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_review.pdf, 審査の要旨 |

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

学位論文審査の要旨

博士の専攻分野の名称　博士 (情報科学)　　氏名　松井　佑介

審査担当者　主　査　教　授　水田 正弘
副　査　教　授　髙井 昌彰
副　査　教　授　棟朝 雅晴
副　査　准教授　南 弘征

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）

本論文は、昨今の技術革新に伴って取得が容易となり膨大に蓄積されるビッグデータの解析にあたって、従来の解析法では直接的な適用が困難であるという点に鑑み、データの多様性を考慮した適切な集約表現を導入し、その解析法について論じている。

従来の多変量解析法では、実数値ベクトルによってデータを扱ってきたが、ビッグデータの特徴とされる大容量、高頻度、多様性に対して、必ずしも適切な表現とはいえない。データを集約化した上で解析する手法として、Diday によるシンボリックデータ解析法 (Symbolic Data Analysis, SDA) が広く研究されている。しかし、そこで扱われている主な集約的データ表現は、多値、区間値、モダル値であり、データに内在する複雑な情報を十分に活用できない場合も考えられる。そこで、著者は、従来の集約的データ表現を拡張し、より複雑な表現に基づく高度集約的シンボリックデータを提案した。特に、分布値、関数値、多次元関数値、尺度混在による表現に対する新たな解析手法について述べている。また、それらの手法を、通信遅延データおよび環境データに適用し、結果と解釈を通じてその有用性を示している。

本論文は 7 章より構成されており、第 1 章では、本論文の導入として、研究背景とその目的について述べている。

第 2 章では準備として、シンボリックデータ解析法の概要について説明し、多値、区間値、モダル値などの基礎集約的シンボリックデータと、本論文で提案している高度集約的シンボリックデータについて論じている。また、次章以下で用いる通信遅延時間データセットおよび環境データセットに関して、それぞれの概要と予備解析の結果を述べている。

第 3 章では、統計的なばらつきを表すことのできる分布値について詳述している。特に、解析対象間の類似性が分布によって表現されている場合を扱い、階層型クラスタリング法を提案している。また、通信遅延時間データセットに提案手法を適用し、特徴的なクラスターを同定するとともに、その解釈を通じて手法の有効性を示している。

第 4 章では、Ramsay により提唱され、時間に依存したばらつきを表現できる関数値について述べるとともに、関数値クラスタリング法について論じている。具体的には、関数値の定義域を連続的に移動させ、時間に依存したクラスター構成の変化について分析を可能とする、移動関数 $k$-means 法を提案している。また、提案手法を環境データセットに適用し、その解釈を通じて有効性を示している。

第 5 章では、さらに複雑なデータ表現への発展として、関数の値域が多次元となるような多次元値関数値を扱い、特異な領域の同定手法を論じている。各次元における関数値に対して $L_0$ 正則化によってエッジを強調した平滑化を行うことで、複数の次元において同時に特異となるような領域を効果的に検出する手法を提案している。また、多数のセンサーによって測定された環境データセットに対して提案手法を適用することで、測定値が同時に異常な状態になるような期間を同定できることを示し、さらに気象データセットをあわせて用いることで詳細な解釈を行い、その有効性を示した。

第 6 章では、異なる尺度が混在している尺度混在型データを扱い、解析対象間の関連性を分析する手法について論じている。尺度混在型データでは変数を対等に扱うことができないため、解析対象間の非類似度を用い、多次元尺度構成法を適用することにより、解析対象をユークリッド空間上へ付置し、関連性を分析する方法を提案している。その応用例として、各地域で複数の方法によって空間線量率を測定した環境データセットに対し、それぞれの測定方法の特性を考慮して、分布値と関数値によるデータ表現を再構成し、提案手法によって分布値と関数値の関連性を分析することで、測定方法間の関係性を分析した結果として示すとともに、その解釈を通じ、手法の有効性を示している。

第 7 章では、2 章から 6 章で得られた研究成果を総括し、今後の展開について論じている。

これを要するに、著者はデータの多様性を考慮した表現法として、高度集約的シンボリックデータとその解析手法を提案するとともに、実データセットへの適用を通じて有効性を実証しており、情報科学に対する貢献は大なるものがある。よって博士 (情報科学) の学位を授与するに値すると認める。