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　 ６ 月 １４ 日現在

機関番号：１７１０２
研究種目：基盤研究(C)（一般）
研究期間：2019～2021
課題番号：１９Ｋ１１８６２

研究課題名（和文）予測モデルのグループ化を目的とするクラスター分析とその応用

研究課題名（英文）Cluster analysis for grouping statistical models and its application

研究代表者
廣瀬　慧（Hirose, Kei）

九州大学・マス・フォア・インダストリ研究所・准教授

研究者番号：４０６０９８０６

交付決定額（研究期間全体）：（直接経費）　　3,400,000円

研究成果の概要（和文）：本研究では、データに多様性を伴う場合に柔軟な予測モデルを構築した。具体的には、単一の回帰分析や判別分析を用いて予測するのでなく、複数の予測モデルを構築し、それらをグループ化することによって、グループごとに予測する方法を考案した。予測モデルをグループ化するためには、クラスター分析を行う必要がある。従来のクラスター分析では、距離行列等を用いてクラスタリングを行うが、本研究では、予測誤差に基づく関数を定義することにより、予測精度を向上を試みた。さらに、クラスタリングを高速に実行するために、効率的な計算アルゴリズムを構築した。

研究成果の学術的意義や社会的意義
近年、ディープラーニングを用いたデータ解析が主流となっているが、ディープラーニングは、画像やテキストなど、サンプルサイズが十分に大きい場合に精度の高い予測モデルが構築できる。一方で、遺伝子データや電力需要量データ、材料データ等、ディープラーニングが実行できるほどの多くの観測が得られないことがある。本研究では、このような場合に、できるだけ精度良く予測できる柔軟なモデルを提案した。また、モデルを新たに構築しただけでなく、高速なアルゴリズムの提案、さらにはRパッケージの公開も行った。

研究成果の概要（英文）：We constructed flexible statistical modeling for capturing complex structures in data. Specifically, instead of using a single regression or discriminant analysis, we constructed multiple statistical models and grouped them; then, a prediction was performed for each group. To group the statistical models, a cluster analysis was performed. Conventional cluster analysis adopted a distance matrix. On the other hand, we defined a function based on the prediction error to improve the prediction accuracy. Furthermore, an efficient computational algorithm to perform clustering was established.

研究分野：統計科学

キーワード：クラスタリング　クロスバリデーション

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景
近年、医学、工学、エネルギー、環境、心理など様々な分野で得られるデータには、サンプルサイズがある程度大きく、かつ多様性を伴うという特徴がある。画像やテキスト等の極めて膨大なデータに対してはディープラーニングを用いればよいが、それほど多くの観測が得られない場合は、多様性を考慮した統計解析手法を構築する必要がある。本研究では、回帰分析や判別問題など「予測」に焦点を当て、多様性を伴った入力・出力データの関係性をつなぐ統計モデルを構築する。

２．研究の目的
通常､柔軟な予測モデルを構築するためには､非線形性を有するモデルを用いる｡しかしながら、急激に構造が変化する等の多様性を伴うデータに対しては, たとえ非線形構造を仮定しても、単一のモデルでは予測精度があまり向上しない場合がある。たとえば、群の数が大きい場合の判別分析を行う場合、一部の群が他の群と比べて大幅に離れていることがある。このような場合は、たとえ複雑な非線形関数を用いたとしても、大幅にかけ離れた他の群に引っ張られてしまい、予測精度が向上しないことがある。回帰分析でも同様の問題が生じる。そこで、予測モデルのクラスタリングを行って予測精度を向上させる。とくに、従来のウォード法のような距離行列に基づいてクラスタリングを行うのでなく、予測精度を向上させるようにクラスタリングを実行する。一旦予測モデルのクラスタリングを実行できたら、新たなデータが得られたとき、まずはそのデータがどの予測モデルに属するかを判定し、次に、選ばれた予測モデルに基づいて予測するという 2 段階の予測を行う。

３．研究の方法
本研究は予測モデルのグループ化を行う。とくに、目的関数を予測誤差とし、目的関数が最小となるように予測モデルをクラスタリングする。この手法の提案を実現するために、階層的クラスタリングを用いる。クラスタリングの指標として、予測誤差の推定量であるクロスバリデーションを用いる。 各群のサンプルサイズが小さい場合にクロスバリデーションが不安定になるのを防ぐため、一個抜きクロスバリデーションを行う。しかし、クラスタリングの各ステップでクロスバリデーションを計算する必要があるため、計算時間がかかる。そこで、クロスバリデーションを高速に計算するアルゴリズムを構築する。
　また､判別分析でなく､回帰分析においても､予測モデルをクラスタリングする方法を考えた。とくに、電力需要予測を行う上で有用なクラスタリング手法を検討した。一般に、電力需要量は高次元時系列データであり、それら高次元データをアグリゲーションしたものを予測することが多い。予測精度を向上させるため、いくつかのグループに分けてグループごとにアグリゲーションして予測する方法を提案した。

４．研究成果
本研究では複数の研究成果がある。
　まず、群の数が大きい場合における判別分析の新たな手法を提案した。群の数が大きいとき、判別しやすい群としにくい群が存在することが多い。そこで、クラスタリングを行って予測精度を向上させることを考えた。まず、ウォード法によりクラスタリングを行い、その結果から群を判別した。このアルゴリズムを実データに適用したところ、提案手法が従来の正準判別よりも予測精度が高くなることを確認した。
　次に､ウォード法でなく､クロスバリデーションを目的関数としたクラスタリングを実行した。クロスバリデーションは予測誤差の一致推定量なので、予測誤差が小さくなるようなクラスタリングが実現できることが期待できる。しかしながら、クロスバリデーションをクラスタリングのステップを行うごとに計算する必要があるため、計算時間がかかる。そこで、クロスバリデーションの高速化を行った。クロスバリデーションの高速化は、重回帰分析でよく知られた方法がある｡そのため､多群判別の重回帰分析による新たな定式化を行うことを考えた｡しかしながら、この方法は、元のデータの計画行列の分散共分散行列の固有値・固有ベクトルに依存し、高速化するには、クロスバリデーションの各ステップで毎回固有値・固有ベクトルを計算しなければならないことがわかった。そこで、固有値・固有ベクトルを１回だけ計算すれば良いようにクロスバリデーションを近似した。近似精度を調べたところ、漸近的には問題ないことが示された。この高速化は、線形の判別分析でなく、カーネル法や他の多変量解析への拡張もできると考えられる。そのため、汎用性の高いクロスバリデーションの高速計算につながる可能性を秘めている。
　提案手法の有効性を示すために、変数間に相関がある場合等様々な状況下で数値シミュレーションを行った。その結果、変数間の相関の大きさに関わらず、提案法は比較的良いパフォーマンスを示すことがわかった。さらに、提案法を 4 つの実データに適用したところ、いずれのデータに対しても、提案法が通常の線形判別よりも予測精度が高くなることが確認できた。さらに、クロスバリデーションの近似についても調べたところ、近似精度が高く、次元が高い場合は、少

なくとも 10 倍は計算スピードが早くなることを確認した。さらに、階層的クラスタリングを並列に計算することにより、より高速に計算することができる。この高速なアルゴリズムを R パッケージにし、github に公開した。この内容は、EcoSta2019 や統計関連学会等の学会で発表し、また、現在論文を投稿し、リバイス中である。

次に、異なるソースの入力情報から、それらの合計値を予測する回帰分析を行った。まず、そもそもクラスタリングすることによって精度向上するか調べた。具体的には、電力需要量のデータに関して、全て個別に予測する場合、合計値を１つのモデルで予測する場合、複数の入力をクラスタリングした場合において、予測精度を調べた。すると、クラスタリングした場合に最も予測誤差が小さくなることが数値的に確認できた。それゆえ、クラスタリングを用いることが予測精度向上につながることがわかった。
また、クラスタリングすることによって予測精度がなぜ高くなるかを考えたところ、クラスター数がモデルの複雑度と対応するためであることがわかった。さらには、クラスタリングにおける新たなバイアス−バリアンス−トレードオフの関係性を見出した。この内容は 2021 年度の統計関連学会連合大会で発表した。

また、上記の回帰分析を行う基盤となる、電力需要予測モデルを構築した。入力情報を過去の需要と予測日の気温として、変化係数モデルによる非線形構造の推定、selective inference による予測区間の構築など、様々な統計の手法を盛り込んだモデリングを行った。この手法の最大の特徴は、推定されたモデルを解釈できるという点にある。この可読性が、デマンドレスポンスなどに応用できると考えられる。さらに、提案手法が従来の機械学習の手法よりも高い精度で予測できることが実データ解析を通じて確認した。特に、提案法のほうがディープラーニングよりも予測精度が高くなることがわかった。その原因は、サンプルサイズが十分に大きくなく、ディープラーニングは学習が不安定になったためだと思われる。この内容は、Frontiers in Energy Research に採択された。

さらに、電力需要予測モデルにおいて、イベント情報を活用した手法を考えた。具体的には、上記の基底展開に基づく電力需要予測モデルに、イベント特有の基底を導入した。さらに、イベントの詳細情報が未知で、基底の構築が難しい場合においても、Generalized Lasso を用いて自動で基底を生成する方法を提案した。提案法をある施設の電力需要量データに適用したところ、12%~20%程度予測精度が向上することを確認した。この内容は雑誌 Energies に掲載された。

また、正則化因子分析の手法である Prenet (Product-based elastic net)を電力需要量データに適用した。すると、L_1 正則化法よりも安定した推定ができることを確認した。また、得られた因子を解釈したところ、新型コロナウイルスが電力需要に少なからず影響を与えていることがわかった。さらに、fMRI データに適用して、次元圧縮して復元したときの復元率を計算したところ、非ゼロ要素の数が小さい場合（スパースな場合）、従来法の Lasso や k-平均法よりも Prenet のほうが高い予測精度をもたらすことを確認した。この内容は、Psychometrika に掲載された。

そのほか、材料データ解析に関する論文が Numerical Methods in Engineering に、植物の遺伝子データ解析に関する論文が Scientific Reports に掲載された。

５．主な発表論文等

〔雑誌論文〕　計7件（うち査読付論文　6件／うち国際共著　0件／うちオープンアクセス　6件）

１．著者名 Hirose Kei、Terada Yoshikazu	４．巻 -
２．論文標題 Sparse and Simple Structure Estimation via Prenet Penalization	５．発行年 2022年
３．雑誌名 Psychometrika	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1007/s11336-022-09868-4	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Hirose Kei、Wada Keigo、Hori Maiya、Taniguchi Rin-ichiro	４．巻 13
２．論文標題 Event Effects Estimation on Electricity Demand Forecasting	５．発行年 2020年
３．雑誌名 Energies	６．最初と最後の頁 5839～5839
掲載論文のDOI（デジタルオブジェクト識別子） 10.3390/en13215839	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Okinaga Yuichi、Kyogoku Daisuke、Kondo Satoshi、Nagano Atsushi J.、Hirose Kei	４．巻 11
２．論文標題 Relationship between gene regulation network structure and prediction accuracy in high dimensional regression	５．発行年 2021年
３．雑誌名 Scientific Reports	６．最初と最後の頁 1-10
掲載論文のDOI（デジタルオブジェクト識別子） 10.1038/s41598-021-90791-6	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Teramoto Keisuke、Hirose Kei	４．巻 123
２．論文標題 Sparse multivariate regression with missing values and its application to the prediction of material properties	５．発行年 2021年
３．雑誌名 International Journal for Numerical Methods in Engineering	６．最初と最後の頁 530～546
掲載論文のDOI（デジタルオブジェクト識別子） 10.1002/nme.6867	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 Hirose Kei	４．巻 9
２．論文標題 Interpretable Modeling for Short- and Medium-Term Electricity Demand Forecasting	５．発行年 2021年
３．雑誌名 Frontiers in Energy Research	６．最初と最後の頁 1-15
掲載論文のDOI（デジタルオブジェクト識別子） 10.3389/fenrg.2021.724780	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

１．著者名 廣瀬慧	４．巻 －
２．論文標題 L1正則化法に基づく因子分析および構造方程式モデリングの最近の展開	５．発行年 2020年
３．雑誌名 計算機統計学	６．最初と最後の頁 －
掲載論文のDOI（デジタルオブジェクト識別子） なし	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 －

１．著者名 廣瀬慧	４．巻 2133
２．論文標題 因子分析モデルにおける構造正則化	５．発行年 2020年
３．雑誌名 京都大学 数理解析研究所 講究録	６．最初と最後の頁 1-10
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 －

〔学会発表〕　計3件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 廣瀬慧
２．発表標題 電力需要予測のための統計モデルとソフトウェア
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 廣瀬 慧, 増田 弘毅
２．発表標題 電力需要の短期予測のための統計モデリング
３．学会等名 ２０１９年度統計関連学会連合大会
４．発表年 2019年

１．発表者名 K. Hirose, K. Miura, A. Koie
２．発表標題 Cluster-based multiclass linear discriminant analysis
３．学会等名 The 3rd International Conference on Econometrics and Statistics (EcoSta 2019)（国際学会）
４．発表年 2019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