2D-07

時系列生体情報を用いた回帰問題のための データバランシング手法の提案

吉川 寛樹 †
大阪大学
大学院情報科学研究科

内山 彰 ‡
大阪大学
大学院情報科学研究科

東野 輝夫 §
大阪大学
大学院情報科学研究科

1 はじめに

近年，機械学習はヘルスケアを含む様々な分野において応用されている．このような応用例では，心拍や体温の時系列情報などの特徴量に基づき心理状態や深部体温など，直接測定することが難しい生体情報を推定する．多くの手法では，機械学習により推定器を構築するため，実環境においてデータを収集する．この際，特にヘルスケア応用においては，データの不均衡が避けられないという大きな課題がある．通常，実環境において希少なデータは学習用のデータセットにもほとんど含まれず，推定値として必要以上に出力されにくくなる．この問題を解決するために，データバランシングと呼ばれる手法が用いられる．データバランシングはオーバーサンプリングやアンダーサンプリングによって，特定の正解ラベルを持つデータの数を増減させることで不均衡を軽減する．既存手法の多くは，分類問題を対象としているのに対して，連続値を推定する問題（回帰問題）を対象としたデータバランシング手法として，SMOTER [1] が提案されている．しかし，ヘルスケア応用においては，心拍や体温などの時系列データが入力として含まれることが多い．また，推定対象も快適度やストレスレベルなどの連続値であることが望ましい場合もある．したがって，時系列データを入力に含む回帰問題に対するデータバランシング手法が必要であるが，そのような手法は我々の知る限り見当たらない．そこで本研究では，SMOTER を基に，時系列データを含むデータセットに対して拡張した手法を提案する．提案手法は Dynamic Time Warping (DTW) 距離を用いて時系列データ間の距離を定義することで，推定に必要な情報を保存しながら時系列データを生成する．

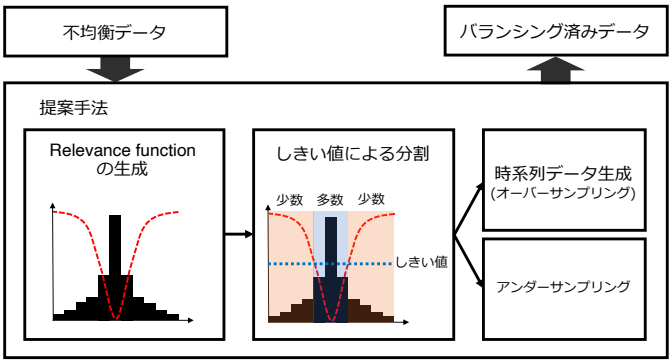

図 1 提案手法の概要.

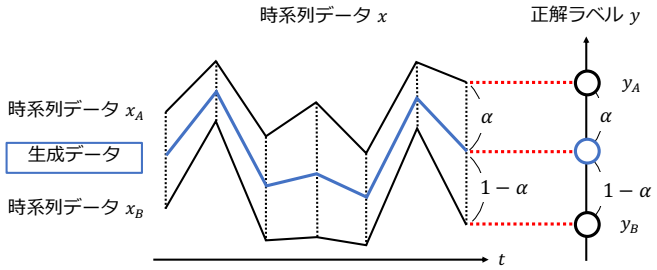

図 2 時系列データ生成手法.

2 提案手法

提案手法の概要を図 1 に示す．入力は時系列データを含む不均衡データセットであり，出力はバランシング後のデータセットである．まず，不均衡データセットにおける正解ラベル y の分布に基づいて relevance function $\phi(y)$ [1] を生成する．ここで，$\phi(y)$ は確率密度関数に基づいて自動的に生成される．次に，不均衡データの分布をしきい値 t_E に基づいて少数データセット D_r と多数データセット D_c に分割する．t_E は対象のデータセットに対し経験的に決定される．分割後，オーバーサンプリングとアンダーサンプリングがそれぞれ D_r と D_c に対して実行される．

オーバーサンプリングは D_r に基づき，時系列データを生成することにより行われる．図 2

A Proposal of a Data Balancing Method for Regression Problems Using Time-Series Physiological Information
† Hiroki Yoshikawa, Graduate School of Information Science and Technology, Osaka University
‡ Akira Uchiyama, Graduate School of Information Science and Technology, Osaka University
§ Teruo Higashino, Graduate School of Information Science and Technology, Osaka University

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.

に提案手法における時系列データ生成のキーアイデアを示す．生成される時系列データは，D_r から距離によって選出される 2 サンプルの加重平均に基づく比 α に基づいて，内挿的に生成される．この生成方法は SMOTER に基づいており，本研究の貢献は SMOTER への DTW 距離の適用である．DTW 距離は最小値や最大値などを保存しながら，時系列データを内挿的に生成することができる [2]．図 2 に示すように，生成される時系列データの各時点での値は DTW によって決定されるペアの間に比 α を用いて内挿される．

3 性能評価

温冷感データセット，深部体温データセットの 2 種類を用いて評価を行った．温冷感データセットでは，温冷感と，温冷感に関わるとされる生体情報を 21 名の被験者から，のべ 1686 回収集した．正解ラベルは Thermal Sensation Vote (TSV) と呼ばれる温冷感の申告値である．被験者は，ASHRAE の 7 段階指標[*1]と呼ばれる, 温冷感を Cold, Cool, Slightly Cool, Neutral, Slightly Warm，Warm，Hot の 7 段階に分割した指標に従って TSV を [-3.5, 3.5] の範囲で申告した．また，生体情報として，心拍数，体表温度，皮膚電気抵抗を腕時計型センサから収集した．深部体温データセットは，13 名の運動時のデータである．正解ラベルは鼓膜温度センサにより収集した深部体温であり，その他推定に用いる生体情報として，胸部センサから心拍数，衣服内温度，腕時計型センサから心拍数，体表温度，皮膚電気抵抗，加速度，環境センサから気温，湿度を収集した．評価指標には文献 [3] で提案されている，回帰問題のための少数データ対する適合率 P_r，再現率 R_r，F 値 F_r と平均絶対誤差 (MAE) を用いる．また，比較手法にはユークリッド距離に基づいて，時系列上の同時刻のペアに対して内挿的にデータを生成する SMOTER を使用する．

表 1 と表 2 はそれぞれ，各データセットの正解ラベルに対し，各バランシング手法を用いて，LSTM 層を含む深層学習モデルによる推定を行ったときの評価結果である．まず，バランシングを行わずに学習した場合は少数データに対し，両データセット共に F_r が最も低い結果となった．これはデータセット全体に対する MAE では観測することができない，少数データに対する誤差が大きいことを示している．この傾向は不均衡データセットを用いたクラス分類問題にも同様にみられる．次に両データセットに対する各指標において，提案手法は SMOTER と比較して，同じか上回る結果となった．これは DTW 距離に基づく時系列データの生成が，推定に必要な情報を保存し，比較的有効に働くことを示している．これらの結果から，提案手法は MAE 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がわかった．

表 1　温冷感データセットを用いた推定結果.

Method	P_r	R_r	F_r	MAE
バランシング無し	0.00	0.18	0.00	**0.52**
SMOTER [1]	0.43	0.31	0.36	0.56
提案手法	**0.75**	**0.32**	**0.44**	0.56

表 2　深部体温データセットを用いた推定結果.

Method	P_r	R_r	F_r	MAE
バランシング無し	0.35	0.51	0.40	0.38
SMOTER [1]	0.45	0.50	0.44	0.38
提案手法	**0.77**	**0.60**	**0.67**	**0.35**

4 おわりに

本稿では，回帰問題を対象とした時系列データを含むデータセットのためのデータバランシング手法を提案した．実データを対象とし評価したところ，提案手法がデータセット全体に対する推定誤差の増加を抑えつつ，希少なデータに対する推定精度を向上させることがわかった．

参考文献

[1] Luís Torgo, Rita Ribeiro, Bernhard Pfahringer, and Paula Branco, “Smote for regression,” *Progress in Artificial Intelligence*, vol. 8154, pp. 378–389, September 2013.

[2] Enrique A. de la Cal, José R. Villar, Paula M. Vergara, Álvaro Herrero, and Javier Sedano, “Design issues in time series dataset balancing algorithms,” *Neural Computing and Applications*, pp. 1–18, January 2019.

[3] Luís Torgo and Rita Ribeiro, “Precision and recall for regression,” *Discovery Science*, vol. 4702, pp. 597 – 604, 2007.

[*1] https://www.ashrae.org/technical-resources/standards-and-guidelines/read-only-versions-of-ashrae-standards

Copyright ©2021 Information Processing Society of Japan.
All Rights Reserved.