学, 期日：５月１７日).
[16] 土田潤, 宿久洋 (2014): F-measure を誤差関数とする潜在クラス 2 項ロジットモデル, 日本分類学会第 32 回大会, (於 首都大学東京秋葉原サテライトキャンパス, 期日：３月１日).

[17] 浅野祐介, 谷岡健資, 宿久洋 (2013): カテゴリカル 3 相 3 元データの分析法について, 日本計算機統計学会 第 27 回シンポジウム講演論文集 p31-34 (於 熊本市民会館, 期日：１１月１５日).
[18] 有重文平, 宿久洋 (2013): 多変量因子回帰分析法の提案, 日本計算機統計学会第 27 回シンポジウム講演論文集 p23-26, (於 熊本市民会館, 期日：１１月１５日).
[19] 谷岡健資, 宿久洋 (2013): 3 元データ

分析のための additive tree 推定と視覚化について，2013 年度統計関連学会連合大会(於 大阪大学，期日：９月８日)．
[20] 山下陽司，宿久洋 (2013)：経時的に得られた非対称非類似度データに対するトレンドとばらつきを考慮した多次元尺度構成法について，第４１回日本行動計量学会(於 東邦大学，期日：９月３日)．
[21] Mitsuhiro, M. and Yadohisa, H. (2013): Simultaneous Fuzzy Clustering with Multiple Correspondence Analysis, the 59th World Statistics Congress of the ISI, Hong Kong, China, August 25.
[22] Terada, Y. and Yadohisa, H. (2013): Reparameterization of Percentile MDS with the non-concentric hyperboxes, Joint Meeting of the IASC Satellite Conference for the 59th ISI WSC and the 8th Asian Regional Section(ARS) of the IASC, Seoul, Korea, August 22.
[23] 光廣正基，宿久洋 (2013)：大学生協食堂の POS データ解析　―ミールカード利用者の食の実態把握をめざして―，2013 PC カンファレンス(於 東京大学，期日：８月３日)．
[24] Yamashita, Y. and Yadohisa, H. (2013): MDS for series data by using candlestick valued dissimilarity measure, the 78th annual meeting of the psychometric society, Arnhem, the Netherlands, July 22.
[25] Umei, T. and Yadohisa, H. (2013): Non-hierarchical clustering algorithm for mixed numerical and categorical three-way three-mode data, International conference of the International Federation of Clasifidation Societies, Tilburg, the Netherlands, July 14.
[26] Tanioka, K. and Yadohisa, H. (2013): Ultrametric tree representation for three-way three-mode data with weights of variables and occasions, International conference of the International Federation of Clasifidation Societies, Tilburg, the Netherlands, July 14.
[27] Mitsuhiro, M. and Yadohisa, H. (2013): Multiple Correspondence Analysis for Mixed Measurement Level Data, European Conference on Data Analysis 2013, Luxembourg, Luxembourg, July 10.
[28] 梅井隆弘，宿久洋 (2013)：量質混在 3 相 3 元データに対する非階層クラスタリング法，日本計算機統計学会 第 27 回大会講演論文集 p105-108 (於 弘前大学，期日：５月１６日)．
[29] 光廣正基，宿久洋 (2013)：対象の分類を伴う多重対応分析法，日本計算機統計学会 第 27 回大会講演論文集 p51-54 (於 弘前大学，期日：５月１６日)．
[30] 山下陽司，宿久洋 (2013)：系列データに対するトレンドとばらつきを考慮した多次元尺度構成法について，日本計算機統計学会 第 27 回大会講演論文集 p35-38 (於 弘前大学，期日：５月１６日)．
[31] 谷岡健資，宿久洋 (2013)：3 相 3 元データに対する変量および条件の重みを考慮した階層的クラスタリング法について，日本分類学会第３１回大会(於 中央大学，期日：３月５日)．
[32] 光廣正基，谷岡健資，宿久洋 (2012)：視覚化法を用いた野球配球の特徴把握　～多元データを用いたアプローチ～，日本計算機統計学会 第 26 回シンポジウム講演論文集，p31-32 (於 東京大学，期日１１月１日)．
[33] 谷岡健資，宿久洋 (2012)：単相３元非対称（非）類似度データに対するクラスタリング法について，日本行動計量学会 第 40 回大会発表論文抄録集，p469-470 (於 新潟県立大学，期日：９月１３日)．
[34] 北野道春，宿久洋 (2012)：符号付有向グラフを用いた非対称データの視覚化，日本行動計量学会 第 40 回大会発表論文抄録集，p281-282 (於 新潟県立大学，期日：９月１３日)．
[35] 谷岡健資，宿久洋 (2012)：正則化に基づく 3 相 3 元部分空間階層的クラスタリング法とその解析結果の視覚化法，2012 年度統計関連学会連合大会講演報告集，p230 (於 北海道大学，期日９月９日)．

６．研究組織

(1)研究代表者
宿久　洋（YADOHISA, Hiroshi）
同志社大学・文化情報学部・教授
研究者番号：５０２４４２２３

(2)研究分担者
波多野賢治（HATANO, Kenji）
同志社大学・文化情報学部・准教授
研究者番号：８０３１４５３２

深川大路（FUKAGAWA, Daiji）
同志社大学・文化情報学部・助教
研究者番号：１０４４２５１８