F-007

ビッグデータ解析における正則化項付き深層学習を用いた要因抽出
Factor Extraction using Deep Learning with Regularization Term for Big Data Analysis

東柚大河†　桑野将司†　南野友香†
Yutaka Azuma　Masashi Kuwano　Yuka Minamino

1. はじめに

ICT や IoT 機器の普及・多様化，官公庁によるデータ利活用の推進などの理由から，膨大なデータが蓄積されるとともに AI 技術が普及した．AI 技術の中でも，機械学習・深層学習手法が特に注目を集めている．深層学習は情報抽出を 1 層ずつ多階層にわたって行う．その複雑な構造と内部計算により，ビッグデータから高度な特徴抽出や複雑なパターン認識を実現する．従来の機械学習では，学習対象となる特徴量を分析者が定義する必要があったが，深層学習は予測したいものに適した特徴量をビッグデータから自動的に学習することができる[1]．特に，画像認識，音声認識，自然言語処理などの分野では，深層学習が優れた性能を発揮し，精緻な予測が可能となった．

深層学習手法は，従来の手法と比べて高精度での予測が可能である一方で，解決すべき課題としてブラックボックス問題がある．深層学習手法では多様なデータを用いて複雑な計算を行うため，内部でどのような計算を経て結果を出力しているかが不明である．そのため，予測結果に影響を与える要因の解釈が困難であり，結果の信頼性が低いことが懸念されている[2]．特に医療分野やマーケティング分野などにおいて，どのような要因が目的変数に影響を及ぼしているかを理解し，そのうえで意思決定を必要とする場合には，モデルの解釈性の向上が重要となる．予測結果に影響を与える要因の解明は，問題への対策や施策検討の観点からも重要である[3].

要因抽出の伝統的かつ代表的な手法として，スパースモデリングが注目されており，情報理論や信号処理，機械学習など工学の様々な分野で活用されている[4]．スパースモデリングは多くのデータ項目の中から重要な項目のみを抽出することにより，複雑で大量のデータがどのような構造かを解明し，わかりやすいモデルとして表現する技法である[5]．スパースモデリングには，少数のデータを扱うため計算コストを削減できる，複雑なデータ構造をわかりやすく表現できるといったメリットがある．嶋村ら[6]は，膨大なデータを持つ気象用レーダの観測データにスパースモデリングの手法の 1 つである圧縮センシングを適用した．膨大なデータを圧縮し，必要なデータのみを復元することで計算コストの削減を達成し，その有用性を示した．

本研究では，深層学習と要因抽出を組み合わせた LassoNet[7]を応用することで，多数存在する説明変数の中から予測結果に影響を与える説明変数を抽出できる予測モデルを構築する．さらに，抽出された説明変数のみを深層学習に適用し予測精度を確認することで，予測結果に影響を与える説明変数を抽出できているかを確認する．構築したモデルの実証分析として，ある銀行の顧客データを用いて，個人の住宅ローン契約確率の予測に影響を与える重要な説明変数を明らかにする．Xiaolu ら[8]は，LassoNet を臨床データに適用し，急性冠症候群の合併症の予測と重要な要因の選択を行った．その結果，年齢をはじめとしたいくつかの重要な要因を抽出し，解釈可能性向上の観点でその有用性を示した．Jiatong ら[9]は，中国の保険データに適用することで，個々のうつ病の影響要因を特定し，要因の重要性の順位や関連する結果を統合することで意思決定者や研究者に重要要因に関する情報を提供することが可能となると述べている．しかし，LassoNet の適用事例は少ない．本研究では，多くの説明変数を有し，サンプル数も多数ある顧客データに LassoNet を適用し，説明変数を多数有するデータを用いた場合においても重要な要因の抽出が可能であるかを検証する．

† 鳥取大学　Tottori University

2. LassoNet

Ismael ら[7]によって考案された LassoNet は Lasso 線形回帰モデルを一般化し，Lasso のスパース性をフィードフォワードニューラルネットワークに拡張したものである．非線形な特徴量を捉えるために，ニューラルネットワークの入力値と出力値の残差における線形成分と非線形成分を同時に最適化する．線形表現である Lasso による説明変数の回帰係数がゼロではない場合，その説明変数は中間層に入力可能となり，スパース性を実現する．

LassoNet は，ニューラルネットワークにおける他の変数選択方法とは異なり，変数選択とパラメータ学習を組み合わせるために，制約条件およびスパース性を有した目的関数を使用する．式(1)に LassoNet の目的関数と制約条件を示す．

$$\min_{\theta,W} L(\theta,W)+\lambda\|\theta\|_1$$
$$subject\ to\ \left\|W_d^{(1)}\right\|_\infty \le M|\theta_d|,\ d=1,\dots,D \qquad (1)$$

ここで，$L(\theta,W)$は残差ニューラルネットワーク，θは線形部分のパラメータ，Wは残差ネットワーク係数，$W_d^{(1)}$は残差ネットワークの第 1 層のd番目の成分，Mは層乗数，λは線形部分の罰則項である．目的関数は，入力値と出力値の残差ニューラルネットワークに L1 正則化項を付与する．制約条件はフィードフォワードニューラルネットワークの各ユニット上の特徴の重みパラメータWがM倍であることを示す．

3. 使用データ

本研究では，ある銀行の顧客データを使用する．データは 2022 年 7 月 31 日に取得されたクロスセクションデータである．予測精度の向上を図るため，新規での契約が見込まれる年齢が 30 代および 40 代のサンプルを分析対象とした．銀行の顧客データには，口座保有者の年代などの「顧客属性情報」，総預金残高などの「預金係数情報」など，顧客ごとの銀行利用状況が蓄積されている．本研究では，表 1 に示す 5 分類 141 変数を説明変数として用いる．

表 1　銀行の顧客データの変数（5 分類 141 変数）

分類	変数
顧客属性情報	年代，性別など 76 変数
契約情報	貸金庫契約有無など 29 変数
実績情報	賞与振込実績有無など 25 変数
預金係数情報	総預金残高，円貨預金残高など 5 変数
融資係数情報	総貸出極度額，消費残高など 6 変数

顧客データには，取引が長期間行われていないサンプルや所得などの理由からローン契約ができないサンプルが存在する．このような契約可能性が極めて低いサンプルを予め除外することでモデルの予測精度が高まると期待できる．しかし，変数が多いことから，契約可能性が低いサンプルを除外する基準を設定することは容易ではない．そこで本研究では，決定木分析を用いて契約可能性が低いサンプルの属性を分析し，その基準によってサンプルを除外した．決定木分析の結果，契約可能性がないと判断したサンプル数は 4,742 件となった．

分析に使用する 97,016 サンプルは，契約サンプルが 5,505 件，未契約サンプルが 91,511 件で構成される．サンプルのクラスが占める割合が不均一な不均衡データに対して機械学習による学習を行うと，予測結果のほとんどを多数派クラスであると判別するモデルが構築され，少数派クラスの分類精度が致命的に低くなる問題が発生する．Wang ら[10]の研究では，オンライン EC サイトでの不正取引の検知において，全体の取引量に対する不正取引の割合が圧倒的に低いため，学習器の実用性が著しく悪化してしまう問題を指摘している．これは，全ての購入データに対して「不正取引ではない」と予測する分類器でも，全体としての分類精度は非常に高いと評価されてしまうためである．本研究では，多数派クラスからサンプリングを行い，少数派クラスとデータ数の比率を調整する手法であるアンダーサンプリングを適用し，データの不均衡性を解消する．ただし，アンダーサンプリングにおいてランダムサンプリングを行うと，多数派クラスが持っている特徴量の分布が崩れる可能性がある．そこで，訓練データの多数派クラスのサンプリングにおいて k-means 法によるクラスタリングを行い，分類された各クラスタのサンプル数に等しい割合で各クラスタからサンプリングを行う手法を採用する．

4. 住宅ローン契約確率の予測結果

図 1 に本研究で構築する LassoNet のモデル概要図を示す．入力層に説明変数として表 1 に示す 141 変数を入力することで，中間層での計算を経て出力層に結果として各サンプルが「契約する」，「契約しない」確率が出力される．本研究では，分析対象サンプルを 7 割の訓練データと 3 割のテストデータに分割する．契約確率の閾値は 0.5 に設定し，出力された確率が 0.5 以上の場合「契約する」と判定する．LassoNet によるモデルの構築では，重みやバイアスのような誤差逆伝播法で勾配を計算され自動的に獲得されるパラメータ以外に，中間層数や学習率，層乗数など手動で調整する必要があるハイパーパラメータが存在する．本研究では，中間層を 3 層に設定し，その他中間層の各層のユニット数をはじめとしたさまざまなハイパーパラメータは Git-Hub[11]に記載されていたデフォルトの値を参考に試行錯誤的に変更し，最も予測精度が高かった組み合わせを採用した．表 2 に構築した LassoNet のモデルの諸条件を示す．

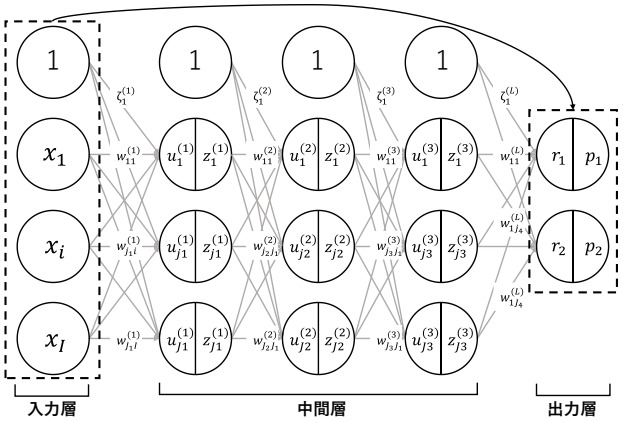

図 1　本研究で使用する LassoNet モデルの概要図

表 2　構築した LassoNet モデルの諸条件

項目	設定した値
中間層（第 1 層，第 2 層，第 3 層）	152，133，254
最適化アルゴリズム	Adam
活性化関数（中間層）	ReLu
活性化関数（出力層）	Softmax
バッチサイズ	92
学習回数（エポック数）	200
学習率	0.001
層乗数	10

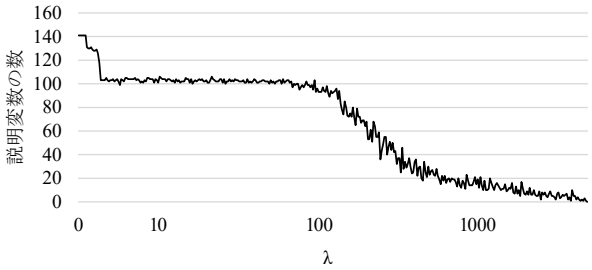

図 2　λに対する説明変数の数

LassoNet のモデルでは，式(1)における罰則項λが大きくなるにつれて，目的変数である「住宅ローン契約確率」への影響が少ない説明変数の係数が 0 となり，その説明変数は削除される．説明変数の数に変更があった際のλの値が出力され，説明変数の数が 0 になった時点で処理を終了するよう設定する．

モデル精度の検証には，観測値と予測確率で算出される再現率を用いる．再現率は，実際に「ローンを契約した」データのうち，正しく「ローンを契約する」と予測されたデータの占める割合を示す．

図 2 と図 3 にλに対する説明変数の数と再現率の結果を示す．図 2 と図 3 より，λの値が 0 の時，141 変数全てが採用されており，再現率は約 0.86 となった．λが大きくなるにつれて説明変数の数は減少し，再現率が低下する傾向がわかる．また，λの値が 800，すなわちモデルに投入する説明変数の数が 20 を下回ると再現率の増減が激しくなっていることから，説明変数の数が少ないとモデル精度が不安定になることがわかる．

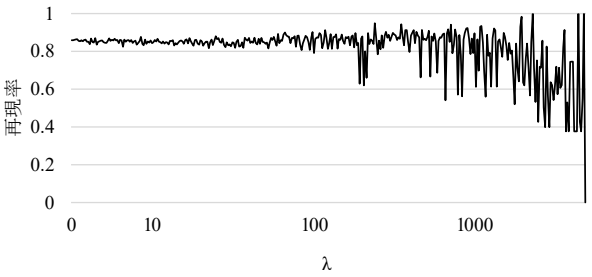

図 3　λに対する再現率

5. 予測に影響を与える説明変数の抽出

本研究では，説明変数の数が変更された数のうち，削除されなかった，すなわち採用された回数の割合が高い説明変数を予測結果に影響を与える重要な説明変数と定義する．

図 4 に各説明変数の採用率およびその順位を，表 3 に各説明変数の採用率の上位 10 変数を示す．図 4 より，採用率の上位 6 変数の採用率が 90%を超えており，予測に大きな影響を与える説明変数であることが示唆された．また，100 位以降の説明変数は大きく採用率が減少しており予測に関係ない説明変数といえる．表 3 の採用率が高い変数に着目すると，例えば「県内居住者ダミー」に関して地域銀行では県外顧客よりも県内顧客の方が住宅ローンを契約する可能性が高いと推察でき，住宅ローンの契約確率に影響を与える変数として納得できる結果である．「水道料振替実績有無」や「電気料振替契約有無」などは，住宅ローン契約を行う顧客の多くが当該銀行をメインバンクとして利用しており，銀行のサービスや商品を利用する可能性が高いと推察できる．

次に，採用率の上位 25 変数のみを用いて深層学習を行った結果，再現率の値は 0.83 となり，141 変数全てを用いた場合の再現率 0.86 と比べて値は大きく低下していないことがわかった．この結果から予測精度を大きく低下させることなく予測結果に影響を与える重要な説明変数を抽出できていることが確認できる．

6. まとめと今後の課題

本研究では，LassoNet を用いて変数抽出の有用性の検証を行った．実証分析として銀行顧客データを用いて住宅ローン契約確率を予測した．構築したモデルによって目的変数への影響が大きいと判定された回数が多い説明変数を予測に影響を与える重要な説明変数と定義し抽出した．抽出された説明変数のみを用いて深層学習を行うことで説明変数の抽出前後で再現率に大きな変化はなく，抽出された説明変数のみでも高い精度での予測が可能であることを示した．以上のことから，高い精度での予測を行いながら，重要な要因の抽出をし，新たな顧客の獲得に向けた知見を得られる分析手法であることを示した．

本研究では，ある時点における顧客のクロスセクションデータを用いて住宅ローン契約の可能性をモデル化した．しかし，クロスセクションデータを用いた分析では，相関関係は解明できても因果関係の特定には至らない．顧客の銀行利用状況を継続的に記録した複数時点のパネルデータを用いることが，ローン契約可能性が高い顧客の特徴分析を行うための今後の重要な課題である．

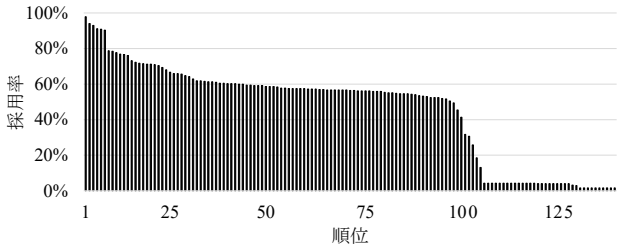

図 4　各説明変数の採用率および順位

表 3　各説明変数の採用率上位 10 変数

説明変数	採用率
水道料振替実績有無	97.849%
県内居住者ダミー	94.086%
消費者ローン契約有無	93.010%
普通預金貸越サービス契約有無	91.129%
電気料振替実績有無	90.860%
性別	90.322%
給与振込金額（円）	78.763%
インターネットバンキングサービス契約有無	78.495%
損害保険振替契約有無	77.688%
クレジット振替契約有無	76.882%

謝辞

本研究はある銀行からデータ提供を賜り実施した．ここに記して謝意を表する．

参考文献

[1] 総務省，“人工知能（AI）の現状と未来”，情報通信白書，pp.232-241，2016. https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/h28/pdf/n4200000.pdf

[2] Cynthia Rudin，Chaofan Chen，Zhi Chen，Haiyang Huang，Lesia Semenova，Chudi Zhong，“Interpretable machine learning: fundamental principles and 10 grand challenges”，Statistics Surveys，Vol.16，pp.1-85，2022.

[3] Christoph Molnar，“Interpretable machine learning: a guide for making black box models explainable”，Independently published，2024. https://christophm.github.io/interpretable-ml-book/

[4] 林直樹，永原正章，“超スマート社会を支える分散スパースモデリング：マルチエージェントネットワーク上のビッグデータ解析”，電子情報通信学会基礎・協会ソサイエティ Fundamental Review，第 13 巻，2 号，pp.95-107，2019.

[5] 日高昇治，松下亮佑，楠田哲也，“スパースモデリングって何だ?：データ構造を解き明かす先端技法”，株式会社カットシステム，2017.

[6] 嶋村重治，菊池博史，松田崇弘，金寛，吉川栄一，中村佳敬，牛尾知雄，“圧縮センシングを用いた気象用レーダの大容量観測データの圧縮”，電気学会論文誌 A（基礎・材料・共通部門誌），第 135 巻，11 号，pp.704-710，2015.

[7] Ismael Lemhadri，Feng Ruan，Louis Abraham，Robert Tibshirani，“LassoNet: a neural network with feature sparsity”，Journal of Machine Learning Research，Vol.22，No.127，pp.5633-5661，2021

[8] Xiaolu Xu，Zitong Qi，Xiumei Han，Yuxing Wang，Ming Yu，Zhaohong Geng，“Combined-task deep network based on LassoNet feature selection for predicting the comorbidities of acute coronary syndrome”，Computers in Biology and Medicine，Vol.170，107992，2024.

[9] Jiatong Han，Hao Li，Han Lin，Pingping Wu，Shidan Wang，Juan Tu，Jing Lu，“Depression prediction based on LassoNet-RNN model: a longitudinal study”，Heliyon，Vol.9，Issue10，e20684，2023.

[10] Shuo Wang，Leandro L Minku，Xin Yao，“A systematic study of online class imbalance learning with concept drift”，IEEE Transactions on Neural Networks And Learning Systems，Vol.29，Issue10，pp.4802-4821，2018.

[11] GitHub，“Welcome to LassoNet’s documentation!” https://lasso-net.github.io//lassonet/api/