적응적 재보정을 통한 표 데이터에 특화된 TabNet 최적화

강민준, 임주완, 이재구*
국민대학교 소프트웨어학부

*jaekoo@kookmin.ac.kr

Optimization of TabNet for Tabular Data through Adaptive Recalibration

Minjun Kang, Juwan Lim, Jaekoo Lee*
College of Computer Science, Kookmin University

요 약

심층학습은 비정형 데이터(사진, 문자, 오디오 등)에 주목할 만한 성능을 이뤘으나, 표와 같은 정형 데이터에는 여전히 결정 트리(decision tree) 등 전통적 기계학습 방법을 활용하는 것에 머물러 있다. 기존 심층학습의 신경망 설계는 정형 데이터에 대한 귀납적 편향이 결여되어 초평면 형태의 결정 경계 근사가 어렵고, 과적합 문제가 있기 때문이다. 하지만, TabNet은 정형 데이터에 효과적으로 작동하도록 결정 트리를 모방하여 정형 데이터에 대해 우수한 성능을 달성하였다. 우리는 TabNet이 단계별 출력 특징을 집계하는 과정에서 각 단계의 출력 특징을 적응적으로 재보정하는 모듈을 제안하였다. 실험 결과 대표적 정형 데이터인 Forest Cover Type과 Adult Census Income 데이터 집합에서 기존 모델과 동등한 성능 달성까지 더 빠른 수렴을 이끌어내며 성공적으로 TabNet의 최적화를 모색하였다.

Ⅰ. 서 론

심층학습은 최근 사진, 문자, 오디오 등 여러 비정형 데이터에서 뛰어난 성능을 보인다[1][2][3]. 하지만, 정형 데이터에는 심층학습보다는 전통적 기계학습 방법을 앙상블 하여 활용하는 것에 머물러 있다. 정형 데이터는 비정형 데이터에 비해 feature 수가 적으며, 일반적으로 초평면 형태의 결정 경계를 가진다[4]. 전통적 기계학습 방법인 결정 트리 모델은 해당 결정 평면을 효과적으로 근사할 수 있는 반면, 기존 심층 신경망 모델은 정형 데이터에 적용하기엔 매개변수가 과하게 많은 경향이 있기 때문이다[4]. 따라서, 기존 심층학습 방법론은 정형 데이터에 과적합이 자주 발생하는 어려움이 있어 활용하기 어렵다. 그러나, 최근 많은 데이터들이 고차원의 정형 데이터로 생산됨에 따라 전문가의 강도 높은 전처리를 요구하는 전통적 기계학습 방법론은 한계점을 갖는다. 따라서, 전통적 기계학습만 쓰이는 정형데이터에 한계를 보완하고자, 데이터 전처리 및 특징 공학 과정이 필요 없는 심층학습 방법론의 필요성이 증가하였고 정형 데이터에 뛰어난 성능을 발휘하는 TabNet[4]이 등장하였다.

TabNet[4]은 결정 트리를 모방하여 설계한 모델로서 정형 데이터에서 부분적으로 SOTA(State-Of-The-Art) 성능을 보이는 심층학습 모델이다. TabNet[4]은 단계별 추출한 출력 특징을 종합(aggregation) 하여 최종 출력을 도출한다. 우리는 출력 특징을 종합하는 과정에서 단계별 출력 특징의 재보정을 유도하여 TabNet[4]을 최적화하는 방법을 제안한다.

Ⅱ. 본론

TabNet[4]은 형상 추출을 위한 학습 가능한 마스크를 생성하는 Attentive Transformer 모듈과 선택한 feature를 추상화하는 Feature Transformer 모듈을 [그림 1]과 같이 순차적으로 배열하여 각 단계(step)를

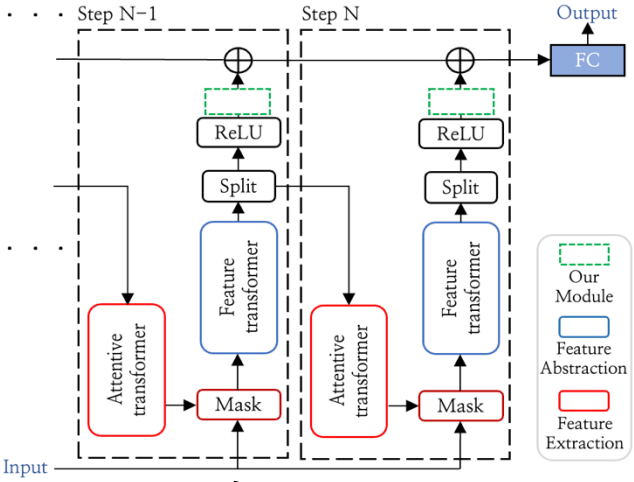

그림 1. TabNet[4] 구조

구성한다. 이러한 모듈들을 통해, 각 단계에서 도출한 출력 특징을 합 연산으로 종합하여 결정 경계의 비율을 결정하는 선형 결합을 유도한다.

제안한 모델 구조는 [그림 2]와 같으며 기존 모델 구조에서 단계별 출력 특징 종합 연산 전에 GAP(Global Average Pooling) 층, sigmoid 층과 element-wise product 연산을 차례로 추가하였다. GAP 층에서는 단계의 출력 특징을 평균 내어 다운 샘플링(down sampling)을 수행한다. 이후, GAP 층을 거쳐 스칼라로 변환한 출력 특징을 sigmoid 층을 통해 0에서 1의 값으로 조정한다. 기존 단계의 출력 특징과 element-wise product 연산을 통해 적응적으로 재조정한다. 이를 통해 각 단계에서 주요도가 낮은 feature들의 영향력을 줄이고 주요한 feature에 더 집중하도록 유도하였다.

Ⅲ. 실험

학습에 사용한 데이터 집합은 대표적 정형 데이터인 Forest Cover Type과 Adult Census Income이다.

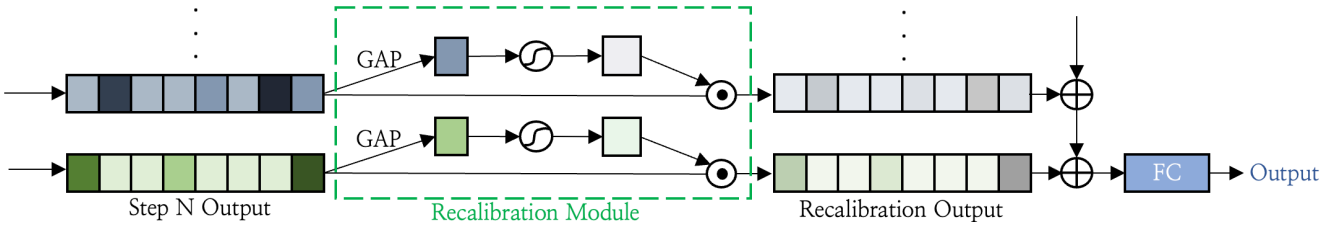

그림 2. 제안한 적응적 재보정 모듈 구조

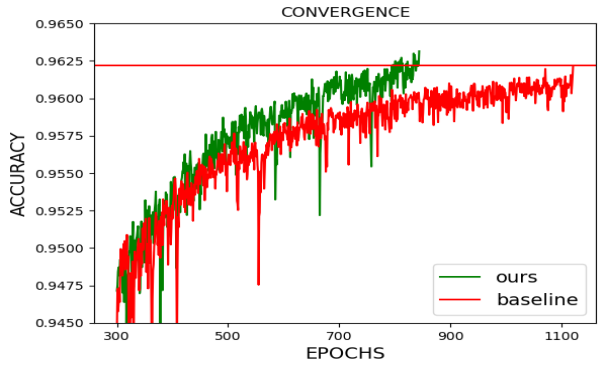

그림 3.(a) Forest Cover Type 학습 수렴 추이

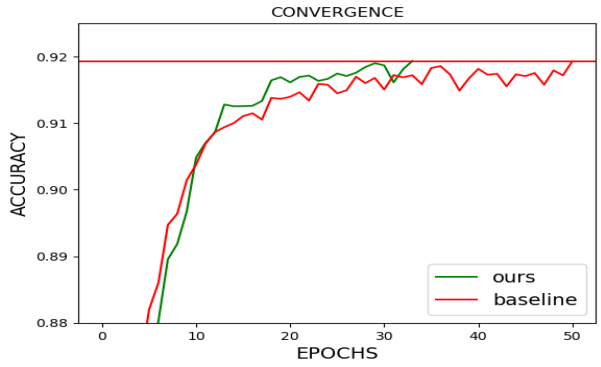

그림 3.(b) Adult Census Income 학습 수렴 추이

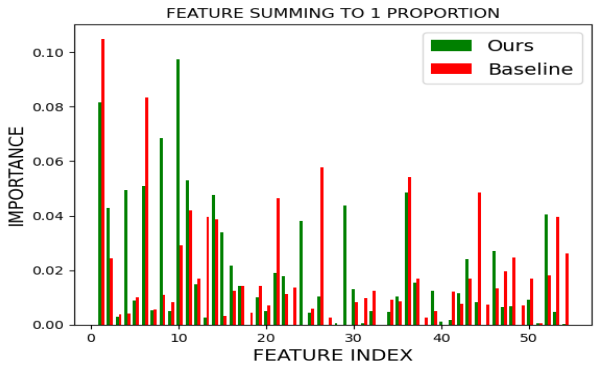

그림 4.(a) Forest Cover Type Global Feature Proportion

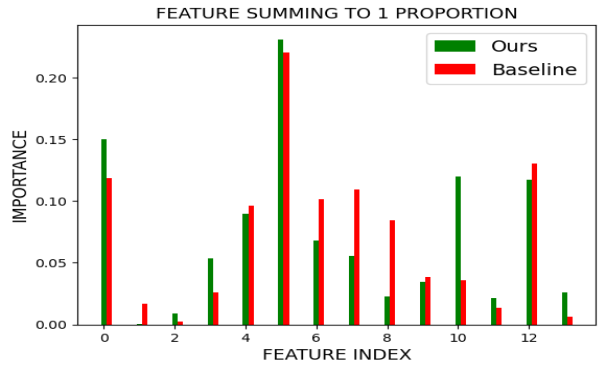

그림 4.(b) Adult Census Income Global Feature Proportion

실험 결과 [그림 3]과 (표 1)에서 볼 수 있듯이, 제안 방법을 통해 두 데이터 집합에서 더 적은 학습으로 기존 모델과 동등한 분류 성능을 달성한 것을 확인할 수 있다. 한편, [그림 4]은 데이터 집합에서 각 모델이 반영한 feature 비율을 나타낸다. 기존 TabNet[4]보다 제안한 방법이 주요하지 않은 feature를 희소하게 반영하며 주요한 feature에 집중한 것을 확인할 수 있다.

IV. 결론

우리는 기존 TabNet[4]의 단계별 출력 종합 과정의 개선점을 인지하였다. 따라서, 기존 모델의 최종 종합 과정 전 각 단계의 출력 특징을 적응적으로 재보정하는 모델을 제안하였다. 그 결과, Forest Cover Type과 Adult Census Income 데이터 집합에 대하여 모델의 수렴을 24.7%, 34%만큼 최적화하였다. 또한, 제안한 모델이 기존 모델 대비 주요도가 적은 feature를 희소하게 반영하면서 주요한 feature에 더 집중하는 학습 경향을 확인하였다.

표 1. 데이터 집합에 대한 각 모델의 성능 비교

Dataset	Baseline		Ours	
	Accuracy	Epoch	Accuracy	Epoch
Forest Cover Type	0.9634	1121	0.9635	844
Adult Census Income	0.8986	50	0.9036	33

ACKNOWLEDGMENT

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(No.RS.2022-0-00516, 국가통계데이터에 적용 가능한 차등정보보호 개념을 도출하고 통계분석의 유용성을 보장해야 하는 문제 해결)

참 고 문 헌

[1] Kenton, Jacob Devlin Ming-Wei Chang, and Lee Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Proceedings of NAACL-HLT. 2019.

[2] He, Kaiming, et al. "Deep residual learning for image recognition." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[3] van den Oord, Aäron, et al. "WaveNet: A Generative Model for Raw Audio." 9th ISCA Speech Synthesis Workshop. 2016.

[4] Arik, Sercan Ö., and Tomas Pfister. "Tabnet: Attentive interpretable tabular learning." Proceedings of the AAAI

[5] Eyre, Francis H. "Forest cover types." Washington, DC: Society of American Foresters. 1980.

[6] https://archive.ics.uci.edu/ml/machine-learning-databases/adult/