예측 모델 지도학습을 위한 적정 데이터양 평가 방법

최국철*, 허석행
LIG넥스원 무인체계연구소.2팀
gukcheol.choi@lignex1.com, seokhaeng.heo@lignex1.com

A Method for assessing the appropriate dataset size for the supervised learning of a predictive model

GukCheol Choi*, SeokHaeng Heo
*LIGNEX1 Co., Ltd

요 약

지도학습 기반 예측 기술은 다양한 분야에서 널리 사용되고 있다. 학습 시 데이터양은 예측 성능에 큰 영향을 미치는 중요한 요소이며, 일반적으로 데이터가 많을수록 뛰어난 예측 성능을 보인다. 하지만 도메인에 따라 데이터 수집에 소모되는 비용이 큰 경우 충분히 많은 양의 데이터 수집이 어려울 수 있다. 적절한 비용으로 유의미한 연구를 하기 위해선 학습이 충분히 가능할 정도의 데이터만을 수집하는 것이 효과적이다. 따라서 본 연구는 도메인에 관계없이 지도학습 시 필요한 적정 데이터양을 평가하는 방법을 제안한다.

Ⅰ. 서 론

머신러닝 기반 예측 기술은 다양한 분야에서 뛰어나 성능을 보이고 있다. 지도학습은 데이터와 레이블 쌍을 이용한 머신러닝 방법 중 하나이며, 지도학습 기반 예측 모델의 목적은 학습 데이터에 기반하여 새로운 데이터에 대해 적절한 레이블 및 분류를 생성하는 것이다[1].

지도학습을 위해선 적절한 데이터양과 질이 요구된다. 일반적으로 데이터의 양이 많을수록 학습 성능이 뛰어나고, 데이터의 양이 적을수록 과적합이 발생할 확률이 높아져 전반적으로 좋은 성능을 기대하기 어렵다[2]. 학습에 필요한 데이터양은 도메인에 따라 상이하며, 데이터 수집에 소모되는 시간과 돈이 큰 도메인의 경우 효과적인 데이터 수집을 위하여 적절한 데이터 수집 계획을 세우는 것이 중요하다.

학습에 필요한 적정 데이터 개수를 평가하는 방법은 아직 정복되지 않은 연구 분야로 남아있다. Alexandre Bailly[3] 등은 대표적인 머신러닝 성능 평가 지표인 Accuracy, F1 score, Area Under the ROC Curve (AUC)를 이용하여 데이터양에 따른 예측 모델의 성능 비교를 수행하였다. 해당 연구는 딥러닝 모델을 포함한 다양한 머신러닝 알고리즘을 기반으로 실험을 진행했으며, 수집한 학습 데이터를 1,000/10,000/100,000개로 나누어 학습 성능을 비교평가 했다. 실험 결과 대부분의 알고리즘이 모든 성능 평가 지표에서 0.01 ~ 0.04 수준의 근소한 성능 차이를 보였다. 따라서 소수점 둘째 짜리 수준의 예측 성능 향상이 목표가 아니라면 해당 도메인의 문제는 1,000개의 데이터만으로 충분히 학습되었다고 평가할 수 있다. 예측 성능 0.01을 올리기 위하여 데이터 수집 비용을 100배 더 사용하는 것은 합리적인 투자가 아닐 수 있다.

본 연구는 적절한 비용으로 유의미한 연구를 수행하기 위해 필요한 적정 데이터양 평가 방법에 대해서 제안한다. 또한, 학습 데이터양이 딥러닝 모델의 예측 성능에 미치는 영향을 함께 알아본다.

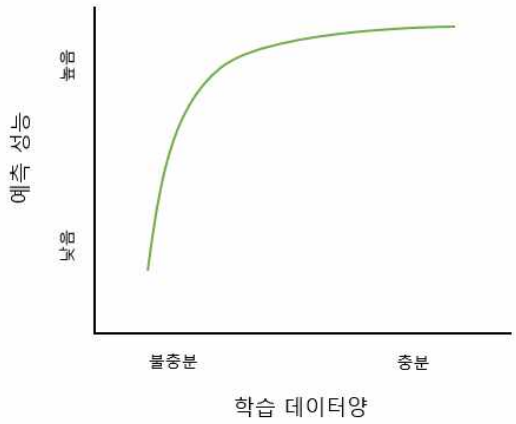

그림 1. 학습 데이터양과 예측 성능 간 상관관계

Ⅱ. 본 론

그림 1은 학습 데이터양과 지도학습 결과 예측 성능 간의 상관관계 그래프이다. 학습 데이터양이 불충분할 경우 데이터양이 증가할수록 예측 성능은 가파르게 향상되며, 충분한 데이터 수량에 근접할수록 예측 성능의 개선은 점차 둔화하는 경향이 있다.

본 연구에서 제안하는 학습 데이터양의 적절성을 평가하는 방법은 수집된 데이터를 나누어 학습 결과의 추이를 관찰하는 것이다. 도메인 및 문제의 복잡도에 따라 학습에 충분한 데이터양은 상이하고 예측이 어렵기 때문에 비용 부담이 크지 않은 수준에서 1차 데이터 수집이 필요하다. 일차적으로 수집된 데이터양이 충분하다면 Alexandre Bailly[3] 등의 연구 결과와 같이 학습 데이터의 일정 수량부터는 예측 성능이 수렴되어 데이터

Train Data Size	Accuracy	Validation Loss
100	71.70	1.186
500	86.78	0.6986
1000	89.50	0.6035
5000	93.18	0.4068
10000	94.70	0.3263
20000	96.38	0.2363
30000	97.22	0.1728
40000	97.90	0.1413
50000	97.96	0.1395
60000	97.88	0.1347

표 1. 학습 데이터양에 따른 Accuracy 및 Validation 데이터셋에 대한 오차

양을 더욱 증가시켜도 성능 개선이 거의 없을 것이다. 만약 데이터양이 부족하다면 전체 데이터를 이용하여 학습할 때까지 예측 성능이 계속 향상할 것이다. 이러한 경우 학습 데이터양을 늘렸을 때 성능 개선의 여지가 있으므로, 추가 데이터 수집 계획을 수립해 볼 수 있다.

본 연구는 학습 데이터양의 적절성 평가 실험을 위하여 MNIST 데이터셋을 사용하였다. MNIST 데이터셋은 손글씨 숫자 분류 학습을 할 수 있는 컴퓨터 비전 분야의 표준 벤치마크이다[4]. 해당 데이터셋은 손글씨 이미지와 정답 쌍 학습 데이터 60,000개와 테스트 데이터 10,000개로 구성되어 있다.

손글씨 이미지 예측을 위한 적절 데이터 개수를 구하기 위해 100/500/1,000/5,000/10,000/20,000/30,000/40,000/50,000/60,000개의 학습 데이터를 시용했을 때의 학습 성능을 비교 평가한다. 테스트 데이터셋은 Validation 데이터셋 5,000개, 테스트 데이터셋 5,000개로 나누어 사용한다.

전반적으로 좋은 성능의 모델을 얻기 위해서는 Validation 데이터셋에 대한 오차가 증가하기 전에 Early stopping을 통해 학습을 멈추는 것이 좋다[2]. Validation 데이터셋에 대한 오차의 증가는 과적합이 발생한 것으로 평가할 수 있기 때문이다. 따라서, 본 연구는 학습 데이터 수에 관계없이 Validation 데이터셋에 대한 오차가 증가할 때 학습을 종료한다. 또한, 공정한 비교 분석을 위하여 딥러닝 모델의 뉴럴 네트워크 구조, 옵티마이저(Optimizer) 등 학습 데이터양 및 학습 횟수를 제외한 모든 요소는 동일한 조건에서 실험한다.

표 1은 학습 데이터양에 따른 손글씨 예측 성능 및 Validation 데이터셋에 대한 오차를 측정한 결과이다. 학습 데이터가 많을수록 Validation 데이터셋에 대한 오차는 줄어드는 것을 관찰할 수 있지만, 학습 데이터양이 늘어날수록 테스트 데이터셋에 대한 정확도는 개선폭이 점차 줄어들다가 60,000개의 학습 데이터를 사용했을 때 오히려 낮아진 것을 관찰할 수 있다. 따라서, 손글씨 예측 모델의 학습을 위해 필요한 적정 데이터양은 예측 정확도가 수렴하다 가장 높은 성능을 보인 50,000개면 충분하다고 평가할 수 있다. 그림 2는 학습 데이터양에 따른 손글씨 예측 성능의 추이를 관찰한 것으로 데이터양이 적을 때는 학습 데이터양의 증가가 큰 성능 개선으로 이어지는 것을 관찰할 수 있으며, 학습 데이터양이 많아질수록 성능 개선이 점차 둔화하는 것을 관찰할 수 있다.

Ⅲ. 결 론

본 연구에서는 지도학습을 이용하여 예측 모델을 학습할 때 필요한 적절 데이터양의 수를 평가하는 방법에 대하여 제안한다. 실험을 위하여

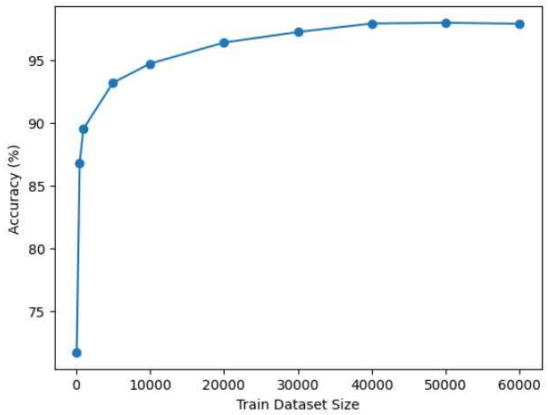

그림 2. 학습 데이터양에 따른 손글씨 예측 성능 추이

MNIST 데이터셋을 이용하였으며, 실험 결과 50,000개 이상의 데이터양은 성능 개선에 크게 도움 되지 않는다는 것을 관찰할 수 있다. 해당 방법은 도메인에 관계없이 적용할 수 있으며, 추후 연구를 통해 수집된 데이터 기반 학습 결과를 이용하여 추가 수집이 필요한 데이터양을 예측하는 연구를 진행할 수 있다.

참 고 문 헌

[1] Abdulraheem, Ajiboye & Abdullah Arshah, Ruzaini & Qin, Hongwu. (2015). Evaluating the Effect of Dataset Size on Predictive Model Using Supervised Learning Technique. International Journal of Software Engineering & Computer Sciences (IJSECS). 1. 75-84. 10.15282/ijsecs.1.2015.6.0006.

[2] Prechelt, L. (2012). Early Stopping – But When?. In: Montavon, G., Orr, G.B., Müller, KR. (eds) Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science, vol 7700. Springer, Berlin, Heidelberg.

[3] Alexandre Bailly, Corentin Blanc, Élie Francis, Thierry Guillotin, Fadi Jamal, Béchara Wakim, Pascal Roy, Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models, Computer Methods and Programs in Biomedicine, Volume 213, 2022, 106504, ISSN 0169-2607.

[4] G. Cohen, S. Afshar, J. Tapson and A. van Schaik, "EMNIST: Extending MNIST to handwritten letters," 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 2017, pp. 2921-2926, doi: 10.1109/IJCNN.2017.7966217.