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：　２６３３０２７１

研究課題名（和文）言語ベースクラスタリング技法の確立－モデルベースからの転換

研究課題名（英文）Establishment of Linguistic clustering - Conversion from Model-based Clustering

研究代表者
遠藤　靖典（ENDO, Yasunori）

筑波大学・システム情報系・教授

研究者番号：１０２６７３９６

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）：本研究課題では、従来のモデルベースによるクラスタリングから転換し、言語ベースによるクラスタリング技法の確立を目的とする。まず、従来のクラスタリングの基となっている数理モデルの抽出を行い、抽出した数理モデルの言語による記述について検討する。次に、それらの検討を踏まえ、クラスタリングにおける言語ベース（言語的記述の推論則）について考察する。その後、検討した言語ベースに基づき、新たな言語ベースに基づいたクラスタリングアルゴリズムを構築する。数値例を通じて、構築したアルゴリズムと既存のモデルベース手法との比較検討を行った後、言語ベースの数理的発展の可能性と、適用可能な実データについての検討を行う。

研究成果の概要（英文）：In this research project, we aim to establish a language-based clustering (linguistic clustering) technique by converting from conventional model-based clustering. First, we extract the mathematical models which are the basis of the conventional clustering algorithms, and discuss the description by the language of the extracted mathematical models. Next, considering the discussion, we consider language-based inference in clustering. Then, based on the language-based inference, some new clustering algorithms based on language-based inference are constructed. After comparing and examining the constructed algorithm with the conventional model-based clustering algorithm through numerical examples, we consider the possibility of mathematical development based on linguistic clustering and the applicable real data.

研究分野：　クラスタリング

キーワード：　クラスタリング　言語ベース　モデルベース　データマイニング　ビッグデータ　ファジィ推論

１．研究開始当初の背景

オンライン上に保管されたデジタル写真、ソーシャル・メディアに掲載された投稿等、多様な形態を持ち、真偽定かでない膨大な情報が、我々の周りを日々高速に行き交っている。そのようなデータをビッグデータと呼ぶが、その中でも、Facebook や Twitter をはじめとする SNS 上のデータやオンラインでの購入履歴などのソーシャルデータは、企業のマーケティングやコミュニティ発見、リスクマネジメントにおいて非常に有用との認識が広がり、そのようなソーシャルデータに対する知識発見・データマイニングの重要性は非常に高い。

データマイニングの手法の中でも、クラスタリングは特に有用な手法として、多くの分野で用いられている。その代表的な技法として、k-平均法、DBSCAN、スペクトラルクラスタリング、階層的クラスタリング等が挙げられるが、それらはすべて何らかの数理モデルを想定し、その数理モデルに沿ってクラスタリングを行っている。例えば k-平均法では予め目的関数を定義し、その目的関数を最小化（最適化）するようにデータのクラスタへの帰属を計算している。

そこで、「前もって何らかの数理モデルを想定している」という意味において、これまで提案されてきたクラスタリングをモデルベースクラスタリングと呼ぶ。モデルベースクラスタリングは、対象データの数理構造がある程度分かっている場合には、分類精度や処理速度の両面で効果を発揮するが、そうでない場合には柔軟に対応することができず、しばしば思いもよらない分類を行ったり、膨大な計算時間がかかる。特に対象がソーシャルデータの場合、明らかにモデルベースクラスタリングでの対応は困難である。

申請者が開発・発展させてきた許容の概念による不確定データに対するクラスタリング技法についても同様で、不確定データや従来は破棄せざるを得なかった欠損データの分類をも可能とした点で優れてはいるものの、不確定データを最適化の枠組みで解析的に扱う許容というモデルを予め想定している点でモデルベースであり、ソーシャルデータを扱うに適しているとは言い難い。そのため申請者は「知識融合最適化によるクラスタ解析手法の開発」を進めてきたが、別の方法論によるクラスタリング技法の確立の必要性を強く感じてきた。

制御の分野では、以前より同様の状況が生じていた。制御における古典手法では、制御対象の数理モデルを記述し、そのモデルに対応する制御系を構成するが、制御対象に予期せぬ不確定性や外乱が混入したり、もともと複雑な制御系の場合、それでは対処できない。例として、申請者らが以前より取り組んでいる鉄道車両ブレーキの滑走防止制御(ABS)を挙げる。鉄道車両ブレーキの滑走現象はモデル化が非常に困難で、同一条件においても再現が不確定であり、そのため、ABS システムの構成はほとんど不可能だった。それを可能にしたのは、滑走現象の数理モデルに基づくモデルベース制御ではなく、言語的記述の推論則（言語ベース）に基づくファジィ制御であり、これは実際に、申請者らによる特許登録と同時に、小田急の特急ロマンスカー 60000 形 MSE に搭載・実用化されている。

ファジィ制御では if-then ルールで構成される言語ベースの推論則により制御を行うので、制御対象の数理モデルを必要としないだけでなく、「スピードがとても速ければブレーキを強くかける」等の人間の直感に合った制御則を与えることができる。そのため、熟練者の制御を反映しやすいという特徴があり、制御対象が複雑・不確定であっても、柔軟な対応が可能となる。

この発想をソーシャルデータに対するクラスタリング技法の開発に応用し、クラスタリングをモデルベースではなく、「データの密度のムラが多ければクラスタの数を多くする」等の言語ベースとして構成すれば、既存のモデルベースクラスタリングではなし得なかった、大規模・複雑かつ不確定なソーシャルデータを処理できる柔軟な解析を可能とすることが期待できる。簡潔に言うと本研究課題の枢要は、クラスタリングにおけるモデルベースから言語ベースへの転換である。

２．研究の目的

そこで本研究課題では、ソーシャルデータのマイニングをターゲットとして、従来のモデルベースによるクラスタリングから転換し、言語ベースによるクラスタリング技法の確立を目的とする。

まず、従来のクラスタリングの基となっている数理モデルの抽出を行い、数理モデルを数通りの類型に分類した上で、それぞれの類型の言語による記述の可能性について検討する。次に、それらの数理モデルを参照しつつ、クラスタリングにおける言語ベース（言語的記述の推論則）の構造について考察する。その際、数あるソフトコンピューティング技法のうち、どれが言語ベースとして適切か、併せて検討する。その後、検討した言語ベースの構造に基づき、適切と判断したソフトコンピューティング技法によって、新たな言語ベースに基づいたクラスタリングアルゴリズムを構築する。ベンチマークデータを通じて、構築したアルゴリズムと既存のモデルベース手法との比較検討を行った後、言語ベースの数理的発展の可能性と、適用可能なソーシャルデータの規模についての検討を行う。

以上を踏まえ、本研究の最終的な自己評価は、

(1) 言語ベースクラスタリング技法がどこまで構築できたか、既存手法と比較した優位性はどこか、
(2) クラスタリング分野における位置付けは

どこにあるか、
(3) どの程度の規模のデータに適用できるか、実際のソーシャルデータへの適用は可能か、

の 3 点から行うこととなる。これら 1 つでも達成できれば、それだけで研究としては十分な成果と言えるが、本研究課題はそれに留まらず、より高次のレベルに目標を設定している。

３．研究の方法

本研究計画では、ソーシャルデータに対するマイニングをターゲットとした言語ベースクラスタリング技法の開発を行う。さらに、言語ベースクラスタリングの理論的発展およびソーシャルデータのマイニングへの実用化を進める。本研究計画の骨子は以下の 7 段階からなる。

(1) 従来のクラスタリングの基となっている数理モデルの抽出および類型分類
(2) クラスタリングと親和性の高い言語ベース（言語的記述の推論則）の構造の検討
(3) 必要に応じて、人間の知識に基づくデータベースの構成
(4) 検討した言語ベースの構造に基づくクラスタリング技法の構築
(5) ベンチマークデータによる開発手法の検証・既存手法との比較検討
(6) 数理的関連性を含む言語ベースクラスタリング技法の包括的発展
(7) 実用化に向けた、開発手法で適切に処理できるソーシャルデータの類型分類

詳細は以下の通りである。

(1) 研究全体の検討：本研究課題全体の詳細計画を確認する。また、協力を仰ぐ専門家、研究補助者への連絡を行い、本年度全体の計画が速やかに行われるための体制を確立する。
(2) 既存のクラスタリング技法の実装と検討：既存の代表的なモデルベースクラスタリングアルゴリズムを実装し、計算時間や有効性の検討を行う。また、それぞれのアルゴリズムにおける問題点を抽出する。
(3) 数理モデルの抽出・類型分類：実装したモデルベースクラスタリングの基となっている数理モデルの抽出を行い、類型に分類する。また、抽出した数理モデルと照らし合わせながら、各アルゴリズムの問題点の源を探る。
(4) 言語ベースの構造に関する検討：クラスタリングと親和性の高い言語ベース（言語的記述の推論則）の構造を検討する。ファジィ制御の場合、ファジィ推論というソフトコンピューティング技法の１つが非常に有効であり、本研究課題でも、言語ベースに適切なソフトコンピューティングの諸技法について検討を行う。
(5) 知識データベースの構成：必要に応じて、人間の知識に基づくデータベースの構成を行う。基本的には知識データベースを必要としないクラスタリング技法を考えているが、場合によっては知識データベースを想定しておいた方が良いケースもあり得る。そのため、知識データベースも本研究課題の考察対象とし、必要な場合には本研究課題以前の研究で得られた知見に基づいて、知識データベースの構築を行う。
(6) 言語ベースクラスタリング技法の開発：検討した言語ベースの構造に基づいて、言語ベースクラスタリング技法の開発を行う。ここで構築する技法とは狭義の意味だけではなく、言語ベースクラスタリングのフレームワークも含めたものとなる。
(7) 開発手法の検証・既存手法との比較検討：ベンチマークデータによる開発手法の検証と、既存手法との比較検討を行う。それにより、開発手法の特徴を把握することができる。もし思ったほどの結果が得られなかった場合、「言語ベースの構造の検討」、「もしくはクラスタリング技法の開発」に戻り、改めて考察し直すことになる。
(8) 言語ベースの観点からのクラスタリング技法の再体系化：開発した言語ベースクラスタリング技法と従来からのモデルベースクラスタリング技法との数理的関連性の考察を行い、言語ベースの観点からクラスタリング技法の再体系化を進める。
(9) 実データのマイニングの実用化：実際のソーシャルデータをはじめとする実データによる検討を通じて、特にデータの規模・複雑さ・不確定性の観点から、開発手法で適切に処理できるソーシャルデータの類型分類を行う。
(10) 最終的な総括・自己評価：本研究課題の最終的な包括を行う。

４．研究成果

本研究における成果は以下の通りである。

(1) 従来のモデルベースではない、言語ベース（言語的記述の推論則）に基づく、以下のような新たなクランスタリングアルゴリズムを開発した。
    ① 力学モデルに基づく階層型言語ベースクラスタリング
    ② 力学モデルに基づく非階層的言語ベースクラスタリング
(2) 言語ベースクラスタリングの開発に関連し、以下の成果を挙げた。
    ① クラスタリングに関連して、各クラスタのサイズが均等になるようなクラスタリングアルゴリズムの構築を行った。
    ② クラスタリングに関連して、ノンメトリック項を導入することによる対制約付きクラスタリングアルゴ

リズムの構築を行った。

③ クラスタリングに関連して、最適化の概念に基づくラフクラスタリングアルゴリズムの構築を行た。

④ 言語的記述の推論側に関連し、これまで進めてきた鉄道車両のブレーキのファジィ制御において、車輪・レール間に働く接線力をファジィ推論によって推定する方法を開発した。

(3) ソーシャルデータをはじめとする大規模なデータの直感的な構造把握において、データの可視化は特に重要である。そこで、本研究課題の実データの応用に関連して、KK 法に基づく新たな可視化手法、特に非対称データについての開発を行った。

最終的に本研究課題で開発した主要な手法は、言語的記述の推論則に基づく言語ベースクラスタリングであり、具体的には、上述の、力学モデルに基づく階層的および非階層的クラスタリングの 2 種のアルゴリズムである。これらのアルゴリズムは、クラスタリングにこれまでなかった言語ベースという概念を取り入れたものであり、その有効性のみならず、数学的興味から、これまでのクラスタリングのあり方を大きく変えるものとして期待されるであろう。

ただし、言語ベースクラスタリングの開発という大枠としてはその目的を達成したが、本研究の目的の１つである、実データへの応用という観点からは必ずしも十分であったとは言いがたい。この点は今後の課題といえる。また、推論側におけるファジィ集合の決め方・演算則の選び方・パラメータチューニング等の推論側に関する詳細を詰める必要があり、また、言語ベースに関する理論的構造把握にも課題が残る。

現在、ラフ集合に基づくクラスタリングの開発を行なっている。ラフ集合はブール集合より柔軟であり、ファジィ集合ほど過情報ではないため、不確実性を扱う新たなツールとして期待されている。そこで、ファジィによる言語ベースに代わり、ラフ集合に基づく言語ベースに基づくクラスタリングアルゴリズムの開発および有効性の検証が必要であろう。また、位相幾何学の概念に基づく新たなクラスタリングアルゴリズムの有効性が注目を集めており、この位相的クラスタリングへの言語ベースの導入も今後取り組むべき課題と言える。

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 19 件）

[1] Naohiko Kinoshita, Yasunori Endo, Akira Sugawara, On Hierarchical Linguistic-Based Clustering, Journal of Advanced Computational Intelligence and Intelligent Informatics , Vol.19, No.6, pp.900-906 (2015.11) (査読有).

[2] Naohiko Kinoshita, Yasunori Endo, Yuchi Kanzawa, Sadaaki Miyamoto, A Note on Non-Hierarchical Linguistic-based Clustering, Proc. of the 12th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2015), USB, pp.25-35 (Skoevde, Sweden, 2015.9.21) (査読有).

[3] Tsubasa Hirano, Yasunori Endo, Naohiko Kinoshita, Yukihiro Hamasuna, On Even-sized Clustering Algorithm Based on Optimization, Proc. of Joint 7rd International Conference on Soft Computing and Intelligent Systems and 15th International Symposium on advanced Intelligent Systems (SCIS & ISIS 2014), TP4-3-5-(3), #69 (Kitakyushu, Japan, 2014.12.4) (査読有).

[4] Yasunori Endo, Naohiko Kinoshita, Kuniaki Iwakura, Yukihiro Hamasuna, Hard and Fuzzy c-means Algorithms with Pairwise Constraints by Non-metric Terms, The 11th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2014), Springer, LNAI 8825, pp.145-157 (Tokyo, Japan, 2014.10.30) (査読有).

[5] Toru Sano, Yasunori Endo, Shin-ichi Nakazawa, Daisuke Hijikata, A Note on Estimating Tangential Force in Brake Systems for Railways, Proc. of The 11th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2014), USB, pp.108-119 (Tokyo, Japan, 2014.10.30) (査読有).

[6] Akira Sugawara, Naohiko Kinoshita, Yasunori Endo, On Linguistic-based Clustering, Proc. of The 2014 IEEE International Conference on Granular Computing (GrC 2014), G273 (Noboribetsu, Hokkaido, Japan, 2014.10.23) (査読有).

[7] Naohiko Kinoshita, Yasunori Endo, Sadaaki Miyamoto, On Some Models of Objective-based Rough Clustering, The 2014 IEEE/WIC/ACM International Conference on Web Intelligence (WIC204) (Warsaw, Poland, 2014.8.11-14) (査読有).

〔学会発表〕（計 6 件）

[1] 伊藤 成彦, 遠藤 靖典, ファジィ推論に

基づく新たな言語ベースクラスタリング, 第 32 回ファジィシステムシンポジウム (FSS2016), TC1-1 (佐賀県佐賀市, 2016.9.1).
[2] 栢分 雄基, 菅原 彬, 木下 尚彦, 遠藤 靖典, 神澤 雄智, 力学モデルに基づく階層型言語ベースクラスタリングについて, 第 31 回ファジィシステムシンポジウム (FSS2015), WB3-3 (東京都調布市, 2015.9.2 (2-4)).
[3] 栢分 雄基, 菅原 彬, 木下 尚彦, 遠藤 靖典, 神澤 雄智, 力学モデルに基づく階層型言語ベースクラスタリング, 第 41 回ファジィワークショップ講演論文集, pp.5-8 (八王子市南大沢, 2015.3.6-7).
[4] Yu Shiraishi, Akira Sugawara, Naohiko Kinoshita, Yasunori Endo, A Note on Visualization of Asymmetric Data, Doctoral Consortium Proc. of The 11th International Conference on Modeling Decisions for Artificial Intelligence (MDAI 2014), USB, pp.19-21 (Tokyo, Japan, 2014.10.30).

〔その他〕
ホームページ等
http://endo.risk.tsukuba.ac.jp/~endo/

６．研究組織

(1)研究代表者
遠藤　靖典　（ENDO YASUNORI）
筑波大学・システム情報系・教授
研究者番号：10267396

(2)研究分担者
なし

(3)連携研究者
なし

(4)研究協力者
なし