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　　６　月　１６　日現在

機関番号：　１２１０２
研究種目：　基盤研究(C)（一般）
研究期間：　2014～2016
課題番号：　２６３３００３３

研究課題名（和文）多次元クラスター尺度構成法によるビックデータ解析とその社会的応用

研究課題名（英文）Big data analytics by multidimensional cluster scaling and its social applications

研究代表者

イリチュ　美佳（佐藤美佳）（SATO-ILIC, Mika）

筑波大学・システム情報系・教授

研究者番号：６０２６９２１４

交付決定額（研究期間全体）：（直接経費）　　3,700,000円

研究成果の概要（和文）：ビックデータ解析に対応する新たな手法として多次元クラスター尺度構成法を開発し、開発した手法の実用化に向けた性能評価を行った。ビックデータの解析の主流をなす方法論は、説明力の乏しいデータを取り除き、データを縮小して、従来型の解析法を適用するという方法論であるが、この方法論では、何をもって説明力がないとするかということの基準によって解析の結果が異なるため、結果の妥当性が明確でないという問題がある。そこで、本研究では、ビックデータそのものの情報はすべて用いるが、データを分類構造という別の尺度で測定される空間で解析する多次元クラスター尺度構成法を提案した。

研究成果の概要（英文）：Multidimensional cluster scaling is developed as a novel method for big data analytics with an evaluation of its performance. The mainstream of the methodology for big data analytics includes the excluding data which has poor explainable power, reducing the data, and applying ordinary analytics. This methodology has a problem in relation to the validity of the result since the result depends on the criterion which determines the explainable power of the data. Therefore, in this study, the multidimensional cluster scaling was proposed in which all of the data information of the big data was used, and the data was analyzed in another space measured by another scale of the classification structure.

研究分野：　統計科学

キーワード：　分類　ビックデータ　尺度構成

１．研究開始当初の背景

近年、大量データの解析が話題を集めている。これは、従来、想定されていなかった程の大量のデータの解析手法とその一連の解決法のことであるが、複雑で大量のデータの解析法の開発は、情報科学の分野のみならず、統計的解析法を扱う統計科学の分野においても重要な問題である。その主たる原因は、複雑で大量のデータ解析の問題は、情報科学の視点から見た“計算機能力が伴わず実質上計算不能であるという問題”のみではなく、統計科学の視点から見た“数理的に従来の統計科学に基づく方法が利用不可能であることが理論的に解明されてきている”ことにある。このような背景の下、統計科学の分野において、ビックデータに対する新たな解析法の確立が急務とされている。

ビックデータの解析には、大きく分けて二つの方法論がある。その一つは、データの中で説明力の乏しいデータや、本来、測定値としての意味をなさない（ノイズとして扱われるべき）データを如何に取り除き、ビックデータを出来るだけ縮小して、従来型のデータ解析法を適用するという方法論であり、他の一つは、データを類似するもの同士にまとめて解析をする方法論である。従来型の多くの統計科学分野のビックデータに対する方法は、損失関数やフィルターの利用など、前者の方法論に沿った方法である。しかし、この方法論では、何をもって説明力がないとするか、あるいは何をもって本来の測定値ではないとするのかということの基準によって解析の結果が異なるため、結果の妥当性が明確でないという問題がある。

そこで、本研究では、ビックデータそのものの情報はすべて用いるが、データを分類構造という別の尺度で測定される空間に変換する“多次元クラスター尺度構成法”を開発し、統計科学的解析を可能とする方法を提案しようとするもので、その位置づけは、後者の方法論の枠組みの中で、新たなビックデータの解析手法を開発しようとするものである。

２．研究の目的

本研究の目的は、ビックデータ解析に対応する新たな手法として「多次元クラスター尺度構成法」を開発することである。

クラスターとは、データを分類した結果、得られるデータのグループのことであるが、従来、多次元データから得られたクラスターを用いて元のデータと融合することによりデータの説明力を高める方法は提案されていた。

しかし、これらの方法では、より精度の高い結果が得られる反面、次元数が大量となるビックデータでは計算量が膨大となることが最大の問題であった。すなわち、ビックデータでは、多次元データについて尺度を構成することそのものが困難な問題であり、これが精度の劣る解を得る最大の原因であった。

そこで、多次元データについて尺度構成をするのではなく、多次元のクラスターについて尺度構成するという考えに思い当たった。この考えに基づいた方法が、多次元クラスター尺度構成法である。この手法は、データの分類とその結果得られたクラスター（グループ）の尺度構成を同時に行う手法である。

この方法により、多次元データの尺度構成に困難があったビックデータの尺度化が可能となり、従来の統計的解析法をそのまま適用することが可能となる。

３．研究の方法

ビックデータに対する多次元クラスター尺度構成法を開発し、実用化に向けて、開発した手法の各種性能を精査した。ビックデータでは尺度化が困難であるという問題を解決するために、平成２６年度は、得られたクラスターについて尺度化を行う方法として、

多次元クラスター尺度構成法を提案し、その性能を評価した。通常、データから得られたクラスターは数学的尺度を持たない。これが、分類結果を利用した解析を困難とする要因であった。提案手法では、得られたクラスターに、ユークリッド空間の次元を数学的に付与することにより、クラスターを尺度として用いる事を可能とした。

次に、平成２７年度は、開発した手法の社会的データに対する応用とその評価を行った。

さらに、平成２８年度においては、平成２６年度、２７年度に開発した方法を拡張し、データが個人や時点毎に得られている場合、個人や時点間の相違性をも可視化する多次元クラスター尺度構成法を開発した。

４．研究成果

（１）研究の主な成果

本研究の具体的成果を要約すれば、ビックデータに対するデータ構造の可視化を目的として、データをグループ（クラスター）にまとめることにより生じた分類構造を尺度としてあつかい、これを有効に利用し、個体の分類と得られたクラスターの尺度構成を同時に行う方法を開発したことである。また、この手法を拡張し、個人や時点間の相違性をも可視化する多次元クラスター尺度構成法を開発したことである。

多次元や大量のデータ構造を低次元で可視化することは、大量のデータ情報を失うことになり、一般的に困難な問題である。

本研究で提案した一連の方法の特徴は、データの潜在構造として得られたクラスターを、データそのものを測る尺度として用いるための尺度構成法を開発したことであり、これにより、多次元や大量データの潜在構造を低次元空間で可視化することを可能とした事である。また、個体とクラスターや、時点の同時尺度化により、個体間・時点間の相違性の解釈を、それらを通じて同一の尺度であるクラスターを介在して計量することにより、可能にした。

また、これらの手法と従来の同時可視化手法との比較研究から、提案手法の理論的優位性を示した。

（２）得られた成果の国内外におけるインパクト

平成２６年度、平成２７年度の２年間の研究成果に対して、米国、フィラデルフィア、サンノゼにおいて開催された２度の国際会議で２年連続の下記の学術的賞を受賞した。

・1st Runner up Theoretical Paper Award, M. Sato-Ilic, Multidimensional Joint Scale and Cluster Analysis, Procedia Computer Sciences, Elsevier, Vol. 61, pp. 11-17, San Jose, CA, USA, 2015.
・1st Runner up Theoretical Paper Award, M. Sato-Ilic and P. Ilic, On A Multidimensional Cluster Scaling, Procedia Computer Sciences, Elsevier, Vol. 36, pp. 278-284, Philadelphia, USA, 2014.

また、国際会議CAS2014で基調講演を行い、パリ大学で招待講演を行った。その詳細は、下記の通りである。

・Clustering Innovations in Data Science, Plenary speech at Complex Adaptive Systems 2014 Conference (CAS2014), Philadelphia, USA, 2014.
・Clustering-based Models from Model-based Clustering, Invited talk at Department of Databases and Machine Learning, LIP6, University of Paris (UPMC), Paris, France in cooperation with the France chapter of the IEEE Computational Intelligence Society, 2014.

さらに、平成２８年度においては、平成２６年度、平成２７年度に開発した方法の社会的データに対する応用と成果報告に力を入れるとともに、個人や時点間の相違性をも可視化する手法に拡張した。これら一連の研究成果については、SOFA2016 国際会議で基調講演を行い、その研究成果について JANOS FODOR Award を受賞した。その詳細は次の通りである。

・JANOS FODOR Award, Soft Data Analysis based on Cluster Scaling, Keynote speech at The 7th International Workshop on Soft Computing Applications, Arad, Romania, 2016.

さらに、３件の章の執筆や、国内外の学会における招待講演も行った。また、研究結果を論文にまとめた。

（３）今後の展望

本研究では、ビックデータの潜在構造を、説明力の乏しいデータを削除することなく、すべて用いて、低次元空間で説明する手法の開発を目的として、データの潜在的特性として得られるクラスターを、そのデータそのものの潜在構造を測るための尺度として構成するための手法を提案した。今後の展望は、この方法の数学的性質を精査すると共に、解の性能向上を目指し、より高次の計量を用いて、尺度構成を行うことである。その一方で、パラメータの設定等に対する結果の変動特性等のシミュレーションに基づく細部の検証が必要である。さらに、ビックデータに対する応用を進める必要がある。

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 24 件）

① M. Sato-Ilic, Individual Compositional Cluster Analysis, Procedia Computer Science, Elsevier, Vol. 95, pp. 254 - 263, 査読有, 2016
② M. Sato-Ilic, Multidimensional Joint Scale and Cluster Analysis, Procedia Computer Sciences, Elsevier, Vol. 61, pp. 11-17, 査読有, 2015 (1st Runner up Theoretical Paper Award 受賞)
③ M. Sato-Ilic, P. Ilic, On A Multidimensional Cluster Scaling, Procedia Computer Sciences, Elsevier, Vol. 36, pp. 278-284, 査読有, 2014 (1st Runner up Theoretical Paper Award 受賞)

〔学会発表〕（計 30 件）

① M. Sato-Ilic, Soft Data Analysis Based on Cluster Scaling, Soft Computing Applications (SOFA2016)（基調講演）, 2016 年 8 月 25 日, Hotel Continental(会議場), アラド, ルーマニア, (JANOS FODOR Award 受賞)
② M. Sato-Ilic, Clustering Innovations in Data Science, Complex Adaptive Systems 2014 Conference (CAS2014)（基調講演）, 2014 年 11 月 4 日, the DoubleTree by Hilton Philadelphia-Valley Forge（会議場), フィラデルフィア, 米国
③ M. Sato-Ilic, Clustering-based Models from Model-based Clustering, Invited talk at Department of Databases and Machine Learning, LIP6, University of Paris (UPMC)（招待講演）, 2014 年 9 月 11 日, パリ大学, パリ, フランス（the IEEE Computational Intelligence Society フランス支部共催)

６．研究組織

(1)研究代表者
イリチュ　美佳（佐藤　美佳）（SATO-ILIC, Mika）
筑波大学・システム情報系・教授
研究者番号：60269214

(2)研究分担者
青嶋　誠（AOSHIMA, Makoto）
筑波大学・数理物質系・教授
研究者番号： 90246679
清水　信夫（SHIMIZU, Nobuo）
統計数理研究所・データ科学研究系・助教
研究者番号：00332130

(3)研究協力者
Christophe Marsala (MARSALA, Christophe)
Department of Databases and Machine Learning, LIP6, University of Paris (UPMC), 教授