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　 ６ 月　 ８ 日現在

機関番号：３４３１０
研究種目：基盤研究(C)（一般）
研究期間：2017～2019
課題番号：１７Ｋ００００６０

研究課題名（和文）複数情報源非類似性データに対するデータマッチング法に関する研究

研究課題名（英文）Data matching method for multiple proximity 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）　　3,400,000円

研究成果の概要（和文）：　分析対象に対して，複数の情報源からデータが得られたとき，そのデータを統合し，対象を表す点を解釈可能な空間へ埋め込む方法の開発を行った．特に，複数の匿名化された（非）類似性データの低次元空間への埋め込み法を提案した．想定している状況は，POSデータやアクセスログデータ，調査データ等の自身が所有しているデータと政府や研究機関などが公開しているオープンデータとの統合利用である．このとき，保持データとオープンデータの情報を併用して，解釈可能な空間上での分析対象の位置を推定する方法を提案した．

研究成果の学術的意義や社会的意義
　複数情報源データの分析は，ビッグデータ分析の一つであり，必要性が認識されているものの，その分析法の開発は進んでいないのが現状である．本研究では，量的データおよび質的データが混在している場面においても適用可能な複数情報源非類似性データの分析法を開発した．
　提案手法により，複数の情報源から得たデータを統合し，解釈することが可能であるになるため，様々なオープンデータの統合，活用に貢献し，加えて，オープンデータへの活用の活発化により，様々なサービスが生まれる一助となると考えられる．

研究成果の概要（英文）：When data are obtained from multiple sources concerning subjects, we develop new methods for integrating and embedding the points into an interpretable low-dimensional space. In particular, we proposed embedding methods for multiple (dis)similarity data from multiple sources and anonymized data. Specifically, our method integrates the data by combining data such as POS data, access log data, and survey data with aggregated data that were published by the government or research institutes. We proposed methods to obtain new knowledge from owned data and open data by using open data to estimate the position of the object in the interpretable space.

研究分野：多変量データ解析

キーワード：ビッグデータ　多次元尺度構成法　正準相関分析法　多ドメインマッチング法

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

類似性データとは，対象間の 2 者関係について記述したデータであり，その値が大きければ 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ようなデータである．類似性データと非類似性データは対応した関係にあり，以降では（非）類似性データと呼ぶことにする．（非）類似性データは心理学，社会学等の様々な分野で観測される．（非）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，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や対象のグループを検出するクラスタリング法などの分析手法が提案されている．このような（非）類似性データについて，複数の情報源から得られたデータのことを複数情報源非類似性データという．

複数のデータセットに対する代表的な多変量解析の手法として，Carroll (1968) によって，データセット間の変量群の関連を表す手法である，Generalized Canonical Correlation analysis (GCANO) が提案されている．変量の数が多い場合，GCANO では解釈が困難になることから Regularized GCANO が Tenenhaus & Tenenhaus (2011) によって提案されている．これらの手法は，複数の情報源からなる同一個体のデータセットについて変量群をもとに，個体を共通の低次元空間へ付置し，解釈を行うものである．

近年，IoT など情報通信技術の発達により得られる複数の情報源のデータにおいては，すべての情報源から同一個体の情報を得るという点という仮定は非常に厳しく，一部個体がある情報源では共通しているという状態が考えられる．この仮定に対し，Shimodaira (2014) では，複数の情報源から，異なる対象のデータが得られたときの次元縮約法として Cross-Domain Matching Correlation Analysis (CDMCA) を提案している．しかし，複数の情報源から得られたデータを分析するために，これらの手法を適用するには以下の 3 点の問題がある．1 つ目は，匿名化され，質的データとなったデータに対し適用が困難であることである．オープンデータは k-匿名化や l-多様性などの方法を用いて，多くの場合，匿名化が施されている．複数情報源すべてのデータを個人単位で取得することは，現実的でなく，多くの場合，オープンデータ等を利用する必要がある．このとき，匿名化されたデータを利用するためには質的データにも適用可能な手法を開発する必要がある．2 つ目の問題点は，関連性を示すデータが既に存在した場合，その情報を利用できないことである．例えば，ユーザーがある商品を購買したかや商品評価の関係が既に存在している場合，商品の特徴，ユーザーの特徴からのみマッチングを行うのではなく，得られている購買，評価の情報を基にマッチングを行う方が現実的である．また，既知の関連性を表すデータをそのデータだけでなく，複数情報源を基に「なぜその関係性が現れたのか」を解釈することは科学一般に置いて重要な課題である．3 つ目の問題点は，データの欠測や異常値，冗長な変量などのノイズを含んでいる場合に誤った解釈をしてしまう場合があることである．複数情報源から得られているデータを想定するため，データが得られている対象が一致しないことに加えて，欠測や異常値が存在するデータが含まれることが多い．これらのデータに対して従来の手法を適用すると，情報の損失や誤った解釈を招く恐れがある．そのため，このような状況に対応できるような手法を開発する必要がある．

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) 複数情報源から得られる類似性データに対する分析手法の総合的調査
(2) 複数情報源から得られる類似性データに対する新たな分析手法の提案
(3) 提案手法の実装・公開
(4) 実世界のデータに対して提案手法を適用することによる，新たな知見の獲得

(1)における複数情報源から得られる類似性データとは，（非）類似度データに加え，多変量データの形も含むデータのことを指す．例えば，k-means 法などのクラスタリング 法は入力データとして多変量データを想定しているが，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，多変量データについても広義の類似性データとみなすこととする．(1)における目的は，これらのデータを対象としている手法について，手法の性質や制約に基づいて各種法を分類し，数理的に特徴づけることである．

(2)については，本研究では，複数の情報源から得られているデータを想定しているため，前述の背景で述べた問題点を考慮した上で特に次の問題点に着目する．とくに，外れ値や意図していないデータの混入によってデータが汚染されている状態，また，データとして観測することが困難である個々の対象についてのノイズが存在している場合に着目し，新たな分析手法の開発を行う．その中で次の 2 点について，手法の開発に取り組む．①質的データにも適用可能な，複数の情報源から得られたデータを共通空間に埋め込む方法の検討及び開発，②一部の（非）類似性データが与えられたとき，複数の情報源から得られたデータを共通空間に埋め込む方法の検討及び開発である．

(3)については，(2)で提案した手法の実装および公開を行う．提案した手法について，ビッグ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，最終的な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提案された新たな分析方法についてビッグデータにも適用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．また開発したアルゴリズムについて，多くの人が容易に使うことができるようにパッケージ化し，CRAN などのレポジトリサイトでの公開を行う．

(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する．これにより，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，先に述べた 4 つの研究目的に沿って説明する．

(1) については，既存の多変量解析法とその制約との関係性の観点において数理的な特徴づけを行う．そのためには，既存手法を比較可能な形で定式化を行う必要がある．既存の多変量解析手法は，得られたデータと分析法のモデルとの誤差をいくつかの制約のもとで最小化する方法として記述可能である．外部から得られているオープンデータなどの匿名化された情報情報をもとにして制約を加えることにより，複数情報源を制約の形で表すことができる．そのため，既存の多変量解析の手法を統一的な方法で記述し，手法間の関連，改善点をまとめ，整理を行う．

(2), (3) については，データの汚染という観点に着目し複数情報源データの分析法の提案を行う．例えば，データの中に外れ値のような汚染が存在する場合には，γダイバージェンスなど，データの汚染に対してロバストなダイバージェンスを用いることにより，外れ値などの影響を取り除いた分析を行うことができる．本研究では，外部の情報を用いた上でノイズとなっているデータの汚染を取り除いた分析を行うことができる手法の提案を行う．そのために分析手法の目的関数自体を他のダイバージェンスを用いるなどして変更した方法，また，別の情報によりバイアスの補正を行うための方法を用いる．これにより，(2) で挙げた問題点の解消を行う．(3) については，実装された提案手法の公開を行う．実装に際して，実際にビッグデータに対して適用することができるように計算の高速化を行う必要がある．例えば，ロバストなダイバージェンスを用いた方法については，majorizing algorithm を導出することにより，計算の高速化を行うことが可能である．そして，計算の高速化を行った手法を実装し，公開を行う．

(4) について POS データやシングルソースデータ，また，調査データなど，マーケティング分野や心理学分野の実データを主に想定して，提案した手法を適用する．そして，実際に応用上問題がないか否か確認する．適用した結果についてデータ提供元と議論を行いながら必要に応じて手法の改良に取り組む．

４．研究成果

本研究では，分析対象に対して，複数の情報源からデータが得られたとき，そのデータを統合し，対象を表す点を解釈可能な空間へ埋め込む方法の開発を行った．特に，複数の匿名化された（非）類似性データの低次元空間への埋め込み法を提案した．想定している状況は，POS データやアクセスログデータ，調査データ等の自身が所有しているデータと政府や研究機関などが公開しているオープンデータとの統合利用である．このとき，保持データとオープンデータの情報を併用して，解釈可能な空間上での分析対象の位置を推定する方法を提案した．

次に先に挙げた研究目的について具体的な成果を述べる．

(1) での研究成果として，既存の多変量解析法について統一的な表記での定式化を行うことができた．具体的には，質的データを想定した制約付き主成分分析の提案を行った．これは，既存の制約付き主成分分析の拡張にあたり，本研究で想定しているような自身が保有しているデータに対して，保有しているデータとは別の情報源から得られる情報を制約として加えることができる分析法である．

(2), (3) での研究成果は，研究目的(2)で挙げた①，②に対応した次の 2 点である．まず，①に対応した 1 点目として，データが汚染されている状況においても対処可能な正準相関分析の提案を行った．この手法は，複数情報源から得られたデータに対して，そのデータに汚染が発生している場合においても解釈可能な低次元空間に対象を埋め込むような手法である．この手法は大きく 2 種類の方法で構成されており，1 つめの方法では，データでは観測されていない変量によって対象がいくつかのクラスに分かれるような状態を想定した上でこれを別のものとして低次元空間に埋め込む方法である．2 つめの方法では，同様に観測されていないクラスによってデータが汚染されている状況であるが，そのうち興味がある対象群について，γダイバージェンスを用いて推定する方法である．次に，②に対応した 2 点目として，順序変量として得られる係留寸描法による回答者間の類似性データとリッカート尺度などで得られる回答者の回答データを統合することによるクラスタリング法を提案した．この手法を用いることで，回答の傾向によるバイアスが存在するようなデータからこのようなバイアスを取り除いたクラスター間の非類似度の埋め込みを行っている．この手法について，R での実装が CCRS というパッケージとして CRAN 上に一般公開されている．

(4) での研究成果は，実データ解析における手法の適用から応用上の観点より，一定の評価を得ることができた．また，研究代表者が取り組んでいる共同研究においても提案した手法を用いることで，従来手法では得られなかった知見を得ることができた．

本研究で得られた以上の成果により，複数の情報源から得たデータを統合し，解釈することが可能であるになるため，様々なオープンデータの統合，活用に貢献し，加えて，オープンデータへの活用の活発化により，様々なサービスが生まれる一助となると考えられる．

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　7件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 Tanioka, K and Yadohisa, H.	４．巻 140
２．論文標題 Asymmetric MDS with Categorical External Information Based on Radius Model	５．発行年 2018年
３．雑誌名 Procedia Computer Science	６．最初と最後の頁 284-291
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.procs.2018.10.318	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 Mitsuhiro, M. and Yadohisa, H.	４．巻 1
２．論文標題 A unified representation of simultaneous analysis methods of reduction and clustering	５．発行年 2018年
３．雑誌名 Japanese Journal of Statistics and Data Science	６．最初と最後の頁 393-412
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s42081-018-0022-6	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 Tanioka, K. and Yadohisa, H.	４．巻 6
２．論文標題 Unfolding models for asymmetric dissimilarity data with external information based on path structures	５．発行年 2018年
３．雑誌名 International Journal of Software Innovation	６．最初と最後の頁 53-66
掲載論文のDOI（デジタルオブジェクト識別子） 10.4018/IJSI.2018070104	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 Tsuchida Jun、Yadohisa Hiroshi	４．巻 114
２．論文標題 Partial Least-Squares Method for Three-Mode Three-Way Datasets Based on Tucker Model	５．発行年 2017年
３．雑誌名 Procedia Computer Science	６．最初と最後の頁 234～241
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.procs.2017.09.065	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 Abe Hiroyasu、Yadohisa Hiroshi	４．巻 32
２．論文標題 A non-negative matrix factorization model based on the zero-inflated Tweedie distribution	５．発行年 2017年
３．雑誌名 Computational Statistics	６．最初と最後の頁 475～499
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s00180-016-0689-8	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 Abe Hiroyasu、Yadohisa Hiroshi	４．巻 29
２．論文標題 AUTOMATIC RELEVANCE DETERMINATION IN NONNEGATIVE MATRIX FACTORIZATION BASED ON A ZERO-INFLATED COMPOUND POISSON-GAMMA DISTRIBUTION	５．発行年 2017年
３．雑誌名 Journal of the Japanese Society of Computational Statistics	６．最初と最後の頁 29～54
掲載論文のDOI（デジタルオブジェクト識別子） 10.5183/jjscs.1608001_233	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 土田　潤、宿久　洋	４．巻 65
２．論文標題 重力モデルを用いたサッカー選手の動きの定量化	５．発行年 2017年
３．雑誌名 統計数理	６．最初と最後の頁 271～286
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 オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計10件（うち招待講演　1件／うち国際学会　10件）

１．発表者名 Morioka, Y., Tanioka, K. and Yadohisa, H.
２．発表標題 Constrained matrix completion algorithm considering individual differences
３．学会等名 1th International Conference of the European Research Consortium for Informatics and Mathematics Working Group on Computational and Methodological Statistics 2018（国際学会）
４．発表年 2018年

１．発表者名
Takasawa, I., Tanioka, K. and Yadohisa, H.

２．発表標題
Constrained LiNGAM approach for tensor data

３．学会等名
Conference of the European Research Consortium for Informatics and Mathematics Working Group on Computational and Methodological Statistics 2018（国際学会）

４．発表年
2018年

１．発表者名
Mizutani,S. and Yadohisa, H.

２．発表標題
Robust canonical correlation analysis via γ-divergence

３．学会等名
The conference of Data Science, Statistics & Visualisation (DSSV 2018)（国際学会）

４．発表年
2018年

１．発表者名
Yamayoshi, M., Tsuchida, J. and Yadohisa, H.

２．発表標題
A Representation of the Relationship Between Variables in Quantitative and Qualitative Mixed Data

３．学会等名
European Conference on Data Analysis (ECDA) 2018（国際学会）

４．発表年
2018年

１．発表者名
Okabe, M., Tsuchida, J. and Yadohisa, H.

２．発表標題
Using Multi-Label Logistic Regression to Maximize Macro F-measure

３．学会等名
European Conference on Data Analysis (ECDA) 2018（国際学会）

４．発表年
2018年

１．発表者名
Yamagishi, Y., Tanioka, K. and Yadohisa, H.

２．発表標題
Constrained Principal Component Analysis for Nonmetric Data

３．学会等名
61th World Statistics Congress, The Palais des Congres, Marrakech, Morocco.（国際学会）

４．発表年
2017年

１．発表者名
Takagishi, M. and Yadohisa, H.

２．発表標題
Visualization of clustering on multiple data

３．学会等名
10th International Conference of the European Research Consortium for Informatics and Mathematics Working Group on Computational and Methodological Statistics 2017（国際学会）

４．発表年
2017年

１．発表者名
Goto, S. and Yadohisa, H.

２．発表標題
Pattern prediction for time series data with change points

３．学会等名
New Zealand Statistical Association and the International Association of Statistical Computing (Asian Regional Section) Joint Conference 2017（国際学会）

４．発表年
2017年

１．発表者名
Tsuchida, J. and Yadohisa, H.

２．発表標題
Canonical covariance analysis for mixed numerical and categorical three-way three-mode data

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

４．発表年
2017年

１．発表者名 Takagishi, M. Velden, M. van de, and Yadohisa, H.
２．発表標題 Clustering Methods for Ordered Categorical Data with Response Style
３．学会等名 Joint Statistical Meeting 2017（国際学会）
４．発表年 2017年

〔図書〕　計0件

〔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