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

令和　２ 年　 ５ 月 ３１ 日現在

機関番号：１２１０２
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００００４３

研究課題名（和文）高次計量による高次元小標本型ビックデータ解析とその社会的応用

研究課題名（英文）High-Dimension Low-Sample-Size Big Data Analysis by Higher Order Metrics

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）　　3,500,000円

研究成果の概要（和文）：高次元小標本型ビックデータの解析には、従来の統計科学に基づく手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。そこで、本研究では、このデータの解析法として、複数のデータを一度に計測し得る共通計量空間における適切な計量を開発するとともに、それに基づくクラスター計量モデルを開発した。さらに、開発した手法の各種性能を精査し、各種のデータに適用することにより、その適用可能性を評価した。

研究成果の学術的意義や社会的意義
一般に、高次元小標本型データが複数得られた場合、典型的な高次元小標本型ビックデータとなる。これらの複数のデータを同時に解析するための新たな方法の開発に取り組んだ。具体的には、複数のデータを通じて共通に得られるクラスターを共通尺度とする計量とそれを利用したモデルの開発を行った。これにより、高次元のデータの動的変動をより低次元の空間で説明することが可能となった。また、共通の部分ベクトル空間への射影を用いることで、種々のデータ構造を比較可能とし、かつ低次元空間に縮約可能とするモデルの開発も行った。この方法は、ビックデータ解析で問題とされる種々のデータの融合法としても有効であることを示した。

研究成果の概要（英文）：It has been theoretically clarified that conventional statistical science-based methods cannot be used to analyze high-dimensional small sample size big data. Therefore, in this research, for the analysis of this data, we developed an appropriate metric in a common metric space that can measure multiple data at once, and developed a cluster metric model based on it. Furthermore, we evaluated the various performances of the developed method, applied it to various data, and evaluated its applicability.

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

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

世界は、今、多種・多様な大量データを生んでいる。しかも、これらのデータは、時々刻々と変化するため、その変化に対応した解析手法の開発が喫緊の課題である。この状況に伴い、これまで､解析不可能とされていたデータの型についても､解析を可能とすることが求められている。その典型的な型の例は､高次元小標本型のデータである｡高次元小標本型のデータとは､変数（次元）の数が個体（標本）の数に比べて、はるかに大きいデータを意味する。高次元小標本データの解析が重要とされる主たる原因は、この問題が､“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。

一般に、多変量データの分析において､“鍵”となることは、多変数間の相関に基づいて、如何に個体間の類似性を測るかということであり、そのための“計量”が重要であるが、高次元小標本データに対しては、それに適した計量が未だに開発されていない。すなわち、高次元小標本データでは、個体が存在する空間の次元があまりにも大きく、既存の計量では、十分な精度をもって個体間の類似性を測ることが出来ない。そこで、まず、高次元小標本データに適した計量を開発することが先決であり、それが成功すれば、これまでの多変量データ解析法の基盤となる理論をそのまま用いて、高次元小標本データを解析することが出来ると考えた。特に、昨今のインターネット社会において、時々刻々と変化する時系列データをリアルタイムで処理する情報技術の創製が望まれているが、通常の時系列解析とは異なり、このデータを、時点を通じて、同じ共通計量空間で計測することで､計算時間の大幅な削減､スケールの動的変動や強いノイズに対する安定性等の汎化能力向上が期待できると考えた。

２．研究の目的

上記の研究背景の下､本研究の目的は、高次元小標本データに適した計量の開発に当たり、まず、当該データ解析の問題の根源である大量のデータを、少数のグループ（クラスター）にまとめ、その分類構造を尺度とし、さらに、まとめることにより失ったデータ情報を、適切な共通計量空間で補完することにより、当該データに適した新たな計量を開発することである。また、この計量に基づくクラスター計量モデルを開発し、開発した手法の各種性能を精査するとともに、各種のデータに適用してその適用可能性を評価することである。

３．研究の方法

上記の目的達成のために、次の方法により研究を進めた。

（１）複数のデータが得られた時に、それらのデータに共通にある潜在的尺度をクラスター構造と射影子の両面から捉え、それにより複数のデータ間の変動を説明し、適切な低次元空間に要約する方法の基礎となる方法論の開発を行う。

（２）異なる性質のデータの融合とクラスターの同定を目的として、それに適用可能なクラスター尺度に基づく計量とモデルの開発を行う。特に、異なる性質の複数のデータの性質を数理的に考慮した上で、データに共通にある潜在的尺度をクラスター構造から捉え、それにより複数のデータを通じてクラスターを同定し得る方法を開発する。

（３）確率的計量を用いた解析法や、複数のデータの分類構造の相違性を求める方法論を開発する。これらの研究は、共通の部分ベクトル空間への射影を用いることで、複数のデータ構造を比較可能とし、かつ低次元空間に縮約可能とする考えに基づき、確率的計量の動的変化量の抽出可能性を探る。

（４）上記の方法に基づき、シミュレーション・実データに適用し、その性能を精査する。

４．研究成果

（１）　研究の主な成果

まず､高次元小標本型データが複数得られた時に､それらを同時に解析するための新たな方法の開発に取り組んだ｡具体的には､複数のデータを通じて共通に得られるクラスターを共通尺度とする計量とそれを利用したモデルの開発を行った。これにより､高次元のデータの動的変動をより低次元の空間で説明することが可能となった。また､共通の部分ベクトル空間への射影を用いることで､種々のデータ構造を比較可能とし､かつ低次元空間に縮約可能とするモデルの開発

も行った。この方法は、ビックデータ解析で問題とされる種々のデータの融合法としても有効であることを示した。
次に、質の異なるデータが複数得られた場合の計量を、分類構造を用いることにより、適切に融合し、複数のデータを通じて分類結果を同定する方法の開発を行った。この研究は、データの性質を考慮した計量をクラスタリング結果から得られた分類構造を用いて定義し、データを測りなおす（再計量する）ことで、適切な融合を図る手法である。数値例より、一定の妥当性を示した。この融合法はビックデータ解析で問題とされる種々のデータの融合法としても有効であると考えられる。また、モデルに基づくクラスタリングとクラスタリングに基づくモデルの二つの概念を定義し、ファジィクラスタリングモデルが類似度データの多様性に対応するように解空間の多様性を取り入れている点や、解の精度向上を目指してクラスターを異なるデータの共通尺度とする多次元クラスター尺度構成法の位置付けを明確にする研究を進めた。
さらに、確率的計量を用いた解析法や、複数のデータの分類構造の相違性を求める方法論の開発に当たった。これらの研究は、共通の部分ベクトル空間への射影を用いることで、複数のデータ構造を比較可能とし、かつ低次元空間に縮約可能とする考えに基づき、確率的計量の動的変化量の抽出可能性を探るものである。数値例より、一定の妥当性を示した。この方法は、ビックデータ解析で問題とされる種々のデータの融合法としても有効であると考えられるが、ビックデータの特性として問題視されているデータの複雑性を念頭に、モデルの汎化性を高めることを目的として、基礎的理論を研究したものである。

（２）得られた成果の国内外における位置づけとインパクト

まず、平成２９年度の研究成果を、米国、シカゴにて開催された CAS2017 国際会議で発表し、発表した論文"Knowledge-based Comparable Predicted Values in Regression Analysis"に対して、1st Runner-Up Theoretical Paper Award を受賞した。さらに、同国際会議にて、"Modeling New Complex Data Structures”と題して基調講演を行うと共に、The 2017 IEEE International Conference on Fuzzy Systems や CFE-CMStatistics2017 での招待研究発表、2017 年度統計関連学会連合大会や The 6th Japanese-German Symposium on Classification、第 33 回ファジィシステムシンポジウム等の研究発表を通じて、これらの手法に対する理論的、応用的研究成果を発表した。
次に、平成３０年度の研究成果を、米国、シカゴにて開催された CAS2018 国際会議で発表し、発表した論文"Homogeneous Cluster Analysis"に対して、2nd Runner-Up Theoretical Paper Award を受賞した。さらに、SCI2018 国際会議にて、"Cluster-Scaled Intelligent Data Analysis”と題して基調講演を行うと共に、COMPSTAT2018 国際会議での招待研究発表、KES-IDT2018 国際会議での発表と論文の出版、2018 年度統計関連学会連合大会や第 34 回ファジィシステムシンポジウム等の研究発表を通じて、これらの手法に対する理論的、応用的研究成果を発表した。
さらに、平成３１年度（令和元年度）の研究成果を、米国、フィラデルフィアで開催された CAS2019 国際会議で発表し、発表した論文"Probabilistic Metric Based Multidimensional Scaling"に対して Best Paper Award を受賞した。さらに、CFE-CMStatistics2019、MIT-Tsukuba Joint-Workshop on Data Systems Science towards Social and Business Innovations での招待研究発表、The 2019 IEEE International Conference on Fuzzy Systems での発表と論文の出版、第 35 回ファジィシステムシンポジウム等の研究発表を通じて、これらの手法に対する研究成果を発表した。

（３）今後の展望

本研究の研究期間全体を通じて、より精度の高い結果を出し得るクラスター計量モデルの開発を行い、その研究成果の有効性が認められて、３度の受賞となった。また、国際会議での基調講演をはじめとして、多数の研究発表を行った。
今後の展望は、開発した計量をさらに高次の計量に発展させ、それらの理論的検証を進めることである。また、それと同時に計量モデルを拡張する必要がある。さらに、開発したモデルを種々のビックデータに適用することで、実用化を図る必要がある。

５．主な発表論文等

〔雑誌論文〕　計12件（うち査読付論文　9件／うち国際共著　0件／うちオープンアクセス　5件）

１．著者名 M. Sato-Ilic	４．巻 168
２．論文標題 Probabilistic Metric Based Multidimensional Scaling	５．発行年 2020年
３．雑誌名 Procedia Computer Science	６．最初と最後の頁 65〜72
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.procs.2020.02.258	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 M. Sato-Ilic	４．巻 1
２．論文標題 Quantification and Visualization for Difference of Fuzzy Clustering Results	５．発行年 2019年
３．雑誌名 The 2019 IEEE International Conference on Fuzzy Systems	６．最初と最後の頁 1-6
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 村山喬則，佐藤美佳	４．巻 1
２．論文標題 高次元データに対するファジィクラスタリングの主成分分析による評価	５．発行年 2019年
３．雑誌名 第 35 回ファジィシステムシンポジウム 講演論文集	６．最初と最後の頁 203-208
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 M. Sato-Ilic	４．巻 140
２．論文標題 Homogeneous Cluster Analysis	５．発行年 2018年
３．雑誌名 Procedia Computer Sciences, Elsevier	６．最初と最後の頁 269-275
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.procs.2018.10.320	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 M. Sato-Ilic	4．巻 7
2．論文標題 Cluster-Scaled Regression Analysis for High-Dimension and Low-Sample Size Data	5．発行年 2018年
3．雑誌名 Advances in Smart Systems Research	6．最初と最後の頁 1-10
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 伊藤佳輝，元田卓，佐藤美佳	4．巻 -
2．論文標題 高次元小標本データに対するT-ノルムに基づくマルチレイヤークラスタリング	5．発行年 2018年
3．雑誌名 第34回ファジィシステムシンポジウム講演論文集	6．最初と最後の頁 480-485
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 M. Aoshima, K. Yata	4．巻 28
2．論文標題 Two-sample tests for high-dimension, strongly spiked eigenvalue models	5．発行年 2018年
3．雑誌名 Statistica Sinica	6．最初と最後の頁 43-62
掲載論文のDOI（デジタルオブジェクト識別子） 10.5705/ss.202016.0063	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 M. Aoshima, D. Shen, H. Shen, K. Yata, Y. Zhou, J.S. Marron	4．巻 60
2．論文標題 A survey of high dimension low sample size asymptotics	5．発行年 2018年
3．雑誌名 Aust. N. Z. J. Stat.	6．最初と最後の頁 4-19
掲載論文のDOI（デジタルオブジェクト識別子） 10.1111/anzs.12212	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 清水信夫，中野純司，山本由和	4．巻 66
2．論文標題 集約的シンボリックデータのカイ2乗統計量を用いた非類似度とその不動産情報データへの適用	5．発行年 2018年
3．雑誌名 統計数理	6．最初と最後の頁 279-294
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 M. Sato-Ilic	4．巻 114
2．論文標題 Knowledge-based Comparable Predicted Values in Regression Analysis	5．発行年 2017年
3．雑誌名 Procedia Computer Science, Elsevier	6．最初と最後の頁 216-223
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.procs.2017.09.063	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

1．著者名 M. Sato-Ilic, P. Ilic	4．巻 －
2．論文標題 Identification and Scaling Methods based on Comparative Quantification for Dissimilarity Data	5．発行年 2017年
3．雑誌名 The 2017 IEEE International Conference on Fuzzy Systems	6．最初と最後の頁 －
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/FUZZ-IEEE.2017.8015443	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

1．著者名 矢吹健二、佐藤美佳	4．巻 －
2．論文標題 ３元マルチソースデータに対する同時ファジィクラスタリング手法	5．発行年 2017年
3．雑誌名 第 33 回ファジィシステムシンポジウム講演論文集	6．最初と最後の頁 441-446
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計23件（うち招待講演　11件／うち国際学会　17件）

１．発表者名
Mika Sato-Ilic

２．発表標題
Statistical Data Science at A Crossroads (基調講演)

３．学会等名
KES-Intelligent Decision Technologies, Smart Innovation, Systems and Technologies（招待講演）（国際学会）

４．発表年
2020年

１．発表者名
Mika Sato-Ilic

２．発表標題
Probabilistic Metric Based Multidimensional Scaling

３．学会等名
Complex Adaptive Systems 2019（国際学会）

４．発表年
2019年

１．発表者名
Mika Sato-Ilic

２．発表標題
Quantification and Visualization for Difference of Fuzzy Clustering Results

３．学会等名
The 2019 IEEE International Conference on Fuzzy Systems（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
Mika Sato-Ilic

２．発表標題
Fuzzy clustering-based non-linear dimensionality reduction

３．学会等名
13th CFE 2019 and 12th CMStatistics 2019（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
Mika Sato-Ilic

２．発表標題
Statistical Data Science Based on Soft Computing

３．学会等名
MIT-Tsukuba Joint-Workshop on Data Systems Science towards Social and Business Innovations（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
村山喬則，佐藤美佳

２．発表標題
高次元データに対するファジィクラスタリングの主成分分析による評価

３．学会等名
第 35 回ファジィシステムシンポジウム

４．発表年
2019年

１．発表者名
Makoto Aoshima

２．発表標題
High-Dimensional Statistical Analysis: Non-Sparsity, Strongly Spiked Noise and HDLSS （基調講演）

３．学会等名
The 7th International Workshop in Sequential Methodologies（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
清水信夫，中野純司，山本由和

２．発表標題
集約的シンボリックデータにおける変数間の相関の指標

３．学会等名
2019年度統計関連学会連合大会

４．発表年
2019年

１．発表者名
M. Sato-Ilic

２．発表標題
Cluster-Scaled Intelligent Data Analysis（基調講演）

３．学会等名
3rd International Conference on Smart Computing & Informatics（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
M.Sato-Ilic

２．発表標題
Soft Clustering-based Models

３．学会等名
23rd International Conference on Computattional Statistics (COMPSTAT 2018)（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
M.Sato-Ilic

２．発表標題
Homogeneous Cluster Analysis

３．学会等名
Complex Adaptive Systems 2018（国際学会）

４．発表年
2018年

１．発表者名
M. Sato-Ilic

２．発表標題
Cluster-Scaled Regression Analysis for High-Dimension and Low-Sample Size Data

３．学会等名
Knowledge-Based and Intelligent Information & Engineering Systems - Intelligent Decision Technologies（国際学会）

４．発表年
2018年

１．発表者名
伊藤佳輝，元田卓，佐藤美佳

２．発表標題
高次元小標本データに対するT-ノルムに基づくマルチレイヤークラスタリング

３．学会等名
第34回ファジィシステムシンポジウム

４．発表年
2018年

１．発表者名
小林大悟，佐藤美佳

２．発表標題
分類構造に基づく異常検知手法

３．学会等名
2018年度統計関連学会連合大会

４．発表年
2018年

１．発表者名
N. Shimizu, J. Nakano, Y. Yamamoto

２．発表標題
Dissimilarity between aggregated symbolic data using chi-squared statistics

３．学会等名
2018 Workshop in Symbolic Data Analysis（国際学会）

４．発表年
2018年

１．発表者名
M. Sato-Ilic

２．発表標題
Modeling New Complex Data Structures (基調講演)

３．学会等名
Complex Adaptive Systems 2017（招待講演）（国際学会）

４．発表年
2017年

１．発表者名
M. Sato-Ilic

２．発表標題
Knowledge-based Comparable Predicted Values in Regression Analysis

３．学会等名
Complex Adaptive Systems 2017（国際学会）

４．発表年
2017年

１．発表者名
M. Sato-Ilic, P. Ilic

２．発表標題
Cluster Identification and Scaling Methods based on Comparative Quantification for Dissimilarity Data

３．学会等名
The 2017 IEEE International Conference on Fuzzy Systems（招待講演）（国際学会）

４．発表年
2017年

１．発表者名
M. Sato-Ilic

２．発表標題
A Fuzzy Clustering based Data Fusion Method

３．学会等名
11th International Conference on Computational and Financial Econometrics and 10th International Conference of the ERCIM Working Group on Computational and Methodological Statistics（招待講演）（国際学会）

４．発表年
2017年

１．発表者名
M. Sato-Ilic

２．発表標題
Cluster-Scaled Forecasting Method For High-Dimension Low-Sample Size Data

３．学会等名
The 6th Japanese-German Symposium on Classification（国際学会）

４．発表年
2017年

１．発表者名
M. Sato-Ilic

２．発表標題
Asymmetric Clustering Methods based on Orthogonal Projector to the Intersection of Subspaces

３．学会等名
2017年度統計関連学会連合大会

４．発表年
2017年

１．発表者名
矢吹健二、佐藤美佳

２．発表標題
３元マルチソースデータに対する同時ファジィクラスタリング手法

３．学会等名
第 33 回ファジィシステムシンポジウム

４．発表年
2017年

１．発表者名
N. Shimizu, J. Nakano, Y. Yamamoto

２．発表標題
Dissimilarities between Groups of Data

３．学会等名
New Zealand Statistical Association and the International Association of Statistical Computing 2017（招待講演）（国際学会）

４．発表年
2017年

〔図書〕　計3件

１．著者名 イリチュ美佳、高木英明	４．発行年 2017年
２．出版社 筑波大学出版会	５．総ページ数 53(2章)
３．書名 データの類似度と多次元尺度構成法（２章）、サービスサイエンスの事訳―データサイエンスと数理科学の融合に向けてー	

１．著者名 Mika Sato-Ilic	４．発行年 2020年
２．出版社 Springer, Switzerland	５．総ページ数 (in press)
３．書名 Fuzzy Clustering Models and Their Related Concepts（１章）, Fuzzy Approaches for Soft Computing and Approximate Reasoning: Theories and Applications	

１．著者名 イリチュ美佳、高木英明	４．発行年 2017年
２．出版社 筑波大学出版会	５．総ページ数 51（３章）
３．書名 分かるために分けるクラスター分析（３章）、サービスサイエンスの事訳―データサイエンスと数理科学の融合に向けてー	

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	青嶋　誠 (Aoshima Makoto) (90246679)	筑波大学・数理物質系・教授 (12102)	
研究分担者	清水　信夫 (Shimizu Nobuo) (00332130)	統計数理研究所・データ科学研究系・助教 (62603)	