博士学位論文審査要旨

２０２４年2月2日

論文題目：　Multivariate Data Analysis Methods using External Label Information
（外部ラベル情報を用いた多変量データ解析法）

学位申請者：　岡部　格明

審査委員：

主　査：	文化情報学研究科	教授	宿久　　洋
副　査：	文化情報学研究科	教授	波多野　賢治
副　査：	文化情報学研究科	教授	鄭　　躍軍
副　査：	文化情報学研究科	教授	田口　　哲也
副　査：	大阪大学大学院人間科学研究科	教授	足立　　浩平

要　　旨：

本論文は，外部ラベル情報を用いた多変量データの分析法に関して，データの縮約という観点から次の３つの問題を取り扱ったものである．１つ目は，二値ラベルが不均衡な場合の分類問題，２つ目は，二値ラベルに誤りが含まれる場合の分類問題，そして，３つ目は，順序ラベルを外部ラベル情報として持つデータに対する次元縮約問題である．本論文では，これらそれぞれの問題に対して，既存の手法を適用する際の問題を解決するための新たな手法を提案している．

第１章では，外部ラベル情報の定義および外部ラベル情報を用いた分析を行う動機と問題点，論文内で用いる諸概念について述べている．第２章では，本論文で扱う問題のうち１つ目の問題である二値ラベルが不均衡な場合の分類問題について，既存のロジスティック回帰モデルを適用する際の問題点に言及した上で，適合率と再現率の調和平均として定義されたF-measureという統計量を最大化するための方法を提案している．また，ここで提案した手法に対して人工データを用いた数値シミュレーションおよび実データ解析によって提案手法の優位性を確認している．第３章では，２つ目の問題である二値ラベルに誤りが含まれる場合の分類問題について，既存の誤りラベルが含まれる場合の分類方法に関する仮定とその特徴を整理しながら，既存の方法の仮定を緩和するための方法を提案し，人工的に誤りを加えたデータに対して適用することにより，提案した手法の有用性を確認している．第４章では，３つ目の問題である順序ラベルを外部ラベル情報として持つデータに対する次元縮約問題について，順序構造が含まれた対象の関係性を制約として用いた多次元尺度構成法を提案し，人工的に生成したデータからその潜在構造を抽出できることを確認し，また，実データ解析によって提案手法の有用性を確認している．

本論文により，外部ラベル情報を用いた多変量解析において，様々な問題に対応した手法が提案され，得られているデータの縮約を元にした分析の可能性を広げた．よって，本論文は，博士（文化情報学）（同志社大学）の学位論文として十分な価値を有するものと認められる．

総合試験結果の要旨

２０２４年2月2日

論 文 題 目： Multivariate Data Analysis Methods using External Label Information
（外部ラベル情報を用いた多変量データ解析法）

学位申請者： 岡部　格明

審 査 委 員：

主　査：	文化情報学研究科	教授	宿久　　洋
副　査：	文化情報学研究科	教授	波多野　賢治
副　査：	文化情報学研究科	教授	鄭　　躍軍
副　査：	文化情報学研究科	教授	田口　　哲也
副　査：	大阪大学大学院人間科学研究科	教授	足立　　浩平

要　　旨：

学位申請者は 2021 年 4 月より本学大学院文化情報学研究科博士課程後期課程に在学しており，国内会議および国際会議での研究発表を通じて研究活動を積極的に行い，それらの成果を，計算機統計学関連の論文誌に 1 本，国際会議 Proceedings に 1 本，人工知能関連の論文誌に 1 本として公刊している．また，英語の語学試験にも合格していることから語学（英語）について十分な能力を有していると認定されている．

2024 年 2 月 1 日木曜日 18:00 から約 1 時間の公聴会と 30 分の審査会において，種々の質疑応答の結果により博士（文化情報学）の学位を有するに十分な学力を有することを確認した．

よって，総合試験の結果は合格であると認める．

博士学位論文要旨
Abstract of Doctoral Dissertation

論文題目：
Title of Doctoral Dissertation
Multivariate Data Analysis Methods using External Label Information
（外部ラベル情報を用いた多変量データ解析法）

氏名：
Name
岡部 格明

要旨：
Abstract

近年，高度な観測技術により，大規模かつ複雑な構造を持つ様々な種類のデータを得ることができるようになった．そのようなデータには，例えば，画像データや遺伝子発現データなどがある．これらのデータには，そのデータが取得された状況を表す外部情報を含んだデータが得られることがある．外部ラベル情報を持つデータは，同一の対象者から得られることが想定されている．この中でラベル情報として得られるデータを外部ラベル情報データと呼ぶ．例えば，遺伝子発現に関するデータでは，ある細胞における遺伝子の発現状況の外部情報として，その細胞が得られた被験者の病態などの情報が外部ラベル情報データとなりうる．このような情報を用いることによって，対象の判別や低次元空間への縮約を通してデータの解釈を行うことができる．

本研究では，外部ラベル情報を用いた多変量データの解析法に関して次の３つの問題を取り扱う．１つ目の問題として，二値ラベルが不均衡な場合の分類問題，２つ目の問題として，二値ラベルに誤りが含まれる場合の分類問題，そして，３つ目の問題として，順序ラベルを外部ラベル情報として持つデータに対する次元縮約問題である．本研究ではこれらの問題それぞれに対して，既存方法の問題を解決するための新たは手法を提案した．

１つ目の問題である二値ラベルが不均衡な場合の分類問題に関して，まず初めに，ラベル不均衡データに対する既存のアプローチの比較とそれらのアプローチの問題点について整理した．ここでは，コスト考慮方学習および判別結果の評価指標の最適化の問題について判別モデルを構築する際の重みづけや閾値の観点から整理した．既存の方法は，ロジスティック回帰のような予測確率をモデリングする方法において閾値の設定を変更することに対応している．予測確率に関して閾値を変更することは，予測確率の推定に対しては影響を与えないが，予測ラベルに対しては影響を与えてしまうことがわかる．予測確率の閾値を変更することにより，予測確率の解釈が困難になってしまうことが問題点として考えられる．そこで，本研究では，予測確率の閾値を変更することなく，予測ラベルの不均衡を考慮した判別モデルを構築することを目的として，F-measure を最大化するようにパラメータ推定を行う方法を提案した．提案手法では，F-measure を最大化するようにパラメータ推定を行うことによって，予測ラベルの不均衡を考慮した判別モデルを構築することができる．この方法は，対象個別への重みづけを行うことに対応している．対象個別の予測スコアに対して，対象個別の F-measure を考え，この指標を relative density ratio として捉えることによって，各対象の重みを推定した．シミュレーション研究と実データ解析を通して，閾値を 0.5 とした状況でのロジスティック回帰モデルにおいて，提案手法は，F-measure の観点で，既存の方法よりも優れたパフォーマンスを示すことを確認した．

２つ目の問題である二値ラベルに誤りが含まれる場合の分類問題に関して，まず初めに，ラベルに誤りが含まれるデータに対する既存のアプローチに関して既存の方法の仮定に関する考察を行った．既存の方法では Balanced Error Rate (BER)の最適化を行うことでラベルに誤りが含まれる場合の分類問題に対処している．しかし，ここで用いられている仮定は，誤りのあるラ

ベルデータに対しては，誤りの発生は特徴量と独立であるという仮定である．例えば，ラベルアノテーションタスクにおいて，ラベルの誤りは，必ずしも特徴量と独立であるとは言えない場合がある．そこで，本研究では，誤りを含むラベルデータに対する分類問題に対して，誤りの発生は特徴量と独立であるという仮定を緩めた判別モデルの学習方法を提案した．ここでは，同一クラスから得られたサンプルは密集していることを仮定して，k 最近傍法を用いた重みづけを行う．これにより，各対象のラベルに対してラベル信頼度を推定することができる．この信頼度を各対象の重みとして判別を行うことで，この問題に対して有効であることをシミュレーション研究によって示した．

３つ目の問題である順序ラベルを外部ラベル情報として持つデータに対する次元縮約問題では，既存のクラス判別方法と次元縮約法の関係について整理した．高次元のデータに対して，外部ラベル情報が順序構造を持つ場合に関して，既存の方法では高次元データと外部ラベル情報を構成する潜在変数との関係に線形性を仮定した上で１次元の変数に縮約している．しかしながら，外部のラベル情報に想定される潜在構造に対して非線形性がある場合には，得られている潜在変数における対象の解釈が困難になる場合がある．一方で，高次元データのみを用いた次元削減法では，外部ラベル情報を用いることができないという問題がある．そこで，本研究では，多次元尺度構成法に対して，外部ラベル情報の順序を考慮した順序距離の互換性を重心に関する制約として導入することによって，ラベル情報も用いた次元削減を行う方法を提案した．シミュレーション研究と実データ解析において，２次元以上の潜在構造によって低次元空間を推定するようなタスクにおいては，提案手法が有効であることが示された．