논문 20-45-10-15 The Journal of Korean Institute of Communications and Information Sciences '20-10 Vol.45 No.10
https://doi.org/10.7840/kics.2020.45.10.1756

동공 크기 변화 기반 효과적 학습 상태 분류를 위한 데이터 재구성 알고리즘

이 정 진•, 조 일 현*, 박 형 곤°

Data Reconfiguration Algorithm for Efficient Learning State Classifications Based on Pupil Sizes

Jungjin Lee•, Il-Hyun Jo*, Hyunggon Park°

요 약

최신 스마트 기기의 발전에 힘입어 웨어러블 장치의 사용이 대중화되었고, 이에 따라 개인의 심박수, 에너지 소비 및 단계 수와 같은 생체 데이터를 원활하게 수집할 수 있게 되었다. 웨어러블 기기에 고성능 및 고정밀 센서가 사용되면서 보다 많은 정확한 생체 데이터를 수집할 수 있게 되었다. 생체 데이터는 인종 및 문화적 차이에 덜 민감하기 때문에 신뢰성이 높은 데이터로 사용할 수 있다. 하지만 데이터 수집 과정에서 포함되는 잡음 때문에 생체 데이터를 효율적으로 처리하고 정확하게 분석하기 위하여 사용하는 머신러닝 기반 알고리즘은 성능의 한계를 갖게 된다. 본 논문에서는 다양한 생체 데이터 중 집중 정도를 잘 나타낸다고 알려져 있는 동공 크기 데이터를 이용한다. 시간에 따라 변화하는 동공 크기의 변화를 분석하기 위하여 머신러닝을 적용하고, 이를 통하여 집중도를 추론하고자 한다. 특히 분류 성능 향상을 위하여 데이터 재구성 전처리 알고리즘을 제안하였고, 시뮬레이션 및 실제 동공 데이터를 이용한 실험을 통하여 제안한 데이터 재구성 알고리즘이 머신러닝 분류 성능을 향상하는 것을 확인하였다.

키워드 : 기계학습, 데이터 통합, 분류모델, 서포트 벡터 머신, K최근접이웃, 시계열 데이터
Key Words : Machine Learning, Data Integration, Classification Model, Support Vector Machine, K-nearest neighbor, Time Series Data

ABSTRACT

With the recent development of smart devices, the use of wearable devices became popular, which enabled us to collect individual biometric data such as heart rate, energy consumption and number of steps smoothly. As high performance and high accurate sensors are equipped in wearable devices, the more biometric data can be collected. Since the biometric data are less susceptible to racial and cultural differences, they can be used as a reliable set of data sources. Because of inevitable noises included in the processes of data acquisition, however, the performance of the machine learning algorithms employed for efficiently processing the biometric

※ 이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구이며(No. 2019-0-00024, 네트워크 자동화를 위한 개방형 네트워크 데이터 분석 기반 지도형 애자일 머신러닝 기술 개발), 2020년도 정부(과학기술정보통신부)의 재원으로 한국연구재단(No. NRF-2020R1A2B5B01002528)의 지원을 받아 수행된 연구임.
• First Author : Ewha Womans University Department of Electronic and Electrical Engineering, jungjin.lee@ewhain.net, 학생회원
° Corresponding Author : Ewha Womans University Department of Electronic and Electrical Engineering, hyunggon.park@ewha.ac.kr, 종신 회원
* Ewha Womans University Department of Educational Technology, ijo@ewha.ac.kr, 교수
논문번호 : 202005-107-C-RN, Received May 13, 2020; Revised July 10, 2020; Accepted July 23, 2020

data and accurately analyzing them can be limited. In this paper, we focus on the sizes of pupil data among a variety of biometric data, which are used as an indicator of concentration. We adopt machine learning based classification algorithms for inferring the degree of concentration by analyzing the time-varying changes of pupil sizes. In order to improve the classification performance, we propose a pre-processing algorithm for data reconfiguration. Our simulation and experiment results confirm that the proposed pre-processing for data reconfiguration improves the data classification performance.

Ⅰ. 서 론

최근 스마트 기기의 발달로 웨어러블 디바이스의 보급이 확대되면서 이를 통해 심장 박동 수, 소비 열량, 걸음 수 등 사용자 데이터를 원활하게 수집할 수 있게 되었다[1]. 혈압, 혈당, 체온이 측정되는 웨어러블 재킷을 통해 의사에게 약물을 처방받기도 하고, 심장 박동 수가 측정되는 시계형 디바이스를 통해 누구나 쉽게 건강관리를 할 수 있다. 4차 산업혁명 시대가 다가오면서 고성능 센서에 ICT(Information and Communications Technologies) 기술이 더해져 원격 진료와 같은 전문적인 분야뿐 아니라, 일상생활까지 활용 분야가 확장되어가고 있다[2].

사람의 대부분 행동과 감정 등의 상태는 다양한 생체 신호를 동반하기 때문에 생체데이터 분석은 그 활용도가 높아 주목받고 있다. 예를 들어 사람이 어떠한 문제에 직면하고 해결에 어려움을 느끼면 스트레스를 받게 된다. 그에 따라 심장 박동 증가, 혈압 증가, 발한, 근육 긴장, 위장 운동 감소, 면역 억제 반응 등과 같은 다양한 생리적 변화가 나타난다[3]. 이러한 생체 신호들은 사용자의 인종, 문화적 차이 등에 대한 영향을 덜 받을 뿐 아니라, 인위적이지 않은 자연 그대로의 데이터이기 때문에 그 신뢰성과 활용도가 높다는 장점이 있다[4].

동공 크기는 문제 해결 난이도에 따라 변화 정도가 다르다고 알려져 있다. 난이도가 다른 수학연산 문제를 피실험자에게 해결하도록 하고 그에 따른 동공 크기를 측정한 결과, 연산 난이도와 동공 크기가 관련이 있는 것을 실험적으로 확인하였다[5]. 또한, 단기 기억 작업 중 동공 크기의 변화에 관한 연구에서도 동공 크기의 변화율은 작업의 어려움과 관련 있다고 보고하고 있다[6]. 공감 유무에 따른 동공 크기에 관한 연구에서는 평상시 동공 크기를 기준으로 비교했을 때 공감하지 않는 경우가 공감하는 경우보다 통계적으로 유의미하게 더 확장된 패턴을 보이는 것을 확인했다[7]. 이와 같이 생체 데이터에 관한 연구가 다양하게 이루어지고 있지만, 기기를 통해 측정된 생체데이터는 그 양이 많을 뿐 아니라 데이터 취득 과정에서 기기 한계 및 생체 반응에 의한 잡음이 포함될 수밖에 없기 때문에 데이터를 효과적으로 처리 및 분석하는데 한계점이 있다[8]. 따라서 생체데이터를 활용하여 데이터의 분석 및 추론을 위해서는 수집된 데이터를 적절히 처리할 수 있는 알고리즘이 필요하다.

본 논문에서는 동영상 학습 과정에 참여한 학습자의 동공의 크기를 활용하여 머신러닝 기반으로 학습자의 학습상태를 효율적으로 분류할 수 있는 데이터 재구성 전처리 방법에 대한 알고리즘을 제시하였다. 일반적으로 학습 과정에서 동공의 확장반응은 과제를 수행하는 노력이 향상할 때 일어난다고 알려져 있으며, 그 자극이 정보로서 가치가 높은 자극일 때 동공 확장이 일어난다고 알려져 있다. 즉, 학습 과정에서 과제를 수행하는 노력이 향상할수록 동공의 크기가 커짐을 의미하므로, 동공의 크기 변화를 기준으로 동공의 크기가 커지는 경우, 높은 집중도가 있어야 하는 학습 과정을, 동공의 크기가 작아지는 경우 낮은 집중도가 있어야 하는 학습 과정으로 해석할 수 있다[9].

본 논문에서는 학습자가 동영상을 이용하여 학습할 때, 동영상 학습 과정에 포함된 두 가지 학습 상태인 동영상 강의를 보며 학습하는 상태와 시험을 보는 상태를 고려한다. 일반적으로 수동적인 학습을 할 때보다 시험을 볼 때 높은 집중도를 필요로 하므로 동영상을 보면서 학습을 할 때보다 시험을 볼 때 동공의 크기가 더 커진다고 할 수 있다. 따라서 동공 크기의 변화로 학습자의 상태를 추론할 수 있다. 하지만, 데이터의 수집과정에서 기기의 측정 오차 및 학습자의 상태에 따라 오차가 잡음의 형태로 포함되기 때문에 수집된 데이터를 그대로 알고리즘의 입력 데이터로 사용하는 경우 머신러닝 기반 분류 알고리즘은 제한적인 성능만을 보여줄 수 있다.

본 논문에서는 분류 알고리즘의 성능을 향상하기 위하여 수집된 데이터를 시계열 클러스터링[10] 방법을 기반으로 데이터의 일부분을 연속적으로 취하는 데이터 재구성 기반 전처리 방법을 제안한다. 제안한 전처리 과정을 통하여 재구성된 데이터 집합을 분류 머신

러닝 알고리즘의 입력 데이터로 사용하는 경우, 서로 다른 클래스에 속한 데이터 사이의 거리가 증가하기 때문에 향상된 분류 성능을 기대할 수 있다. 본 논문에서는 제안한 알고리즘으로 데이터를 재구성한 경우 서로 다른 클래스에 속한 데이터 사이의 거리를 증가시키기 때문에, 머신러닝 분류 알고리즘의 성능이 향상됨을 증명하였다. 그리고 이를 시뮬레이션 데이터와 실제 실험자로부터 취득한 동공 데이터로부터 검증하였다.

본 논문의 구성은 다음과 같다. Ⅱ장에서는 관련 연구에 대해 소개를 하고 Ⅲ장에서는 문제 정의 및 제안 알고리즘을 설명한다. Ⅳ장에서는 제안하는 알고리즘을 구현하고 시뮬레이션 결과 및 실험 결과를 제시한다. 마지막으로 Ⅴ장에서는 본 논문에 대한 결론을 맺는다.

Ⅱ. 관련 연구

생체 데이터를 이용하여 사용자의 상태나 감정을 분류하는 연구는 다양하게 진행되고 있다. 비침습 형 4채널 뇌파 측정 장비와 시각 자극 기반의 연속수행검사(Simplified Continuous Performance Test, SCPT)를 분석하여 객관적인 집중도를 파악하는 방법을 제시하는 연구는 명상할 때와 시각 연속수행검사(Visual SCPT) 콘텐츠를 수행할 때의 주의집중도 차이를 비교·분석하기 위해 뇌파검사(Electroencephalography, EEG) 평가 지수와 SCPT 검사 변수들의 평갓값들을 종합하여 집중도 지수를 산출하였다[11]. 비침습 형 뇌파 측정의 경우 잡음이 심하고, 뇌파 측정 부위가 사람마다 차이가 있으므로 정확한 측정이 어렵다는 단점이 있다. 지속적인 자극이 주어질 때 동공 확장에 대한 교감신경 및 부교감신경 경로의 기여도를 환경 및 약물 처리를 통해 알아보는 연구도 진행되었다[12].

최근 들어 생체데이터에 머신러닝 알고리즘을 이용하여 분석하는 연구도 널리 진행되고 있다. 유전자 발현 데이터, 자기공명영상(Magnetic Resonance Imaging, MRI) 데이터, 뇌파 데이터 등 실제 수집된 생체 데이터를 목적에 맞게 분류하기 위하여 적절한 알고리즘의 종류, 파라미터의 값 등을 결정하는 연구가 진행되어 왔다[13-16]. 뇌파 데이터에 주성분 분석(Principal Component Analysis, PCA)과 시간 지연 임베딩(time delay embedding)을 통해 다섯 가지의 다른 클래스로 분류할 수 있음을 보여주었다[17]. 이 연구에서는 분석에 시간적 정보를 추가하기 위해 각 샘플을 시간에 따라 다른 샘플로 증가시키는 시간 지연 임베딩을 사용하였다. 이와 비슷하게 시계열 데이터의 예측과 관련해서 데이터를 재구성하여 예측 모델을 생성하는 연구도 있다. 데이터를 재구성함으로써 학습 데이터 중에서 가장 가까운 패턴을 찾아내고 이를 통해 미래의 값을 찾아내는데 K-최근접 이웃 알고리즘(K-Nearest Neighborhood, KNN)을 사용하였다[18]. 비슷하게 데이터를 재구성하여 나온 위상공간에 대한 결합 확률밀도함수를 추정하여 재구성한 데이터를 통계적으로 모델링한다. 추정된 결합 확률밀도함수의 분포를 가우시안 혼합 모델(Gaussian Mixture Model, GMM)을 통해 이 분포를 반영하는 모델을 다시 모델링 한 후 신호 분류를 위해 이 가우시안 혼합 모델을 사용하여 특징을 추출하고 베이지안 최대우도와 인공지능신경망(Artificial Neural Network, ANN)으로 분류함으로써 성능향상을 보였다[19].

다양한 채널을 통해 데이터 수집이 가능한 뇌파에 관한 연구도 활발히 이루어지고 있다. 뇌파를 수집하기 위해서는 머리에 전극 부착용 전도성 겔을 사용하여 전극을 부착하게 되는데 전극 부착 시간이 긴 것과 측정 부위를 세척해야 하는 등 데이터의 수집 환경이 간단하지 않다는 단점이 있다[20]. 그러나 뇌파와 같은 데이터는 여러 채널을 사용하여 데이터를 수집하기 때문에 피처가 하나뿐인 데이터와는 다르게 다양한 특성을 추출할 수 있다. 또한 추출한 특성들을 대표하는 좋은 피처의 선택이 가능하다[21]. 하지만 피처가 하나뿐인 데이터를 분류하기 위해서는 좋은 피처의 선택이 불가능하기 때문에 수집한 데이터를 최대한 활용하는 것이 필요하다.

동공의 크기 변화는 일반적으로 인간의 집중 정도를 반영하는 것으로 알려져 있다. 이는 뇌간에 위치한 청반의 활성화와 깊은 연관이 있다. 집중력이 일정 수준 이상 소모되어 청반이 활성화되면 시각과 관련된 부분이 활성화되고 동공이 확장되기 때문이며[22], 안구의 운동과 위치변화 그리고 동공의 크기변화는 자극의 정보에 대한 주의를 기울이고 뇌의 인지적인 과정 혹은 감정적인 과정을 거치고 있음을 보였다[23]. 따라서 동공의 확장반응은 과제를 수행하는 노력이 향상할 때 일어나며 그 자극이 정보로서 가치가 높은 자극일 때 동공 확장이 일어난다고 할 수 있다.

Ⅲ. 문제 정의 및 제안 해결 방법

3.1 동공 데이터 설명 및 분류 문제 정의

서로 다른 n개의 실험 조건을 나타내는 n개의 클

래스 Y_i $(i=1,\cdots,n)$를 고려하자. 이상적인 환경에서 i번째 실험 조건인 클래스 Y_i에서 생성되는 동공 크기 샘플 데이터 x_t는 값 m_i를 갖는다고 할 수 있다. 그리고 이렇게 생성된 데이터를 수집하는 과정에서 잡음이 포함될 수 있기 때문에, 수집된 데이터는

$$x_t = m_i + n_t \tag{1}$$

로 나타낼 수 있으며, 이때 n_t는 수집과정에서 포함되는 잡음에 대한 분포를 나타낸다. 본 논문에서는 $n_t \sim N(0,\sigma^2)$의 정규 분포를 따르고 독립적이라 가정한다. 즉, 샘플 데이터 x_t는 확률변수(random variable) X의 실현 값(realization)으로 볼 수 있으며, 확률변수 X는 평균 m_i와 분산 σ^2를 갖는 분포를 따른다고 할 수 있다. 클래스 Y_i에 대하여 수집된 샘플 데이터의 확률변수 X의 확률밀도함수 $f_{X|Y_i}(x)$는

$$f_{X|Y_i}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{(x-m_i)^2}{-2\sigma^2}} \tag{2}$$

로 나타낼 수 있다. 본 논문에서는 수집된 T개의 데이터 샘플 집합 D,

$$D = [x_1,\ x_2,\ \ \ldots,\ x_T] \tag{3}$$

를 이용하여, 해당 데이터를 생성한 클래스 Y_i를 찾아내는 일반적인 분류 문제

$$Y_i^* = \mathrm{argmax}_{\{Y_i,\ i=1,2,\ldots n\}} P(Y_i|D) \tag{4}$$

를 해결하고자 한다. 식 (4)로 표현된 분류 문제는 다양한 머신러닝 기법을 이용하여 해결할 수 있다[24].

3.2 데이터 재구성 전처리

수집된 데이터 집합 D는 수집과정에서 잡음이 포함되었기 때문에 식 (4)를 해결하기 위하여 기존에 알려진 분류를 위한 머신러닝 알고리즘을 바로 적용하는 경우, 제한적인 분류 성능만을 보여줄 수 있다. 따라서 분류 머신러닝 알고리즘의 성능을 향상하기 위하여 본 논문에서는 수집된 데이터를 시계열 클러스터링[10] 방법을 기반으로 데이터의 일부분을 연속적으로 취하는 데이터 재구성 기반 전처리 방법을 제시한다. 제안한 전처리 과정을 통하여 재구성된 데이터 집합을 분류 머신러닝 알고리즘의 입력 데이터로 사용하는 경우, 서로 다른 클래스에 속한 데이터 사이의 거리가 증가하기 때문에 분류 성능이 향상될 수 있다.

수집된 데이터 집합 $D = \{x_1, x_{2,}\ \ldots, x_d, x_{d+1}, \ldots, x_T\}$에 대하여 다음과 같이 길이 d인 서브 데이터의 집합인 $S_D = [D(1), D(2), \ldots, D(T-d+1)]$로 재구성한다.

$$\begin{aligned} D(1) &= [x_1,\ x_2,\ \ \ldots, x_d] \\ D(2) &= [x_2,\ x_3,\ \ \ldots, x_{d+1}] \\ &\ldots \\ D(T-d+1) &= [x_{T-d+1},\ x_{T-d+2},\ \ \ldots, x_T] \end{aligned} \tag{5}$$

이는 그림 1과 같이 나타낼 수 있다.

본 논문에서 제안한 데이터 재구성 방법을 통한 서브 데이터 집합 S_D에 속하는 데이터들 사이의 클래스 간 거리가 원래의 데이터 집합 D보다 커지기 때문에 분류 머신러닝 알고리즘의 성능을 높일 수 있다.

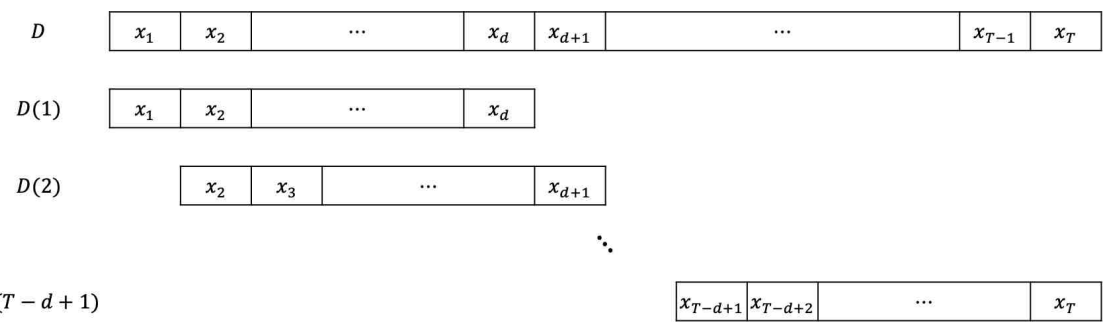


그림 1. 서브데이터 재구성
Fig. 1. Sub-data reconfiguration

클래스 Y_i와 클래스 Y_j에서 발생한 데이터를 수집한 데이터 집합을 각각 $D_i = \{x_{i1}, x_{i2}, ..., x_{iT}\}$과 $D_j = \{x_{j1}, x_{j2}, ..., x_{jT}\}$라 하자. 그리고 제안한 데이터 재구성 방법을 통한 서브 데이터 집합을 $S_{D_i} = [D_i(1), ... D_i(T-d+1)]$와 $S_{D_j} = [D_j(1), ... D_j(T-d+1)]$라 하자. 이때 각 서브 데이터 집합에 속한 두 임의의 원소 사이의 제곱 거리는

$$| D_i(k) - D_j(l) |^2 = \sum_{m=0}^{d-1} (x_{i(k+m)} - x_{j(l+m)})^2 \tag{6}$$

로 나타낼 수 있으며 D_j 및 D_j에 속한 각 원소는 독립이기 때문에 식 (6)은

$$\begin{aligned} | D_i(k) - D_j(l) |^2 &= \sum_{m=0}^{d-1} (x_{i(k+m)} - x_{j(l+m)})^2 \\ &= d(x_{ik'} - y_{jl'})^2 \end{aligned} \tag{7}$$

로 나타낼 수 있다. 이때 $x_{ik'} \in D_i(k)$이고 $x_{jl'} \in D_j(l)$인 임의의 원소를 나타낸다. 식 (7)에서 $(x_{ik'} - x_{jl'})^2$은 원 데이터 집합인 D에 포함된 임의의 두 데이터 샘플의 제곱 거리이므로 제안한 재구성 방법을 통한 서브 데이터 집합의 데이터 사이의 거리는 $\sqrt{d}$ 만큼 증가한다는 것을 알 수 있다. 즉, 원래의 데이터 집합인 D를 그대로 분류 머신러닝 알고리즘의 입력으로 하였을 때보다 제안한 데이터 재구성 방법을 하여 전처리된 데이터 집합인 S_D를 입력을 하였을 때 분류 성능 향상을 기대할 수 있다.

하지만, 이렇게 제안한 데이터 재구성 방법은 동일한 클래스의 데이터 사이의 거리와 분산도 증가시키는 경향이 있다. 이를 정량적으로 판단하기 위하여 재구성된 서브 데이터 집합에 포함된 데이터의 제곱 거리에 대한 평균과 표준편차의 비율인 변이 계수(coefficient of variation, CV)[25]를 사용하고자 한다. 즉, 클래스 Y_i의 길이 d인 서브 데이터의 집합인 $S_D = [D(1), D(2), ..., D(T-d+1)]$에 대하여 변이 계수는

$$\mathrm{CV} = \frac{\sqrt{Var[|D(k) - D(l)|^2]}}{E[|D(k) - D(l)|^2]} \tag{8}$$

와 같이 정의되며, $E(\cdot)$와 $Var(\cdot)$는 평균 및 분산을 나타낸다. 이때,

$$\begin{aligned} &E[|D(k) - D(l)|^2] \\ &= E[d(x_k - x_l)^2] \\ &= d \cdot E[(m_i + n_k - m_i - n_l)^2] \\ &= d \cdot E[(n_k - n_i)^2] \\ &= d \cdot E[n_k^2 + n_l^2 - 2n_k n_l] \\ &= 2d\sigma^2 \end{aligned} \tag{9}$$

이고,

$$\begin{aligned} &Var[|D(k) - D(l)|^2] \\ &= d^2 \cdot Var[n_k^2 + n_l^2 - 2n_k n_l] \\ &= d^2 \{ Var(n_k^2) + VAR(n_l^2) - 2\,Var(n_k)\,Var(n_k) \} \\ &= d^2 \{ 3\sigma^4 + 3\sigma^4 - 2\sigma^2\sigma^2 \} \\ &= d^2 \cdot 4\sigma^4 \end{aligned} \tag{10}$$

이므로, 동일한 클래스의 서브 데이터의 집합에 대하여 변이 계수

$$\mathrm{CV} = \frac{\sqrt{Var[|D(k) - D(l)|^2]}}{E[|D(k) - D(l)|^2]} = \frac{2d\sigma^2}{2d\sigma^2} = 1 \tag{11}$$

로 일정함을 알 수 있다. 즉, 제안한 데이터 재구성 방법을 통한 서브 데이터 집합을 이용한 경우, 클래스가 다른 데이터에 대해서는 데이터 사이의 거리가 커지지만, 같은 클래스에 속한 데이터에 대해서는 데이터 사이 거리에 대한 평균 증가량과 분산 증가량이 일정함을 알 수 있다.

Ⅳ. 구현 및 검증

4.1 실험 설정 및 학습 상태에 따른 동공 크기 데이터 수집

본 실험은 동영상 학습 과정에 포함된 두 가지의 학습 상태, 동영상 강의를 보며 학습하는 상태와 시험을 보는 상태로 나누어진다. 두 가지 상태에서 수집된 동공 크기 데이터를 각각 다음과 같이 D_1과 D_2로 나타낸다.

$$D_1 = \{x_{11}, x_{12}, \cdots, x_{1T}\}. \quad (12)$$

$$D_2 = \{x_{21}, x_{22}, \cdots, x_{2T}\}. \quad (13)$$

T와 T'는 각각 수집된 데이터의 크기를 뜻하며, D_1과 D_2를 레이블 -1, +1로 할당한다.

본 연구에서는 학습자들의 동공 크기를 측정하기 위해 아이트래커 Tobii Pro X2-30을 사용하였다. 측정 빈도가 30Hz인 이 장비는 길이 184mm, 무게 200g 정도의 스틱형이며, 타임스탬프, 동공 크기(mm), 안구 운동의 종류, 시선 위치 등의 데이터를 추출할 수 있다 [26].

아이트래커를 이용하여 수집한 대학교 2학년 35명의 동공 크기 데이터를 사용하였다. 수집된 데이터에는 참가자 코드, 좌우 동공의 크기, 학습 상태로 구성되어 있다. 좌우 동공의 크기의 경우 비슷한 값으로 측정되었기 때문에 좌우 동공의 크기의 평균을 이용하였다. 상황별로 수집된 동공 크기의 평균은 $m_1 = 2.906$mm, $m_2 = 3.005$mm이고 표준편차는 $\sigma_{D_1} = 0.227$mm, $\sigma_{D_2} = 0.159$mm이다. 참가자별로 동공의 크기 변화가 크게는 약 0.3mm, 작게는 0.01mm 정도의 변화가 있었다. 이는 그림 4에 나타내었다.

실험 참가자별로 동영상을 보는 중과 시험을 보는 중의 학습 상태에 따라 동공 크기 데이터의 레이블을 할당하였다. 3장에서 제안한 데이터 재구성 방법을 통

그림 2. 동공크기 측정 시스템 및 측정 과정
Fig. 2. Pupil size measurement system and measure process

	A	B	C	D	E
1	ParticipantName	Time	MediaName	PupilLeft	PupilRight
2	P01	2017.3.7 11:56:20	http://localhost/ewha/lear2_1 (CRC)	3.61722	3.56778
3	P01	2017.3.7 11:56:21	http://localhost/ewha/lear2_1 (CRC)	3.62143	3.60472
4	P01	2017.3.7 11:56:22	http://localhost/ewha/lear2_1 (CRC)	3.65088	3.49606
5	P01	2017.3.7 11:56:23	http://localhost/ewha/lear2_1 (CRC)	3.67677	3.51238
6	P01	2017.3.7 11:56:24	http://localhost/ewha/lear2_1 (CRC)	3.56621	3.47621
7	P01	2017.3.7 11:56:25	http://localhost/ewha/lear2_1 (CRC)	3.54486	3.45946
8	P01	2017.3.7 11:56:26	http://localhost/ewha/lear2_1 (CRC)	3.65536	3.51435
9	P01	2017.3.7 11:56:27	http://localhost/ewha/lear2_1 (CRC)	3.57321	3.37481
10	P01	2017.3.7 11:56:28	http://localhost/ewha/lear2_1 (CRC)	3.54029	3.45771

그림 3. 동공크기 측정 데이터의 예
Fig. 3. Example of pupil size measurement data

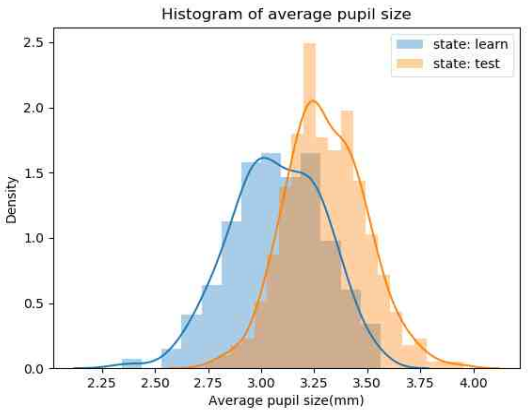


그림 4. 동공 크기 히스토그램
Fig. 4. Histogram of pupil sizes

하여 생성된 S_{D_1}과 S_{D_2}를 분류 머신러닝 알고리즘의 입력으로 사용하고, 성능 평가 기준으로는 전체 판단 결정 중 올바르게 판단한 비율인

$$\frac{P(TP)+P(TN)}{P(TP)+P(TN)+P(FP)+P(FN)} \quad (14)$$

를 사용한다. 여기서 TP와 TN은 각각 True Positive, True Negative를 뜻하며 실제 값과 예측 값이 같은 경우의 수를 의미하고, FP와 FN은 False Positive, False Negative를 뜻하며 실제 값과 예측값이 다른 경우를 뜻한다. 분류 머신러닝 알고리즘으로는 SVM(Support Vector Machine)과 KNN 방법을 사용하였고, 모델 검증 알고리즘으로는 K겹 교차 검증(K-fold Cross Validation)을 이용하였다.

4.2 시뮬레이션 데이터 생성 및 결과

본 절에서는 제안하는 데이터 재구성 방법의 성능을 수학적 모델 기반으로 생성된 데이터를 이용하여 검증하고자 한다. 제안하는 데이터 재구성 방법이 다양한 통계적 특성을 가진 데이터에 대해서도 성능향상을 보이는지 검증하기 위하여 본 시뮬레이션에서 사용한 수학적 모델은 동공 크기 데이터를 모사할 수 있는 균등분포모형, 자기회귀모형, 그리고 정규분포모형을 이용하였다.

동공 크기를 가장 간단하고 단순하게 모델링하기 위하여 균등분포모형을 사용하여 데이터를 생성할 수 있다. 두 값 a, b 구간 내에서 균등분포모형으로 생성된 데이터 x의 확률밀도함수는 다음과 같다.

$$f_X(x) = \begin{cases} \frac{1}{b-a} & (a < x < b) \\ 0 & \text{otherwise} \end{cases}. \quad (15)$$

a와 b는 각각 하한선, 상한선을 의미한다.

또한 동공 크기는 시계열 데이터로 과거의 값으로부터 규칙을 발견하고 미래에도 같은 규칙이 계속 될 것이라는 기대 하에 자기회귀모델을 이용하여 데이터를 생성할 수 있다. 일반적으로 p 자기회귀모형은 다음과 같은 관계식을 가진다.

$$x_t = \sum_{i=1}^{p} \phi_i x_{t-i} + \epsilon_t \quad (16)$$

여기서 x_t는 시점 t에서의 동공의 크기를 말한다. p는 자기회귀모형의 차수로 과거 p개의 값을 이용한다는 의미이며, ϕ_i는 자기상관계수, ϵ_t는 백색잡음이다.

본 실험에서 사용한 데이터는 1차 자기회귀모형을 기반으로 생성하였으며, 자기상관계수는 $\phi = 0.6$ 정하여 그림 5와 같이 데이터 D_1과 D_2를 생성하였다. D_1과 D_2는 각각 레이블 -1, +1에 해당하며, 각각 500개의 샘플 데이터를 포함하고 있다. 이 때 D_1과 D_2의 평균값의 차이인 $c = m_1 - m_2 \in \{0.3, 0.5, 0.7\}$로 설정하였다. 균등분포모형과 정규분포모형의 데이터는 표 1에 나타낸 것과 같이 자기회귀모형으로부터 계산된 최소값, 최대값, 평균값, 표준편차인 $AR_{\min}$, $AR_{\max}$, AR_{mean}, AR_{std}을 이용하여 500 샘플씩 생성하였다.

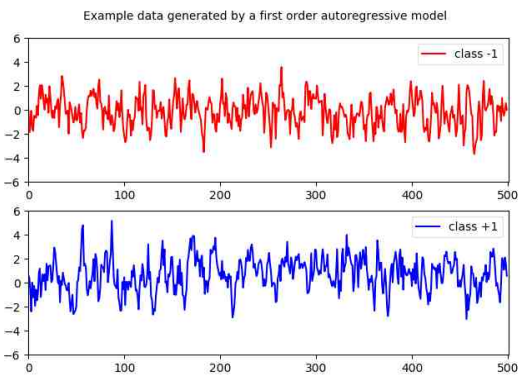


그림 5. $\phi = 0.6$, $c = 0.7$인 1차 자기회귀모형으로 생성한 데이터 예시
Fig. 5. Example of data generated by the first order autoregressive model with $\phi = 0.6$ and $c = 0.7$

표 1. 데이터 생성 모형 별 파라미터 설정
Table1. Hyper-parameters for data generation models

자기회귀모형(AR)	$p = 1$, $\phi = 0.6$
균등분포모형(UNIF)	하한선: $AR_{\min}$ 상한선: $AR_{\max}$
정규분포모형(NORM)	$\mu = AR_{mean}$ $\sigma = AR_{std}$

표 2. 데이터 생성 모형 별 적용 머신러닝 알고리즘
Table2. Machine learning algorithms for data generation models

		알고리즘	
		SVM	KNN
모형	자기회귀모형	AR_SVM	AR_KNN
	균등분포모형	UNIF_SVM	UNIF_KNN
	정규분포모형	NORM_SVM	NORM_KNN

다양한 머신러닝 기반 분류 알고리즘 중에서 일반적으로 좋은 성능을 보인다고 알려져 있는 SVM과 KNN의 두 알고리즘을 사용하였고[27], 이를 표 2에 나타내었다.

시뮬레이션은 D_1과 D_2 데이터를 훈련데이터로 사용하여 분류모델을 생성한 후, 테스트데이터를 생성한 분류모델에 적용하여 분류 정확도를 비교한다. 두 상태에 대한 차이인 $c = m_1 - m_2 \in \{0.3, 0.5, 0.7\}$와 데이터 재구성 시 길이인 $d(1 \le d \le 60,\ d \in R)$의 값에 변화를 주었다. 그리고 c값별로 1,000번씩 반복하여 머신러닝 알고리즘의 평균적 분류 성능을 비교하였다. SVM은 선형커널을 사용하였고, KNN은 $K = 5$와 유클리디안 거리를 이용하여 알고리즘을 적용하였다. 시뮬레이션은 Python 3.6을 이용하였으며, Windows 7 Enterprise 64bit, Intel® Core i7 860 CPU, RAM 8GB 환경에서 진행하였다.

그림 6은 SVM, 그림 7은 KNN 알고리즘으로 분류한 결과로 자기회귀모형, 균등분포모형, 정규분포모형으로 생성된 데이터에 대하여 1,000번의 평균 정확도를 데이터 재구성 길이인 d에 대하여 나타내었다. 자기회귀모형, 균등분포모형, 정규분포 모형의 모든 경우에서 생성된 데이터에 대하여 SVM또는 KNN의 분

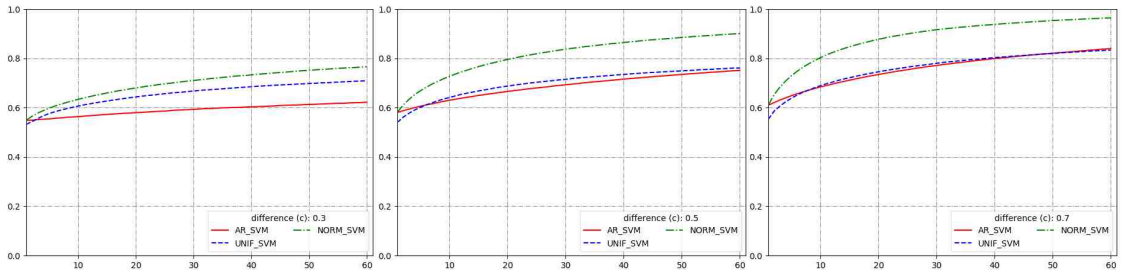


그림 6. 분류 정확도(SVM)
Fig. 6. Accuracy of classification(SVM)

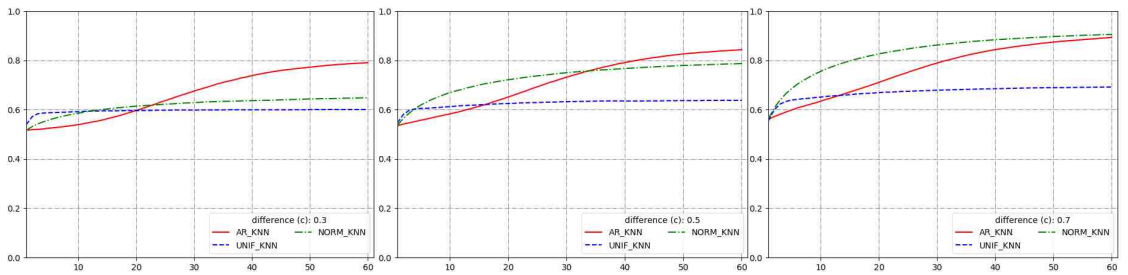


그림 7. 분류 정확도(KNN)
Fig. 7. Accuracy of classifiacation(KNN)

류 알고리즘을 적용한 경우, 정확도의 성능향상 형태는 상이하나 제안한 데이터 재구성 알고리즘의 길이인 d가 증가할수록 분류 정확도가 향상되는 것을 확인할 수 있다. 즉, 제안하는 데이터 재구성 방법은 분류하고자 하는 데이터의 특성과 상관없이 성능 향상을 가져올 수 있음을 확인할 수 있다.

4.3 실험 데이터 분류 결과

그림 8과 9는 실제 수집한 실험 데이터에 본 논문에서 제안하는 데이터 재구성 알고리즘을 적용하고, 각각 SVM 및 KNN 알고리즘으로 분류한 결과이다. 35명의 실험 참가자 데이터를 사용한 분류 알고리즘별로 50% 신뢰구간을 포함한 평균 분류 정확도를 제시하였다.

본 논문에서 제안하는 방법을 적용하여 분류하면 $d=60$ 기준으로 전체 평균 분류 정확도가 SVM은 0.64에서 0.75로 약 17%, KNN은 0.64에서 0.76으로 약 17.5%의 성능향상을 보여줬다. 제안한 데이터 재구성 전처리 알고리즘을 실제 데이터에 적용한 때도 모델 기반으로 생성한 임의 데이터 결과와 마찬가지로 분류 성능이 향상되는 것을 확인할 수 있다. 그림 10과 그림 11은 데이터 길이 d에 따른 분류 정확도 변화량을 나타낸다. 분류 정확도 변화량은 $d+\alpha$일 때의 분류 정확도와 d일 때의 분류 정확도 차이를 계산하였다. 본 결과는 $\alpha=5$일 때의 값이다. 사용한 분류 알고리즘인 SVM과 KNN 모두에서 일정 데이터 길이까지는 분류 정확도 변화량은 높으나 그 이후로는 분류 정확도가 크게 변하지 않는 포화 상태가 되는

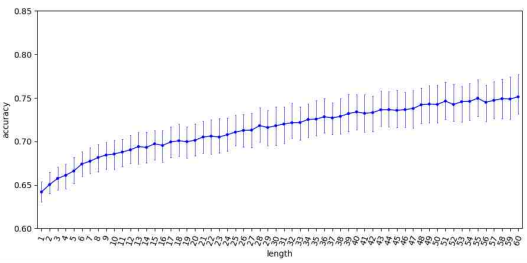


그림 8. 실험 참가자 분류 정확도 (SVM)
Fig. 8. Classification accuracy of participants (SVM)

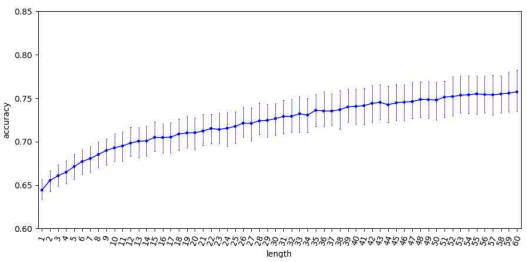


그림 9. 실험 참가자 분류 정확도 (KNN)
Fig. 9. Classification accuracy of participants (KNN)

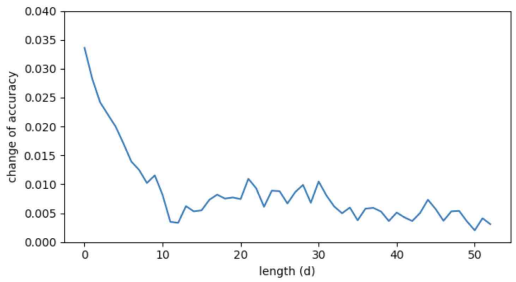


그림 10. 분류 정확도 변화량 (SVM)
Fig. 10. Change of accuracy (SVM)

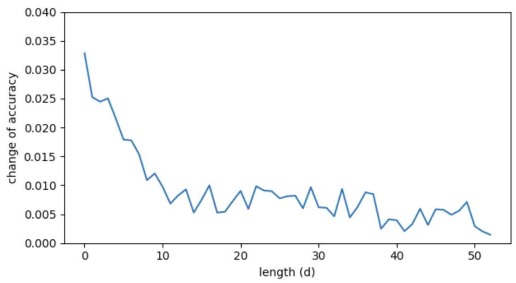


그림 11. 분류 정확도 변화량 (KNN)
Fig. 11. Change of accuracy (KNN)

것을 확인할 수 있다. 하지만 데이터 길이 d가 길어질수록 학습에 필요한 복잡도가 높아지기 때문에 데이터 재구성 알고리즘을 적용할 경우 분류 정확도 변화량이 0에 가까워지는 데이터 길이 d를 사용하는 것이 가장 효율적이다.

References

[1] D. Choi and S. Kang, "Software architecture of a wearable device to measure user's vital signal depending on the behavior recognition," *J. KICS*, vol. 41, no. 3, pp. 347-358, 2016.

[2] C. Lim, "A study on the analysis of technology and service issues for wearable devices and future development direction," *J. KINGC*, vol. 13, no. 4, pp. 81-89, 2017.

[3] E. Jang, A. Kim, and H. Yu, "Relationships of psychological factors to stress and heart rate variability as stress responses induced by cognitive stressors," *Korean J. Sci. Emotion & Sensibility*, vol. 21, no. 1, pp. 71-82, 2018.

[4] E. Jang, Y. Eum, S. Kim, and J. Sohn, "Discrimination of three emotions using parameters of autonomic nervous system responses," *J. Ergonomics Soc. Korea*, vol. 30, no. 6, pp. 705-713, 2011.

[5] E. H. Hess and J. M. Polt, "Pupil size in relation to mental activity during simple problem-solving," *Science*, vol. 143, no. 3611, pp. 1190-1192, 1964.

[6] D. Kahneman and J. Beatty, "Pupil diameter and load on memory," *Science*, vol. 154, no. 3756, pp. 1583-1585, 1966.

[7] D. W. Lee, et al., "Pupil data measurement and social emotion inference technology by using smart glasses," in *Proc. KIBME*, pp. 1-4, 2019.

[8] M. Saecker and V. Markl, "Big data analytics on modern hardware architectures: A technology survey," in *Proc. Eur. Business Intell. Summer School*, pp. 125-149, 2012.

[9] Y. Oh and I. Jo, "The effects of mathematics anxiety on performance efficiency: Focused on the pupil size of college students," *J. Edu. Technol.*, vol. 33, no. 3, pp. 653-680, 2017.

[10] S. Zolhavarieh, S. Aghabozorgi, and Y. W. Teh, "A review of subsequence time series clustering," *The Scientific World J.*, 2014.

[11] J. Park, S. Kang, B. Lee, U. Kang, and Y. Lee, "Design of user concentration classification model by EEG analysis based on Visual SCPT," *J. Korea Soc. Comput. and Info.*, vol. 23 no. 11, pp. 129-135, 2018.

[12] S. R. Steinhauer, G. J. Siegle, R. Condray, and M. Pless, "Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing," *Int. J. Psychophysiology*, vol. 52, no.1, pp. 77-86, 2004.

[13] X. W. Wang, D. Nie, and B. L. Lu, "Emotional state classification from EEG data using machine learning approach," *Neurocomputing*, vol. 129, pp. 94-106, 2014.

[14] N. Y. Liang, P. Saratchandran, G. B. Huang, and N. Sundararajan, "Classification of mental tasks from EEG signals using extreme learning machine," *Int. J. Neural Syst.*, vol.

16, no. 1, pp. 29-38, 2006.

[15] M. Khondoker, R. Dobson, C. Skirrow, A. Simmons, and D. Stahl, "A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies," *Statistical Methods in Med. Res.*, vol. 25, no. 5, pp. 1804-1823, 2016.

[16] Y. H. Li, F. You, K. Chen, L. Huang, and J. Xu, "A real-time system for monitoring driver fatigue," *Transp. Planning and Technol.*, vol. 39, no. 8, pp. 779-790, 2016.

[17] C. W. Anderson, J. N. Knight, T. O'Connor, M. J. Kirby, and A. Sokolov, "Geometric subspace methods and time-delay Embedding for EEG artifact removal and classification," *IEEE Trans. Neural Syst. and Rehabilitation Eng.*, vol. 14, no. 2, pp. 142-146, 2006.

[18] G. Bontempi, S. B. Taieb, and Y. A. Le Borgne, "Machine learning strategies for time series forecasting," *Eur. Business Intell. Summer School, Springer*, pp. 62-77, 2012.

[19] R. J. Povinelli, M. T. Johnson, A. C. Lindgren, F. M. Roberts, and J. Ye, "Statistical models of reconstructed phase spaces for signal classification," *IEEE Trans. Sign. Process.*, vol. 54, no. 6, pp. 2178-2186, 2006.

[20] H. Eun, "Basics of electroencephalography for neuropsychiatrist," *J. Korean Neuropsychiatric Assoc.*, vol. 58, no. 2, pp. 76-104, 2019.

[21] F. C. Morabito, D. Labate, F. L. Foresta, A. Bramanti, G. Morabito, and I. Palamara, "Multivariate multi-scale permutation entropy for complexity analysis of Alzheimer's disease EEG," *Entropy*, vol. 14, no. 7, pp. 1186-1202, 2012.

[22] B. Wahn, D. P. Ferris, W. D. Hairston, and P. König, "Pupil sizes scale with attentional load and task experience in a multiple object tracking task," *PLoS ONE*, vol. 11, no. 12, 2016.

[23] G. Sperling and E. Weichselgartner, "Episodic theory of the dynamics of spatial attention," *Psychological Rev.*, vol. 102, no. 3, pp. 503-532, 1995.

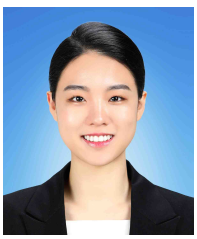
[24] C. M. Bishop, "Pattern recognition and machine learning," *Springer*, 2006.

[25] B. Everitt, "*The Cambridge dictionary of statistics*," Cambridge University Press, 2006.

[26] "*Tobii Pro X2-30 screen-based eye tracker*," Tobiipro.com, 2019. [Online] Available: https://www.tobiipro.com/product-listing/tobii-pro-x2-30/. [Accessed: 19- Jun- 2019].

[27] A. Atla, R. Tada, V. Sheng, and N. Singireddy, "Sensitivity of different machine learning algorithms to noise," *J. Computing Sci. in Colleges*, vol. 26, no. 5, pp. 96-103, 2011.

이 정 진 (Jungjin Lee)



2014년 2월 : 이화여자대학교 전자공학과 졸업
2019년 8월 : 이화여자대학교 전자전기공학과 석사
2010년 2월~현재 : 이화여자대학교 전자전기공학과 연구원
<관심분야> 머신러닝 기반 네트워크 데이터 분류
[ORCID:0000-0002-2422-1153]

조 일 현 (Il-Hyun Jo)

1987년 : 서울대학교 농경제학과 졸업
1994년 : 연세대학교 산업교육학 석사
2001년 : Florida State University 교육공학과 Ph.D.
2008~현재 : 이화여자대학교 사범대학 교육공학과 교수
<관심분야> 에듀테크 교수설계, 학습분석학, HRD

박 형 곤 (Hyunggon Park)

2004년 2월 : 포항공과대학교 전자전기공학과 졸업

2006년 3월 : University of California, Los Angeles (UCLA) M.S.

2008년 12월 : University of California, Los Angeles (UCLA) Ph.D.

2010년~현재 : 이화여자대학교 전자전기공학과 부교수

<관심분야> 멀티에이전트 네트워크 시스템, 머신러닝 기반 분산적 의사 결정 전략, 게임이론 기반 네트워크 분산적 자원 관리, 네트워크 코딩

[ORCID:0000-0002-5079-1504]