様式 C-19

科学研究費補助金研究成果報告書

平成２１年５月２２日現在

研究種目：基盤研究（C）
研究期間：2006～2008
課題番号：１８５００１７０
研究課題名（和文）知識融合最適化による不確実データに対する
クラスタリングアルゴリズムの開発
研究課題名（英文）Development of Clustering Algorithm for Uncertainty Data
Using Knowledge-fused-optimization
研究代表者
遠藤　靖典（ENDO YASUNORI）
筑波大学・大学院システム情報工学研究科・准教授
研究者番号：10267396

研究成果の概要：本研究では、不確実性のモデルの構築を行い、実際のデータと比較検討することにより、モデルの妥当性を検証した。次に、構築したモデルがデータ解析、特にクラスタリングアルゴリズムに援用可能であるかの検討を行い、現実問題に適応可能な不確実性を扱うことのできるモデルを開発した。開発したモデルのうち主なものは、許容範囲という概念と、確率類似度という測度に基づくものである。さらに、そのモデルをベースにした、不確実性を扱えるアルゴリズムの構築を行った。

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００６年度	1,300,000	0	1,300,000
２００７年度	1,000,000	300,000	1,300,000
２００８年度	1,300,000	390,000	1,690,000
年度			
年度			
総　計	3,600,000	690,000	4,290,000

研究分野：ソフトコンピューティング
科研費の分科・細目：情報学、感性情報学・ソフトコンピューティング
キーワード：ファジィ理論、クラスタリング

１．研究開始当初の背景

データ解析、その中でも特にクラスタリングに関する研究は、古くから多くの検討が行われてきたし、また現在でも、多くの研究者の興味の対象となっている（例えば文献[1～3]）。クラスタリングとは、与えられたデータをいくつかのグループに分類する手法の一つであり、パターン認識をはじめとするさまざまな場面で活用されてきた。しかし、我々は、かつてとは比較にならないほど多様化した価値観の中で、膨大な量のデータを処理しなければならない状況に立たされているにも関わらず、クラスタリングに関する研究は、以前ほどの盛隆を見せていない。

一つには、新しいクラスタリングアルゴリズムを開発せずとも、現状で十分であるという理由もあろうし、また一つには、ある程度のアイディアは出尽くし、新しいアルゴリズムを構築することに困難が伴っているという理由もあろう。しかし、現在我々が手にしているアルゴリズムでは、現在の複雑かつ大規模なデータの処理には限界があり、また、多

くの人間の主観や、リスク・チャンス発見といった、さまざまな様相を持つ概念に対応することにも無理がある。特に、不確実性を持つような情報・データに関しては、これまでのクラスタリングアルゴリズム（だけではなく、古典的な学問体系自体もそうであるが）では、実用的で適切な処理をすることはできないと本申請者は考えている。

不確実性に関する研究も古くから議論されてきたが（例えば文献[4,5]）、急速に脚光を浴びるようになってきたのはここ数年のことである。不確実性にはさまざまなフェーズがあり、そのため、不確実性の定義自体あいまいなのが現状だが、古典的な方法論だけではなかなか説明のつかない、確定的でない事象（特に人間が介在している場合が多い）に対して、それらを解析・処理する、何らかの新しい方法論の開発が望まれており、これもまた、多くの研究者の興味を引くところとなっている。ファジィ・ニューロといったソフトコンピューティングも、もともとはそれらの解析・処理を対象として生まれてきたともいえる。これら人間の持つ知識を効果的に利用することにより、不確実性への効率的なアプローチが可能になるであろう。そして、人間の持つ知識を抽出・定式化することによって得られたデータベース（知識ベースと呼ぶ）と最適化手法を融合させた知識融合最適化によってデータ処理のモデルを構成し、構成されたモデルに基づいてアルゴリズムを組めば、不確実性を伴うデータ（不確実データ）に対して、これまでよりも柔軟な解析が可能になると思われる。

２．研究の目的

そこで、本研究課題では、まず、データやそれを扱う人間に含まれる不確実性をモデル化し、それらの不確実性を効果的に扱うための知識ベースの構築を通じて、知識融合最適化によるデータ処理のモデルを構成し、構成したモデルに基づいて、クラスタリングアルゴリズムを中心とするデータ解析のためのツールの開発を行うことを目的とする。

まず、不確実性の概念を検討しなおす。従来から提案されているモデルのどれが適切かを議論し、適切なモデルがない場合には、現実に即した不確実性のモデルを構築する。

次に、検討した不確実性のモデルをもとに、不確実な事象に対して人間はどのように判断し、対処しているのかを検討する。アンケート・インタビュー・文献調査等を通じて、不確実性に対して人間が持っている知識を抽出し、さまざまなフェーズ毎に分類する。そして、それを知識ベースとしてデータベース化する。

次に、さまざまな最適化手法と知識ベースを組み合わせ、知識ベースと融合させるにはどのような最適化手法が適しているかを検討する。最適化手法には、古くは線形計画法から、NP困難な問題を解くための焼きなまし法、ローカル･サーチ、遺伝的アルゴリズム等、多くの方法が存在する。これらの手法と知識ベースとをどのように組み合わせるかについて議論する。

以上の検討をもとに、クラスタリングアルゴリズムについて検討を行う。これまで提案されてきたクラスタリングアルゴリズムを援用することによって構成できるかの検討を行い、援用できない場合には、全く新たなアルゴリズムを構成する。

また、実際のデータを用いて、構築したアルゴリズムの検証を行う。

３．研究の方法

以下に挙げる各研究内容について、番号順に実施する。

（１） 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） 高速のデータ解析用コンピュータを複数台購入し、すでに提案されている不確実性を伴うデータに対する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対する問題点を抽出する。数値計算に当たっては、言語処理プログラムC++を主に使用して行う予定である。また、コンピュータはLANを利用して有機的に結合し、シミュレーションを行う。

（３） 不確実性のモデルを検討しなおす。不確実性のモデルにはさまざまな考え方があるが、それらがすべて現実に即したモデルとは言えない。そこで、関連分野の専門家の指導を受け、現実に即したモデルを検討する。

（４） 検討したモデルのうち、どのモデルがデータ解析、特にクラスタリングアルゴリズムに援用可能であるかの検討を行う。また、適切なモデルがない場合には、現実に即した不確実性のモデ

ルを構築する。

（５） データ解析において人間が利用する知識に関する調査を開始する。アンケート・インタビュー・被験者によるデータ分類テスト等を通じて、不確実性の処理に対して人間が持っている知識を抽出し、さまざまなフェーズ毎に分類する。得られた知見をもとに、知識ベースとしてデータベース化する。

（６） 知識ベースの開発は、現在、webテクノロジーと密接に結びついており、データベース管理システムを適切に設定することと、XML（eXtensible Markup Language）の枠組みを効果的に利用することが必要となる。そこで、Microsoft Access等のデータベース開発ソフトウェアを利用して、知識ベースの開発を開始する。

（７） 前年度に引き続き、データ解析時の人間の知識に関する調査および知識ベースの開発を継続して行う。

（８） 知識ベースを最適化手法の枠組みの中で利用するためには、知識の変換と融合が必要となる。平成19年度以降は、知識の変換・融合方法についての実行可能性を考察する一方、マニュアルでの変換、最適化における融合を開始する。

（９） データ解析に適切な最適化手法の検討を開始する。最適化手法には、古くは線形計画法から、NP困難な問題を解くための焼きなまし法、ローカル・サーチ、遺伝的アルゴリズム等の発見的手法等、多くの方法が存在し、これまで非階層的クラスタリングのベースとなってきた数理計画法にとどまらない。そこで、利用可能な最適化手法に関する文献調査を行い、どの最適化手法が不確実データを柔軟に扱えるかについて、数値計算を通じて検討する。

（１０） 検討したさまざまな最適化手法と知識ベースを組み合わせ、知識ベースと融合させるにはどのような最適化手法が適しているかを検討する。最適化手法には、これらの手法と知識ベースとをどのように組み合わせるかについて議論する。

（１１） 以上の考察をもとに、知識融合最適化による不確実データに対するクラスタリングアルゴリズムの開発を行う。これまで提案されてきたクラスタリングアルゴリズムを援用することによって構成できるかの検討を行い、援用できない場合には、全く新たなアルゴリズムを構成する。

（１２） 実際のデータを用いて、構築したアルゴリズムの検証を行う。

４．研究成果

まず、2006年度初めに、本研究課題全体の計画の確認を行った。また、協力を仰ぐ研究者への連絡を行い、本年度の計画の速やかな実施のための体制の確立を行った。

続いて、計算機上で、すでに提案されている不確実性を伴うデータに対するクラスタリングアルゴリズムを実装し、計算時間や計算時間や有効性の検討を行った。また、それぞれのアルゴリズムに対する問題点を抽出した。その結果、従来のアルゴリズムに関して、

（１）データ間の類似度・非類似度とは別に、不確実性を扱うための類似度・非類似度を用意する必要がある。

（２）不確実性をあらわす領域の境界しか対象としていない。

（３）一般に計算時間が大きい。

等の問題点があることがわかった。

以上のもとに、不確実性のモデルの検討を行い、特に実際のデータのクラスタリングに援用可能なアルゴリズムの構築を行った。

まず、許容ベクトルという新たな概念を導入することによって、不確実性を伴うデータに対するクラスタリングの定式化を行った。これは、不等式制約化での非線形目的関数の最適化問題に帰着する。次いで、この問題を解くことによって得られた解をもとに、アルゴリズムを新たに構築した。特に注目すべきは、これまで円形でしか表現できなかった不確実性を、矩形での表現も可能にした点にある。それにより、実データに対する応用範囲が驚異的に広がった。

また、データ解析における人間の知識に関する調査を行い、知識ベースとしてのデータ化を開始した。特に、貨物の積み付け問題に対してクラスタリング手法の援用を行い、その中で、クラスタリングが組み合わせ問題の中でどのように有効に働くかの検証を行った。

2007年度は、まず、2006年度に引き続き、関連分野の専門家のアドバイスを受け、前年度において構築した不確実性のモデルのうち、どのモデルがデータ解析、特にクラスタリングアルゴリズムに援用可能であるかの検討を行い、修正・新たなモデルの提案を行った

。特に、不確実性を扱う場合、前年度提案したモデルでは、データの各属性毎における不確実性を独立して扱うことができなかったため、そのような形のデータも扱えるように、モデルの修正を行った。

また、2006年度に引き続き、データ解析において人間が利用する知識に関する調査を継続して行った。アンケート・インタビュー・被験者によるデータ分類テスト等を通じて、不確実性の処理に対して人間が持っている知識を抽出し、さまざまなフェーズ毎に分類した。しかし、人間の持っている知識は非常に多様であり、判断も人によって大きく異なるため、この問題については引き続き研究が必要である。

さらに、知識ベースを最適化手法の枠組みの中で利用するためには、知識の変換と融合が必要となるので、2007年度は、知識の変換・融合方法についての実行可能性を考察する一方、マニュアルでの変換、最適化における融合を開始した。
また、データ解析に適切な最適化手法の検討を開始した。特に、知的情報処理で用いられるコサイン相関と、一般の非類似度である自乗距離との長所を併せ持つ、射影相関という独自の非類似度を開発し、階層的クラスタリングに適用させることにより、独自の最適化手法の開発を行った。

2008 年度は、まず、前年度に引き続きデータ解析において人間が利用する知識に関する調査を行った。ウェブやメディア上で公開されている情報等を通じて、不確実性の処理、特にリスク認知に対する人間の処理モデルの構築を行い、実際のデータと比較検討することにより、モデルの妥当性を検証した。次に、構築したモデルがデータ解析、特にクラスタリングアルゴリズムに援用可能であるかの検討を行い、現実問題に適応可能な不確実性を扱うことのできるモデルを開発した。
これらのモデルを用いて、アルゴリズムのプロトタイプを構築した。さらに、そのモデルをベースにした、不確実性を扱えるアルゴリズムの構築を行った。これらの成果を、コンピュータによって、実データを用いた構築アルゴリズムの有効性を検証したとともに、国内外の研究会・国際会議、ジャーナルへの投稿を通じて発表した。

最終的に本研究で開発したモデルの主要なツールは、許容範囲という概念と、確率類似度という測度に基づくものである。許容範囲とは、許容ベクトルという概念をクラスタリングの基本となる目的関数に組み込むことにより、データの持つ不確実性を最適化の枠組みで議論することを可能にした新たな手法であり、確率類似度とは、確率距離をもとにして構成された、データ間の類似度を確率の枠組みで扱うことのできる新たな測度である。これらの新概念は、その有効性のみならず、数学的興味から、これまでの不確実性の扱いを大きく変えるものとして期待されている。

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３２件）

① Yasunori Endo, Yasushi Hasegawa, Yukihiro Hamasuna, Sadaaki Miyamoto, *Fuzzy c-means for Data with Rectangular Maximum Tolerance Range,* Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol.12, No.5, pp.461-466 (2008) (査読有).
② Yuchi Kanzawa, Yasunori Endo, Sadaaki Miyamoto, *Fuzzy c-Means Algorithms for Data with Tolerance based on Opposite Criterions,* IEICE Trans. Fundamentals, Vol.E90-A, No.10, pp.2194-2202 (2007) (査読有).
③ 遠藤 靖典, 半澤 光希, 濱砂 幸裕, *グループ化を用いた貨物積み付けにおけるメタ戦略アルゴリズム,* 日本知能情報ファジィ学会誌, Vol.18, No.6, pp.859-866 (2007) (査読有).
④ Ryuichi Murata, Yasunori Endo, Hideyuki Haruyama, Sadaaki Miyamoto, *On Fuzzy c-Means for Data with Tolerance,* Journal of Advanced Computational Intelligence and Intelligent Informatics (JACIII), Vol.10, No.5, pp.673-681 (2006) (査読有).

〔学会発表〕（計９件）

① 遠藤 靖典, 濱砂 幸裕, *許容範囲付きデータに対する階層的クラスタリングについて,* 第 24 回ファジィシステムシンポジウム (大阪府、阪南大学, 2008.9.4).
② 濱砂 幸裕, 遠藤 靖典, *Tolerant Fuzzy c-Means について,* 第 24 回ファジィシステムシンポジウム (大阪府, 阪南大学, 2008.9.3).
③ 濱砂 幸裕, 遠藤 靖典, 宮本 定明, 長谷川 康, *許容範囲付きデータに対するハードクラスタリングについて,* 第 23 回ファジィシステムシンポジウム (名

古屋，名城大学，2007.8.30).

④ 長谷川 康，遠藤 靖典，濱砂 幸裕，宮本 定明，*超直方体で定義された許容範囲付きデータに対するファジィ c-平均法について*，第 23 回ファジィシステムシンポジウム（名古屋，名城大学，2007.8.30).

〔図書〕（計１件）

① 遠藤 靖典，岡本 健，掛谷 英紀，岡島 敬一，庄司 学，伊藤 誠，*リスク工学の基礎*，コロナ社，pp.1-24（2008).

６．研究組織

(1)研究代表者

遠藤　靖典（ENDO YASUNORI）

筑波大学・大学院システム情報工学研究科・准教授

研究者番号：10267396

(2)研究分担者

なし

(3)連携研究者

なし