博士学位論文審査要旨

２０１９年7月１日

論 文 題 目：Clustering and visualization for enhancing interpretation of categorical data
「カテゴリカルデータの解釈容易性を向上させるためのクラスタリングと視覚化法について」

学位申請者：　髙岸　茉莉子

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　川﨑　廣吉
副　査：　文化情報学研究科　教授　金　明哲
副　査：　文化情報学研究科　教授　鄭　躍軍
副　査：　東京大学大学院教育学研究科　准教授　岡田　謙介

要　　旨：

本論文では，近年増え続けている大規模カテゴリカルデータに対し，その解析結果の解釈の場面で生じる問題を考えている．データが大規模な場合，そのままでは解釈が困難になるため，クラスタリングや次元縮約を用いた視覚化などで，データの潜在的な構造を調べる方法が有用とされるが，既存の方法を用いた際に様々な問題が生じる．本論文では，その中でも特にカテゴリカルデータをクラスタリングする際に生じる2つの問題に対して，問題解決のための方法を提案している．

1つ目は質問紙調査で生じる回答スタイルの問題である．回答スタイルとは質問項目の内容に関係ない回答者特有の回答傾向を意味し，本論文では，回答スタイルにバイアスがあるデータに対して，内容ベースクラスタリングを適用する方法である，Correcting and Clustering Response Style (CCRS)を提案し，シミュレーション及び実データ解析によりその有用性を示している．

続いて2つ目の問題としてカテゴリカルデータの視覚化について述べている．本論文では，外部情報クラスごとに複数のクラスターを抽出し，それら全てを共通の低次元空間上に同時布置するMultiple Set Cluster Correspondence Analysis (MSCCA)と呼ばれる方法を提案している．MSCCAの利点は，異なる外部情報クラスのクラスターを全て同じ空間上に布置することで，外部情報として複数の変量を用いることも可能となり，更に異なる外部情報クラス間の関係も視覚的に解釈できることである．

本論文により，今まで解釈困難だったカテゴリカルデータの解析結果に対しても解釈容易な結果が得られるようになり，新たな知見を得るための可能性を広げた．よって本論文は，博士（文化情報学）（同志社大学）の学位を授与するにふさわしいものであると認められる．

総合試験結果の要旨

２０１９年7月１日

論 文 題 目：Clustering and visualization for enhancing interpretation of categorical data
「カテゴリカルデータの解釈容易性を向上させるためのクラスタリングと視覚化法について」

学位申請者：　髙岸　茉莉子

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　川﨑　廣吉
副　査：　文化情報学研究科　教授　金　明哲
副　査：　文化情報学研究科　教授　鄭　躍軍
副　査：　東京大学大学院教育学研究科　准教授　岡田　謙介

要　　旨：

学位申請者は，2015 年度 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いる．在学中は国内会議，国際会議での研究発表を精力的に行い，それらの成果を，国際会議 Proceedings に 1 本，計量心理計学関連の論文誌に 1 本，行動計量学関連の論文誌に 1 本公刊している．また，TOEIC も文化情報学研究科の学位取得基準である，850 点を超えた TOEIC のテストスコアを保持していることから，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．

2019 年 6 月 29 日土曜日 10:30 からの約 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）（同志社大学）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

論文題目：Clustering and visualization for enhancing interpretation of categorical data
（カテゴリカルデータの解釈容易性を向上させるための
クラスタリングと視覚化法について）

氏　　名：高岸　茉莉子

要　　旨：

本論文では，近年増え続けている大規模カテゴリカルデータの，データ解釈の場面で生じる問題を考えた．データが大規模な場合，データの解釈が困難になるため，クラスター分析や次元縮約を用いた視覚化などで，データの潜在的な構造を調べる方法が有用とされる．しかしその際にも様々な問題が生じる．そこで本論文では，カテゴリカルデータの潜在的な構造を調べる手法に生じる２つの問題を挙げ，問題解決のための方法を提案した．

１つ目は大規模質問紙調査で生じる，回答スタイルの問題について検討した．回答スタイルとは質問項目の内容に関係ない回答者特有の回答傾向を意味し，例えば極端なカテゴリばかり選ぶ対象が多ければ，クラスター分析を行う際，回答スタイルベースでなく内容ベースのクラスターが抽出されることがある．回答スタイルを補正するための方法として，Constrained Dual Scaling（以下CDS）がある．これは双対尺度法（Dual Scaling）の拡張手法であり，対象間に潜在的にあるとされる複数の回答スタイルのパターンを低次元空間で検知することを目的としている．更にCDSでは，その低次元空間における（回答スタイルの特徴を表す）数量をスプライン関数で平滑化することにより，結果として得られた連続関数を回答スタイルの補正に用いている（具体的には，低次元空間上で推定した連続関数を用いて，回答スタイルの影響を受けたデータから補正されたデータへの変換を行う）．しかしこの方法では関数の推定を低次元空間で行うことになるため，関数の表現幅に制限があり，結果補正の際に対象固有の（回答スタイルを取り除いた）実際の回答意図に関する情報が失われる．このような理由から，CDSによる補正では，クラスター分析と組み合わせても回答スタイルベースのクラスター結果しか得られない．従って本論文の第２章では，この問題を解決するための方法，Correcting and Clustering Response Style（以下CCRS）を提案する．CCRSではCDSの「連続関数を用いたデータ変換により回答スタイルを補正する」のアイデアは保ちつつ，次元縮約を必要としない関数データモデルとして，その補正に用いる連続関数を推定する．更にその関数の推定とクラスター分析を同時に行うことにより，補正による対象の個人の回答意図の情報損失を防ぐ．

第２章は以下のように構成されている．まず2.1章で問題点について述べ，2.2章では，双対尺度法に依存しない形で，回答スタイルを補正するための関数データモデルを提案する．そして2.3章では，そのモデルとクラスター分析法 k-means を組み合わせた方法を提案する．さらに2.4章のシミュレーションを通じて既存のCDSとCCRSのクラスター分析の精度を比較した．そして，2.5章のアジア間の国際比較調査のデータ解析を通じて，CCRSは，既存手法よりも解釈しやすい結果が得られることを示す．

続いて２つ目の問題例としてカテゴリカルデータの視覚化法について言及する．カテゴリカルデータが大規模な時，データの次元を減らし，データの構造を視覚化できるMultiple Correspondence Analysis（以下MCA）がよく用いられる．MCAでは低次元空間上の対象，カテゴリの座標が推定され，座標間の距離に基づき関連の強さを解釈する．このような図はバイプロットとも呼ばれる．MCAのバイプロットを解釈容易にするために，バイプロットに外部情報（座標

の推定には用いないが，バイプロットにそれに関する解釈を加えたい場合に使う情報）を加える方法がある．本論文では，外部情報として性別，国籍などのクラス情報を表すカテゴリカル変量を想定する．そして対象を外部情報ごとにクラス分けし，対象とカテゴリの関係の，クラスごとの特徴を視覚的に解釈することを目的とする．上記目的を達成する単純な方法として，対象の座標をクラスごとに平均をとり，それを１つの座標点とする方法が考えられる．しかしこの方法ではクラス内の大多数の人が同じ傾向を持つ場合は，その傾向はバイプロット上でも解釈しやすいが，そうでない場合は解釈が難しくなる．例えばクラス内で複数の傾向に等分に分かれている，また同じ傾向を持つ対象の中でも少数派のクラスが含まれている（例，ある傾向を持つのは大半が若者だが，多少の高齢年齢層もいる），などの状況は，外部情報ごとの特徴を知る上では有益な情報になりうるが，平均をとる方法では視覚的に解釈することは難しい．そこで本論文の３章では，上記のように外部情報クラス内で複数の異なる傾向がある場合，これらの傾向を１つのバイプロット上で解釈するための方法を提案する．具体的には，外部情報クラスごとに複数のクラスターを抽出し，それら全てを共通の低次元空間上に同時布置する．ここでクラスタリングすることは，各クラス内で似た傾向を持つ対象のみに人数を絞った上での，相対的なカテゴリとの関連の強さを見ることを意味する．これにより，例え少人数のみが持つ傾向であっても，関連の強さの情報が保たれ，視覚化結果にも反映されやすくなる．また異なる外部情報クラスのクラスターを全て同じ空間上に布置することで，外部情報として複数の変量を用いることも可能となり，更に異なる外部情報クラス間の関係も視覚的に解釈できる．

３章は以下のように構成されている．まず3.1章で問題点について述べ，3.2章では，提案手法，Multiple Set Cluster Correspondence Analysis（以下MSCCA）を定義し，既存手法（Cluster CA,行線形制約のアプローチ）などとの関係を述べる．そして3.3章では，シミュレーションを通じて，外部情報のクラスとしてどのような変量を選ぶかによりクラスタリングやバイプロットの精度は変動するかどうかを評価した．結果外部情報クラスの選択により，バイプロットの精度は大きく変わるが，クラスタリング結果はあまり変わらないことがわかった．また結果として，外部情報はクラスの人数の均衡が取れており，かつカテゴリ数が少ないものを選ぶことが良いことも分かった．最後に3.4章では交通事故に関する実データを解析し，既存手法とMSCCAを比較する．そこで既存手法ではクラスについての大まかな傾向しかわからなかったものが，一方のMSCCAでは各クラスの細かな状況の違いも解釈できることを示す．

これらの新たな手法の提案より，今まで解釈困難だった大規模カテゴリカルデータに対しても解釈容易な結果が得られるようになり，新たな知見を得ることを助けることが期待できる．