웨어러블 기기에서 데이터수 기반 마하라노비스 군집화 연합학습을 통한 스트레스 및 감정탐지

윤태환[1], 최봉준[2]
[1]숭실대학교 컴퓨터학과 석사과정
[2]숭실대학교 컴퓨터학과 교수
{dbs1045, davidchoi}@soongsil.ac.kr

Stress Affect Detection At Wearable Devices Via Clustered Federated Learning Based On Number of Samples Mahalanobis Distance

Tae-Hwan Yoon[1], Bong-Jun Choi[1]
[1]School of Computer Science, and Engineering, Soongsil University

요 약

웨어러블 디바이스에서는 사용자의 다양한 메타데이터를 수집할 수 있다. 그러나 이런 개인정보를 함유하고 있는 데이터를 수집하는 것은 사용자에게 개인정보침해 위협을 야기한다. 때문에 본 논문에서는 개인정보보호를 통한 웨어러블 디바이스 데이터활용방안으로 **연합학습**을 채택하였다. 다만 기존 연합학습에서도 해결해야할 문제점들이 있다. 우리는 그중에서도 데이터이질성(Data Heterogeneity) 문제해결을 위해 **군집화(Clustering)** 방법을 활용하였다. 또한 기존의 코사인유사도 기반 군집화에서 파라미터중요도가 반영되지 않는다는 문제점을 해결하고자 **데이터수 기반 마하라노비스거리(Number of Samples Mahalanobis Distance) 군집화** 방법을 제시하였다. 이를 통해 WESAD(Werable Stress Affect Detection)데이터에서 피실험자의 데이터 이질성이 존재하는 상황에서 기존 연합학습보다 학습 안정성 측면에서 좋음을 보여주었다.

1. 서론

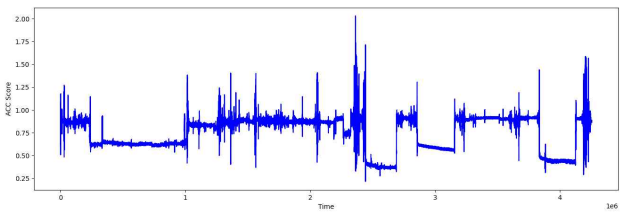
최근 웨어러블 기기에서는 다양한 센서를 이용하여 사용자의 정보들을 수집할 수 있다. 이런 메타정보들은 개인정보를 함유하고 있다. 때문에 웨어러블 기기데이터를 개인정보 보호와 함께 활용할 수단을 찾아야 한다. 본 논문에서는 이런 솔루션으로 **연합학습(Federated Learning)**을 채택하였다. 연합학습은 사용자의 실제데이터의 공유 없이 오직 파라미터만을 특정 서버와 공유하여 실제 모델을 학습시키는 방법이다. 연합학습을 통해 우리는 개인정보를 보호함과 동시에 모델학습 및 활용이 가능하다[1].

그러나 기존 연합학습에서도 해결하여야 할 문제들이 있다. 그중에서도 우리는 데이터 이질성(Data Heterogeneity) 문제를 해결하기 위해 **마하라노비스거리(Mahalanobis Distance) 및 샘플수를 고려한 군집화(Clustering)** 방식을 사용하였다. 이를 통해 데이터의 이질성을 보완할 수 있게 되었다[2]. 우리는 WESAD(Werable Stress Affect Detection) 데이터셋을 활용하여 기존의 연합학습과 비교해 군집화된 연합학습의 성능을 실험하였다.

2. 본론

2.1 WESAD 데이터셋

WESAD 데이터셋은 특정 실험자들을 대상으로 실험을 통해 탐지한 스트레스 수치 및 메타데이터를 함유 하고 있다. 메타데이터 수집은 손목과 가슴에서 측정한 웨어러블 기기를 통해 수집하였다. 가슴에서 측정한 메타데이터의 종류로는 (ACC: 3-axis Accelerometer, ECG: Electrocardiogram, EMG: Electromyogram, EDA: Electrodermal Activity, TEMP: Skin Temperature, RESP: Respiration)가 있다[3]. 특정 실험자의 데이터셋을 기준으로 가슴에서 측정한 데이터의 분석을 시도하였다.



<그림 1> : 특정 피실험자의 ACC 시계열 데이터

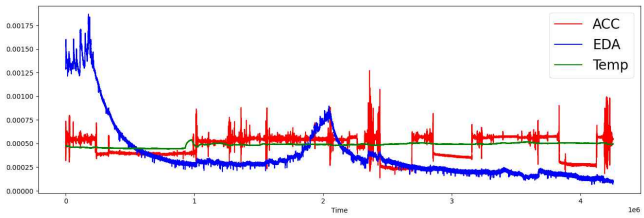
나머지 메타데이터도 <그림1>과 유사한 시계열적인 데이터 특징을 가지고 있다. 여기서 딥러닝을 위한

독립변수를 설정하기위해 종속변수인 스트레스 수치를 가지고 각 변수들의 상관지수를 분석하였다.

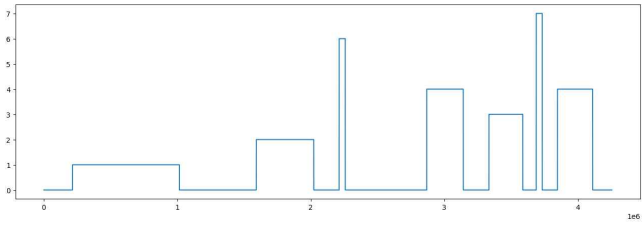
독립변수	상관지수(correlation)
ACC	-0.24731
ECG	0.00006
EMG	0.00557
EDA	-0.31643
Temp	0.37716
Resp	0.00233

<표 1> : 스트레스지수와 각 메타데이터와의 선형적인 상관지수

<표 1>을 통해 스트레스 탐지에 주된 유의미한 독립변수 3개(ACC, EDA, Temp)를 선정하였다. 3개의 데이터의 시계열 데이터의 특징은 다음 그래프와 같다.



<그림 2> : 시간대별 피실험자의 ACC, EDA, Temp 수치(독립변수)



<그림 3> : 시간대별 피실험자의 스트레스 수치(종속변수)

<그림 2,3>에서 ACC, EDA의 증가는 스트레스의 감소에 영향을 Temp 수치의 증가는 스트레스의 증가에 영향을 끼치는 양상임을 볼 수 있다.

2.2 스트레스 및 감정 탐지 모델

본 논문에서는 시계열적인 데이터의 특징을 살리기 위해 시계열 데이터 분석 모델에서 자주 사용되어지는 **LSTM(Long Short-Term Memory) 모델**을 사용하였다. LSTM은 순환 신경망(RNN)의 한 종류다. 때문에 순차적인 데이터(예: 문장, 시계열 데이터 등)를 처리하기 위한 신경망 구조를 가지고 있다. 이전의 RNN 모델보다 성능 안전성을 확보하였다고 볼 수 있다. 또한 LSTM은 정보를 오랫동안 기억하고 필요에 따라 삭제하거나 업데이트할 수 있는 편의 메커니즘을 기용하였다[4][5].

2.3 군집화 연합학습 알고리즘

군집화 연합학습은 기존 연합학습에서 데이터의이질성이 존재하는 상황에서 어떻게 성능을 향상시킬지에 대한 솔루션으로 제시되었다. 데이터의 이질성이 발생할 수 있는 Non-IID 상황은 다음 <표 2>와 같이 크게 5가지로 나누어 볼 수 있다.

Non-IID case	Description and examples
Feature distribution skew	Marginal distributions of data features differ ex) Even if two individuals wear the same smartwatch model and exercise for the same time duration, the features of measured values are unique due to the personal characteristics difference, such as their gait
Label distribution skew	Marginal distributions of data labels differ ex) Frostbite is a disease that frequently occurs in cold areas because it is caused by exposure to severe cold resulting in tissue damage to body parts. Therefore, it is rare in places with relatively warm temperatures
Same label but different features	Conditional distributions of data features differ ex) Medical devices are used to measure healthcare data such as neuro images and biomarkers of patients. However, hospitals do not use the identical medical device brands
Same feature but different labels	Conditional distributions of data labels differ ex) Lung imaged by the recent pandemic COVID-19 virus are difficult to distinguish from the pneumonia because they have similar features in many lesions
Quantity skew	Amount of each patients/hospital data differs ex) Suppose five times more patients have visited hospital A than hospital B. The quantity of data each hospital has will also significantly differ

<표 2> : Non-IID 5가지 상황 및 묘사

[출처]

Yoo, Joo Hun, et al. "Open problems in medical federated learning." International Journal of Web Information Systems 18.2/3 (2022): 77-99 [6].

본 논문에서는 데이터 이질성 환경을 설정하기 위해 특징 편중분포(Feature distribution skew)의 Non-IID 상황을 채택하였다. 이를 통해 최대한 실생활에 근접한 연합학습 환경을 구성함으로써 군집화 알고리즘의 성능을 평가할 수 있다.

본 논문에서는 기존의 군집화 연합학습 알고리즘을 <표 3>같이 적용하였다. 일반적으로 목표(target) 파라미터와 특정 클라이언트가 함유하고있는 로컬(local) 파라미터 사이의 코사인 유사도(Cosine Similarity)를 통해 클러스터링을 수행하였

다.

Algorithm: Clustered Federated Learning

input: initial parameters θ_0, number of local iterations epochs N, K−means clustering $Kmeans()$
output: improved parameters on every client θ_i
init: Selected clients C={1,..,n}, set initial models $\theta_i \leftarrow \theta_0$ $\forall i$=1,..,N, set initial update $\Delta\theta_c \leftarrow 0$, $\forall c \in$C.
while *not converged* **do**
  for i =1,..,N **in parallel do**
    Client i does:
      $\theta_i \leftarrow \theta_0$
      $\Delta\theta_i \leftarrow \mathrm{SGD_n}(\theta_i, D_i) - \theta_i$
  end
  Server does:
  C={1,..,n} **(select clients)**
  for c∈C **do**
    $\Delta\theta_t \leftarrow \frac{1}{|c|}\sum_{i\in c}\Delta\theta_i$
    (Clustering Part)
    $\alpha_{t,j} \leftarrow \frac{<\Delta\theta_t, \Delta\theta_j>}{\|\Delta\theta_t\|\|\Delta\theta_j\|}$ $(j \in c)$
  end
  $c_1, \ldots, c_j \leftarrow Kmeans(\alpha_{t,j}, \alpha_{t,i+j})$
  (K=2, $i+j \in c$ $j \neq i+j$)
  $Cl_1 \leftarrow (\forall c_j = 1)$, $Cl_0 \leftarrow (\forall c_j = 0)$
  if $n(cl_1) > n(cl_0)$:
    $\Delta\theta \leftarrow \frac{1}{|s|}\sum_{i\in Cl_1}\Delta\theta_i$
  else:
    $\Delta\theta \leftarrow \frac{1}{|s|}\sum_{i\in Cl_0}\Delta\theta_i$
  end(epoch)
end(round)
return θ

<표 3> : 군집화 연합학습 알고리즘 명세

[출처]

Sattler, Felix, Klaus-Robert Müller, and Wojciech Samek. "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints." IEEE transactions on neural networks and learning systems 32.8 (2020): 3710-3722[7].

2.4 제안하는 NSMD(Number of Samples Mahalanobis Distance) 유사도

식 (1)같이 코사인 유사도를 활용하여 클러스터링을 하였을때 모든 클라이언트의 파라미터 중요도가 같다는 가정에서, 문제가 되지 않는다.

$$d(\Delta\theta_i, \Delta\theta_j) = \frac{<\Delta\theta_i, \Delta\theta_j>}{\|\Delta\theta_i\|\|\Delta\theta_j\|} \quad (1)$$

식 (1)에서 $\Delta\theta_i$와 $\Delta\theta_j$는 클러스터의 모델 파라미터이다. $d(\Delta\theta_i, \Delta\theta_j)$는 두 모델 파라미터의 코사인 유사도 수치이다. 그러나 일반적으로 각 클라이언트의 파라미터 중요도가 클라이언트가 가지고 있는 데이터의 수에 따라 달라질 수 있다. 이런 경우 단순 유사도를 통해서는 정확한 데이터의 유사도를 계산하는 것에 한계점이 있다. **마하라노비스 거리**는 분산의 정규화라는 과정을 통해 변수들의 상관성을 함유하는 전처리를 수행한 후 L^1 혹은 L^2 혹은 L^∞ 거리를 구하는 공식이다. 다음과 같이 식 (2)로 표현할 수 있다.

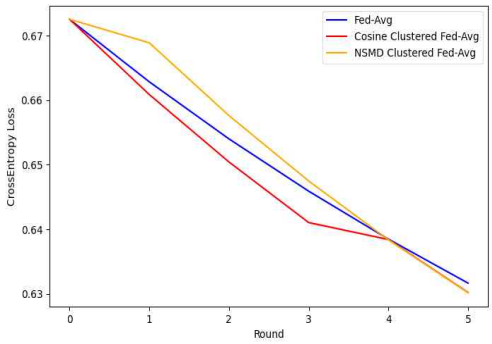
$$d(\Delta\theta_i, \Delta\theta_j) = [(\Delta\theta_i - \Delta\theta_j)S^{-1}(\Delta\theta_i - \Delta\theta_j)^T]^{1/2} \quad (2)$$

식 (3)와 같이 본 논문에서는 식 (2)의 마하라노비스거리에서 각 클라이언트의 데이터 수의 따른 파라미터 중요도를 반영하였다.

$$d(X, Y) = [(X-Y)S^{-1}(X-Y)^T]^{1/2} \quad (3)$$
$$(X = \frac{c_i}{C}(\Delta\theta_i),\ Y = \frac{c_j}{C}(\Delta\theta_j))$$

식 (3)에서 c_i와 c_j는 i,j클라이언트의 데이터 수이다. C는 클러스터가 함유하고 있는 모든 클라이언트의 데이터 수 이다.

2.5 실험 및 결과



<그림 4 >기존 연합학습(Fed-Avg) vs 군집화된 연합학습(Clustered Fed-Avg)

<그림 4>에서 보여주듯이 학습 안정성측면에서 기존 연합학습보다 군집화된 연합학습 알고리즘이 데이터 이질성이 있는 상황에서 성능이 더 좋아질 가능성이 높다. 또한 제시한 NSMD 클러스터링의 경우 데이터의 이질성이 심하고, 연합학습에 참여하는 클라이언트들이 가지고 있는 샘플개수에 차이가 클 때 더욱 효과적이다.

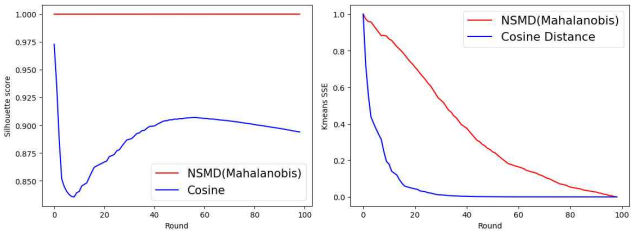
Stress vs Non-stress

CL or FL	Accuracy	F1-score	Loss
CL(Centralized Learning)	76.50%	43.34%	0.5397
Fed-Avg [1]	76.48%	43.33%	0.6095
Clustered Fed-Avg [7]	**76.49%**	**43.34%**	**0.6077**
NSMD Fed-Avg (Proposed)	**76.49%**	**43.34%**	**0.6077**

<표 4> : FL(Fed-Avg) 와 제안하는 CFL(Clustered Fed-Avg)에서 예측 성능 비교

<표 4>에서 기존 연합학습(Fed-Avg)은 중앙집중형 학습(CL)과 비교하였을때, 에러율(Loss)에서의 차이가 있다. 반면에 **군집화 연합학습 알고리즘 (Clustered Fed-Avg)**을 적용했을 때 데이터 이질성이 보완되어 Fed-Avg보다 성능이 조금 향상되었음을 볼 수 있다. 이런 결과들을 통해 데이터 이질성이 심한 상황일 때 군집화를 활용한 연합학습이 효과적임을 알 수 있다.

다음으로는 클러스터링의 성능을 기준으로 제안 알고리즘(NSMD)과 기존 Cosine을 기반으로한 알고리즘을 평가해보았다.



<그림 5 > 제안하는 NSMD vs Cosine Distance
(좌: Silhouette Score 우: 표준화된 Sum Squared Error)

<그림 5> 좌측에 있는 그래프를 통해 제안하는 NSMD 기반 클러스터링이 Cosine 기반 클러스터링보다 Silhouette 점수가 99%를 유지하는 더 좋은 성능을 보여준다.

3. 결론

본 논문에서는 WESAD 데이터셋을 통해 개인정보 보호와 데이터 활용을 위한 **군집화 연합학습** 알고리즘을 통해 기존 연합학습 알고리즘보다 학습 안정성을 향상시켰다. 또한 기존의 클러스터링 방식의 파라미터중요도 누락의 문제점을 **마하라노비스** 유사도를 응용 및 보완하여 클러스터링의 효율을 향상시켰다. 앞으로의 연구에서는 데이터 수 이외에 다른 모델피처들이 있을지에 대한 연구를 통해 성능 향상을 기대해 볼 예정이다.

사사문구

본 성과는 과학기술정보통신부의 재원으로 한국연구재단의 지원을 받아 수행된 연구이며(NRF-2022R1A2C4001270), 또한 과학기술정보통신부 및 정보통신기획평가원의 대학ICT연구센터육성지원사업의 연구결과로 수행되었음(IITP-2022-2020-0-01602).

참고문헌

[1] McMahan, et al. "Communication-efficient learning of deep networks from decentralized data." Artificial intelligence and statistics. PMLR, 2017.
[2] Jiang, et al. "Federated clustered multi-domain learning for health monitoring." Scientific Reports 14.1 (2024): 903.
[3] Schmidt, et al. "Introducing wesad, a multimodal dataset for wearable stress and affect detection." Proceedings of the 20th ACM international conference on multimodal interaction. 2018.
[4] Hochreiter, et al. "Long short-term memory." Neural computation 9.8 (1997): 1735-1780.
[5] omerbsezer.LSTM_RNN_Tutorials_with_Demo.*GitHub,* (2018),https://github.com/omerbsezer/LSTM_RNN_Tutorials_with_Demo?tab=readme-ov-file
[6] Yoo, Joo Hun, et al. "Open problems in medical federated learning." International Journal of Web Information Systems 18.2/3 (2022): 77-99.
[7] Sattler, et al. "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints." IEEE transactions on neural networks and learning systems 32.8 (2020): 3710-3722.