The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

脳波特徴の階層性を利用した効率的なシャープレイ値推定による脳波識別の根拠の定量化

Quantification of contribution of features to EEG classification using the efficient estimation of Shapley Value based on a hierarchy of EEG features

立川 和樹 *1 Kazuki Tachikawa　河合 祐司 *1 Yuji Kawai　朴 志勲 *1 Jihoon Park　浅田 稔 *1 Minoru Asada

*1大阪大学大学院工学研究科
Graduate School of Engineering, Osaka University

Understanding how black-box classifiers predict is important in many applications, especially in medical diagnosis systems. We propose a method to quantify contribution of features to EEG classification using efficient estimation of the Shapley Sampling Value (SSV). EEG data have hierarchical features: an electrode signal, signals in various frequency-bands, and amplitude and phase. If contribution of a feature at a higher level (e.g., a signal of an electrode) is very small, contribution of features at the lower levels of the feature (e.g., signals of frequency-bands of the electrode) should be also small. The method prunes such features at lower levels to reduce computational complexity. We verified the usability of the method in two datasets for EEG classification. The result showed the method could reduce computational complexity of the SSV by one third, while maintaining high accuracy of the conventional SSV.

1. はじめに

深層学習の応用範囲は画像や音声，自然言語だけではなく，脳波へも広がっている [Schirrmeister 17]．このような脳波識別器を医療診断に応用する際には，その診断根拠を示すことが重要となる．しかし，深層学習を含め，ほとんどの機械学習モデルによる識別はブラックボックス化されており，識別根拠を人が理解することは難しい．そのため，学習器の識別根拠を理解するための様々な手法が提案されている [Sundararajan 17, Lundberg 17, Zhang 18]．これらの手法の中でも，シャープレイ値に基づくものは，各特徴の貢献を正確に計算するために理想的な公理を満たす唯一の方法であるため，識別根拠を理解するために理論上優れているといわれている [Lundberg 17]．しかし，シャープレイ値の計算量は，特徴数に対して階乗的に増加するため，これをより少ない計算量かつ少ない誤差で近似する手法が望まれる．そのため，Shapley Sampling Value (SSV) [Štrumbelj 14] が，識別モデルを限定しない近似手法として提案されている．

本研究では，SSV を用いることで，より正確，かつ，人に理解しやすい形式で，脳波の識別モデルの識別根拠を明らかにできることを示す．また，SSV の計算量を削減するために，脳波は複数の周波数帯成分の重ね合わせで構成され，それぞれの周波数帯成分に複数の特徴が含まれるという特徴の階層性を利用して，一度の計算に必要な特徴の数を減らすことで計算量を削減する手法を提案し，その有効性を実験的に示す．

2. 関連研究

2.1 シャープレイ値

シャープレイ値は協力ゲーム理論において協力によって得られた利得を公正に分配するための唯一の手法として Shapley により提案された [Shapley 53]．識別モデルを f とし，入力を x，入力する特徴の数を M，それぞれの特徴が識別へ貢献した度合いを ϕ とする．また，S は i 番目の特徴 x_i を除く x の部分集合である．モデル f と入力 x についての識別における i 番目の特徴の貢献は，次式で与えられる．

$$\phi_i(f,x) = \sum_{S \subseteq x \setminus i} \frac{|S|!(M-|S|-1)!}{M!}[f_{S\cup i}(S\cup i) - f_S(S)] \quad (1)$$

上式は，識別モデルへの入力 x に，ある特徴 x_i がある場合とない場合でのモデルの出力の差分を，必ずしもすべての可能な特徴の組み合わせで計算し，その差分に応じて，特徴に貢献を割り当てることを意味する．

この組み合わせを厳密に計算するためには，その組み合わせごとにモデルを学習させる必要がある．さらに，計算量は特徴の数 n に対して $\mathcal{O}(n!)$ であるため，計算量が膨大になるという問題がある．SSV は，各特徴が互いに独立に分布していることを仮定し，さらに，モンテカルロサンプリングにより，シャープレイ値を推定したものである．それによって，組み合わせごとの学習が不要になり，計算量が少なくなることが示されている [Štrumbelj 14]．

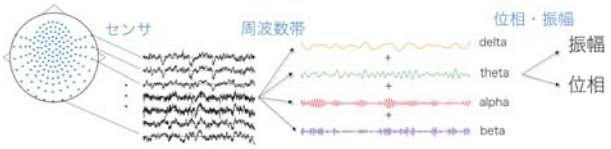

図 1: 脳波特徴の階層性

2.2 脳波識別において識別根拠を明らかにする方法

図 1 に示すように，脳波識別において識別根拠を明らかにする方法は，特徴の詳細さに応じてセンサレベル，周波数帯レベル，そして，振幅・位相レベルの三つの階層に分けられる．センサレベルでは，どのセンサの信号がどの程度識別に貢献したかを示し，周波数帯レベルでは，センサごとにどの周波数帯がどの程度識別に貢献したかを示す．さらに，振幅・位相レベルでは周波数帯ごとに位相・振幅がそれぞれどの程度識別に貢献したかを示す．

Sturm et al. は，Layer-wise Relevance Propagation (LRP) を用いた手法を提案した [Sturm 16]．LRP は識別モデルが

連絡先: 立川 和樹，kazuki.tachikawa@ams.eng.osaka-u.ac.jp

ニューラルネットの場合に，入出力の勾配を利用して識別根拠を明らかにする方法である [Ancona 17] ．この手法はシャープレイ値の公理系を満たさないが，それらを満たすように修正された Integrated Gradients (IG) 法が提案されている [Sundararajan 17] ．しかし，振幅や位相は入力空間にはない特徴であるため，この手法では，このレベルでの貢献を直接求められない．

Schirrmeister et al. は，入力する波の振幅にノイズを加え，そのノイズの大きさと出力の変化の相関を計算することにより，識別器の挙動を明らかにする方法である Input-perturbation network-prediction correlation map (IPNPCM 法) を提案した [Schirrmeister 17]．この方法は，振幅レベルでモデルの挙動を可視化できる．しかし，この方法は，波の振幅の大きさの変化とそれに伴う識別器の出力の変化の関係が線形である度合いを計算するものであり，これらの非線形な関係や他の変数との交互作用を考慮していないため，厳密な識別根拠を示すことは難しい．

3. 提案手法

各電極で計測された波をバンドパスフィルタにより任意の周波数帯ごとに分け，それぞれのバンドパスされた波を FFT により，さらに振幅成分と位相成分に分けたものを特徴とし，SSV によりそれぞれの特徴の貢献を求め，識別根拠を明らかにする手法を提案する．また，その計算量を削減する方法も提案する．図 1 のように，脳波識別の根拠を説明する方法には三つのレベルがあり，センサレベルのような低いレベルでの説明では，波を周波数帯ごとに分けないため，特徴の数は比較的少ない．また，低いレベルでの説明で識別に貢献していないと判断された特徴については，それをさらに詳細にした高いレベルでの説明においても貢献は低いと考えられる．そこで，まずセンサレベルで SSV を計算し，貢献の小さいセンサを除いて周波数帯レベルで SSV を計算し，さらに，貢献の小さい周波数帯を取り除き，最後に振幅・位相レベルで SSV を計算するという階層的な手法を提案する．このように，一度に貢献を計算する特徴の数 n を減らすことで，計算量を大きく削減できることが期待される．

4. 実験

4.1 データセット

二つの公開データセットを用いて提案手法の有効性を示す．

一つ目のデータセットは，PhysioNet で公開されている睡眠ポリグラフデータセットである [Goldberger 00]．このデータセットは，健常な成人男性 10 名，成人女性 10 名の計 20 名の 1〜2 晩分の睡眠データを 100Hz のサンプリング周波数で計測したものであり，30 秒を 1 エポックとしている．このデータセットには，Fpz-Cz と Pz-Oz に配置された電極から得られた脳波データや眼電データ (EOG) など，および，それらのデータに対する睡眠段階を Wake (W) , REM (R) , N1, N2, N3 に分類したラベルデータが付与されている．これらの睡眠段階は，脳波や EOG などに基づいて決められている．例えば，ラベルが N3 である基準は，2Hz 以下の低周波活動の振幅が 75μV 以上である状態が 1 エポック中で 20%以上続くことである．そのため，N3 についての識別においては，低周波活動の振幅に貢献が大きくなることが予想される．

二つ目のデータセットは UCI Machine Learning Repository で公開されている EEG Database である [Asuncion 07]．このデータセットは 256Hz-64 チャンネル脳波計で，被験者に視覚刺激を提示した際の 1 秒分の脳波を計測したものである．被験者は 122 人であり，その内 77 人はアルコール中毒であり，45 人は健常である．

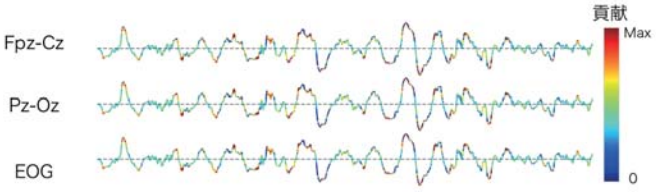

図 2: IG による識別根拠の可視化結果

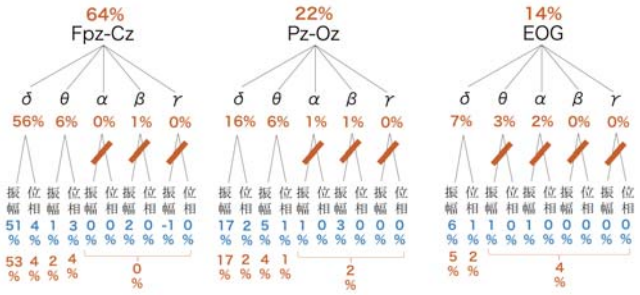

図 3: 提案手法による識別根拠の説明．青は SSV による貢献の推定，オレンジは提案法による貢献の推定の結果を表す

4.2 実験設定

PhysioNet のデータセットについては，3 層の畳み込み層と 1 層の全結合層を持つ識別モデルで睡眠段階を 5 クラス分類した．10 分割交差検証でのテストデータの識別率は約 81%であった．この識別器を用いて，ランダムに選ばれた N3 クラスの一つのデータについて IG [Sundararajan 17] と提案手法により識別根拠を求めた．また，提案手法による計算量の削減量と近似誤差を評価するために，ランダムに選ばれた 30 データに対して，SSV を各特徴に対して 1000 回サンプリングしたものを貢献の真値として，その真値と提案手法による推定値との差の絶対値を各データに対して 10 回ずつ計算した．今回，モンテカルロサンプリングの回数を {5, 10, 20, 50, 100, 200, 500} のいずれかに設定し，これらの回数に対する手法の性能を評価した．なお，貢献の値が最大値の 1/5 以下の場合に，それ以下のレベルの貢献は計算しないようにした．

UCI のデータセットについても同じく，3 層の畳み込み層と 1 層の全結合層を持つ識別モデルを用いて，アルコール中毒の有無について 2 クラス分類した．10 分割交差検証でのテストデータの識別率は約 75%であった．なお，電極位置が標準 10-20 法にはない 3 チャンネル (X, Y, nd) については位置が不明のため，使用していない．IPNPCM 法 [Schirrmeister 17] と提案手法を，ランダムに選ばれた一つのデータに対して適用した．

なお，どちらの実験においても，周波数帯の区分を，δ 波 <2Hz <θ 波 <7Hz <8Hz <α 波 <13Hz <β 波 <30Hz <γ 波とした．なお，どちらの識別器も畳み込み層はそれぞれのセンサで記録された波を時間方向に畳み込み，空間方向へは畳み込まないため，識別器の構成上，空間的な情報である位相特徴の貢献は低くなると考えられる．

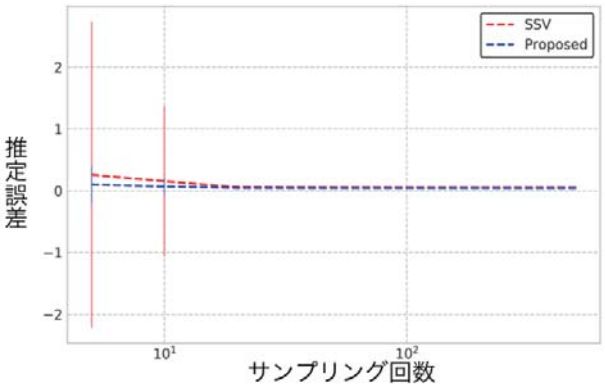

図 4: 提案手法と SSV による推定誤差の平均．エラーバーは標準偏差を示す．

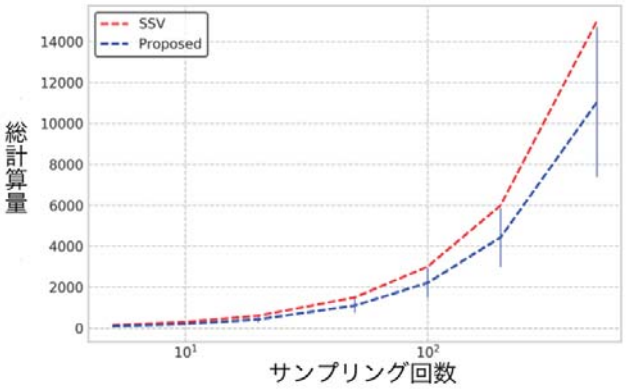

図 5: 提案手法と SSV による計算量の平均．エラーバーは標準偏差を示す．

4.3 実験結結果

睡眠ポリグラフデータセットでの識別について，図 2 に従来手法である IG により識別根拠を可視化した結果を，図 3 に提案手法による結果を示す．IG による可視化では信号の何が識別に貢献しているのかが，直感的にわからず，さらなる解析が必要であるといえる．提案手法では δ 帯の振幅特徴に着目していたことが定量的に示されている．また，この結果は 4.1 節で述べたクラス N3 の定義（低周波の振幅）と整合しており，この識別根拠が正しいことを示す．次に，この課題における SSV と提案手法の比較結果を図 4 と図 5 に示す．これらの図より，提案手法は少ないサンプリング回数で，より正確に貢献の値を推定できていることがわかる．

EEG Database での識別について，図 6 に IPNPCM 法 [Schirrmeister 17] による可視化結果を，図 7 に提案手法による可視化結果を示す．なお，提案手法において，負の貢献を持つ特徴は，そのクラスの識別関数の値を下げることで識別に貢献する．IPNPCM 法は，ほとんどの特徴に 1 か -1 かの極端な値を与えていることに対して，提案手法は特定の特徴に大きな貢献を与えている．これにより，一部の脳領域における θ 帯域の振幅が識別に特に大きく貢献していることがわかる．

5. 結論

本稿では，脳波を周波数帯ごとに分解し，それぞれの波を FFT により振幅成分と位相成分に分解したものを特徴とし，SSV によりそれぞれの特徴の識別への貢献を定量的に求める方法を提案した．それによって，脳波識別における識別根拠を従来手法よりも直感的に，かつ，定量的に示すことができた．また，脳波特徴の階層性を利用した枝刈りにより SSV の計算量を削減する方法を提案し，実際の脳波データを用いた実験により，SSV による推定と比べて，より少ない計算量でほとんど誤差なく推定可能であることを示した．

謝辞

本研究の一部は国立研究開発法人科学技術振興機構 (JST) の研究成果展開事業「センター・オブ・イノベーション (COI) プログラム」の支援によって行われた．

参考文献

[Ancona 17] Ancona, M., Ceolini, E., Öztireli, C., and Gross, M.: A unified view of gradient-based attribution methods for Deep Neural Networks, *arXiv preprint arXiv:1711.06104* (2017)

[Asuncion 07] Asuncion, A. and Newman, D.: UCI machine learning repository (2007)

[Goldberger 00] Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K., and Stanley, H. E.: Physiobank, physiotoolkit, and physionet, *Circulation*, Vol. 101, No. 23, pp. e215–e220 (2000)

[Lundberg 17] Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, in *Advances in Neural Information Processing Systems*, pp. 4768–4777 (2017)

[Schirrmeister 17] Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., and Ball, T.: Deep learning with convolutional neural networks for EEG decoding and visualization, *Human Brain Mapping*, Vol. 38, No. 11, pp. 5391–5420 (2017)

[Shapley 53] Shapley, L. S.: A value for n-person games, *Contributions to the Theory of Games*, Vol. 2, No. 28, pp. 307–317 (1953)

[Štrumbelj 14] Štrumbelj, E. and Kononenko, I.: Explaining prediction models and individual predictions with feature contributions, *Knowledge and Information Systems*, Vol. 41, No. 3, pp. 647–665 (2014)

[Sturm 16] Sturm, I., Lapuschkin, S., Samek, W., and Müller, K.-R.: Interpretable deep neural networks for single-trial EEG classification, *Journal of Neuroscience Methods*, Vol. 274, pp. 141–145 (2016)

[Sundararajan 17] Sundararajan, M., Taly, A., and Yan, Q.: Axiomatic attribution for deep networks, *arXiv preprint arXiv:1703.01365* (2017)

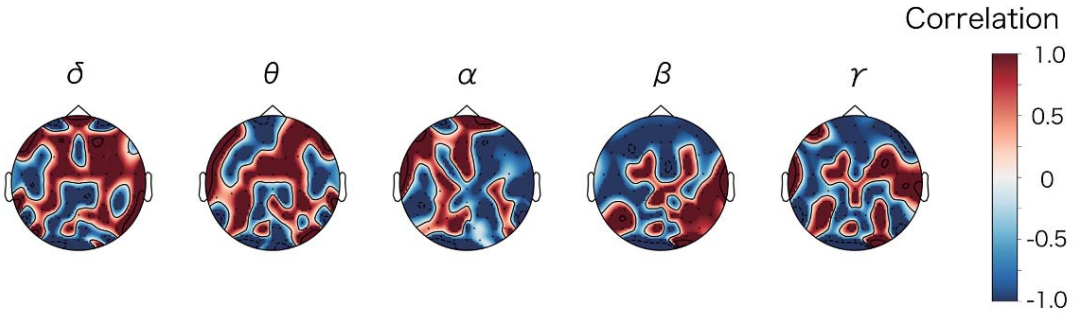

図 6: IPNPCM 法 [Schirrmeister 17] による識別根拠の可視化結果

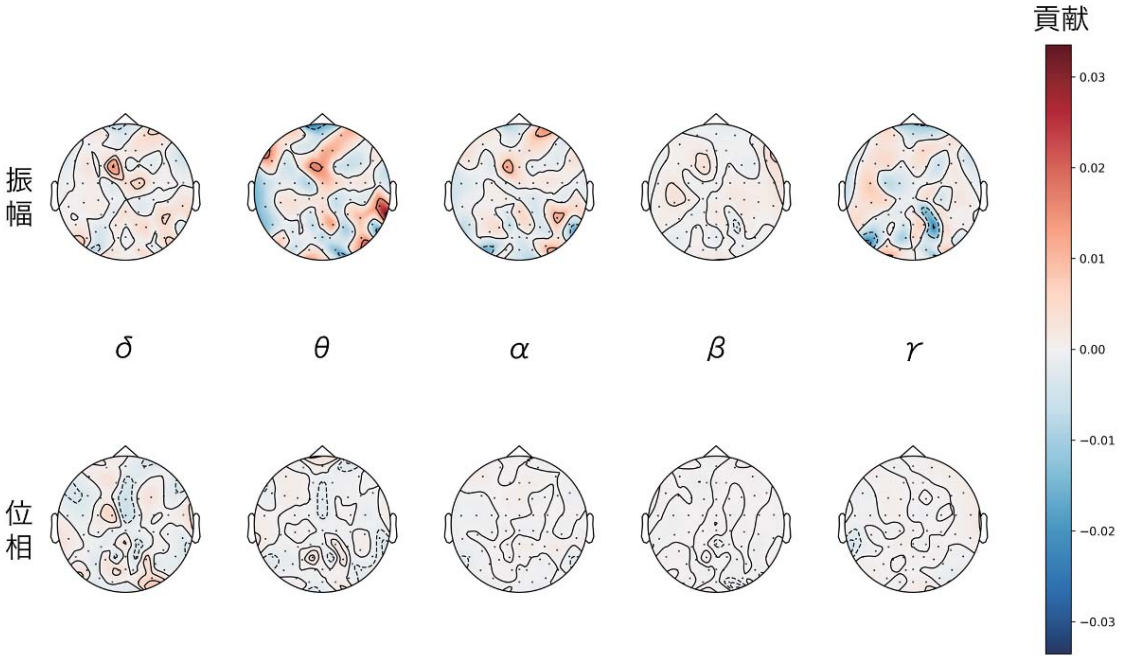

図 7: 提案手法による識別根拠の可視化結果

[Zhang 18] Zhang, Q. and Zhu, S.-C.: Visual Interpretability for Deep Learning: a Survey, *arXiv preprint arXiv:1802.00614* (2018)