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　３０　年　　６　月　　６　日現在

機関番号：　３２６１９
研究種目：　基盤研究(C)（一般）
研究期間：　2015～2017
課題番号：　１５Ｋ００３４８

研究課題名（和文）不完全データに対する球面ファジィクラスタリング技法の確立

研究課題名（英文）Fuzzy Clustering Methods for Incomplete Spherical Data

研究代表者
神澤　雄智（Kanzawa, Yuchi）

芝浦工業大学・工学部・教授

研究者番号：００２９８１７６

交付決定額（研究期間全体）：（直接経費）　　3,600,000円

研究成果の概要（和文）：クラスタリングは特に重要なデータ解析手法として、多くの分野で用いられている．本研究課題では，これまでに研究を推し進めてきた球面ファジィクラスタリングを実世界，実社会の現象や事象への適用を可能とするために，不完全データのための球面ファジィクラスタリング技法を確立することを目的にしている．
平成27年度は，完全データに対する球面ファジィクラスタリング手法群を整備し、データの不完全部分を削除する方法の性能を評価した．平成28年度は，人工不完全データを用いて，開発した手法と従来法の精度を比較した．平成29年度は，リニア統計における完全情報最尤推定法に対応する球面完全化手法群を開発した．

研究成果の概要（英文）：Clustering is applicable in many fields as an important data analysis tool. The objective of this research is constructing fuzzy clustering methods for incomplete spherical data, which will make fuzzy clustering methods for spherical data applicable to practical situations.
In 2015 fiscal year, we arranged fuzzy clustering methods for complete spherical data, and evaluated their performance for incomplete data by whole data strategy. In 2016 fiscal year, we compared the proposed methods with the conventional methods in terms of clustering accuracy using artificial data sets. In 2017 fiscal year, we constructed imputation methods for incomplete spherical data, which is corresponding with the complete information maximum likelihood estimation for linear statistics.

研究分野：　ファジィクラスタリング

キーワード：　ファジィクラスタリング　球面データ　不完全データ

１．研究開始当初の背景

大規模データ社会において，ユーザが膨大なデータから有用な情報を独力で抽出することは不可能であり，データ群の構造化は必須である．クラスタリングはデータを外的基準なしに自動的に分類する手法で，データ構造化の有用な技術として注目を集めている．現実のデータでは，個体が唯一のクラスタに属するよりも，複数のクラスタに適当な割合で属すると解釈すべき場面が多くあり，帰属の曖昧さを扱うファジィ理論に基づいたファジィクラスタリングが活発に研究されている．

ファジィ c-平均法などの基礎的ファジィクラスタリング手法が対象とする各個体はベクトルである一方で，個体の大きさが同じ場合にはそれらは超球面上に存在し，これを球面データという．球面データは．地球表面上の気候データや眼球運動などの，3 次元空間上の 2 次元球面上に位置するデータだけではない．文書データの Bag of Words 表現，画像データの Bag of Keypoints 表現，カーネルデータ解析を通じた全ての類似性データは(高次元) 球面データと看做すことができる，このように，数多くのデータが球面データとして扱えるにも関わらず，球面データに対するファジィクラスタリングに関する研究は未熟であった．以下，球面データをクラスタリングする手法を球面クラスタリングといい，これらと区別するために，通常のベクトルデータをリニアデータ，リニアデータをクラスタリングする手法群をリニアクラスタリングという．

申請者はこれまで，球面ファジィクラスタリング技法において，初期値依存性による局所収束性，クラスタ数の自動推定の問題点をそれぞれ解決してきた。そのため申請者は，いかなるデータも球面化してクラスタリングすべきと考えている．

しかし，いざ実問題に取り組もうとすると，ほとんどのデータには欠損部分があってそのままでは開発手法を適用できない．このようなデータを不完全データといい，リニア統計では不完全データを完全化(リニア完全化)する手法が整備されている．安直に欠損を持つ個体を対象から外して球面クラスタリングを適用すると，既存の手法と組み合わせてリニアクラスタリングを適用した場合に比べて著しく精度が低くなってしまうケースが多く見られた．申請者の調査によれば、球面データにおける不完全データの完全化に関する方法論的基盤は見当たらず，リニア完全化手法群をそのまま球面不完全データに適用できるかどうかも明らかではない．

２．研究の目的

そこで本研究課題では，これまでに研究を推し進めてきた球面ファジィクラスタリングを実世界，実社会の現象や事象への適用を可能とするための次の研究課題として，不完全データのための球面ファジィクラスタリング技法を確立することを目的にした．具体的には、(1-1) 申請者が球面完全データに対する技法を確立した際に用いた方法論に基づいて，リニア完全化手法群のそれぞれに対応する球面完全化手法群を開発し，(1-2) 開発手法群の周辺分野との理論的類似性を明らかにしつつ，(2-1) 各開発手法の性能を実験を通じて定量的に評価し，(2-2) 開発手法と申請者のこれまでの研究成果を組み合わせ，様々なデータに対する実験を通じて，リニアクラスタリングを精度面で凌駕する，不完全データのための球面ファジィクラスタリング技法を確立することを目的とした．

３．研究の方法

申請者が開発してきた球面クラスタリング技法を不完全データにも応用できるようにして実用に供する，という目的に照らして，球面不完全データの完全化手法群の開発とそれらの理論的特徴を解明しつつ球面クラスタリング性能を実験的に評価する．開発手法群は大きく，1)．データの不完全部分を削除する方法群，2)．不完全部分を初等的に予測して完全化する方法群，3)．不完全部分を予測する際に確率密度分布を用いる手法群，に分けられ，概ねこの順序で開発・評価・比較する．実験については大きく，(1) 開発手法単体での性能評価のための，欠損数とクラスタリング精度の定量的関係の解明，(2) 不完全データに対して，開発手法に基づく球面クラスタリングと既存の完全化手法群を用いたリニアクラスタリングとの精度比較，に分けられ，各手法毎にこれら実験を行う．

４．研究成果

初年度は，不完全データに対応するための準備として先ず，これまで開発してきた，完全データに対する球面ファジィクラスタリング手法群を整備し，そして，データの不完全部分を削除する方法の性能を評価することを目標とした．前者については，概ね良好に研究を遂行でき，幾つかの査読付き論文を出版し，1 件の国際会議にて成果を発表した．特に，国際会議で発表した手法は，これまで他の研究者によって提案されてきた多くの手法群を統一的に扱えるものである．元々は多くの各手法について個別に，不完全データに対応する手法を検討していくことを考えていたが，提案した手法を基にして研究を遂行することによって，多くの手法の不完全データへの対応をまとめて統一的に扱うことができることになった．後者については，前者の手法の球面不完全データに対する定量的性能を実験的に評価することを試みた．具体的には，人工的に生成した球面完全データから徐々に欠損箇所を増やしていくことによって得られる不完全データに対してこの手法群を適用し，欠損比率とクラスタリング精度との関係を数値的実験によって得ようとした．しかしながら、その関係は欠損箇所に大きく依存し，欠損比率を増やしていくとクラスタリング精度の分散が爆発的に増え

てしまい，定量的な関係を見出すことはできなかった．ただし，ここでは定量的な関係を見出すことそのものが目的なのではなく，今後に開発していく不完全データへの新たな手法の有効性を最も基本的な手法と比較検討していくための基準が得られたという意味では，提案手法の特性検証を次年度に行っていくための準備が整った．

次年度は，開発した手法と従来法の精度を定量的に比較するために，ベンチマーク用人工完全データ群から徐々に欠損箇所を増やしていくことによって得られる不完全データに対して，リニアクラスタリングと球面クラスタリングについて，欠損比率とクラスタリング精度との関係を明らかにすると共に，欠損箇所とクラスタリング精度との関係についても検証した．これによって，不完全データに対して，開発手法群を用いた球面クラスタリングとリニアクラスタリングとの精度面での優位性を明らかにできた．また，球面三角法に基づいて欠損値を予測する手法について開発し，球面主成分分析と多次元尺度構成法を，研究代表者がこれまで球面クラスタリング開発に用いてきた Young-Householder 変換に基づいて組み合わせ，類似度完全データを球面に直接附置するための球面構成法を開発した上で，不完全な類似度関係性データを球面に附置して得られる球面不完全データに対して，上記の球面完全化手法群を適用する手法を開発した．

最終年度は，リニア統計における完全情報最尤推定法に対応する球面完全化手法群を開発した．事前に欠損値の確率密度分布を決める必要があり，リニア統計では正規分布が仮定される．そこでまずは，球面統計において正規分布に対応する von Mises-Fisher 分布を仮定して球面完全化法を開発した．開発手法群の定量的性能を評価するために，手順 1 に基づいた実験を，開発手法群と従来法を定量的に比較するために手順 2 に基づいた実験を行った．完全情報最尤推定法は EM アルゴリズムに基づいていて，クラスタリングもまた混合密度分布の混合係数が未知という意味で EM アルゴリズムの枠組みで論じられることがある．その意味では，完全化とクラスタリングを同時に行って，欠損値と混合係数を合わせた未知変数を解く手法も考え，研究代表者が開発した球面クラスタリング技法群を EM アルゴリズムの枠組みで見直して，欠損値も同時に得るアルゴリズムを開発した．von Mises-Fisher 分布に基づく球面クラスタリングが外れ値に影響を受けるのに対して，Pearson VII 型分布に基づいて外れ値に頑健な球面クラスタリング開発を試みた．集中度パラメータを固定した場合のアルゴリズムを導出することはできたが，集中度パラメータの推定を陽に行うことが非常に難しいことが判明したため，何らかの数値解法を構築する必要があることが分かった．

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　6 件）

① Yuchi Kanzawa, A Maximizing Model of Spherical Bezdek-Type Fuzzy Multi-Medoids Clustering, Journal of Advanced Computational Intelligence and Intelligent Informatics, 査読有り, Vol.19, pp.738-746 , https://www.fujipress.jp/JACIII/

② Yuchi Kanzawa, Bezdek-Type Fuzzified Co-Clustering Algorithm, Journal of Advanced Computational Intelligence and Intelligent Informatics, 査読有り, Vol.19, pp.852-860 , https://www.fujipress.jp/JACIII/

③ Yuchi Kanzawa, A Maximizing Model of Bezdek-Like Spherical Fuzzy c-Means, Journal of Advanced Computational Intelligence and Intelligent Informatics , Vol.19, pp.662-669 , https://www.fujipress.jp/JACIII/

④ Yuchi Kanzawa, Power-Regularized Fuzzy c-Means Clustering with a Fuzzification Parameter Less than One, Journal of Advanced Computational Intelligence and Intelligent Informatics , Vol.20, pp.561-570 , https://www.fujipress.jp/JACIII/

⑤ Yuchi Kanzawa, Semi-supervised fuzzy c-means algorithms by revising dissimilarity/kernel Matrices , Studies in Computational Intelligence, Vol.671, pp.45-61 , https://www.springer.com/series/7092

⑥ Yuchi Kanzawa, Power-Regularized Fuzzy Clustering for Spherical Data, Journal of Advanced Computational Intelligence and Intelligent Informatics , Vol.22, pp.163-171 , https://www.fujipress.jp/JACIII/

〔学会発表〕（計　9 件）

① Yuchi Kanzawa , On Possibilistic Clustering Methods Based on Shannon/Tsallis-Entropy for Spherical Data and Categorical Multivariate Data, MDAI2015, 2015 年 9 月 21 日, シェブデ(スウェーデン)

② Yuchi Kanzawa, Fuzzy Clustering based on alpha-Divergence for Spherical Data and for Categorical Multivariate Data, FUZZ-IEEE2015, 2015 年 8 月 2 日, イスタンブール(トルコ)

③ Yuchi Kanzawa , On Bezdek-Type Possibilistic Clustering for Spherical Data, Its Kernelization, and Spectral Clustering Approach, MDAI2016, 2016 年 9 月 20 日, サンジュリアデロリア(アンドラ)

④ Yuchi Kanzawa , On Possibilistic Clustering Algorithms based on Noise Clustering, SCIS&ISIS2016, 2016 年 8 月 25 日, 北海学園大学(北海道・札幌市)
⑤ 神澤雄智, von Mises Fisher 分布と Polya 分布に基づくファジィクラスタリングについて, 第 26 回インテリジェント・システム・シンポジウム, 2016 年 10 月 27 日, 大阪大学(大阪府・大阪市)
⑥ 徳山晴紀、 遠藤靖典、 神澤雄智, 数値例による球面 k-平均法++の検討, 第 32 回ファジィシステムシンポジウム, 2016 年 8 月 31 日, 佐賀大学(佐賀県, 佐賀市)
⑦ Tadafumi Kondo, Yuchi Kanzawa , Comparison of Fuzzy Co-Clustering Methods in Collaborative Filtering-based Recommender System, MDAI2017, 2017 年 10 月 19 日, 九州工業大学(福岡県・北九州市)
⑧ Masayuki Higashi, Yuchi Kanzawa , Comparison of Vectorial and Spherical Methods in Fuzzy Clustering-based Classifier, MDAI2017, 2017 年 10 月 19 日, 九州工業大学(福岡県・北九州市)
⑨ 東正雪, 神澤雄智, 球面ファジィクラスタリングに基づく文書識別器の比較, 第 33 回ファジィシステムシンポジウム, 2017 年 9 月 13 日, 山形大学(山形県・米沢市)

〔図書〕(計　0 件)

〔産業財産権〕

○出願状況(計　0 件)

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況(計　0 件)

名称：
発明者：
権利者：
種類：
番号：
取得年月日：
国内外の別：

〔その他〕
ホームページ等

６．研究組織
(1)研究代表者
神澤　雄智(KANZAWA, Yuchi)
芝浦工業大学・工学部情報通信工学科・教授
研究者番号：０９０２９８１７６

(2)研究分担者
なし　(　)

研究者番号：

(3)連携研究者
なし(　)

研究者番号：

(4)研究協力者
なし(　)