HOKKAIDO UNIVERSITY

Title	高度集約的シンボリックデータに対する解析手法に関する研究［論文内容及び審査の要旨］
Author(s)	松井, 佑介
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第11605号
Issue Date	2014-12-25
Doc URL	https://hdl.handle.net/2115/57732
Rights(URL)	https://creativecommons.org/licenses/by-nc-sa/2.1/jp/
Type	doctoral thesis
File Information	Yusuke_Matsui_abstract.pdf, 論文内容の要旨

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

学 位 論 文 内 容 の 要 旨

博士の専攻分野の名称　博士（情報科学）　氏名　松井　佑介

学 位 論 文 題 名

高度集約的シンボリックデータに対する解析手法に関する研究
（A Study on Analysis of Advanced Aggregated Symbolic Data）

情報技術の著しい発達に伴い、さまざまな分野における収集方法の多様化から、莫大なデータの蓄積が可能となっている。例えば、インターネットを流れる通信に関する膨大なログデータや、センサーネットワークから高頻度で得られるセンシングデータなどが代表的である。このようなデータは、容量、更新頻度、多様性などを主な特徴とするビッグデータとして、注目を集めている。しかし、統計科学において代表的な解析法をビッグデータへそのまま適用することを想定すると、時としてこれらの特徴が弊害となる可能性が示唆され、新たなアプローチが必要である。そこで、本論文では、これまでの統計解析において扱われることが少なかった、多様性に着目した新たなデータ表現とその解析法について提案する。

データをどのような形で表現するかは、解析を行うために重要な要素の一つである。従来の多変量解析では、各解析対象を実数値多次元ベクトルによって表す記述が用いられてきたが、大規模データに対しては、そのまま用いることが適切ではない場合も多い。Diday が提唱したシンボリックデータ解析法 (SDA) において、解析対象とそれらを表す記述を一体として、内部変動を含むオブジェクトとして捉え、それを対象とする新たなデータ解析の枠組みが示されている。これを用いることで、内部変動を含みうる集約的な形にデータ表現を拡張することが可能となる。従来の SDA の研究で用いられるデータ表現として、主に多値や区間値、モダル値が扱われ、これらを対象とする回帰分析法や判別分析法、主成分分析法、クラスター分析法、多次元尺度構成法などの多変量解析手法が拡張されてきた。しかし、多様な大規模データに鑑み、より複雑な集約的データ表現に対する解析法が必要と考えられる。

本論文では、従来の SDA で主に扱われてきた多値データ、区間値データなどのシンボリックデータを、基本的な集約的表現と位置づけ、基礎集約的シンボリックデータと呼ぶ。さらに、より複雑な集約的表現を、高度集約的シンボリックデータとして新たに提案し、それらに対する解析法を実現する。具体的には、分布値データ、関数データ、さらに尺度混在型データに対する高度集約的シンボリックデータ解析法を提案する。

本論文の構成は以下の通りである。まず第 1 章では、本研究の背景とその目的について述べる。第 2 章では準備のため、SDA の基本的概念と基礎集約的シンボリックデータに対する解析法を概説するとともに、本論文を通じ、実解析例として具体的に扱う、複数のサイト間におけるインターネット上の通信遅延時間を測定したデータ、および、東京電力福島第一原子力発電所で発生した事故を原因として環境中に放出された放射性物質の空間線量率を測定したデータについて、概要を説明する。

第 3 章では、解析対象毎に複数の測定値が観測されている場合に、統計的なばらつきを併せて表すことのできる分布値データを対象とした新たな解析法を提案する。具体的には、集約的シンボリックデータ間の関係性を分布値によって表す分布値非類似度を導入し、関連性の高いデータ群を同定するための階層型シンボリッククラスター分析法を提案する。シミュレーションによる既知の手法との対比を通じて本手法の優位性を示すとともに、実際の通信遅延データへ本手法を適用し、その結果と解釈から本手法の有用性を示す。

第 4 章では、Ramsay らにより提唱され、解析対象毎に連続的に測定される大量のデータを関数の形で表現する関数データを対象とした、高度集約的シンボリックデータ解析法について論じる。具体的には、部分的な定義域を移動させながら関数クラスター分析を行う移動関数 k-means 法を提案し、従来は捉えにくかった部分的な特徴を反映した形で、クラスター構成の変化を分析できることをシミュレーションにより示す。さらに、第 2 章で触れた、複数のセンサーにより一定期間、短い時間間隔で測定された放射性物質の空間線量率データに対して提案手法を適用した結果から、クラスター構成が大きく変化するような期間の同定が可能となることを示した上で、地理的情報を考慮した解釈の検討を通じ、本手法の有用性を示す。

第 5 章では、値域が多次元である関数データを多次元値関数データとして定義し高度集約的シンボリックデータとして捉え、同時に異常とみなしうる領域を同定する解析手法を提案する。各関数データにおける、部分的な領域での異常に対して、L_0 正則化を用いた平滑化処理を施すことで効率的に検出する方法を示し、局所的なノイズに影響されにくく、大域的な異常領域を検出できるなど、シミュレーションを通じて本手法の有効性を示す。さらに、第 2 章で触れた実際の大規模環境データにおいて、多数のセンサーで同時に発生した異常領域を同定する問題に本手法を適用し、同定された期間と気象データを用いた解釈を通じて、本手法の有用性を示す。

第 6 章では、各解析対象の特徴が複数の記述により表される尺度混在型データを、高度集約的シンボリックデータとして扱う。異なる集約的表現で記述されたデータを直接比較することは一般に難しいが、様々な集約的表現に対して定義が可能である非類似度に着目し、特徴毎に多次元尺度構成法を適用することによって、集約的シンボリックデータを同一の空間に布置し、その上で、データの特徴間の関連性を分析する手法を提案する。応用例として、第 2 章で触れたデータを地域毎に構成した高度集約的シンボリックデータを扱う。データセットが複数の測定方法毎に得られているため、それらを考慮した形で改めて高度集約化データを再構成し、その高度集約化シンボリックデータ間の関連性について分析する。得られた結果の解釈を通じて、本手法が有用であることを示す。

第 7 章では、2 章から 6 章で得られた研究成果を総括するとともに、今後の展開について述べる。