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　６ 年　 ６ 月　 ５ 日現在

機関番号：３３９１６
研究種目：研究活動スタート支援
研究期間：2022～2023
課題番号：２２Ｋ２１１８６

研究課題名（和文）Development of a novel method for prediction using artificial image and image identification

研究課題名（英文）Development of a novel method for prediction using artificial image and image identification

研究代表者

ＨＥ　ＹＵＰＥＮＧ（He, Yupeng）

藤田医科大学・医学部・助教

研究者番号：００９５３２６７

交付決定額（研究期間全体）：（直接経費）　　1,000,000円

研究成果の概要（和文）：本研究では、モデルの精度を向上させるため、人工画像を用いた新手法を開発した。この概念は画像識別から得られた。デジタル画像のピクセルは識別モデルの訓練時に特徴として使用される。同様に、疫学データにも特徴の順序に関係があると仮定した。特徴をピクセルに変換し、ピクセルの順序を入れ替えた拡張された人工画像サンプルセットを用いてモデルを訓練した。予備実験では、10,000個の人工画像サンプルセットをランダムに選定し、複数のモデルを訓練し、精度（ROC曲線下面積の値）は鐘形分布を示した。特徴の順序がモデル性能に強く影響を与えることを示している。新手法はモデルの予測精度を向上させる可能性を示唆している。

研究成果の学術的意義や社会的意義
従来の疫学研究でよく使われる線形モデルと比較して、本研究で開発した新手法は、特徴を2次元人工画像の形式で配置することで、１）モデルの精度を向上させる。２）複数の特徴間の交絡要因を究明できる。３）ブラックボックスのような機械学習モデルを視覚的に説明できる。４）特徴の位置を使用して特徴の重要性を説明する。５）疫学調査以外の順序不特定のデータの分析に活用できる。

研究成果の概要（英文）：A novel method using artificial image was developed to enhance the model precision in epidemiology study. The concept was inspired from image identification. Pixels in digital images are used as features when training the identification model. The order-related relationship is assumed to exist in epidemiological data. Given a certain dataset, features are transformed to pixel values for generating artificial images. Orders of pixels are randomly permutated and the model is trained using pixel-permutated artificial image sample sets. In the preliminary experiment, one binary response was designed to be predicted by 76 features. 10,000 artificial image sample sets were randomly selected for model training. Models' performance (area under the receiver operating characteristic curve values) depicted a bell-shaped distribution. Namely, feature order information had a strong impact on model performance. Our novel model construction strategy has potential to enhance model predictability.

研究分野：Public Health

キーワード：Artificial image　人工画像　疫学研究　予測モデル　機械学習

様　式　C－１９、F－１９－１（共通）

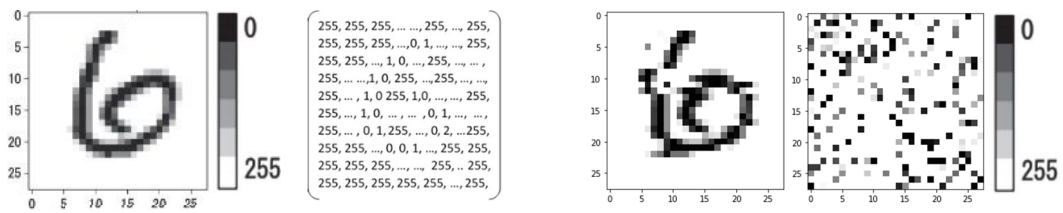
１．研究開始当初の背景

線形モデルは疫学研究でよく使う統計方法である。それらのモデルを利用し、因子と疾患の関連性の推定や、個体の疾患の発生リスクの予測などの問題を解決できす。線形モデルには、線形回帰、ロジスティック回帰、コックス回帰などの一般化線形回帰の中のさまざまな手法を含む。先行研究では、線形モデルは非線形関連や複雑な相互作用の解析などの限界があることを報告しており、それらの問題を解決するために非線形モデルを採用した。例えば、人工ニューラルネットワークという非線形モデルを疫学研究に導入したことがある。

モデルの精度を向上させるために、先行研究では次の解決策を実施した。(1)特徴量を増やし、サンプルサイズを拡大する。(2)パラメータの最適化を行う。(3)複数の機械学習手法を比較し、最も高いパフォーマンスを示す手法を選定する。例えば、非線形モデル (neural networks、support vector machines など) を使用し、特徴間の多重共線性や不十分な特徴選択によって引き起こされる線形方程式の弱点に対処し、より複雑な情報を抽出することが可能である[1-3]。

画像認識技術にインスパイアされたもので、優れた特徴の構造がモデルトレーニングにおける重要な要素である。例として手書き数字認識を挙げると（図 1）、デジタル画像はピクセルで構成されている（図 1-a）。ピクセル値は 0 から 255 の範囲にあり（図 1-b）、色は黒から白へと変化する（図 1-a）。各ピクセルは画像認識モデルのトレーニングで特徴として使用される（図 1-b）。しかし、ピクセルの順序を任意に変更すると（図 1-c）、手書き数字は正しく認識されなくなる。

一方、ニューラルネットワークに関する先行研究では、ディープラーニングに基づく画像認識技術によって卓越した精度を達成しており、多くの場合 98%を超えている。線形モデルに基づく疫学研究では、予測精度が 90%以上であるという報告は少ない。したがって、疫学調査データの項目をピクセルとみなすと、それらのピクセルを一定の順序で並べ、人工画像を形成することができる。さらに、この高精度の画像認識技術を利用して新たな手法を構築し、疫学研究に活用できると考えている。



S1(a)　　S1(b)　　S1(c)

図 1

２．研究の目的

人工画像によりモデルの精度を向上させる新手法を開発する。

３．研究の方法

この部分で新手法の開発の過程を説明する。

3.1 人工画像の形成：変数のピクセル化とピクセルの配列

3.1.1 変数のピクセル化

グレースケール画像では、ピクセルは 0 から 255 までの値を示す 8 ビット整数で表される（図 S1-c)。疫学調査で収集されたデータを 8 ビットの整数に変換しないと、人工画像を形成することができない。したがって、下記の関数 rescale function $\mathbb{P}$を作成した。

$$\mathbb{P}(X_n) = \left\lfloor \left(\frac{X_n - \min(X_n)}{\max(X_n) - \min(X_n)} \right) \times 255 + 0.5 \right\rfloor, n = 1, \ldots, f$$

疫学調査データの各特徴について、リスケール関数$\mathbb{P}$ を適用し、特徴の値を 0 から 255 の範囲内で正規化する。たとえば、5 つのサンプルを含むデータの場合、特徴の 1 つは年齢：20 歳、30 歳、40 歳、50 歳、60 歳。これらの年齢値は、0、64、128、191、255 のグレーレベルに正規化し、ピクセルに変更する。

3.1.2 ピクセルの配列

ある疫学調査データでは f個特徴を含めている。f個の特徴を f個の画素に変換した場合、$f!$個の可能な画素順序が存在する。今回は研究方法を単純化するために、画像の回転(90°, 180°, or 270°)や反転（vertical, horizontal, or diagonal）を考慮せず、ピクセルを正方形の配列に整理することを選択した。その結果、画像に f個の特徴が与えられる可能性 Fは $1/8f!$に等しくなる。図 2 は、あるサンプルのある画素順序の一つの人工画像の例である。

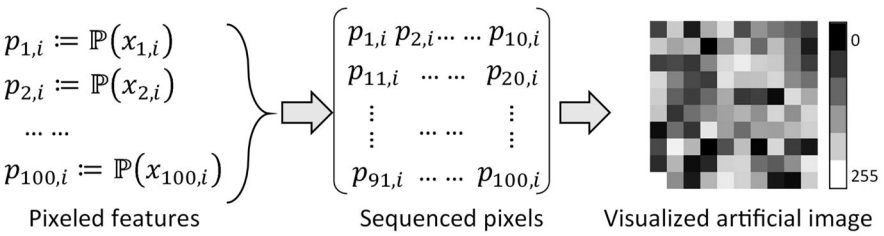


図 2

3.2 データの拡張

f個の特徴から F種類の人工画像を生成することで、元のデータセットを F種類の異なるデータセットに拡張できる。そこで、元のデータを $S_{original}$とし、拡張されたデータセットを S_1、S_2、…、S_Fと表す。これらの拡張データセットを *Candidate* データセットとする。(図 3).

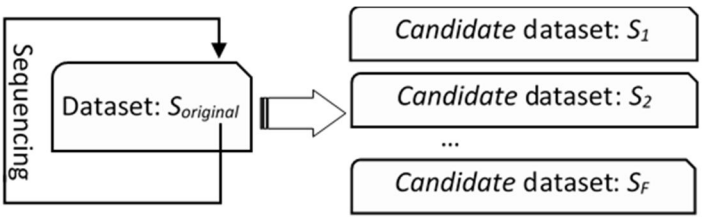


図 3

3.3 データプロセシング

F 個の *Candidate* データセットのそれぞれに、同一の処理（プロセシング）を行っう。まず、各 *Candidate* データセットを 70:10:20 の比率でランダムにトレーニング、検証、テストの

三つのセットに分ける。バランスの取れた学習を確保するため、Synthetic Minority Oversampling Technique（SMOTE 手法）をトレーニングセットに適用する[4]。モデルは、トレーニングセットと検証セットで画像識別手法を使用し訓練される。 具体的には、モデルの学習プロセスはトレーニングセット内で実行され、学習効果を評価するための継続的な評価を行い、検証セットで評価された損失関数の値が増加しなくなった時点で訓練が終了する。次に、モデルをテストセットに適用してパフォーマンスを評価し、結果を記録する。(図 4)。

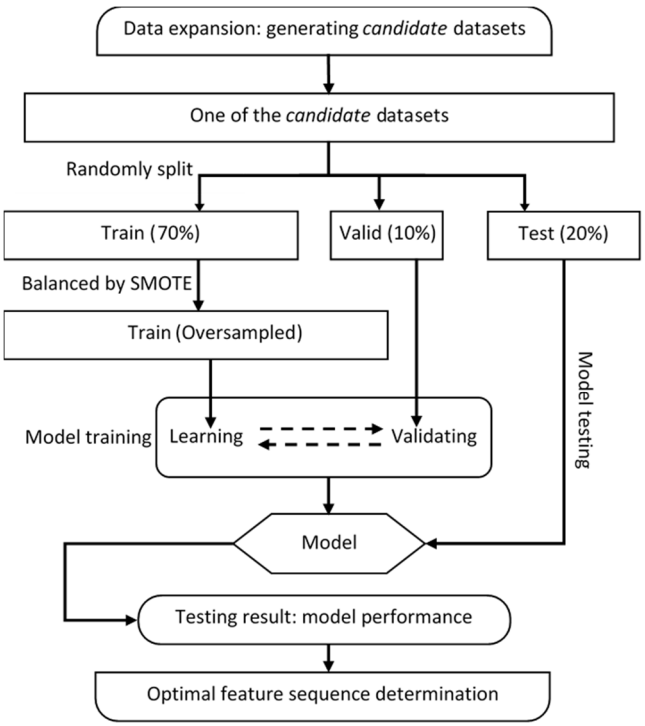


図 4

最高のパフォーマンスを示したモデルが最適な予測モデルとみなされる。したがって、特定の特徴順序から生成された人工画像を含むこの *Candidate* データセットは、モデル構築に最適なデータセットであると考えられる。これらの人工画像は、特徴と応答の間の複雑な関係を効果的に捉えている。

４．研究成果

新手法の実現可能性を証明するため、次の実験を実施した。

4.1 予備実験

このセクションでは、前述で開発した手法を使用し、統合失調症判別モデルの構築を例として、この手法の有効性を検証する。

統合失調症の有無を分類するためのモデルを作成する。インターネット調査機関の pooled panel によって実施されたオンライン調査からデータを収集した[5]。データセットは、223 人の統合失調症患者と 1,776 人の健康な対照者を合わせて構成した。各サンプルについて、これらの情報を調査から抽出した：統合失調症の有無と、人口統計、健康関連背景、身体的併存疾患、精神医学的併存疾患、社会的併存疾患を含む 76 の特徴である。サンプリングのの詳細と特徴の定義については他の場所で公開されている[6]。

モデルの構築は人工ニューラルネットワークを使用して行った。具体的には、5 つの隠れ層(各層のニューロン数：128-64-32-16-8)、HeNormal 重み初期値、隠れ層の ReLU 活性化関数、出力層の Sigmoid 活性化関数で構成された。学習率 0.01 と設定し、5 回の連続更新でパフォーマンス値は 0.001 の増加を達成しない場合の早期停止にした。モデルのパフォーマンスは、受信機動作特性曲線下面積（AUC）を使用して評価された。（AUC の閾値は：0.5＝識別なし、0.5～0.7＝識別不良、0.7～0.8＝許容可能な識別、0.8～0.9＝優れた識別、0.9 超＝卓越した識別。）*Candidate* データセットの数は巨大ので、すべての可能性をあげるのは困難であったため、本予備実験ではランダムに選択した 10,000 の *Candidate* データセットで実験を行った。統計分析は Python 3.8 と Jupyter Notebook を Coding book として使用した。

4.2 結果

予備実験に基づいて、10,000 回の実験にわたる AUC スコアの分布を示している。ほとんどの

モデルは、約 0.88 の AUC スコアを達成し、優れた識別力を示している。ただし、一部のモデルは 0.5〜0.7 の AUC スコアを達成し、さらに少数のモデルは 0.93 を超えるスコアを達成し、優れた識別力を示した （図 5)。つまり、新手法によりモデルの精度を向上させることを示唆している。

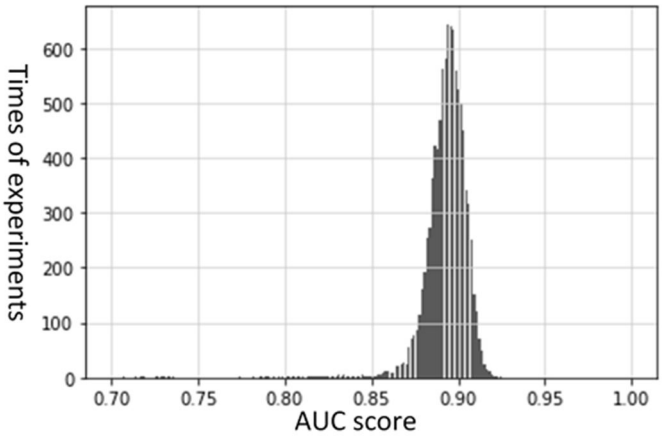


図 5

4.3 考察

この実験では、訓練されたモデルの精度は人工画像内の特徴の位置によって異なることを示した。数が少ないデータセットによって、高精度のモデルを作成できる。つまり、新手法がモデルの精度を向上できること、新手法によって最適のモデルと最良の人工画像を得る可能性があることを証明した。

新手法の強みがあるにもかかわらず、完全な実装にはいくつかの課題に対処する必要がある。まず、この手法はかなりの計算能力を必要とする。次に、予備実験では、データセットには十分な数の特徴が欠けていた。そのため、モデルの訓練においては、畳み込みニューラルネットワークのような成熟した画像認識技術を使用する条件を満たさない。今後の実験では、少なくとも 400 の特徴を導入することを予定する。また、特徴と応答の関係は無視できない。特徴と応答の間に関係がないか、関係が弱い場合、最先端の手法であっても高精度の予測モデルを構築するのは困難である。

参考文献

1. Masegosa AR, Cabañas R, Langseth H, Nielsen TD, Salmerón A.. Probabilistic models with deep neural networks. Entropy (Basel). 2021;23(1):117.
2. Greenland S, Daniel R, Pearce N.. Outcome modelling strategies in epidemiology: traditional methods and basic alternatives. Int J Epidemiol. 2016;45(2):565-575.
3. Guyon I, Elisseeff A.. An introduction to variable and feature selection. J Mach Learn Res. 2003;3:1157-1182.
4. Chawla, N, Bowyer K, Hall L, Kegelmeyer, W. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002 Jun 2;16: 321-357.
5. 大規模疫学研究データと診療報酬明細書（レセプト）データを用いた一般住民における入院外統合失調症及び統合失調症関連障害の有病率推定方法の開発. https://mhlw-grants.niph.go.jp/project/169571
6. Matsunaga M, Li Y, He Y, et al. Physical, Psychiatric, and Social Comorbidities of Individuals with Schizophrenia Living in the Community in Japan. Int J Environ Res Public Health. 2023; 20(5):4336.

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

１．著者名 He Yupeng、Sun Qiwen、Matsunaga Masaaki、Ota Atsuhiko	４．巻 7
２．論文標題 Can feature structure improve model ' s precision? A novel prediction method using artificial image and image identification	５．発行年 2024年
３．雑誌名 JAMIA Open	６．最初と最後の頁 1-4
掲載論文のDOI（デジタルオブジェクト識別子） 10.1093/jamiaopen/ooae012	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 －

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　1件）

１．発表者名 He Yupeng
２．発表標題 Does the Feature Order Affect the Performance of Artificial Neural Network Model? A classifier for the existence of schizophrenia based on a Japanese online survey（特徴量の順序は人工ニューラルネットワークモデルの性能に影響するか？ 日本のオンライン調査に基づく統合失調症有無の分類）
３．学会等名 第9回藤田医科大学学内研究シーズ・ニーズ発表交流会
４．発表年 2023年

１．発表者名 He Yupeng, Matsunaga Masaaki, Ota Atsuhiko
２．発表標題 Development of a novel method for prediction using artificial image and image identification
３．学会等名 第34回日本疫学会学術総会（国際学会）
４．発表年 2024年

〔図書〕　計0件

〔出願〕　計1件

産業財産権の名称 人工画像データ生成装置、予測装置、人工画像データ生成方法、予測方法、及びプログラム	発明者 He Yupeng	権利者 藤田学園
産業財産権の種類、番号 特許、Ｊ５８９９１Ａ１	出願年 2023年	国内・外国の別 国内

〔取得〕　計0件

〔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