様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　４ 年　 ５ 月 ３０ 日現在

機関番号：１５３０１
研究種目：若手研究(B)
研究期間：2017～2021
課題番号：１７Ｋ１２６４８

研究課題名（和文）予測可能なクラスター構造の推定方法の開発と臨床医学への応用

研究課題名（英文）A new estimation method for predictable cluster structure and its application to clinical medicine

研究代表者
山本　倫生（Yamamoto, Michio）

岡山大学・環境生命科学学域・准教授

研究者番号：５０７２１３９６

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：複数の目的変数の特徴を反映したクラスター構造の予測を、ある説明変数群によって行うことを目的とする状況を考える。そのような場合、一般には、まず目的変数だけをデータとしてクラスター分析を適用し、目的変数間に内在するクラスター構造を抽出する。続いて、それをラベルとして、説明変数を用いた予測式の推定を行うことが多い。本研究では、このような従来の逐次的な方法ではなく、目的変数のクラスター構造の抽出と、説明変数によるそのクラスターラベルの予測を同時に行う方法を提案した。提案方法を用いることにより、取り扱っている現象に関連の強いクラスター構造が得られることが期待される。

研究成果の学術的意義や社会的意義
臨床医学における典型的な研究として、まず（1）クラスター分析などの教師なし学習により、疾患の重症度などを用いて症例のクラスタリングを行い、次に（2）得られたクラスターをラベルとして判別分析などの教師あり学習を用い、バイオマーカーによるサブタイプの予測や予測に重要なバイオマーカーの特定を行うものがある。このようなアプローチでは、段階ごとに異なる目的関数の最適化を行うため、真のクラスター構造と、それを予測可能な説明変数群の特定に失敗してしまう。本研究では、この問題を解決するために、教師なし学習と教師あり学習の両方の目的を同時に達成するための新たな統計解析の枠組みを提案することとなる。

研究成果の概要（英文）：Due to the recent advances in data collection and storage, data sets for statistical analysis have become complex and enormous. In the analysis of repeated measures data, for example, the data are often considered as a certain function, and such an analysis is called functional data analysis. In this study, I developed a new clustering method that conducted clustering and dimension reduction of multivariate functional objects simultaneously. Related to the method, I developed another clustering method with dimension reduction for multivariate binary data. In addition, I developed a new clustering method that identified a cluster structure of outcome variables and predicted cluster memberships of future individuals based on explanatory variables.

研究分野：統計科学

キーワード：クラスタリング　次元縮小

１．研究開始当初の背景

臨床医学における典型的な研究として、図 1 に示すように、まず（1）クラスター分析などの教師なし学習により、疾患の重症度など（目的変数）をデータとして症例のクラスタリング（疾患のサブタイプ探索）を行い、次に（2）得られたクラスターをラベルとして判別分析などの教師あり学習により、バイオマーカー（説明変数）によるサブタイプの予測や予測に重要なバイオマーカーの特定を行うものがある。このような逐次的な解析方法を Clustering-Prediction (CP) アプローチと呼ぶ（Yamamoto et al., 2016）。しかし、CP アプローチでは、段階ごとに異なる目的関数の最適化を行うため、データの背後に存在する真のクラスター構造と、その真のクラスター構造を予測可能な説明変数群の特定に失敗してしまうことが知られている。

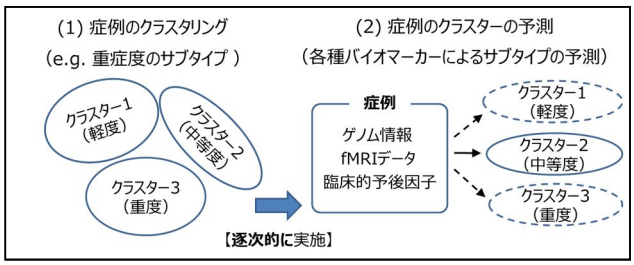

図 1：一般的な臨床医学研究における、教師なし学習と教師あり学習の逐次的な利用（CP アプローチ）

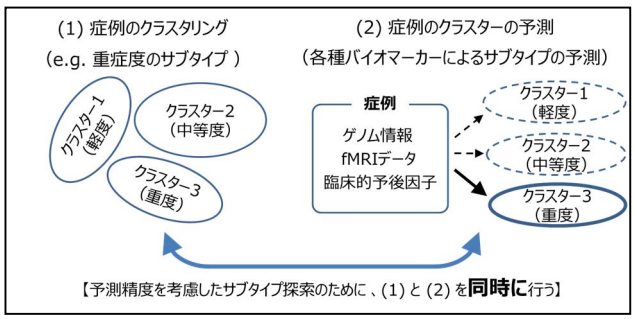

図 2：提案する同時分析アプローチ

CP アプローチの問題を解決するために、疾患の重症度と関連のあるバイオマーカーをまずは特定し、得られたバイオマーカーを特徴量としてクラスター分析を行うことで、バイオマーカーによって予測可能な疾患のサブタイプを推定する方法が提案されている。しかし、この方法もまた 2 段階の逐次的な方法であり、上記の CP アプローチの本質的な問題点は解決されていない。実際に、必ずしも重症度を反映するようなサブタイプが得られないことが予備的な検討から判明している。

そこで、CP アプローチの問題点を本質的に解決するための方法として、段階的に教師なし学習と教師あり学習を利用するのではなく、目的変数と説明変数の共分散情報そのものを利用して症例のサブタイプを推定する、教師なし学習・教師あり学習の同時分析法が申請者らにより開発されている（Yamamoto et al., 2016）。この方法では、クラスター間の違いが目的変数と説明変数の共分散で表現されることから、目的変数と説明変数の両方に関連するクラスター構造が推定可能である。この方法により、CP アプローチがうまく機能しない場合であっても、真のクラスター構造を推定可能であることが数値実験や実データ解析により示されている。

２．研究の目的

申請者らによる提案方法には、依然として以下の 3 つの問題点がある。

【1】目的変数のクラスター構造と説明変数の直接的な関連が不明であるため、得られたクラスター構造が説明変数によって精度よく予測できない可能性がある。

【2】クラスター数やその他のチューニングパラメータの選択方法が不十分である。特に、クラスター数の選択について、モデルの複雑さが原因となり、一般的なクラスター数の選択方法が正しく機能する保証がない。

【3】アルツハイマー病などを対象とする場合、説明変数として教育歴や年齢などの背景情報だけでなく、脳情報や遺伝子情報などをバイオマーカーとして利用する必要がある。現状では、このように異なる複数のデータソースを統一的に扱うことができない。

そこで、本研究では「オミクスデータや脳情報を含む種々のデータソースを説明変数候補とし、それらによって予測可能な疾患のサブタイプを推定する統計解析法」という観点から、目的変数のクラスターが説明変数によって予測可能であり、クラスター数の選択などのモデル選択が可能な、複数のデータソースを統一的に利用できる方法の開発を目的とする。

３．研究の方法

上記の目的の達成のために、以下の 3 つの研究を行う。

（1）予後因子によって予測可能な臨床アウトカムのクラスター構造の推定方法を開発する。具体的には、説明変数がクラスターの予測に対して直接的な関連をもつ統計モデルを導入する。本研究では可能な限りシンプルなモデルを導入し、その理論的性質の検討を行う。クラスタリング研究で重要な、損失関数とクラスター中心それぞれの推定量の一致性、平均二乗誤差に対する大

偏差不等式など、漸近的・非漸近的な性質の検討を行う。さらに、シミュレーションによって、説明変数の一部が目的変数の真のクラスターに関連をもつ場合に、提案方法によって正しく真のクラスターを推定できるかどうかを評価する。
（2）提案手法におけるモデル選択方法として、教師なし・教師あり学習の同時分析アプローチで適用可能なモデル選択方法を開発する。特に、クラスター数の決定方法として、教師なし学習で利用される Clustering Stability をベースとしたモデル選択方法を開発する。モデル選択に対する一致性を示すとともに、シミュレーションによってモデル選択方法の性能評価を行う。
（3）複数のデータソースを説明変数として利用できるように提案モデルを拡張する。さらに、研究（2）で開発するモデル選択方法も、提案モデルの拡張に合わせて再検討する。提案方法を用いて脳情報・遺伝子情報・臨床バイオマーカーを用いた疾患サブタイプの探索を行う。

４．研究成果

（1）クラスタリングと予測に関する損失関数の凸結合によって構成される新たな損失関数を導入し、さらに、説明変数に対する重み係数も含めることで、クラスターの予測に影響を与える因子の特定を試みた。具体的には、説明変数$\boldsymbol{X} \in \mathbb{R}^p$による目的変数$\boldsymbol{Y} \in \mathbb{R}^q$の K 個のクラスターの予測を直接的に表現するために、クラスタリングの良さと予測精度のバランスを制御するパラメータ$\alpha \in [0,1]$を与えたもとで、損失関数を以下のように定義した。

$$L(C, \mathbf{A}) = \int \min_{\boldsymbol{f} \in C} \left[\alpha \|\boldsymbol{Y} - \boldsymbol{f}_Y\|^2 + (1-\alpha)\|\boldsymbol{X} - \mathbf{A}\mathbf{A}'\boldsymbol{f}_X\|^2 \right] \mathbb{P}(d(\boldsymbol{Y}, \boldsymbol{X})).$$

ここで、$\boldsymbol{f}_Y \in \mathbb{R}^q, \boldsymbol{f}_X \in \mathbb{R}^p$ は $\boldsymbol{Y}, \boldsymbol{X}$ に対するクラスター中心を表しており、C は K 個のクラスター中心 $\boldsymbol{f}_k = \{\boldsymbol{f}_{Yk}, \boldsymbol{f}_{Xk}\}, (k=1,...,K)$ からなる集合である。また、$\mathbf{A} \in \mathbb{R}^{p \times m}$ は $\boldsymbol{X}$ に対する列直交な係数行列であり、クラスターの予測に重要な説明変数を示すパラメータである。被積分関数に含まれる和の第 1 項は $\boldsymbol{Y}$ のクラスタリングの良さを示す指標であり、一般的なクラスタリング手法の損失関数と考えることができる。

提案方法では、クラスター中心集合と重み係数行列を同時に推定する必要があるが、以前開発した Yamamoto and Hwang (2014) の推定方法を参考に、固有値分解と K-means アルゴリズムを組み合わせた推定アルゴリズムを開発した。また、クラスタリングの統計的性質として、損失関数の一致性およびパラメータ（クラスター中心集合、重み係数行列）の推定量の一致性を検討した。提案手法の損失関数は、Reduced K-means (RKM) 法の特殊な形として表現できることから、RKM 法における証明方法を参考に、損失関数の一致性およびパラメータの推定量の一致性を示した。

開発した手法の性能を検討するために、目的変数の真のクラスター構造が説明変数によって予測可能である場合を想定したシミュレーションを実施した。その結果、既存のアプローチに比べて提案手法がより正確に真のクラスター構造を再現でき、かつ、得られたクラスター構造が説明変数によって予測可能であることを確認した。

（2）上記の提案手法では 2 つの損失関数の凸結合を損失関数としている。しかし、事前の検討により部分最小二乗回帰（partial least squares regression; PLS 回帰）を利用することによって、目的変数のクラスタリングと説明変数による予測を同時に達成する方法によっても、提案手法と同様の目的を達成することが可能であることがわかっている。そこで、PLS 回帰を応用する形で目的変数のクラスター構造の探索と説明変数によるラベルの予測が可能なモデルの定式化を行った。数値実験により、(1) で提案していた手法との比較を行った結果、2 つの方法が同等の性能を示すことが判明していた。しかし、理論的性質などは未解明のままであり、今後さらに検討していく必要がある。

（3）上記の研究におけるモデルの定式化や推定アルゴリズムから派生して、多変量カテゴリカルデータのクラスタリングの新たなモデルを開発した。提案手法では、単にクラスタリングを行うだけでなく、各特徴量がクラスター構造に与える影響度を定量化し、推定されたクラスター構造の解釈を行うことが可能である。また、数値実験および実データ解析によって、単なるクラスタリング手法としての性能も、既存の方法と比べて同等かそれ以上であることが確認された。

＜参考文献＞

Yamamoto, M., Hwang, H. (2014). “A general formulation of cluster analysis with dimension reduction and subspace separation”. *Behaviormetrika*, 41: 115-129.
Yamamoto, M., Kawaguchi, A., Hwang, H. (2016). “Predictive clustering using a component-based approach”, The 22nd International Conference on Computational Statistics (COMPSTAT 2016).

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　5件／うち国際共著　2件／うちオープンアクセス　1件）

１．著者名 Michio Yamamoto、Heungsun Hwang	４．巻 34
２．論文標題 Dimension-Reduced Clustering of Functional Data via Subspace Separation	５．発行年 2017年
３．雑誌名 Journal of Classification	６．最初と最後の頁 294～326
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s00357-017-9232-z	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 該当する

〔学会発表〕　計18件（うち招待講演　10件／うち国際学会　9件）

１．発表者名 山本倫生，寺田吉壱
２．発表標題 スパースな経時測定データにおけるクラスタ構造の推定
３．学会等名 日本行動計量学会 第49回大会
４．発表年 2021年

１．発表者名 Yamamoto, M., Terada, Y.
２．発表標題 K-means clustering for sparsely sampled longitudinal data
３．学会等名 13th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2020)（招待講演）（国際学会）
４．発表年 2020年

１．発表者名 山本倫生，寺田吉壱
２．発表標題 スパース経時測定データに対する関数クラスタリング
３．学会等名 第25回 情報・統計科学シンポジウム（招待講演）
４．発表年 2020年

１．発表者名
山本倫生，寺田吉壱

２．発表標題
スパースな経時測定データに対するK-means法

３．学会等名
2020年度統計関連学会連合大会

４．発表年
2020年

１．発表者名
Yamamoto，M.，Terada，Y.

２．発表標題
Functional canonical correlation analysis for multivariate stochastic processes

３．学会等名
The 3rd International Conference on Econometrics and Statistics (EcoSta 2019)（招待講演）（国際学会）

４．発表年
2019年

１．発表者名
山本倫生，寺田吉壱，谷岡健資

２．発表標題
クラスタリング法のcheat sheet

３．学会等名
日本行動計量学会第47回大会

４．発表年
2019年

１．発表者名
山本倫生，寺田吉壱

２．発表標題
多変量関数データに対する正準相関分析の定式化と解の存在性について

３．学会等名
2019年度統計関連学会連合大会（招待講演）

４．発表年
2019年

１．発表者名
山本倫生

２．発表標題
多変量カテゴリカルデータに対するクラスター構造の推定とその可視化について

３．学会等名
「複雑多変量データの解析法に関する研究」研究会

４．発表年
2018年

１．発表者名
Yamamoto, M.

２．発表標題
A component-based approach for the clustering of multivariate categorical data

３．学会等名
The 2nd International Conference on Econometrics and Statistics (EcoSta 2018)（招待講演）（国際学会）

４．発表年
2018年

１．発表者名
Michio Yamamoto

２．発表標題
Clustering of multivariate categorical data with dimension reduction via nonconvex penalized likelihood maximization

３．学会等名
The 2017 conference of the International Federation of Classification Societies (IFCS 2017)（招待講演）（国際学会）

４．発表年
2017年

１．発表者名
山本倫生

２．発表標題
関数データのクラスタリングとクラスター構造の可視化について

３．学会等名
統計学・機械学習若手シンポジウム「大規模複雑データに対する統計・機械学習のアプローチ」

４．発表年
2017年

１．発表者名
山本倫生

２．発表標題
多変量カテゴリカルデータに内在する低次元クラスター構造の推定

３．学会等名
行動計量学岡山地域部会第64回研究会（招待講演）

４．発表年
2017年

１．発表者名
Michio Yamamoto

２．発表標題
Clustering of multivariate categorical data via penalized latent class analysis with dimension reduction

３．学会等名
2017 Hangzhou International Statistical Symposium（国際学会）

４．発表年
2017年

１．発表者名
Michio Yamamoto

２．発表標題
Model-based clustering for multivariate categorical data with dimension reduction

３．学会等名
The 10th Conference of the IASC-ARS（招待講演）（国際学会）

４．発表年
2017年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関			
カナダ	McGill University			