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　５ 年　 ５ 月 １５ 日現在

機関番号：３２６１９
研究種目：基盤研究(C)（一般）
研究期間：2019～2022
課題番号：１９Ｋ１２１２６

研究課題名（和文）ノイズを含む超球面データのためのクラスタリング方法論の確立

研究課題名（英文）Clustering Methodology for Hyper-Spherical Data with Noise

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）　　3,300,000円

研究成果の概要（和文）：本研究では、超球面上に附置されるデータに対してノイズが含まれる場合にクラスタリングするための方法論を確立することを目的として、アルゴリズムの開発に取り組んだ。
ノイズ個体を吸収するためのクラスタを仮定する手法についてアルゴリズムを開発した．また，その数理的特性を実験的に確認した。推薦システムへの応用を踏まえて，球面クラスタリングと協調フィルタリングを組み合わせたアルゴリズムを開発し，実データに適用した結果，従来提案されていた単体データのためのクラスタリングと協調フィルタリングを組み合わせたアルゴリズムに比べて高い推薦精度を達成した．

研究成果の学術的意義や社会的意義
本研究課題では、大規模データのマイニングを目的に、超球面上に付置されたデータにノイズを際のクラスタリング手法の開発に取り組んだ。開発したアルゴリズムの特性を実験的に確認すると共に，その推薦システムへの応用を見据えて，開発したクラスタリングアルゴリズムと協調フィルタリングを組み合わせた手法を提案した．これらの成果により、大規模データに隠された構造を抽出クラスタリングとその推薦システムへの応用実現に向けた方法論的基盤を築いた。

研究成果の概要（英文）：The purpose of this study was to establish a methodology for clustering data placed on a hypersphere that contains noise. We worked on developing an algorithm for this purpose. We developed an algorithm for assuming clusters to absorb noisy individuals. Additionally, we experimentally confirmed its mathematical characteristics.
Based on its application to recommendation systems, we developed an algorithm that combines spherical clustering and collaborative filtering. When applied to actual data, this algorithm achieved a higher recommendation accuracy compared to the algorithm that simplexial clustering and collaborative filtering proposed for individual data.

研究分野：ファジィクラスタ分析

キーワード：ファジィクラスタリング

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

ビッグデータの持つ知の有用性とその顕在化手法であるデータ・マイニングが、社会共通の理解になると同時に、データ・マイニング自体の困難さも強く認識されるようになってかなりの年月が経っている。その中でも、 IoT の普及に伴って特に近年注目されているのがソーシャル・データ・マイニングである。データ・マイニングの中でもクラスタリングは、「外的基準を必要としない」、即ち膨大な学習データが不要な手法であり、ビッグデータマイニングには非常に効率的と言えるので、クラスタリングによるソーシャル・データ・マイニングは近年特に注目されている。ところで、データは基本的には有限次元ユークリッド空間の元、すなわち、ベクトルとして表現されるが、ソーシャルデータをはじめとして、地球表面上の気候データ、眼球運動、文書データの Bag of Words 表現、画像データの Bag of Keypoints 表現など、長さ（ノルム）が一定となるデータは枚挙にいとまがない。ノルムが一定の場合、データは球面に分布することから、それらのデータは球面データと呼ばれる。即ち、球面データ解析は直接的にソーシャルデータ解析をはじめとした数多くのデータ解析とリンクするため、球面データに対する解析手法の開発は可及的速やかに行う必要があった。

２．研究の目的

本研究課題では、球面データのため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、球面データにノイズが存在する場合に適切にクラスタリングするための技法を確立することを目的にすると共に、確立した技法を推薦システムに応用して実用に供することを目指す。これを踏まえて、次の課題を設定した:

(1) リニアデータでは、ノイズを吸収するクラスタを仮定したアルゴリズムが提案されている。これと同じ趣旨で、球面上に付置されたデータにノイズが伴う場合にも同様のアルゴリズムを開発することができるか？
(2) リニアデータでは、混合 t-分布モデルなど、混合分布の要素分布に裾野が重いものを活用してノイズ耐性のあるクラスタリングアルゴリズムが提案されている。これと同じ趣旨で、球面上に付置されたデータにノイズが伴う場合にも同様のアルゴリズムを開発することができるか？
(3) リニアデータでは、可能性クラスタリングという、帰属度の総和が必ずしも１でなくてもよい緩い制約を活用して、逐次的にクラスタを抽出するアルゴリズムが提案されている。これと同じ趣旨で、球面上に付置されたデータにノイズが伴う場合にも同様のアルゴリズムを開発することができるか？
(4) 応募者がこれまで取り組んできた知見を活かして、何か独自の観点から目的を達成するアルゴリズムを開発できないか？
(5) 開発したアルゴリズムを協調フィルタリングと組み合わせて、高精度な推薦を達成する手法を開発できないか？

３．研究の方法

次の観点からアルゴリズムの開発に取り組んだ:

(A) ノイズを吸収するクラスタを仮定する手法: 通常のクラスタの他にノイズが取り込まれることを期待するクラスタを仮定し、通常のクラスタがノイズから影響を受けないようにする手法を開発する。ノイズの一種に外れ値があり、外れ値は正常なデータが附置する箇所から遠く離れている。このことから、外れ値は通常のデータに比べて無限遠点に近いと考え、無限遠点をクラスタ中心とするようなクラスタを形成するようにアルゴリズムを開発する。
(B) 裾野の重い確率密度分布を用いる手法: リニアデータにおいて t-分布が正規分布に比べて外れ値に対して頑健であるのはその裾野が重いからである。このことを踏まえて、球面上の確率分布において裾野が重い確率分布の混合分布モデルに基づくアルゴリズムを開発する。
(C) クラスタが偏在する部分超球面を抽出する手法: k-平均法や混合正規分布は予め定めたクラスタ数分のクラスタにデータを分割する。一方で可能性クラスタリングは予め分割数を定める代わりに、逐次的にクラスタを抽出する手法である。全てのデータを抽出し終わる手前でアルゴリズムを終了することによってノイズをクラスタから除外することが期待できる。これを踏まえて球面データに対する可能性クラスタリングアルゴリズムを開発する。
(D) Gaussian カーネルなどから得られる特徴空間上の点は、データ数を次元とする空間上の球面上に附置され、空間の各軸が各個体に対応する。共クラスタリング手法は個体だけでなく特徴をもクラスタリングし、特徴集約による次元削減（ クラスタ化を阻害する特徴を無視する ）をもたらす。この機構をカーネルデータに施すことはクラスタ化を阻害する個体を無視することになることを踏まえて、ノイズ個体を排除しながらクラスタリングするアルゴリズムを開発する。
(E) 開発したクラスタリングアルゴリズムを協調フィルタリングと組み合わせて、高精度な推薦を達成する手法を開発する。

４．研究成果

上記で述べた研究方法ごとに，得られた成果を以下に示す．

(A) ノイズを吸収するクラスタを仮定する手法: アルゴリズムを開発し，その数理的特性を明らかにした．さらに、その数理的特性を実験的に確認した。実データ実験を通して雑誌論文への投稿を進めていく．

(B) 裾野の重い確率密度分布を用いる手法: 球面上の裾野が重い分布として Pearson VII 型分布があるが、この規格化定数は幾何級数で構成されていて、単一の分布における規格化定数を算出するだけでも難しいにも関わらず、その混合分布における要素分布の規格化定数を算出することは困難であった。数値的に算出することも試みたが計算量と精度のバランスを勘案して、この観点からアルゴリズムを開発することを断念した。

(C) クラスタが偏在する部分超球面を抽出する手法: 球面上で可能性クラスタリングを実行するアルゴリズムを開発した．試験的に協調フィルタリングと組み合わせた推薦システムを開発して一部の実データに対して従来法よりも高精度な結果が得られた．これを国際会議で発表した．しかし，クラスタリング・タスクについて本格的に実験を進めたところ，ノイズと通常の個体を区別するための条件を設定することが非常に難しいことが分かった。ノイズの割合、データの次元、クラスタ数、個体の分布状況など、データに依存して設定条件が大きく異なることが分かったため、実用化に向けた発展が難しいことが分かった。

(D) クラスタリングにおいてノイズ個体を排除する機構を実現する次元削減とクラスタリングを同時に実現する手法を幾つか開発したが、それらをカーネルに適用する段階にまでには至らなかった。なお，カーネルデータ解析と次元削減を組み合わせる手法開発の準備として，リニアデータに立ち戻り，次元削減するための複数の新たな手法を複数開発して，人工データ実験を通してその有効性を示すことができた成果については，国内研究会１件，国際会議１件で発表した．

(E) 開発したクラスタリングアルゴリズムを協調フィルタリングと組み合わせた結果，従来法よりも高精度な推薦を達成した．これを雑誌論文に掲載した．

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 Toshiki Ishii, Yuchi Kanzawa	４．巻 13199
２．論文標題 On Some Fuzzy Clustering Algorithms with Cluster-Wise Covariance	５．発行年 2022年
３．雑誌名 Lecture Notes in Computer Science	６．最初と最後の頁 191-203
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 Yuchi Kanzawa, Tadafumi Kondo	４．巻 -
２．論文標題 Collaborative filtering with q-divergence-based fuzzy clustering for spherical data	５．発行年 2021年
３．雑誌名 Journal of Ambient Intelligence and Humanized Computing	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） 10.1007/s12652-021-03128-6	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 －

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 石井 俊希，神澤 雄智
２．発表標題 クラスタごとにデータの分散を考慮した 幾つかのファジィクラスタリングについて
３．学会等名 第 37 回ファジィシステムシンポジウム 講演論文集
４．発表年 2021年

１．発表者名 森岡良介，神澤雄智
２．発表標題 高次元データに対して次元削減を行うファジィクラスタリング
３．学会等名 第36回ファジィシステムシンポジウム
４．発表年 2020年

１．発表者名 Yuchi Kanzawa
２．発表標題 On Collaborative Filtering with Possibilistic Clustering for Spherical Data Based on Tsallis Entropy
３．学会等名 MDAI2019（国際学会）
４．発表年 2019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

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関