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２４年　５月２８日現在

機関番号：34310
研究種目：基盤研究（C）
研究期間：2009～2011
課題番号：21500284
研究課題名（和文）　関連性データ解析に関する数理的研究

研究課題名（英文）　Mathematical approach for reference data analysis

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

研究成果の概要（和文）：

本研究では，様々な関連性データの解析法について，既存の方法の体系化を行った．さらに，新たな解析法および解析結果の評価法の提案を行った．主な研究成果は，大規模データに対応した多次元尺度構成法，クラスター分析法およびネットワーク分析法を提案したことである．ここでは，大規模データをシンボリックデータ（多値，区間値，分布値などの値をもつデータ）としてとらえ，データの縮約による情報損失を抑えながら，解析するためのいくつかの方法を提案した．

研究成果の概要（英文）：

In this study, we categorized various types of existing methods for relational data and proposed new methods for analysis of the data and evaluation of the results. As a main result, new methods for multidimensional scaling, clustering algorithm and network analysis have been proposed to deal with the large and complex data. More precisely, the large and complex data are regarded as symbolic data, which consists of multi-valued data, interval-valued data and distribution-valued data, and several methods that reduce the loss of information have been proposed based on the notion of symbolic data.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
２００９年度	1,500,000	450,000	1,950,000
２０１０年度	1,000,000	300,000	1,300,000
２０１１年度	900,000	270,000	1,170,000
総　計	3,400,000	1,020,000	4,420,000

研究分野：総合領域
科研費の分科・細目：情報学・統計科学
キーワード：クラスター分析，多次元尺度構成法，シンボリックデータ，大規模複雑データ

１．研究開始当初の背景

社会現象や心理現象を扱う様々な学問分野あるいは応用分野で，２者関係を示すデータは多種多様に存在し，それらを分析する需

要は多い．２つの対象が心理的に似ているほど，両者を間違えたり，混同したりしやすいし，また一者から他者を連想しやすい．２者間の類似性，混同率，連関性，心理的距離を示すデータは，総称として「関連性データ」（または「類似性データ」）と呼ばれる．社会心理学では，集団の成員間の相互作用や親近性，また社会的資源（地位，権力，金銭など）の交換を分析するが，これらも関連性データである．社会学では，世代間の職業移動や地域間の移住，通婚圏などの社会移動を分析したり，ネットワーク（個人間，組織間，国家間）を扱うが，いづれも２者間の社会的距離を反映すると考えられ，関連性データとみなせる．市場調査で扱うブランド交換や商品代替，生態学では２種の植物の同時繁茂率，経営工学では２要因の交互作用，情報検索学では雑誌間の相互引用などを分析するが，それらも関連性データとみなせる．

このように，様々な分野で取り扱われる関連性データであるが，分野によってはデータの持つ性質が全く異なる．当然ながら，量的な場合もあれば質的な場合もあるし，対称な場合もあれば非対称な場合もある．さらに最近では，多元であるもの，多相であるもの，非常にスパースなもの等，複雑データに対する解析法の需要や，非常に大きなサイズのデータ（大規模データ）に対する解析法の需要がでてきている．

このような流れの中で，様々な種類の関連性データを解析するために，クラスター分析法や多次元尺度構成法をはじめ，多変量データ解析の手法の開発提案が行われている．しかしながら，古典的な手法の多くは取り扱うデータに様々な仮定をおいている．例えば，量的であること，対称であることなどである．近年では，より複雑なデータに適用するための手法の提案がなされているが，一般の応用分野に普及しているとは言いがたく，さらに，最近の応用分野の解析ニーズに答えているとは言いがたい状況である．

２．研究の目的

本研究では，

1. 様々な関連性データ解析法の理論的特徴付け
2. 既存の関連性データ解析法の総合的な調査
3. 関連性データの新解析法の提案
4. 関連性データの解析結果の評価法の提案
5. 新たな応用分野への手法の適用可能性の検討

を目的として研究を行った．

既述のように，関連性データは通常の（サンプルを変量ごとに観測した）多変量データと異なり様々な（取り扱いし辛い）性質を持つ．このことは，研究対象としての興味深さはあるものの，どうしても個別に特化した研究が多くなり，研究の体系的な整理や一般の応用分野への普及は困難なものとなる．本研究では様々な関連性データに関して，その解析手法を整理・体系化し，さらに新たなタイプのデータへ適応可能な手法の開発を試みたい．本研究で特に着目した関連性データの特長としては，「質的・量的データ」，「対称・非対称データ」，「多元データ」，「多相データ」，「シンボリックデータ」である．

研究代表者は，一貫して関連性データの解析法に関する研究に取り組んできた．その成果は，「Data Analysis of Asymmetric Structures - Advanced Study of Computational Statistics (Marcel Dekker Inc., New York)」としてまとめている．この本は，関連性データの非対称性に着目し，研究代表者らが取り組んできた非対称データを扱う解析法をはじめ様々な解析法をまとめている．また，「関連性データの解析法－多次元尺度構成法とクラスター分析法（共立出版，東京)」は関連性データの解析手法である多次元尺度構成法，クラスター分析法についての一般的な事項をまとめたものである．これらの本をまとめるにあたり，既存の類似性データ解析法の問題点，制限事項などを再確認し，これらの手法の拡張の必要性を強く感じたことが本研究を行う１つの動機となっている．

本研究では，今までの研究成果を踏まえ，上記の様々な観点から既存の手法を整理体系化するとともに，今までは適用できなかったタイプのデータ（特に大規模かつ複雑な性質をもつデータ）へ適応可能な新手法の開発を行った．

３．研究の方法

本研究では，以下の課題に取り組んだ．

1. 様々な関連性データの理論的特徴付け

関連性データには様々な種類があり，多くの応用分野で用いられている．そこで，それらのデータについて，数理的に特徴付けるとともに，利用可能な解析法について整理行った．

2. 既存の関連性データ解析法の総合的な調査及び体系化

関連性データの解析法について，既に様々な結果が公表されている．そこで，文献を収集し，既存の手法について体系的に取りまとめる．特に，着目した手法は，「多次元尺度構成法」，「クラスター分析法」，「ネットワーク分析法」である．

3. 関連性データ解析の新規手法の提案

研究代表者が取り組んできた研究を踏まえ，他のタイプの関連性データ解析法の拡張，精密化，一般化による実用化に取り組んだ．

4. 関連性データ解析の解析結果の評価法の調査
既存の評価法について，文献を収集し整理を行った．

5. 感性メディアデータ，Ｗｅｂデータ，テキストデータからの特徴抽出
感性メディアデータ，Ｗｅｂデータやテキストデータのもつ情報から，関連性データの作成を試みる．既存の手法の適用可能性を確認し，必要であれば既存の手法の改良に取り組む．

４．研究成果

多次元尺度構成法，クラスター分析法，ネットワーク分析法それぞれについて，既存の解析手法を拡張した新たな解析手法の提案を行った．以下に，その成果を解析手法毎にまとめる．

多次元尺度構成法:

パーセンタイル値とよばれるシンボリックデータに対する，超球モデルを仮定した多次元尺度法と超直方体モデルを仮定した多次元尺度法を新たに提案した．区間値に対する多次元尺度構成法は，Denoeux and Masson (2000)やGroenen, et al(2006)で提案されていたが，区間値は集団の最大値と最小値を用いる為，外れ値や誤差の影響を受けやすい．そのため，パーセンタイル値を用い，各パーセント点を入れ子状に表現するモデルを仮定することで，その問題点を解消した．

クラスター分析法:

データサイズが大規模な類似性データが与えられた際に，従来のクラスタリング法ではクラスターが超球状を仮定しており，任意の形状のクラスターは検知できない．特に多変量データのように座標が与えられているデータに対しては kernel 法などを用いて検知することが可能であるが，類似性データは座標を定義できないため困難な問題となる．Ester, et al.(1996), Guha, et al.(1999), Karypis, et al.(1999)などでは任意の形状であるクラスターを検知する方法を提案している．特にシミュレーションにおいてKarypis, et al.の手法は任意の形状を検知することに優れていることが知られているが，群平均法を基にした手法であるため，単連結法で検知できるような形状のクラスターは検知することができない．そこで，そのような問題点を指摘し，(非対称)類似性データが与えられた際に，Guha, et al.(1998)で用いられているサンプリング法及び Karypis, et al.で採用されているバラツキの測度を用いて Karypis,et al. の手法では検知できない形状であるクラスターを検知する階層的クラスタリング法を提案した．本手法は外れ値の影響も考慮することができ，単連結法の欠点である空間濃縮の問題を抑えることができる．

ネットワーク分析:

複雑なネットワークを表現する一つの方法として，符号付グラフがある．符号付グラフは，各辺に正または負の符号が付与されたグラフであり，対象間の positive, negative または nonadjacent の関係性を表現することができる．そして，符号付グラフを用いた分析手法のひとつとして，コミュニティ（集合間に negative な関係が，集合内に positive な関係が相対的に多い対象の集合）の検出があるが，ある組織の友好・敵対関係両方を考慮したコミュニティを検出する方法，つまり，符号付きグラフに対するコミュニティの検出法の研究は進んでいない．そこで，符号付きグラフに対する，新しいコミュニティの検出法を提案した．この提案手法は，単語の極性を考慮したテキストマイニング等，多くの分野での適用が考えられ，実際に新聞記事のテキストデータに対する適用も行い，有用な結果がえられた．

５．主な発表論文等

〔雑誌論文〕（計７件）

1. Takagi, I. and Yadohisa, H., Correspondence analysis for symbolic contingency tables based on interval algebra, Procedia Computer Science, 6, (2011), 352-357. (査読有)
2. Tamura, K., Hatano, K. and Yadohisa, H., Calculating query likelihoods based on web data analysis, Intelligent Decision Technologies, 10, (2011), 707-718. (査読有)
3. Tanioka, K. and Yadohisa, H., Hierarchical clustering algorithm with combined criteria for large and complex similarity data, International Journal of Knowledge Engineering and Soft Data Paradigms, 3(2), (2011), 121-131. (査読有)
4. Terada, Y. and Yadohisa, H., Multidimensional scaling with the nested hypersphere model for percentile dissimilarities, Procedia Computer Science, 6, (2011), 364-369. (査読有)

5. Terada, Y. and Yadohisa, H., Multidimensional scaling with hyperbox model for percentile dissimilarities, Intelligent Decision Thechnologies, 10, (2011), 779-788. (査読有)
6. Tamura, K., Hatano, K. and Yadohisa, H., A characterizing method of web pages based on query likelihood of neighbor pages, Proceedings of the 5th International Conference on Digital Information Management (ICDIM 2010), (2010), 392-397. (査読有)
7. Terada, Y. and Yadohisa, H., Non-hierarchical clustering for distribution-valued data. COMPSTAT 2010: Proceedings in Computational Statistics, (2010), 1653-1660. (査読有)

〔学会発表〕(計 40 件)

1. Takagi, I. and Yadohisa H., A feature extraction method for mixed feature-type symbolic data, Joint Meeting of the Taipei International Statistical Symposium and 7th Conference of the Asian Regional Section of the IASC (招待講演), December 17, 2011, Taipei, Taiwan.
2. Kitano, M. and Yadohisa, H., Text mining with extraction of similar expression patterns by using signed bipartite graph, Joint Meeting of the Taipei International Statistical Symposium and 7th Conference of the Asian Regional Section of the IASC, December 16, 2011, Taipei, Taiwan.
3. Ishii, R. Kitano, M. and Yadohisa, H., Characterization and comparison of Japan Professional Football Clubs based on attack patterns, Joint Meeting of the Korea-Japan Conference of Computational Statistics and the 25th Symposium of Japanese Society of Computational Statistics, November 12, 2011, Busan, South Korea.
4. Koike, K. Abe, H. and Yadohisa, H., Pitching prediction by multinomial logit model in Nippon Professional Baseball, Joint Meeting of the Korea-Japan Conference of Computational Statistics and the 25th Symposium of Japanese Society of Computational Statistics, November 12, 2011, Busan, South Korea.
5. Yamashita, Y. and Yadohisa, H., Similarity measure for candlestick chart variable, Joint Meeting of the Korea-Japan Conference of Computational Statistics and the 25th Symposium of Japanese Society of Computational Statistics, November 11, 2011, Busan, South Korea.
6. Abe, H., Terada, Y. and Yadohisa, H., Measuring an equivalence of purchase intervals using a revised Gini index, The 58th World Statistics Congress of the International Statistical Institute, August 25, 2011, Dublin, Ireland.
7. Terada, Y. and Yadohisa, H., Modal interval-valued dissimilarity between histogram- valued data. 3rd German-Japanese Workshop "ADVANCES IN DATA ANALYSIS AND RELATED NEW TECHNIQUES AND APPLICATIONS", July 20, 2010, Karlsruhe, German.
8. Saito, Y. and Yadohisa, H., Visualization of asymmetric clustering result with digraph and dendrogram, GfKl 2010, July 22, 2010, Karlsruhe, German.
9. Tanioka, K. and Yadohisa, H., Effect of data standardization on the result of k-means clustering, GfKl 2010, July 22, 2010, Karlsruhe, German.
10. Terada, Y. and Yadohisa, H., Kernel methods for analyzing symbolic data. GfKl 2010, July 23, 2010, Karlsruhe, German.
11. Terada, Y. and Yadohisa, H., Principle component analysis for histogram-valued data, The 10th China-Japan Symposium on Statistics, October 16, 2010, China.

〔図書〕(計 2 件)

1. 宿久洋, 村上亨, 原恭彦, 確率と統計の基礎 1 増補版, 2011 年, ミネルヴァ書房, 京都.
2. 大森崇, 阪田真己子, 宿久洋, R Commander によるデータ解析, 2011 年, 共立出版株式会社, 東京.

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