

分布値に対するシンボリックデータ解析法に関する研究

学位論文内容の要旨

近年の情報環境の発展により、収集されるデータが複雑化しつつあり、そこから有益な情報を適切に抽出するための解析手法の開発が期待されている。しかし既存の方法では複雑なデータを適切に扱うことが難しい事例も多い。

そのような背景を契機として、シンボリックデータ解析が Diday (1988) により提唱された。シンボリックデータ解析法における重要な特徴は、解析対象の記述の柔軟性である。一般的な多次元データ解析において、解析の対象となるデータは複数の個体であり、各個体は単一の値、もしくはベクトルの形で記述されることが多い。これに対して、シンボリックデータ解析法では個体の集まりそのものを実際の解析対象とし、その記述として、区間、分布など、従来の手法では直接扱えない形を考えることができる。

シンボリックデータ解析法における多くの先行研究において、種々の多次元データ解析手法が拡張されてきた。しかし、その多くは主に区間で記述された解析対象を扱っている。区間の場合、最大値と最小値のみに注目することとなり、内在する情報を適切に扱えないことが多い。例えば区間同士の非類似度として提案されているハウスドルフ距離では、区間の記述が等しければ、区間内のデータの内容によらず、距離は 0 とされる。同様にシンボリック主成分分析において、最大値と最小値をそのまま用いる頂点法や、最大値と最小値から求めた midpoint のみを用いる midpoint 法などが提案されているが、解析の途中で区間幅が大きくなり、適切な次元縮小とならないことがある。

また複雑なデータの典型例として、個人の年収や消費について、個人別に見るのではなく、住んでいる都市別にとらえることとし、さらに都市を要素として都道府県を構成するようなケースを考えることができる。このような、個体により構成されているグループが、さらにいくつかの集まりを構成するような場合を、本論文ではグループ構造と呼ぶ。グループ構造を対象とした先行研究では各グループに対して同一の条件をおいているものが報告されているが、解析対象に対する柔軟性という観点から考えた、新たな展開が必要である。

以上の背景から、本論文では、区間で記述されている解析対象ではなく、区間内の情報も含めうる記述として、分布を対象としたシンボリックデータ解析手法の構築を目的とする。具体的には、データマイニング等でのニーズが高い、分類と次元縮小について重点を置くこととし、分布を要素としたシンボリック階層的クラスタリングと、シンボリック主成分分析法の展開形の一つとして、分布で記述された個体に基づくグループ構造に対する解析法を提案する。

本論文は、5つの章から構成される。各章の概要は以下の通りである。

第1章では、本論文の背景、目的、および構成について説明している。

第2章では、本研究の基礎となるシンボリックデータ解析法について、その概念と用語の定義を与え、解析対象の記述について説明している。具体的には、従来の多次元データ解析法において一般的な解析対象である個体を第1レベル、個体の集まりを第2レベルとし、これをコンセプト(クラス、カテゴリー)と呼びあらし、先行研究として、区間で記述されたコンセプトに対するクラスタリングと主成分分析について説明している。また、事前情報に基づきグループ構造を持っていることが分かっている解析対象に対する次元縮小の先行研究として、個体を解析対象とし、そのグループ構造を考慮しながら次元縮小を行う手法である共通主成分モデル(Flury, 1988)について説明している。

第3章では、コンセプトが分布で記述されている場合のクラスタリングを提案している。具体的には階層クラスタリング手法を拡張し、分布間の非類似度をシンメトリックカルバックライブラー情報量で定義し、非類似度の小さいものから順次結合させていくアルゴリズムを提案している。このとき、結合されたクラスターによる新たな記述は混合分布となるが、初期状態として与えられた解析対象が正規分布で記述されている場合、クラスター結合後の非類似度として混合正規分布間のシンメトリックカルバックライブラー情報量を求める必要がある。しかしながらこれは解析的に求めることができないため、一般的にはモンテカルロ法などによるところ、本論文では計算コストの観点から Unscented Transform 法(Goldberger, 2004)を用いて近似値による定義を与えている。提案した手法の有効性を検証するためにエジプト人の頭がい骨データ、アイリスデータ、スイス銀行偽札データに基づいた分布によるデータセットに対して提案手法を適用することにより、妥当性を確認している。

第4章では、正規分布を要素とするグループ構造に対する主成分分析を提案している。先行研究である共通主成分モデルでは、すべてのグループに対して、主成分ベクトルは同一であるという条件のもとで次元縮小を行っていたが、現実的には非常に厳しい条件と考えられる。そこで、複数のグループにおいて、各グループの分散共分散行列に対しての制約を置かないモデルを考え、さらに先行研究におけるデータの基本構成要素である個体を分布で与えるような拡張モデルを提案する。また、提案モデルに基づいて得られた主成分を用いてグループ間の非類似度を定義し、多次元尺度構成法を適用することで、低次元での可視化を行う。さらに提案モデルにおける尤度関数を導出し、最尤法により尤度方程式を導出する。この尤度方程式は F-G アルゴリズム(Flury and Neuenschwander, 1995)によって解析的に解くことが可能である。これによって、主成分スコアや寄与率に関する推定や検定を行うことができる理論的な基盤を構築した。数値実験として、15個の4次元正規分布が5つのグループを構成しているデータセットに対し、提案モデルを適用することにより、これらの妥当性を確認した。

最後に第5章で、本論文の総括とともに成果について述べている。

学位論文審査の要旨

主 査 教 授 水 田 正 弘

副 査 教 授 栗 原 正 仁

副 査 教 授 赤 間 清

副 査 准教授 南 弘 征

学 位 論 文 題 名

分布値に対するシンボリックデータ解析法に関する研究

データが多量化, 複雑化する現代において, そこから有益な情報を適切に抽出することは重要な課題である。しかし古典的な多次元データ解析では, 複雑なデータを適切に扱うことが困難である事例も多い。

複雑なデータに対する解析法として, シンボリックデータ解析法が Diday (1988) により提唱されている。シンボリックデータ解析法は, 解析対象の記述の柔軟性が大きな特徴である。一般的な多次元データ解析において, 解析の対象となるデータは複数の個体であり, 各個体は単一の値, もしくはベクトルの形で記述されることが多い。これに対して, シンボリックデータ解析法では個体の集まりそのものを実際の解析対象とし, その記述として, 区間, 分布など, 従来の手法では直接扱えない形を考えることができる。しかし, シンボリックデータ解析法における先行研究の多くは, 主に区間で記述された解析対象を扱っており, 特徴を充分活用しているとは言い難い。

このような背景から, 本論文では, 区間で記述されたデータではなく, 区間内の情報も含めうる分布で記述されたデータの解析法を扱っている。特に, 分類と次元縮小について新たな解析法を提案している。具体的には, 分布を要素としたシンボリック階層的クラスタリングと, 個体が分布で記述され, かつグループ構成を有するデータに対するシンボリック主成分分析法の提案を行っている。

本論文の成果は以下のとおりである。

解析対象が分布で記述されているデータに対し, 分布間の非類似度をシンメトリックカルバックライブラー情報量で定義し, 非類似度の小さいものから順次結合させていく階層的シンボリッククラスタ分析のアルゴリズムを提案している。ここで, 結合されたクラスターを表す新たな記述は混合分布となるが, 初期状態として与えられた解析対象が正規分布で記述されている場合, クラスタ結合後の非類似度として混合正規分布間のシンメトリックカルバックライブラー情報量を求める必要がある。しかしながらこれは解析的に求めることができないため, 本論文では計算コストの観点から Unscented Transform 法 (Goldberger, 2004) を用いて近似値による定義を与えている。ま

た、数値実験として、エジプト人の頭がい骨データ、アイリスデータ、スイス銀行偽札データに基づき、分布によるデータセットと区間によるデータセットを作成し、提案手法と先行研究である区間値クラスタリングの結果について比較検討を行うことにより、提案手法の有効性を確認している。

また、次元縮小に関しては、解析対象が分布で記述され、複数のグループを構成している場合について、各グループの分散共分散行列に対する制約を置かないモデルを考え、さらにデータの基本構成要素である個体を分布で与えるような主成分分析の提案を行っている。加えて、提案モデルに基づいて得られた主成分スコアを用いてグループ間の非類似度を定義し、多次元尺度構成法を適用することで、低次元での可視化を行っている。また、提案モデルにおける尤度関数を導出し、最尤法により尤度方程式を導出することにより、主成分スコアや寄与率に関する推定や検定を行うための理論的な基盤を構築している。数値実験として、4次元正規分布が複数のグループを構成しているデータセットに対して、提案モデルを適用し、これらの有効性を確認している。

これを要するに、著者は、分布値を対象とした複数のシンボリックデータ解析手法を提案、発展させることにより、大量かつ複雑なデータから有益な情報を抽出するための手法について有効な知見を得たものであり、情報科学ならびに計算機統計学に貢献するところ大なるものがある。よって著者は北海道大学博士(情報科学)の学位を授与される資格あるものと認める。