科学研究費助成事業　　研究成果報告書

平成　２６　年　　５　月　３０　日現在

機関番号：　１２１０２
研究種目：　基盤研究(C)
研究期間：　2011～2013
課題番号：　２３５００３４３

研究課題名（和文）高次シンボリックデータに対するクラスターワイズ手法の開発とその応用

研究課題名（英文）Development of clusterwise methods for high dimensional symbolic data and its applications

研究代表者
イリチュ　美佳（佐藤美佳）（SATO-ILIC, MIKA）

筑波大学・システム情報系・教授

研究者番号：６０２６９２１４

交付決定額（研究期間全体）：（直接経費）　　3,900,000円、（間接経費）　　1,170,000円

研究成果の概要（和文）：得られたデータの次元が標本数に比べて遥かに大きいデータ（高次元小標本データ）に、既存の統計手法を適用すると、有効な結果が得られないという問題がある。この問題を解決するための解析手法は、シンボリックデータに対しては、未だ開発されていない。そこで、高次元小標本シンボリックデータに対する解析手法を開発し、データの多様性を考慮した新たな知識発見手法の提案と、実用化に向けた性能評価を行った。

研究成果の概要（英文）：A problem with conventional statistical analyses for high dimension low sample-size data is that an efficient solution will not be obtained. Methods to solve this problem for symbolic data have not been proposed. Therefore, this research develops methods for high dimension low sample-size symbolic data and proposes a new knowledge discovery method considering a variety of data with an evaluation of the performance of this method.

研究分野：　統計科学

科研費の分科・細目：　情報学・統計科学

キーワード：　分類　パターン認識　シンボリックデータ　高次元小標本データ

１．研究開始当初の背景

計算機技術の発展が、より大規模なデータの解析を可能としてから、データ解析手法の開発は、目覚ましく進展してきた。しかし、近年においては、計算機を駆使した統計科学から、より広範囲なデータの型を扱う知識発見の分野へのパラダイムシフトがなされている。この新たなパラダイムは、主として、機械学習やデータマイニングの実用的技術の基盤となる考えとして進展してきたが、近年になり、これらの技術を取り入れ、データの背景にある現象に統計的な仮定をした新たな研究が“Statistical Learning”の名の下に開発されている。シンボリックデータ解析は、その中核を担う解析手法である。シンボリックデータ解析の本質は、従来のデータの型からより多様なデータの型を扱う点であり、データの概念を多層化した構造によって解釈することよって、これまで解析の対象とされていなかった種類のデータをも解析可能としていることにある。

一方、近年、多変量解析の分野において、得られたデータの次元が標本数に比べて遥かに大きいデータ（高次元小標本データ）に対する解析が問題とされている。これらのデータは、遺伝子解析におけるマイクロアレイデータや脳科学分野における脳波データ等、幅広く存在する。このデータに、既存の統計手法を適用すると、いわゆる“次元の呪い“により、有効な結果が得られないという問題がある。この問題を解決するために、様々な解析法が提案されている。

しかし、これらの解析法は、従来のデータの型に対する手法にとどまり、シンボリックデータに対しては、未だ開発されていない。

そこで、本研究では、高次元小標本シンボリックデータに対する解析手法を開発しようとするものであり、その位置づけは、データの多様性を考慮した新たな知識発見手法の提案と、実用化に向けた性能評価である。

２．研究の目的

本研究の目的は、高次元小標本シンボリックデータに対する解析手法を提案することである。まず、ファジィクラスタリングを適切にシンボリックデータに適用することにより得られた分類構造を導入した次元縮約法を提案する。次に、これらの提案手法を実データに適用することにより、その性能を評価し、実用化を図る。

シンボリックデータの表現法は、ノイズを伴うデータの不確定性の扱いについて、汎用性が高いことが知られている。また、本研究で用いるファジィクラスタリングは、ロバスト性が高いことが知られている。このため、これらの利点を融合した本手法は、これまでのクラスタリングに比べて、よりロバストで汎用性の高い解が得られることが期待される。更に、このクラスタリングを利用し次元縮約法を展開することにより、この性能を有する次元縮約法の開発が期待でき、縮約に伴う計算時間の縮小が期待される。

３．研究の方法

高次元小標本のシンボリックデータに対するクラスターワイズ手法を開発し、開発した手法の各種性能を精査するとともに、実用化を図った。高次元小標本データに、既存の統計手法を適用すると、有効な結果が得られないという問題を解決するために、平成２３年度は、主に、データの次元（属性）を分類して次元を縮小する方法を開発し、その性能の評価を行った。この方法の特性は、シンボリックデータの考えを導入し、これまで単一の次元として考えていた解析に、カテゴリー（シンボル）としての次元表現を取り入れ、その解析法を開発したことである。

次に、平成２４年度は、元のデータ構造を分類構造へ変換することにより次元の縮小を図る方法を開発し、その評価を行った。その特性は、ファジィクラスタリング結果から得

られる分類構造間の相関の概念を導入することで、高次元小標本データの変数間の新たな相関構造を定義し、従来の手法の適用を可能にしたことである。

さらに、平成２５年度においては、平成２３年度、２４年度に開発した方法の応用と成果報告に力を入れた。

４．研究成果

（１） 研究の主な成果

本研究の具体的成果を要約すれば、これまで、高次元小標本データでは、固有値の一致性の問題から解が得られないという問題について、クラスター構造間の相関を導入することにより、この問題を解決したことにある。

また、クラスター構造の類似性による変数選択基準を提案し、それにより、変数の部分空間によるクラスタリング手法を開発したことにある。これまでの変数の扱いは、変数のカテゴリーという考えが含まれていない。その主な原因は、従来のデータ解析においては、変数をベクトル空間の次元とみなして数学的解析を行ってきたことにある。しかし、次元が非常に大きい場合には、それをカテゴリー化して次元を縮約する操作をする必要がある。そこで、シンボリックデータの考えを導入し、これまで単一の次元として考えていた解析に、カテゴリー（シンボル）としての次元表現を取り入れ、その解析法を開発したことである。

上記の内容を実現するために、本研究では、二つの次元縮小法を提案した。一つは、データの次元（属性）を分類して次元を縮小する方法であり、他方は、元のデータ構造を分類構造へ変換することにより次元の縮小を図る方法である。

その結果、理論的実績として、ファジィクラスタリング結果から得られる分類構造間の相関は、研究代表者らがすでに提案したファジィ自己類似性の特性を使って説明できることを明らかにした。この特性とは、提案した相関は、データの類似性に存在するノイズを学習して除去し、データの分類による説明力を取り入れることである。また、数値例から、遺伝子に関する実データで、有効な結果を得た。

（２） 得られた成果の国内外におけるインパクト

平成２３年度、平成２４年度の２年間の研究成果に対して、米国、シカゴ、及びワシントンD.C.において開催された２度の国際会議で２年連続の下記の学術的賞を受賞した。

・ Best Theoretical Paper Award , M. Sato-Ilic, On Fuzzy Clustering Based Correlation, Procedia Computer Sciences, Elsevier, Vol. 12, pp. 230-235, Washington, D.C., USA, 2012.

・ Best Theoretical Paper Award, M. Sato-Ilic, Symbolic Clustering with Interval-Valued Data, Procedia Computer Sciences, Elsevier, Vol. 6, pp. 358-363, Chicago, USA, 2011.

また国際会議KES-IDT2012で基調講演を行い、パリ大学で２度の招待講演を行った。その詳細は、下記の通りである。

・Cluster-based Scaling for Symbolic Data and its Applications in Decision Making, Keynote speech at 4th International Conference on Intelligent Decision Technologies (KES-IDT2012), Gifu, Japan, 2012.

・Cluster Harnessing Analyses for High Dimension Low Sample-Size Data, Invited talk at Department of Databases and Machine Learning, LIP6, University of Paris (UPMC),

Paris, France, 2012.
・Intelligent Symbolic Clustering through High Dimensional Space, Invited talk at Department of Databases and Machine Learning, LIP6, University of Paris (UPMC), Paris, France, 2011.

さらに、平成２５年度においては、平成２３年度、平成２４年度に開発した方法の応用と成果報告に力を入れた。特に、協働学習に関するＷｅｂログデータの解析についての応用研究で、米国、バルチモアにおいて、下記の学術的賞を受賞した。

・ Best Theoretical Paper Award, M. Sato-Ilic, P. Ilic, Fuzzy Dissimilarity Based Multidimensional Scaling and Its Application to Collaborative Learning Data, Procedia Computer Sciences, Elsevier, Vol. 20, pp. 490-495, Baltimore, USA, 2013.

さらに、２件の章の執筆や、国内外の学会における招待講演も行った。また、研究結果を論文にまとめた。

（３） 今後の展望

本研究では、クラスタリングに基づく相関やカテゴリーとしての次元等の新たな数学的概念を定義し、これらに基づいた新しいデータ解析法を提案した。今後の展望は、この方法論の数学的性質を精査し、一般化手法を提案することである。その一方で、パラメータの設定等に対するモデルの挙動の評価等、各種のシミュレーションに基づく細部の検証が必要である。

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 22 件）

① M. Sato-Ilic, P. Ilic, Fuzzy Dissimilarity Based Multidimensional Scaling and Its Application to Collaborative Learning Data, Procedia Computer Sciences, Elsevier, Vol. 20, pp. 490-495, 査読有, 2013 (Best Theoretical Paper Award受賞)
② M. Sato-Ilic, On Fuzzy Clustering Based Correlation, Procedia Computer Sciences, Elsevier, Vol. 12, pp. 230-235, 査読有, 2012 (Best Theoretical Paper Award受賞)
③ M. Sato-Ilic, Symbolic Clustering with Interval-Valued Data, Procedia Computer Sciences, Elsevier, Vol. 6, pp. 358-363, 査読有, 2011 (Best Theoretical Paper Award受賞)

〔学会発表〕（計 31 件）

① M. Sato-Ilic, Cluster-based Scaling for Symbolic Data and its Applications in Decision Making, 4th International Conference on Intelligent Decision Technologies (KES-IDT2012)（基調講演）, 2012年5月25日, 長良川国際会議場, 岐阜県岐阜市
② M. Sato-Ilic, Cluster Harnessing Analyses for High Dimension Low Sample-Size Data, Invited talk at Department of Databases and Machine Learning, LIP6, University of Paris (UPMC)（招待講演）, 2012年3月22日, パリ大学, パリ, フランス
③ M. Sato-Ilic, Intelligent Symbolic Clustering through High Dimensional Space, Invited talk at Department of Databases and Machine Learning, LIP6, University of Paris (UPMC)（招待講演）, 2011年4月8日, パリ大学, パリ, フランス

６．研究組織

(1)研究代表者
イリチュ　美佳（佐藤　美佳）（SATO-ILIC MIKA）
筑波大学・システム情報系・教授
研究者番号：60269214

(2)研究分担者
青嶋　誠（AOSHIMA MAKOTO）
筑波大学・数理物質系・教授
研究者番号： 90246679
清水　信夫（SHIMIZU NOBUO）
統計数理研究所・サービス科学研究センター・助教
研究者番号：00332130

(3)連携研究者
田中　一男（TANAKA KAZUO）
電気通信大学・情報工学研究科・教授
研究者番号：00227125