단일 채널 뇌파 데이터에 대한 LIME 분석의 적용을 통한 XAI모델 설계

백수환†, 유현수†, 이지운, 박철수*
광운대학교 컴퓨터정보공학부
*e-mail : lunit@kw.ac.kr, byeng3@kw.ac.kr, webmaster@kw.ac.kr, cheolsoopark@kw.ac.kr **

Design of Explainable AI Model with LIME for Single Channel Electroencephalogram

Su-Whan Baek†, Hyun-Soo Yu†, Ji-Woon Lee, Cheol-Soo Park*
School of Computer Engineering
Kwangwoon University

Abstract

To generate a machine learning model to solve real-world problems, analysis of how the model is actually learned and predicted from data is now essential for solving practical problems. In this paper, we propose a model based on single-channel EEG(Electroencephalogram) inputs that satisfy the XAI features. For this implementation, two key methods are proposed which are the component decomposition of EEG data and the application of LIME (Locally Interpretable Model Agnostic Expiations) to each decomposed EEG band to proceed. With this method, we can design an EEG signal based XAI model that can be inferred reasonable reason for their prediction.

I. 서론

의사결정지원시스템(Decision Supporting System) 혹은 그 이상의 판단을 하기 위해 제안된 인공지능 알고리즘들에게 있어 생성된 모델이 어떤 방식으로 학습하고 추론하였으며, 데이터의 어떤 부분에서 주요하게 학습을 하였는지를 분석하는 것은 이제 실제적인 문제를 해결하는데 있어서 필수적이라고 할 수 있다[1]. 특히 최근 인공지능 기술이 비약적으로 발달함에 따라 운전, 의술, 기타 위험 산업 등 인공지능 모델의 오 판단 자체가 큰 문제를 일으킬 수 있는 분야에 대해서 이러한 경향을 더욱 더 뚜렷하게 나타나지만[2] 이러한 시장의 요구에도 불구하고 인공지능 모델 자체의 설명력 (Explainability)은 최근 들어 높아진 모델들의 예측 성능에 비해 낮은 수준이며, 실례로 이러한 오판단으로 인한 인명사고와 이러한 사고 이후의 인공지능 모델에 대한 분석 어려움은 여러차례 보고된 바 있다[2,3]. 이러한 문제를 해결하기 위해 최근에는 Attention 알고리즘 등을 통해 설명가능 성을 충족하는 일부 알고리즘이 제안되기도 하였으나[4], 이러한 분석 방법은 모델에 지나치게 종속적이기 때문에 Attention 적용이 불가능한 경우 활용 될 수 없다는 단점 등이 존재한다. 모델에 덜 종속적이며, 효과적으로 모델의 추론 과정을 분석할 수 있는 LIME(Locally Interpretable Model Agnostic Expiations)[5] 알고리즘은 단순하고 효과적인 분석 방법을 가졌으며, 이러한 특징을 활용해 다양한 형태의 데이터와 모델에 적용 될 수 있는 많은 종류의 알고리즘[6]이 제안된 바 있다. 하지만 이러한 분석 방법들의 제안에도 불구하고, 시계열 기반 신호를 입력으로 하는 모델들은 다른 이미지를 기반으로 한 데이터를 입력으로 하는 모델과 비교하여 아직까지 매우 부족한 수준의 접근만이

제안되고 있는데, 이러한 이유는 크게 두가지로 볼 수 있다. 첫째로, LIME 알고리즘은 기본적으로 이미지에서는 '슈퍼 픽셀' 이라고 하는 샘플 퍼뮤테이션(sample permutation) 과정을 통해 이미지를 복수의 조각으로 나누어 해당 조각을 조합하는 방식을 통해 동작하는데, 시계열 신호에 있어 신호의 특정 영역을 유의미한 특성을 갖는 단위로 세그먼트(segment)하는 과정 자체가 이미지에 비해 더욱 어렵다는 점이다. 신호의 특성상 연속적인 값들의 변동이 유의미한 정보를 담고 있는 경우가 많으며, 어떤 영역이 어떤 특성을 갖는지 판단하기 어렵기 때문에 보통 사람이 직접 세그먼트 하기도 쉽지 않으며, LIME 분석을 통해 최종적으로 분석결과를 확인한다고 해도 해당 신호의 영역이 어떤 정보를 가지고 있는지 해석하기 어렵다. 둘째로, 세그먼트 된 영역의 일부를 조합해 LIME 을 통한 모델의 추론 분석을 위한 신호를 만들어 내는 방식 자체가 이미지에 비해 훨씬 접근하기 어렵기 때문이다. 이러한 요인들로 인해 시계열 신호 데이터에 대한 LIME 의 적용은 쉽지 않으며, 분석된 결과를 통해 모델의 결정 과정에 대해 합리적으로 추론 하기에 어려움이 따른다. 본 논문에서는 이러한 문제를 해결하고자 단일 채널 뇌파 데이터를 별도의 필터 설계가 필요 없는 EMD (Empirical Mode Decomposition)[7]을 통해 분해하여 추론된 결과에 대해 보다 쉽게 담고 있는 정보를 확인 할 수 있도록 하고, HMM 기반 신호 세그먼트 알고리즘[8]을 통해 보다 유용한 성분을 유지한 상태로 LIME 분석을 진행 할 수 있도록 하였다. 또한 신호를 조합하는 방식에 있어서 효과적인 Background 를 형성 할 수 있도록 생성될 신호를 바탕으로 적절한 화이트 노이즈를 형성하는 방법을 통해 LIME 분석을 위한 신호가 모델의 추론 과정 분석에 가장 적은 영향을 미치는 방법을 고안하였다.

II. 본론

2.1 EMD (Empirical Mode Decomposition)

뇌파 신호 (EEG)의 경우 하나의 채널에서 여러 신호 성분이 섞인 대표적인 Multimodal 한 특성을 가진 신호라고 할 수 있다. 따라서 신호의 분해 과정이 필수적이다. 본 논문에서는 사전적인 필터 설계가 필요하지 않고, 신호 본연의 정보를 바탕으로 신호를 분해해 내는 BEMD[7]를 통해 신호를 분해한다.

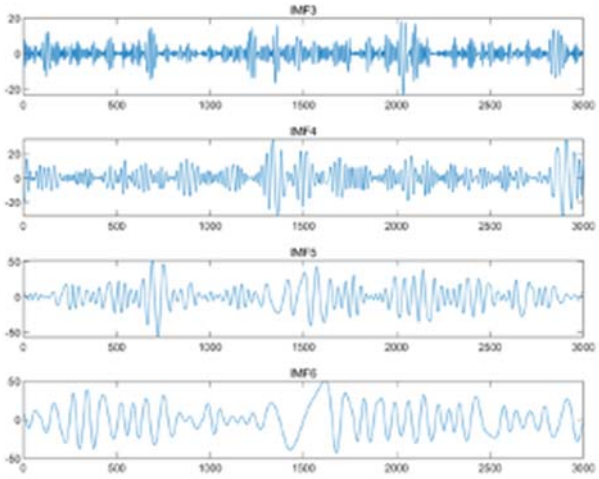

그림1. EMD를 통해 IMF로 분해된 뇌파 성분

EMD를 통해 분해된 신호는 각각의 다른 주파수 성분 밴드를 가진 Intrinsic Mode Functions (IMF)로 분해 되며, 분해된 신호는 원본 신호에 비해 정제된 특성을 갖기 때문에 주파수 밴드에 따라 나타나는 다른 신호의 특성 혹은 정보를 원본 신호에 비해 시각적으로 판단하기 쉽게 되며, LIME 분석 결과에 대해 보다 결정을 쉽게 추론 할 수 있도록 한다.

2.2 HMM Signal Segment

LIME 알고리즘을 통해 모델을 분석하기 위해서는 입력되는 데이터에 대한 적절한 세그먼트 과정이 필수적이다. 이러한 segment에서 주의할 점은 정보의 유의성이 손실되지 않는 최소 단위의 local neighborhood가 가장 손상되지 않게 남아 있을 수 있는 구간을 바탕으로 세그먼트 되어야 한다는 점이다. 이러한 세그먼트를 위해 HMM (Hidden Markov Model)[8] 알고리즘을 활용 하였다. 해당 알고리즘을 통해 분해된 신호는 이후 LIME을 통한 분석 과정에서 이미지 기반 LIME에서와 동일하게 '슈퍼 픽셀 (Super Pixel)'[5] 역할을 하게 된다. 본 논문에서는 보다 쉽게 이러한 구간을 정의하기 위해 이러한 세그먼트 된 신호의 영역을 '슈퍼 블록 (Super Block)'이라고 정의하였다.

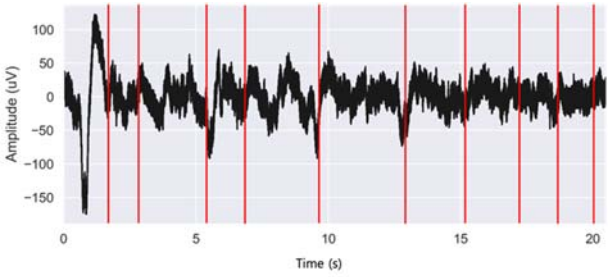

그림2. HMM을 통해 세그먼트 된 신호

2.3 Super Block Combination

LIME 분석을 위해서는 세그먼트 된 신호 구간 즉 '슈퍼 블록' 들을 조합하여 원본 모델이 대상을 잘 분류 할 수 있는 적합한 신호를 생성하는 절차가 필요하다. 이미지의 경우 일반적으로 나뉘어진 슈퍼 픽셀의 일부 만을 같은 좌표에 할당 시키고 이외의 영역에 검정, 혹은 평균적인 픽셀의 RGB값을 계산하여 할당하는 Background의 적용을 통해 구현하지만, 신호의 경우 적합한 Background의 할당이 같은 방식으로 불가능하다. 신호의 경우 단순히 이외의 영역을 0으로 치환하는 방식도 신호 값의 변동에 큰 영향을 주기 때문에 모델의 정상 입력 영역에 대한 예측 자체에 영향을 미칠 수 있으며 이를 회피하기 위해 화이트 노이즈를 Background로 활용하는 방법, Blur 과정을 통해 해당 부분의 신호의 정보를 약화 시키는 방법 등이 제안되었으나 [9,10], 여전히 신호의 크기와 주파수에 영향을 받아 적절한 노이즈를 반영하기 어렵다는 문제가 따른다. 이러한 문제를 해결하기 위해 본 논문에서는 각 슈퍼 블록 마다 평균 주파수와 분산을 계산하여, 각 세그먼트에 적합한 화이트 노이즈를 생성하여 슈퍼 블록 조합을 생성 하도록 하였으며, 해당 과정을 통해 최종적으로 LIME을 위한 샘플 퍼뮤테이션 (sample permutation) 과정을 완료한다.

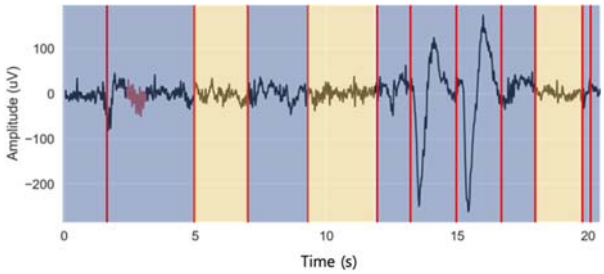

그림3. 퍼뮤테이션이 완료된 뇌파신호, 노란 부분은 화이트 노이즈가 추가된 슈퍼 블록, 파란 부분은 원본 신호를 가진 슈퍼 블록

2.4 LIME

위의 과정을 거쳐 샘플 퍼뮤테이션을 완료한 데이터를 바탕으로 LIME[5]의 적용을 통해 모델에 대한 Explanation을 얻어낸다. f 는 머신러닝 모델, m_x는 슈퍼 블록 π_x에 대한 조합, g는 슈퍼 블록 π_x의 조합 정보인 m_x를 입력으로 받는$f(\pi_x)$와 같은 값을 함께 출력하는 XAI모델이라고 할 때. 이러한 Explanation은 아래 식과 같이 정의 될 수 있다.

$$Explanation(x) = \underset{g \in G}{argmin} L(f, g, \pi_x)$$

수식1. LIME에서 Explanation의 정의[5]

식과 같이 LIME알고리즘은 손실함수가 가장 낮은 값을 가지는 슈퍼 블록 조합을 찾는 과정으로 정의 될 수 있다. 모든 슈퍼 블록의 조합에 대한 모델의 최종 분류와 발생한 슈퍼 블록 조합 즉, 샘플 퍼뮤테이션 결과에 대한 XAI 모델의 출력 값이 가장 작도록 함으로서 모델에 영향을 가장 크게 미치는 슈퍼 픽셀을 찾아 각 슈퍼 블록의 중요도를 정의하게 된다. 이러한 중요도를 분석 함으로서 모델이 학습 과정에서 가장 주요하다고 판단한 영역을 분석할 수 있으며, 모델의 결정과정을 합리적으로 추론 할 수 있다. 이러한 LIME 알고리즘을 EMD를 통해 분해된 각각의 신호 주파수 밴드 성분에 적용 함으로서 우리는 설명가능한 인공지능 (XAI) 모델을 설계 할 수 있다.

Ⅲ. 구현

본 논문에서 제안된 방법을 활용한 단일 채널 EEG를 입력으로 받는 XAI모델을 설계 하였다. 본 모델은 단일 채널 EEG를 입력으로 받아 30초 간의 EEG 윈도우 마다 해당 신호 내의 수면 방추(Sleep Spindle) [11]의 유무를 찾는 딥러닝 기반의 단일 분류기 모델이다.

Order.	Layer	# of Filter	Kernel Size Or Pool Size	Output shape
1	Input	-	-	10 x 6000 x 1
2	1D convolution	64	97	10 x 1968 x 64
3	MaxPooling	-	2	10 x 948 x 64
4	Dropout (Rate = 0.5)	-	-	10 x 984 x 64
5	1D convolution	128	17	10 x 968 x 128
6	1D convolution	128	19	10 x 950 x 128
7	MaxPooling	-	2	10 x 475 x 128
8	1D convolution	128	11	10 x 465 x 128
9	1D convolution	128	13	10 x 453 x 128
10	MaxPooling	-	2	10 x 226 x 128
11	1D convolution	128	5	10 x 222 x 128
12	1D convolution	128	7	10 x 216 x 128
13	MaxPooling	-	2	10 x 108 x 128
14	1D convolution	32	2	10 x 107 x 32
15	1D convolution	32	3	10 x 105 x 32
16	GlobalAverage Pooling	-	-	10 x 32

표1. 딥러닝 모델의 구조

제안된 모델이 올바르게 학습하고 있다면 전체적인 주파수 밴드에서 수면 방추가 많이 발생한다고 알려진 적은 주파수 밴드에서 보다 많은 슈퍼 블록이 높은 중요도를 가질 것이며, 라벨링 된 수면 방추가 있는 슈퍼 블록이 그렇지 않은 블록보다 더 높은 중요도를 보일 것이다.

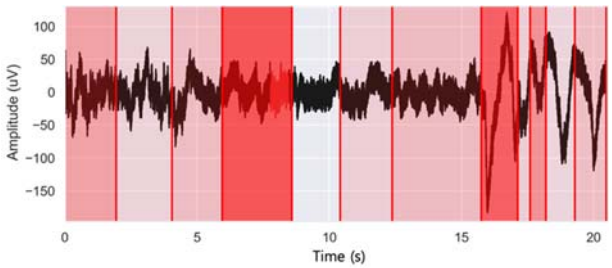

그림4. 수면 방추 추정 XAI 모델의 데이터 중요도 분석

30초간의 EEG데이터에 대한 LIME 분석 결과는 그림4 와 같으며 보다 붉은 색이 높은 중요도를 보인다. 수면 방추가 있는 것으로 모델 외부에서 사전에 라벨링 된 부분의 중요도가 더욱 높은 것을 확인 할 수 있다.

Ⅳ. 결론 및 향후 연구 방향

본 논문에서는 시계열 뇌파 신호에 대해 신호 분해 과정과 LIME분석법의 적용을 통해 충분한 설명력을 갖춘 모델을 제안하였으며, 해당 모델을 실제로 구현하여, 모델이 데이터의 어떤 특성을 통해 분류를 진행하고 학습하였는지를 확인 할 수 있는 XAI 모델을 제안하였다. 해당 모델을 통해 설계된 의사결정지원 시스템은 사용자가 보다 납득할 수 있을 만한 수준의 결정과정을 보여 줄 수 있을 것으로 기대해 볼 수 있다. 하지만 본 논문에서의 Neighborhood 기반의 세그먼트가 올바르게 진행되었는지에 대해서는 아직 모델의 설명력이 충분히 주어지지 못했다는 점, Background 이상적인 설계여부에 대한 평가가 진행되지 못했다는 점, LIME의 고질적인 문제인 non-deterministic한 분석 문제를 해결하지 못했다는 점 등의 문제가 남아있다. 해당 문제는 보다 발전된 신호기반 세그먼트 알고리즘의 적용과 다양한 신호 기반 Background 생성 방법에 대한 모델 평가 등을 통해 해결 할 수 있을 것이다. 또한 본 논문에서 제안한 알고리즘을 발전시킨다면 뇌파 뿐만이 아닌 다양한 Multivariate/Multimodal 신호에 적용 할 수 있는 LIME 알고리즘 또한 제안 할 수 있을 것으로 기대된다.

Acknowledgement

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2019-0-00426, IoT 기반 이식-침습형 고위험 의료장치를 위한 능동형 킬 스위치 및 바이오 마커 활용 방어 시스템 개발)

참고문헌

[1] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. Science Robotics, 4(37), eaay7120.

[2] Banks, V. A., Plant, K. L., & Stanton, N. A. (2018). Driver error or designer error: Using the Perceptual Cycle Model to explore the circumstances surrounding the fatal Tesla crash on 7th May 2016. Safety science, 108, 278-285.

[3] Loyola-Gonzalez, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. IEEE Access, 7, 154096-154113.

[4] Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019, October). Explainable AI: A brief survey on history, research areas, approaches and challenges. In CCF international conference on natural language processing and Chinese computing (pp. 563-574). Springer, Cham.

[5] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).

[6] Mathews, S. M. (2019, July). Explainable artificial intelligence applications in NLP, biomedical, and malware classification: a literature review. In Intelligent computing-proceedings of the computing conference (pp. 1269-1292). Springer, Cham.

[7] Baek, J., Lee, C., Yu, H., Baek, S., Lee, S., Lee, S., & Park, C. (2022). Automatic Sleep Scoring Using Intrinsic Mode Based on Interpretable

Deep Neural Networks. IEEE Access, 10, 36895-36906.

[8] Tóth, László, and András Kocsor. "A segment-based interpretation of HMM/ANN hybrids." Computer Speech & Language 21.3 (2007): 562-578.

[9] Zhang, W., Ge, P., Jin, W., & Guo, J. (2018, July). Radar signal recognition based on TPOT and LIME. In 2018 37th Chinese Control Conference (CCC) (pp. 4158-4163). IEEE.

[10] Barus, D. T., Masri, F., & Rizal, A. (2020, October). NGBoost Interpretation Using LIME for Alcoholic EEG Signal Based on GLDM Feature Extraction. In Proceedings of the Computational Methods in Systems and Software (pp. 894-904). Springer, Cham.

[11] De Gennaro, L., & Ferrara, M. (2003). Sleep spindles: an overview. Sleep medicine reviews, 7(5), 423-440.