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　５ 年　 ６ 月 １７ 日現在

機関番号：３４３０９
研究種目：研究活動スタート支援
研究期間：2021～2022
課題番号：２１Ｋ２１２９７

研究課題名（和文）高次元データを含む不均衡データを用いた回帰問題のためのデータバランシング手法

研究課題名（英文）Data balancing for regression using imbalanced dataset

研究代表者
吉川　寛樹（Yoshikawa, Hiroki）

京都橘大学・工学部・助教R

研究者番号：１０９０５３５０

交付決定額（研究期間全体）：（直接経費）　　2,400,000円

研究成果の概要（和文）：回帰問題，分類問題，それぞれに対し推定値の不均衡を解消するための手法を提案した．1つ目は時系列データを説明変数とする回帰問題のためのデータバランシング手法である．この手法ではデータセットから抽出した2つのサンプルから，内挿的に時系列データを生成することで新たなサンプルを生成する．性能評価では平均絶対誤差の増加を抑えつつ少数データに対する推定精度を向上させることが可能であることを確認した．2つ目の手法は，条件付き敵対的生成ネットワークを用いた，分類問題のためのデータバランシング手法である．性能評価ではオープンデータセットを用いて評価を行い均衡の取れた推定器の訓練が可能となることを確認した．

研究成果の学術的意義や社会的意義
利用者が気づきにくい不均衡データによる推定値の偏りを軽減する手法を提案し，様々な機械学習との組み合わせ・応用を可能とする点が本研究の社会的意義である．特に近年ではセンシングデバイスの小型化・低価格化が進み，機械学習の科学・医療など様々な分野への応用手法が開発されていることから，今後ますますモバイル・ユビキタス分野において機械学習は利用されることが予想される．そのような応用事例において本研究は大きな役割を果たすと申請者は考える．

研究成果の概要（英文）：We propose methods to address the imbalance of estimated values in regression and classification problems, respectively. The first method is a data balancing technique for regression problems using time series data as explanatory variables. This method generates new samples by interpolating time series data from two extracted samples in the dataset. Through performance evaluation, we found that it is possible to improve the estimation accuracy for minority data while suppressing the increase in mean absolute error. The second method is a data balancing technique for classification problems using conditional generative adversarial networks. Through performance evaluation using open datasets, we found that the proposed method achieved training a well-balanced estimator.

研究分野：情報ネットワーク

キーワード：機械学習　不均衡データ　データバランシング　分類問題　回帰問題

１．研究開始当初の背景

近年では深層学習を用いて複雑な処理を伴う手法が多く開発されているが，その処理がパッケージ化されることで専門的な知識無しに利用が可能となっている．しかしながら，訓練のされ方によらず推定器は何らかの推定値を出力するため，利用者は推定器がもっともらしい値を出力していれば正しく動作していると思い込んでしまう．気づくことが難しい問題の一つに，不均衡データの訓練による推定値の偏りがある．不均衡データと呼ばれる，正解値の分布に偏りがあるデータを訓練に用いると，推定器は多数派のデータを必要以上に推定値として出力しやすくなり，本来少数派のデータに対しても多数派と誤推定してしまう．

連続値を推定対象とする回帰問題においてもこの問題が見られる．しかし実世界においては，異常時の状態を正確に推定したい問題が多く存在する．例えば，発熱者の病状の深刻さの指標とするために体温推定器を機械学習によって構築する場合には，発熱状態にある異常時のデータに対し正確に推定する必要がある．

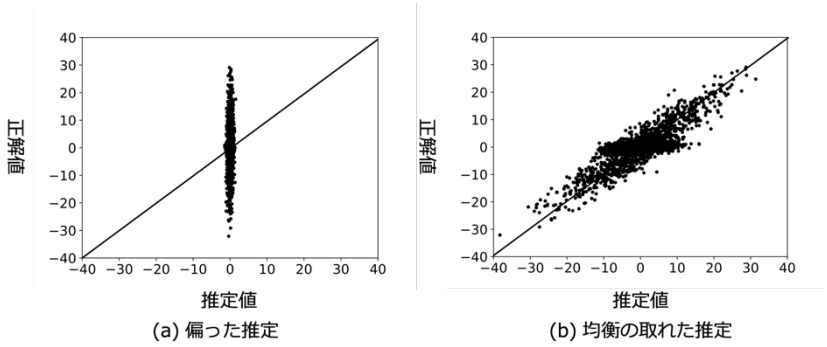

図 1: 推定値の偏りの例

このような推定値の偏りは，本質的には正解値と推定値の絶対誤差のみに基づいて推定器を訓練することに起因する．例えば，一見訓練が失敗している図 1(a)の例に対し，図 1(b)の例は正解値と推定値に相関があり，一見訓練が成功しているように見える．しかし一般的に用いられる評価指標である推定値の平均絶対誤差（MAE）を算出すると，図 1(a), (b)ともに 3.2 と同じ値になり，MAE だけでは図示したような違いを見つけることが難しい．訓練データの分布を自由に変更できるような収集環境においては問題とはならないが，物理的あるいは倫理的側面から不均衡データとならざるをえない状況があり，特に人の生体情報から訓練データを収集する際には発生しやすい問題である．

２．研究の目的

「１．研究開始当初の背景」で述べた機械学習における不均衡データを用いて訓練を行う際の推定値の偏りを軽減する手法の提案が本研究の目的である．

図1に示すような不均衡データを訓練に用いると本来少数派のデータに対しても多数派と推定しやすい推定器が訓練される．本研究では連続値を推定対象とする回帰問題において，分類問題で一般的に用いられる手法であるデータバランシングを応用し，推定値が偏る問題の解決を目指す．

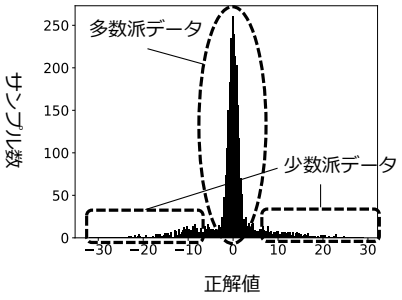

図 2: 不均衡データの分布例

３．研究の方法

本研究では前述の課題に対し，データバランシングによる解決を目指す．提案手法の概要を図 3 に示す．本研究では，敵対生成ネットワークの派生であり，ラベル付きデータを生成する条件付き敵対生成ネットワーク（cGAN）を応用する．cGAN は，既に分類問題において高い性能の向上を実現しており，横断面データに

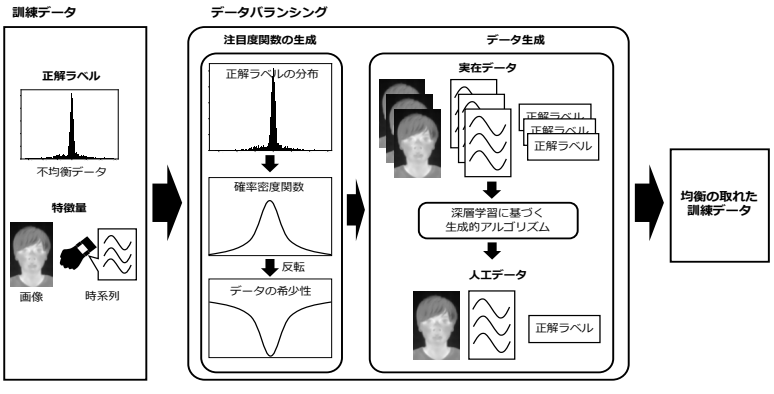

図 3: 手法の概要

限らず，時系列や画像などさらに高次元のデータの生成が可能となる．cGAN を回帰問題に対して応用するために，注目度関数（データの希少性を示す関数)を確率密度関数に基づいて生成する．これにより実在データを重み付きで学習することで正確に人工データの生成が可能となる．さらに性能評価として被験者から画像や時系列を含む生体データを収集し，測定の難しい人体の深部体温などの推定器を用いて検証を行うことで実データに対する有効性を検証する．

４．研究成果

研究目的の達成のため，回帰問題，分類問題，それぞれに対し 2 通りのアプローチを用いて推定値の不均衡を解消するための手法をそれぞれ提案した．1 つ目は時系列データを説明変数とする回帰問題のためのデータバランシング手法である．この手法ではデータセットから抽出した 2 つのサンプルから，内挿的に時系列データを生成することで新たなサンプルを生成する．生体情報を収集した 2 つの異なるデータセットを用いた性能評価では，提案手法により生成されたデータセットを用いて推定器を訓練し，目的変数の推定を行なった．この評価実験から，提案手法により生成されたデータセットを用いることで，平均絶対誤差の増加を抑えつつ少数データに対する推定精度を向上させることが可能であることを確認した．2 つ目の手法は，条件付き敵対的生成ネットワーク(cGAN)を用いた，分類問題のためのデータバランシング手法である．こちらの手法では，不均衡なデータセットを用いて訓練した生成モデルから生成されるデータの偏りの解消を目的としている．性能評価では，3 つのオープンデータセットを含む 4 つのデータセットに対し手法を適用することで評価を行い，訓練時にデータセットに含まれるデータの分布を元に損失関数の出力に重みを与えることで，均衡の取れた推定器の訓練が可能となることを確認した．

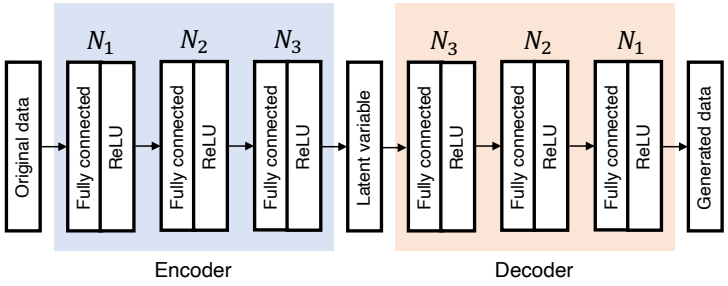

図 4: 提案手法の構成

さらに不均衡データセットにより引き起こされる推定値の偏りを軽減するための手法であるデータバランシングにおいて，元のデータセットに人々のプライバシーに関わる情報が含まれることを想定し，深層学習を用いた生成モデルであるオートエンコーダを用いて人を特定可能な情報を除去する手法を提案した（図 4）．性能評価では人の温冷感を収集したデータセットに含まれる手首装着型センサから取得した心拍数データやサーモグラフィから取得した人の体温データに含まれる人の特定に関わる情報の除去を試みた．

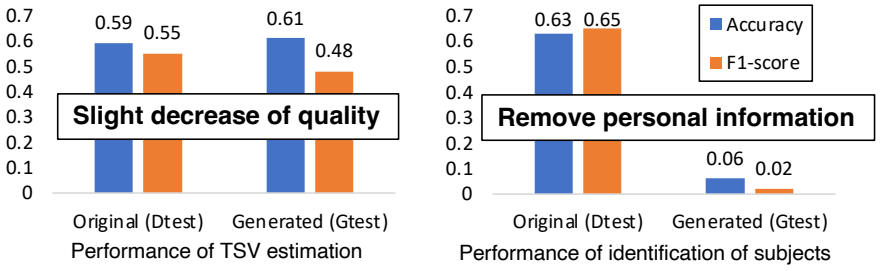

図 5: 温冷感推定結果と被験者 ID 推定結果

図 5 に示す評価結果によると提案手法が，前述の情報を削減しながら，元のデータセットを用いた訓練を行ったときと同程度の推定器の性能を引き出すことが可能であることを示した．

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 Hiroki Yoshikawa, Akira Uchiyama, Teruo Higashino	４．巻 9
２．論文標題 TSVNet: Combining Time-Series and Opportunistic Sensing by Transfer Learning for Dynamic Thermal Sensation Estimation	５．発行年 2021年
３．雑誌名 IEEE Access	６．最初と最後の頁 102835 - 102846
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/ACCESS.2021.3097882	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6件（うち招待講演　2件／うち国際学会　3件）

１．発表者名 Hiroki Yoshikawa, Akira Uchiyama, Teruo Higashino
２．発表標題 Data Balancing for Thermal Comfort Datasets Using Conditional Wasserstein GAN with a Weighted Loss Function
３．学会等名 The 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys 2021) Workshops（国際学会）
４．発表年 2021年

１．発表者名 Hiroki Yoshikawa, Akira Uchiyama, Teruo Higashino
２．発表標題 Time-Series Physiological Data Balancing for Regression
３．学会等名 The 2021 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA 2021)（国際学会）
４．発表年 2021年

１．発表者名 吉川 寛樹，内山 彰，東野 輝夫
２．発表標題 不均衡データセットを用いた回帰問題における損失関数の検討
３．学会等名 情報処理学会MBL研究会第99回研究発表会
４．発表年 2021年

１．発表者名
Hiroki Yoshikawa, Akira Uchiyama, Teruo Higashino

２．発表標題
Privacy-Preserving Data Augmentation for Thermal Sensation Dataset Based on Variational Autoencoder

３．学会等名
The 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation (BuildSys 2022)（国際学会）

４．発表年
2022年

１．発表者名
吉川 寛樹

２．発表標題
時系列生体データを用いた機械学習による温冷感推定

３．学会等名
情報処理学会MBL研究会第105回研究発表会（招待講演）

４．発表年
2022年

１．発表者名
吉川 寛樹

２．発表標題
不均衡なデータセットのためのデータバランシングと時系列生成

３．学会等名
電子情報通信学会ソサエティ大会（招待講演）

４．発表年
2022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