Check for updates

일반논문 (Regular Paper)
방송공학회논문지 제29권 제5호, 2024년 9월 (JBE Vol.29, No.5, September 2024)
https://doi.org/10.5909/JBE.2024.29.5.654
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

네트워크 품질 예측을 위한 상관관계 기반 정형 데이터 증강 기법

오 성[a], 박 지 연[b], 김 정 식[b], 이 나 래[b], 김 명 호[b], 배 성 호[a]‡

A Correlation-based Tabular data Augmentation method for Network QoS Prediction

Sung Oh[a], Ji-Yeon Park[b], Joung-Sik Kim[b], Na-Rae Yi[b], Myungho Kim[b], and Sung-Ho Bae[a]‡

요 약

최근 심층신경망(deep neural network)은 높은 모델 용량(model capacity)를 기반으로 다양한 예측 문제에 활용되어 탁월한 성능을 보이고 있다. 본 논문은 Network QoS (Quality of Service) 예측 문제에 있어 심층신경망 일반화 성능을 최대화하기 위한 새로운 데이터 증강(data augmentation)기법을 제안한다. 일반적으로, Network QoS 데이터는 정형 데이터로서, 각 특징이 목표값에 미치는 영향의 편차가 큰 특성을 보인다. 이를 반영하여 본 논문은 각 특징 간 목표 값과의 상관관계를 고려하여, 상관관계 점수가 낮은 특징을 마스킹(masking)함으로써 데이터를 증강하는 데이터 증강 기법을 제안한다. 이를 통해 제안 방법은 원본 데이터의 표현력을 유지하면서 모델의 일반화 성능을 높일 수 있는 양질의 증강 데이터를 생성할 것으로 기대하였다. 실험 결과, 제안 방법은 베이스라인 대비 BerlinV2X Network QoS 데이터셋에 대하여 RMSE 관점에서 최대 6.6%p의 성능 향상을 보였다.

Abstract

Recently, deep neural networks (DNNs) have demonstrated excellent performance across various prediction tasks, leveraging their high model capacity. In this paper, we propose a novel data augmentation technique for enhancing the generalization performance of DNNs in the context of Network Quality of Service (QoS) prediction. Typically, Network QoS data is structured, and each feature exhibits significant variability in its impact on the target value. Taking this into account, our approach considers the correlation between each feature and the target value. By masking features with low correlation scores, we generate augmented data that maintains the expressive power of the original data while improving the model's generalization performance. Experimental results show that our proposed method achieves up to a 6.6% improvement in RMSE compared to the baseline on the BerlinV2X Network QoS dataset.

Keyword : Deep learning, Network dataset, Tabular data, Regression task, Dataset augmentation

a) 경희대학교 컴퓨터공학과(Kyung Hee University)
b) 한화시스템(Hanwha Systems)
‡ Corresponding Author : 배성호(Sung-Ho Bae)
E-mail: shbae@khu.ac.kr
Tel: +82-31-201-2593
ORCID: https://orcid.org/0000-0002-3389-1159

※이 논문은 2022년도 정부(방위사업청)의 재원으로 국방기술진흥연구소의 지원을 받아 수행된 연구임(KRIT-CT-22-077, 전장 적응형 다계층 통신을 위한 통합 통신단말 및 네트워크 기술 개발)
· Manuscript June 25, 2024; Revised August 22, 2024; Accepted August 22, 2024.

Copyright Ⓒ 2024 Korean Institute of Broadcast and Media Engineers. All rights reserved.
"This is an Open-Access article distributed under the terms of the Creative Commons BY-NC-ND (http://creativecommons.org/licenses/by-nc-nd/3.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited and not altered."

Ⅰ. 서 론

심층 신경망 기술(deep neural networks)의 발전으로 딥러닝의 활용 분야는 컴퓨터 비전과 자연어 처리의 영역을 넘어 다양한 영역으로 확산하고 있다. 그 중 통신 데이터셋의 일반적 형태인 정형 데이터 처리 영역에서도 딥러닝 기술의 사용이 확산되고 있다. 기존에는 정형 데이터를 처리하는 데에 Random Forest[1], XGBoost[2], LightGBM[3], CatBoost[4] 등과 같은 머신러닝 기법이 주로 활용되었다. 그러나 최근에는 TabNet[5], MLP, ResNet[6]과 같은 딥러닝 기반의 정형 데이터 처리 모델들이 기존의 머신러닝 방법을 뛰어넘는 성능을 보이고 있다.

딥러닝을 통한 통신 데이터 처리는 고려해야할 특징(feature)이 복잡하고, 대량의 데이터를 처리해야 할 경우에 매우 유용하다. 예를 들어 전장(battle field)과 같은 특수 상황에서 수많은 통신 파라미터를 고려하여 안정적인 네트워크를 유지하기 위해서는 현재의 네트워크 품질(Quality of Service; QoS)을 예측하고 통신 링크를 적절한 시점에 핸드오버 하는 기술이 필요한데, 딥러닝 기반의 모델을 통해 이를 달성할 수 있다.

딥러닝 모델은 데이터 양이 많을수록 성능이 향상되는 경향이 있으나 일부 데이터는 수집에는 큰 비용이 발생하여 데이터 부족 문제가 존재한다. 따라서 비용을 추가로 발생시키지 않으면서 데이터를 최대한 활용하는 방법이 필요한데, 이를 위해 사용될 수 있는 효과적인 방법 중 하나로 데이터 증강 기법[7]이 있다.

정형 데이터의 데이터 증강 기술로는 SMOTE (Synthetic Minority Over-sampling Technique)[8]가 제안되었다. 이 방법은 목표값을 기준으로 간헐적 샘플 분포가 존재하는 영역의 샘플이 모델의 일반화 성능에 악영향을 준다고 보고, 해당 영역의 샘플에 대해 특징 값 보간(interpolation)을 이용하여 데이터 부족 및 불균형 문제를 완화하였다. 그러나 이 방법은 데이터 불균형이 심각한 과업에 적합할 수 있으나 네트워크 QoS 예측과 같이 샘플별 불균형 보다 샘플 내 각 특징별 중요도의 불균형이 심각한 과업에서는 효과적으로 활용될 수 없다.

본 연구에 사용된 BerlinV2X Network QoS 데이터셋[9]은 특징(feature)별로 목표값과의 상관계수 편차가 큰 특성이 있다. 그림 1은 BerlineV2X 데이터셋의 목표값과 각 특징 간(설계행렬의 열벡터) 피어슨 상관계수(Pearson Linear

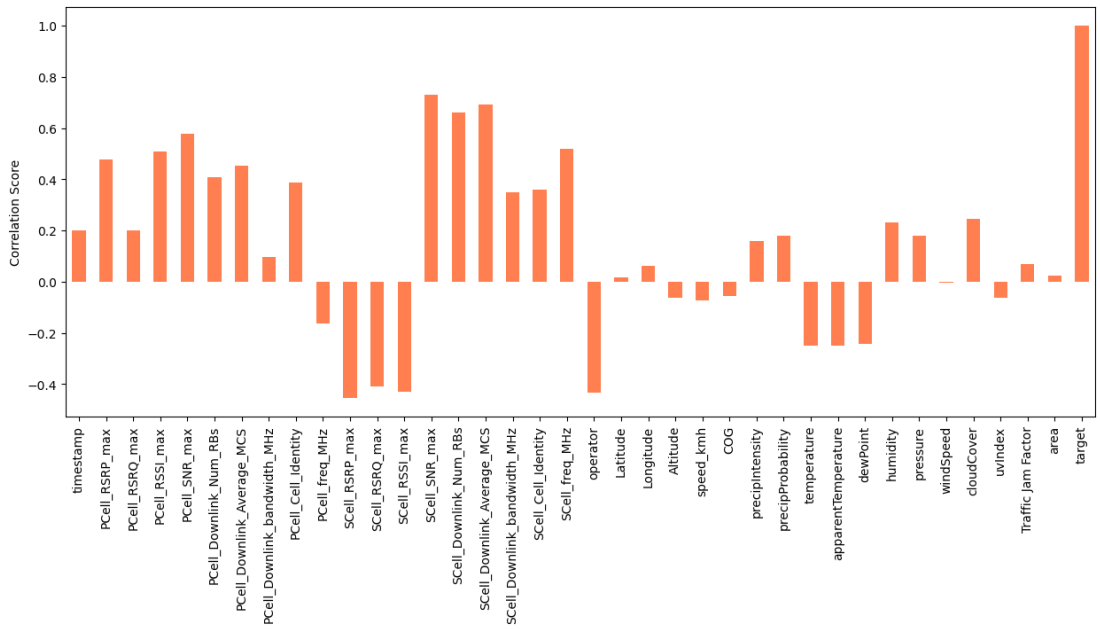

그림 1. Target feature와 다른 feature 간의 Correlation Scores
Fig. 1. Correlation Scores between target feature and other feature

Correlation Coefficient, PLCC)를 보인다. 그림 1에서 보이듯이, 각 특징과 목표값 간 상관계수의 편차가 상당히 큰 것을 알 수 있고, 이는 데이터 증강 시 특징별로 상관계수를 반영하는 증강 방법이 효과적일 수 있다는 가능성을 보여준다.

따라서 본 논문에서는 특징별 중요도를 데이터의 목표값과 단일 특징(design matrix의 column vector)간 상관계수(correlation coefficient)로 정의하고 특징 별 중요도에 따라 선별적으로 특징을 변형하여 데이터를 증강하는 방법을 제안한다. 즉, 특징별 중요도가 높은 특징은 보존하고, 중요도가 낮은 특징을 변형하여 데이터를 증강하는 새로운 방법을 제안한다. 실험 결과, 제안한 방법은 데이터 증강을 사용하지 않은 베이스라인에 비해 최대 6.6%p 더 낮은 RMSE 성능을 보여주며 우리의 방법이 효과적임을 증명하였다.

II. 본 론

1. 관련연구

본 연구에서는 TabNet[3]을 baseline으로 사용한다. TabNet은 최초로 제안된 딥러닝 기반의 정형 데이터 처리 모델로 기존의 머신러닝 방법의 성능을 뛰어넘는 성능을 보여준다. TabNet은 간헐 특징 선택법(sparse feature selection)을 활용하여 모델의 예측에 중요한 특징을 파악하고 해당 특징들을 사용하여 예측을 수행하는데, 본 연구에서 제안하는 방법도 비슷한 맥락으로, 목표값과 관련성이 높은 특징을 유지하고, 관련성이 낮은 특징을 변형하기 때문에 TabNet의 예측 성능 향상에 도움이 될 것으로 기대한다.

또 다른 딥러닝 기반의 정형 데이터 처리 모델에는 MLP, ResNet 모델이 있다. MLP 기반의 정형 데이터 처리 모델은 기존의 Multi-Layer Perceptron[6]을 정형 데이터를 처리할 수 있도록 응용한 모델이다. ResNet 기반 모델[6]도 마찬가지로 기존의 컴퓨터 비전 영역에서 사용되는 ResNet 모델[9]의 구조를 응용한 모델이다.

데이터 증강 기법은 수집된 훈련 데이터를 최대한 활용해 데이터의 다양성을 증가시켜 모델의 예측 성능을 향상시키는 방법이다. 정형 데이터를 증강하는 데이터 증강 기법들도 개발되었는데, 정형 데이터 증강기법인 SMOTE[8]는 훈련 데이터의 클래스 불균형 문제를 해결하기 위해서 제안된 방법이다. SMOTE는 샘플의 수가 적은 클래스의 데이터를 선택하고, 그 샘플과 가까운 이웃 사이에 새로운 데이터 샘플을 생성한다. 이를 통해서 소수 데이터 분포를 크게 벗어나지 않는 새로운 데이터 샘플을 생성한다. 그러

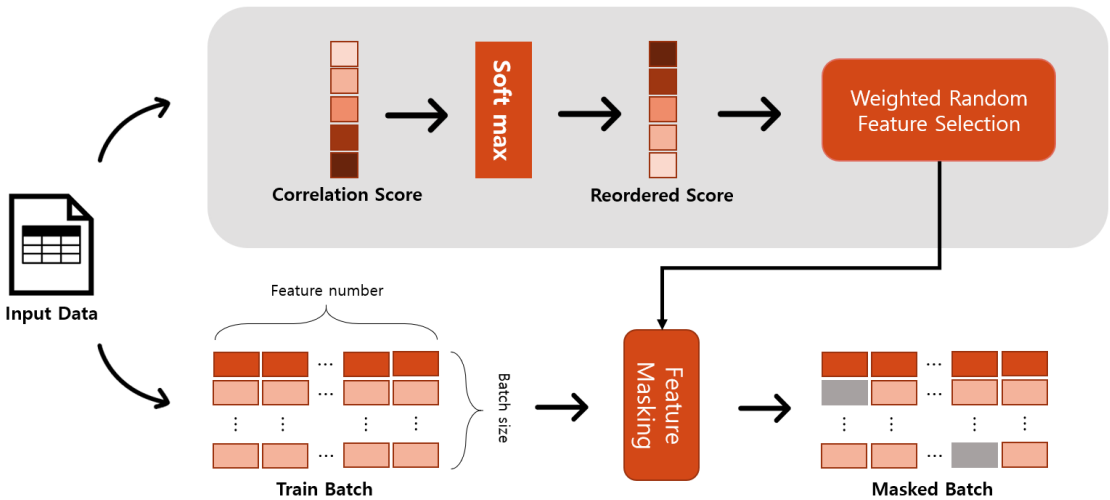

그림 2. 네트워크 품질 예측을 위한 상관관계 기반 정형 데이터 증강 기법 개요도
Fig. 2. Overview of the Correlation-based Tabular data Augmentation method for Network QoS Prediction

나, 해당 방법은 클래스 불균형 문제를 해결하는데 집중하여 특징 별 중요도를 데이터 증강 기법에 반영하지 못한다.

2. 네트워크 품질 예측을 위한 상관관계 기반 정형 데이터 증강 기법

제안하는 방법의 개요도가 그림 2에 나와있다. 먼저 입력 데이터에 대해서 배치별로 목표값과 특징들 사이의 상관관계 점수를 구한다. 그런 다음, 상관관계 점수를 소프트맥스 함수를 통과하여 재배열한 후, 재배열된 상관관계 점수를 기반으로 점수가 낮은 특징을 선택하여 마스킹(masking)을 하여 데이터를 변형한다. 최종적으로 변형된 데이터를 네트워크 훈련을 위한 데이터로 사용한다.

상관관계 점수를 얻기 위해 우리는 목표값과 다른 특징들 사이의 상관관계(PLCC) 점수를 계산한다. 이 점수는 목표값과의 관련성이 높을수록 높은 값을 가지게 된다. 마스킹 할 대상 특징을 선택하기 위해서 우리는 가중 랜덤 함수(weighted random function)를 사용하는데, 이 랜덤 함수는 가중치 값이 높을수록 더 높은 확률로 해당 특징을 선택한다. 그러나 우리가 원하는 것은 관련성이 낮은 특징들을 변형하고, 관련성이 높은 특징들은 유지하며 데이터를 증강하는 것이다. 따라서 상관관계 점수의 값을 재배열(reordering)해야 한다. 이를 위해 온도(temperature)를 도입한 소프트맥스 함수(Φ)를 사용한다.

$$\Phi(x_i) = exp^z / \sum_{i=1}^{n} exp^z, z = -x_i / T \quad (1)$$

여기서 n은 특징의 개수를 나타내며, x_i는 각 특징의 상관관계 점수를 의미한다. 식 (1)의 T는 온도를 나타낸다. 높은 온도 값은 상대적으로 더 평평한 분포를 만들어주고, 낮은 온도 값은 상대적으로 뾰족한 분포를 만들어준다. 결과적으로, 제안 방법은 중요도에 반비례해서 변형할 특징의 확률을 할당하며, 이때 T를 통해 분포의 모양을 조절한다.

한편, 기존의 마스킹 방법과 같이 선택된 특징의 값을 0으로 설정하면 성능이 크게 저하되는 현상을 관찰했다. 이는 큰 값을 가지고 있는 특징의 마스크 값을 0으로 설정하게 되면 증강된 데이터가 원래 데이터 분포에서 크게 멀어지게 되어 노이즈로 작용하기 때문이다. 따라서 본 연구에서는 마스크의 값을 0이 아닌 각 특징의 배치별 평균값(mean) 혹은 중위값(median)을 사용한다. 이를 통해 증강된 데이터가 원본 데이터 분포를 크게 벗어나지 않게 유지될 수 있다. 마스킹 비율은 데이터 증강 대상으로 선택된 샘플에서 상관관계 점수에 따라 선택된 1개의 특징만 마스킹 된다.

최종적으로, 배치별로 증강할 데이터의 비율(ratio)을 따라 입력 데이터에 마스크를 씌워 증강된 데이터를 훈련 데이터로 사용한다.

3. 실험 결과

실험에는 네트워크 통신 품질을 예측하는 회귀 과업 데이터셋인 BerlinV2X Network QoS 데이터셋이 사용되었다. BerlinV2X 데이터셋은 40개의 특징을 가지는 34,274개의 네트워크 데이터 샘플로 이루어져 있다. 또한 일반 회귀 과업 데이터셋의 결과와 비교하기 위해 같은 회귀 task 데이터셋인 캘리포니아 주택 가격 예측 데이터셋을 사용하였다. 캘리포니아 주택 가격 예측 데이터셋의 경우 10개의 특징을 가진 20,640개의 데이터 샘플로 이루어져 있다. 데이터 전처리로는 Nan 값을 0으로 대체하고, 카테고리별 특징들에 대해서는 Ordinal Encoder[11]를 사용해서 인코딩 하였다. 실험 결과는 모델이 예측한 값과 목표값 사이의 오차를 RMSE (Root Mean Square Error)로 평가한다. Berlin V2X 데이터셋의 경우 목표값의 크기가 값이 커서 학습이 잘 되지 않는 현상을 보였다. 따라서 목표 값 평준화(target value normalization)을 적용해 훈련을 진행하였다. 각 실험은 5번의 실험을 평균한 결과를 제시한다.

본 연구에선 상관관계 점수를 재배열하기 위해 소프트맥스 함수의 온도를 조절해 특징별로 선택되는 샘플의 분포를 조절한다. 각 온도에 따른 특징별 샘플의 분포가 Appendix 1에 나와있다. 샘플 분포를 보면 온도 값이 0.2 ~ 0.25의 값일 때, 원본 분포와 가장 비슷한 분포를 가진다.

표 1에는 여러 딥러닝 모델에 대한 BerlinV2X 데이터셋의 실험 결과가 정리되어 나와있다. RMSE 값이 Mbit/s로 표기된 이유는, Berlin V2X 데이터셋의 타겟값인 datarate의 단위가 Mbit/s이기 때문이다. 표 1의 Ours에는 데이터 증강 비율(ratio) 5%에 온도는 0.2로 설정한 결과를 나타냈

표 1. Berlin V2X 데이터셋의 여러 딥러닝 모델에 대한 RMSE 스코어 비교표
Table 1. RMSE score comparison for multiple deep learning models on the Berlin V2X dataset

Dataset	Model	Method	RMSE(Mbit/s)
Berlin V2X	TabNet	Baseline	10.6550
		SMOTE	10.6711
		Ours	10.0821
	MLP	Baseline	4.3202
		SMOTE	4.3355
		Ours	4.2534
	ResNet	Baseline	4.0430
		SMOTE	4.3546
		Ours	3.8675

다. 실험 결과 기존의 정형 데이터 증강 기법인 SMOTE는 데이터 증강 기법을 적용하지 않은 Baseline에 비해 성능이 떨어지는 결과를 보여주었으나, 우리가 제안한 방법은 더 낮은 RMSE를 보여주는 것을 확인할 수 있다. 다른 딥러닝 기반 모델인 MLP 모델과 ResNet 모델의 결과에서도 우리가 제안한 방법이 가장 좋은 성능을 보여주는 결과를 확인할 수 있다. 더 다양한 ratio와 온도에 따른 실험 결과는 그림 3의 그래프에서 확인할 수 있다. 그림 3 왼쪽 그래프를 통해 데이터 증강 비율이 증가할수록 RMSE가 낮아지는 것을 확인할 수 있는데, 이는 제안하는 데이터 증강 기법이 효과적임을 보인다. 또한 그림 3 오른쪽 그래프를 통해 원본 분포와 유사한 분포를 만들어주는 온도(0.2)에서 가장 좋은 성능을 보이는 것을 확인할 수 있다.

표 2. 캘리포니아 데이터셋의 RMSE 스코어
Table 2. RMSE score for the California dataset

Dataset	Model	Method	RMSE
California	TabNet	Baseline	6.3481
		SMOTE	6.8237
		Ours	6.2672

표 2는 Berlin V2X 데이터셋과 동일한 회귀 작업 데이터셋인 캘리포니아 데이터셋의 실험 결과를 나타낸다. 캘리포니아 데이터셋에서도 마찬가지로 SMOTE는 성능을 떨어트렸으나, 우리가 제안한 방법은 성능이 향상되었음을 보여줌으로써 본 제안 방법이 다른 회귀 데이터셋에서도 유용함을 보인다.

Ⅲ. 결 론

본 연구에서는 목표값과 다른 특징들의 상관관계를 구하고, 상관관계 점수를 기반으로 데이터를 증강하는 새로운 데이터 증강 기법을 제안한다. 제안하는 방법을 통해 우리는 네트워크 QoS 예측 데이터셋에 대해서 목표값과 관련성이 높은 특징은 유지하고, 관련성이 낮은 특징을 변형하여

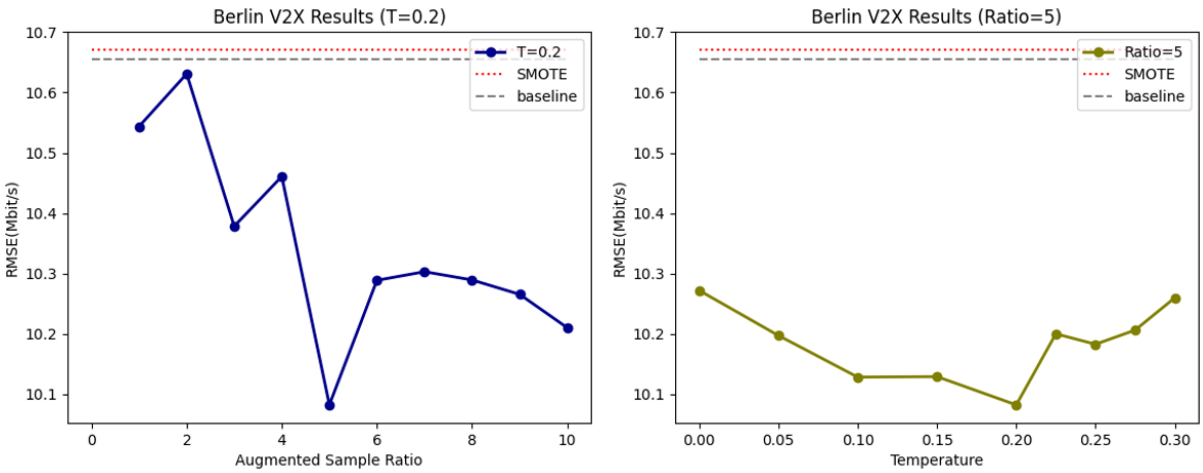

그림 3. Temperature와 Ratio에 따른 성능 비교 그래프
Fig. 3. Performance comparison graph according to Temperature and Ratio

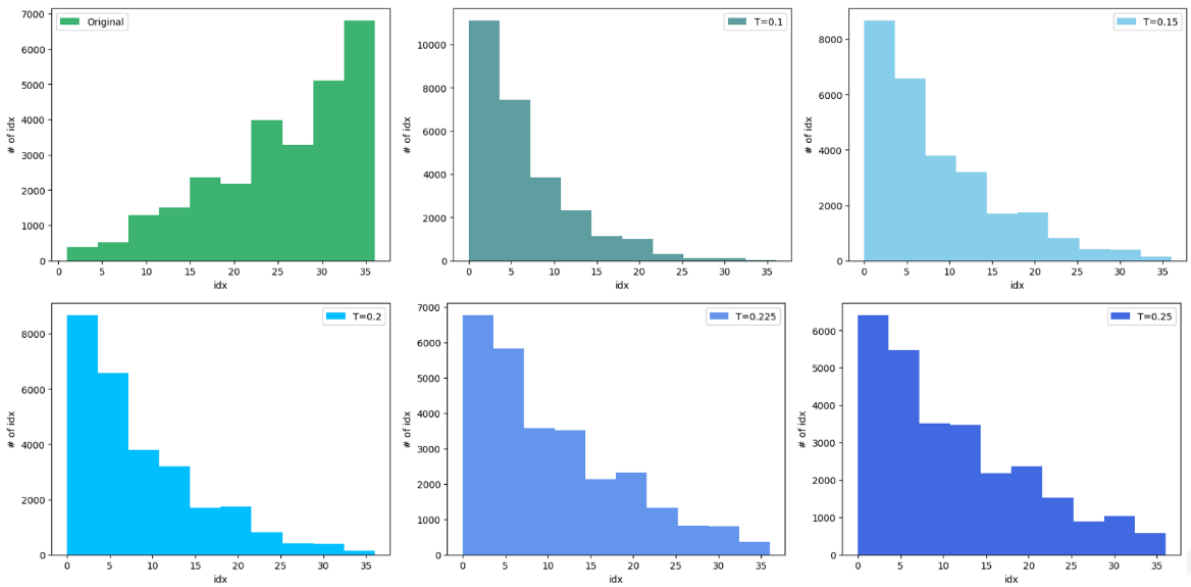

그림 4. 온도에 따른 특징 별 마스크된 샘플의 분포
Fig. 4. Distribution of masked samples by feature according to temperature

데이터의 다양성을 효과적으로 증가시켰다. 실험 결과 다른 데이터 증강 기법은 베이스라인 대비 성능을 떨어트리는 반면 우리는 해당 방법을 통해 TabNet 모델의 예측 성능을 베이스라인 성능 대비 최대 6.6%p 향상시켰다.

Ⅳ. 부록

1. 온도에 따른 특징 별 masked 샘플의 분포

상관관계 점수의 재배열을 위해 도입한 소프트맥스 함수의 온도에 따른 특징 별 선택된 샘플의 분포가 그림 4에 나와있다. 온도는 각 0.1, 0.15, 0.2, 0.225, 0.25로 설정하였다. 그래프를 보면, 온도가 0.2 ~ 0.25의 값을 가질 때 Original 분포의 역순과 가장 유사한 분포를 보이는 것을 확인할 수 있다.

참 고 문 헌 (References)

[1] Leo Breiman. “Random Forests.” Machine Learning 45, 5-32 (2001).
doi: https://doi.org/10.1023/A:1010933404324

[2] Tianqi Chen, Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System.” arXiv:1603.02754 (2016).
doi: https://doi.org/10.48550/arXiv.1603.02754

[3] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu. “LightGBM: A Highly Efficient Gradient Boosting Decision Tree.” NIPS 2017.

[4] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, Andrey Gulin. “CatBoost: unbiased boosting with categorical features.” arXiv:1706.09516 (2017).
doi: https://doi.org/10.48550/arXiv.1706.09516

[5] Sercan O. Arik, Tomas Pfister. “TabNet: Attentive Interpretable Tabular Learning.” AAAI (2021).
doi: https://doi.org/10.48550/arXiv.1908.07442

[6] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, Artem Babenko. “Revisiting Deep Learning Models for Tabular Data.” NeurIPS (2021).
doi: https://doi.org/10.48550/arXiv.2106.11959

[7] Luis Perez, Jason Wang. “The Effectiveness of Data Augmentation in Image Classification using Deep Learning.” arXiv:1712.04621 (2017).
doi: https://doi.org/10.48550/arXiv.1712.04621

[8] Nitesh V. chawla et al. “SMOTE: Synthetic Minority Over-sampling Technique.” Journal of Artificial Intelligence Research 16 (2002) 321-357.
doi: https://doi.org/10.48550/arXiv.1106.1813

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. “Deep Residual Learning for Image Recognition.” CVPR 2016.
doi: https://doi.org/10.48550/arXiv.1512.03385

[10] Rodrigo Hernangómez, Philipp Geuer, Alexandros Palaios, Daniel Schäufele, Cara Watermann, Khawla Taleb-Bouhemadi, Mohammad Parvini, Anton Krause, Sanket Partani, Christian Vielhaus, Martin

Kasparick, Daniel F. Külzer, Friedrich Burmeister, Frank H. P. Fitzek, Hans D. Schotten, Gerhard Fettweis, Sławomir Stańczak. “Berlin V2X: A Machine Learning Dataset from Multiple Vehicles and Radio Access Technologies.” arXiv: 2212.10343
doi: https://doi.org/10.48550/arXiv.2212.10343

[11] Pedregosa, F. and Varoquaux, G. and Gramfort, A. Michel, V. Thirion, B. and Grisel, O. and Blondel, M. Prettenhofer, P. Weiss, R. Dubourg, V. Vanderplas, J. Passos, A. Cournapeau, D. Brucher, M. Perrot, M. Duchesnay, E. “Scikit-learn: Machine Learning in {P}ython.” Journal of Machine Learning Research 12(2011) 2825-2830

저 자 소 개

오 성

- 2013년 3월 ~ 2021년 2월 : 경희대학교 전자정보대학 전자공학과 학사
- 2022년 3월 ~ 현재 : 경희대학교 컴퓨터공학과 석박통합과정
- ORCID : https://doi.org/0009-0004-3416-9426
- 주관심분야 : 딥러닝 기반 데이터 압축, 컴퓨터 비전, 지식 증류

박 지 연

- 2017년 3월 ~ 2021년 2월 : 한양대학교 전자공학부 학사
- 2021년 3월 ~ 2023년 2월 : 한양대학교 전자공학과 석사
- 2023년 2월 ~ 현재 : 한화시스템 연구원
- ORCID : https://doi.org/0000-0001-7620-1499
- 주관심분야 : 딥러닝, 전술통신, 시계열 데이터 처리

김 정 식

- 1998년 3월 ~ 2005년 2월 : 경북대학교 컴퓨터공학과 학사
- 2005년 3월 ~ 2007년 2월 : 경북대학교 컴퓨터공학과 석사
- 2011년 2월 ~ 현재 : 한화시스템 수석연구원
- ORCID : https://doi.org/0009-0006-0951-5766
- 주관심분야 : 네트워크, 전술통신, 인공지능

이 나 래

- 2009년 3월 ~ 2013년 2월 : 가톨릭대학교 정보시스템공학과 학사
- 2022년 9월 ~ 현재 : 한화시스템 전문연구원
- ORCID : https://doi.org/0009-0004-6520-0625
- 주관심분야 : 네트워크, 5G, 전술통신, 인공지능

저 자 소 개

김 명 호

- 1996년 3월 ~ 2002년 2월 : 경희대학교 전자공학과 학사
- 2002년 3월 ~ 2004년 2월 : 경희대학교 전자공학과 석사
- 2004년 3월 ~ 현재 : 한화시스템 수석연구원
- ORCID : https://doi.org/0009-0002-2689-9336
- 주관심분야 : 5G/B5G통신, MIMO, NTN, 인지무선시스템

배 성 호

- 2004년 3월 ~ 2011년 2월 : 경희대학교 전자정보대학 컴퓨터공학 및 전자공학 공학사 (복수전공)
- 2011년 2월 ~ 2016년 8월 : KAIST 전기 및 전자공학과 공학박사
- 2016년 7월 ~ 2017년 8월 : MIT Computer Science and Artificial Intelligence Laboratory (CSAIL) 박사 후 연구원
- 2017년 9월 ~ 현재 : 경희대학교 전자정보대학 컴퓨터공학과 조교수
- ORCID : https://doi.org/0000-0002-3389-1159
- 주관심분야 : AI 데이터/모델 압축, 이미지/비디오 압축, 영상처리