노이즈 데이터 정제를 통한 분류모델 성능 향상

정운국[1], 강승식[2]
[1]국민대학교 소프트웨어융합대학원 석사과정
[2]국민대학교 인공지능학부 교수
jukyellow@ktnet.co.kr, sskang@kookmin.ac.kr

Enhancing Classification Model Performance through Noise Data Refinement

Unkuk Jeong[1], Seungshik Kang[2]
[1]Graduate School of Software Convergence, Kookmin University
[2]Dept. of Artificial Intelligence, Kookmin University

요 약

자연어 기반의 분류모델을 개발할 때 높은 성능을 획득하기 위해서는 데이터의 품질이 중요한 요소이다. 특히 무역상품 국제 분류체계 HS-CODE에서 상품명을 기반으로 HS코드를 분류할 때, 라벨링된 데이터의 품질에 의해서 분류모델의 성능이 좌우된다. 하지만 현실적으로 확보 가능한 데이터셋에는 데이터 라벨링 오류나 데이터로 활용하기에 특징점이 부족한 데이터들이 다수 존재하기도 한다. 본 연구에서는 분류모델 학습 데이터의 정제 방법론으로, 딥러닝 기반 노이즈 검출 알고리즘을 제안한다. 분류 대상의 특징점이 분류 경계값 주변에 존재한다면 분류하기 모호한 노이즈 데이터일 가능성이 높다고 가정하고, 해당 노이즈 데이터를 검출하는 방법으로 딥러닝 기술을 활용한다. 해당 경계값 노이즈 검출 알고리즘으로 데이터를 정제한 뒤 학습모델의 성능비교 결과, 기존 대비 우수한 분류 정확도를 기록하였다.

1. 서론

수출입 무역,물류 업무에서 HS코드 분류는 상품의 과세를 결정하는 관세행정상 중요한 요소이다. 세계관세기구(WCO)에서는 국제공통코드 HS코드 6단위를 5년주기로 개정하고 있고, 수출입 무역 상품의 HS코드 분류는 전무가(관세사)의 업무 영역이기도하다. 컴퓨터 사이언스와 인공지능 기술의 발전 이래로 무역 상품의 HS코드 자동분류를 위한 노력은 지속되어 왔다. 최근 국제 전자상거래의 활성화, 국제 물류 풀필먼트 서비스의 활성화를 계기로 디지털 전환(DX)과 자동화를 위한 HS코드 자동분류의 필요성이 대두되고 있고, [1,2]와 같이 관련 연구가 활발히 진행되고 있다. 이를 위해 AI 기술을 활용한 상품명-HS코드 자동분류 시스템 개발이 늘어나고 있고, HS코드 자동분류의 정확도를 담보하기 위해서는 라벨링된 데이터의 확보가 핵심 과제이다. 일반적으로 수출입 무역기업 또는 관세사에 의해서 작성된 수출입신고서의 데이터를 활용해 상품명-HS코드를 확보하여 AI 모델 학습을 위한 데이터를 구축한다. 하지만 수출입 신고된 데이터라고 할지라도 라벨링 오류 데이터가 존재하거나(HS코드 기재오류), 상품명 하나의 특징만으로는 HS코드 분류가 모호한(분류를 위한 특정점을 찾을 수 없는) 데이터가 다수 포함되어 있어서, 분류모델의 정확도를 담보할 수 없다. 본 연구에서는 딥러닝 기반 경계값 노이즈 검출 알고리즘을 제안하고, HS코드 분류모델의 정확도를 정제 전후 데이터를 사용해 비교 검증하였다. HS코드 분류 연구에 있어서 [3,4]와 같은 선행연구에서는 데이터를 효과적으로 활용하기 위한 모델구조와 방대한 데이터를 기반으로 딥러닝 모델을 활용하는 사례가 제시되고 있다. 하지만 앞서 언급했듯이 확보한 데이터의 품질을 보증할 수 없기 때문에 데이터 정제가 필수적이며, 전문가에 의해서 작성된 데이터(수출신고서)라고 할지라도 노이즈 데이터가 존재하고 노이즈 정제를 통해 HS분류모델의 정확도가 개선됨을 확인하고자 한다.

2. 노이즈 데이터 검출 방안

선행연구에서는 노이즈 데이터에 강건하게 학습하는 모델을 구성하기 위해 다양한 연구가 진행되고 있다[6,7,8]. 본 논문에서는 경계값 근처에 존재하는 데이터 데이터를 검출하고 정제하는데 포커스를 맞췄다. (그림 1)과 같이 분류대상에 있어 그룹을 구분

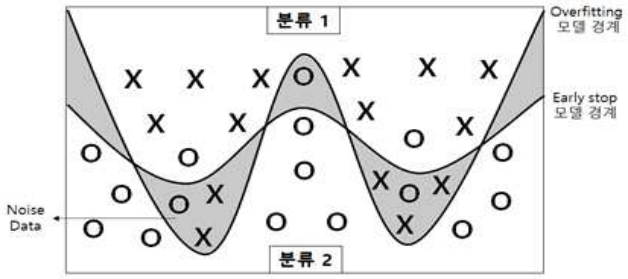
한다는 것은, 대상의 특징을 기준으로 구분하는 경계선을 긋는 작업과 같다고 할 수 있다. 하지만 이 기준선 주변에는 일반화를 위한 기준선에 부합하지 않는 데이터가 존재할 수 있다. 모든 데이터를 만족하는 기준선을 긋는다면 오버피팅 문제가 발생할 수 있기 때문에, 분류가 잘못되는 데이터가 발생할지라도 일반화를 위한 기준선을 잡는 것이 더 효과적이라고 할 수 있다.

(그림 1) 분류 경계와 노이즈 데이터

이 경계선 주변의 데이터를 관찰해보면, 분류관점에서 오답 데이터가 존재할 가능성이 높다. 당연하게도 오버피팅 시키지 않았기 때문에 일반적으로 학습 데이터에서 분류가 더 잘 되는 (정답과 오답의 차이가 줄어드는) 방향으로 모델 학습을 진행하기 때문이다. 학습관점에서 보면 해당 데이터들은 학습의 오차를 발생시켜 학습의 성능을 저해하는 노이즈 데이터라고 정의할 수 있다.

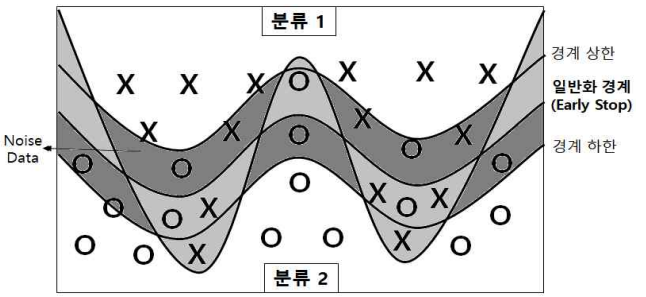
그림 1의 사례에서는 O/X 데이터는 대상을 구분하기에 명확한 특징을 가지고 있으므로, 엄밀히 말하면 노이즈 데이터라고 할 수 없지만, 현실 세계에서는 노이즈 데이터라고 할 수 있는 사례가 많다. 이진분류 문제에서 가장 흔한 예시로 들 수 있는 강아지와 고양이 분류문제를 예로 들 수 있다. 일반적으로 강아지와 고양이의 특징은 명확하지만, 강아지와 고양이의 특징을 모두 가지는 품종이 있다면 사람은 구분하기 힘들 것이다. 여기서 노이즈 데이터라 함은 사물을 구분하기에 특징이 모호한 또는 특징점이 부족한 데이터라고 할 수 있다.

그림 1의 대상을 분류하기 위한 경계선 주변의 노이즈 데이터를 검출하기 위해 아래 방법을 제안한다. 해당 노이즈 검출 방법은 2단계 프로세스로 구성되며, 1단계에서는 일반화 경계와 오버피팅 경계 사이에 존재하는 데이터를 검출 한다. (그림 2)와 같이 일반화 경계는 early-stop을 적용한 모델로 찾고, 과적합 경계는 오버피팅시킨 모델로 해당 경계선을 찾는다. 이렇게 하면 두 분류 모델의 분류가 서로 다르다고 평가한 대상이 경계선 주변에 있는 데이터로 간주할 수 있다. 왜냐하면 분류가 확실한 대상이라면 기하학적인 관점에서 분류 경계선 주변에서 멀리 떨어진 데이터일 것이고, 그렇다면 두 모델이 같은 부류로 평가했을 가능성이 높다.



(그림 2) 노이즈 검출 1단계

노이즈 검출 2단계에서 (그림 3)과 같이, 분류 경계선 근처의 데이터를 검출한다. 분류 경계선에서 멀어진 데이터일수록 분류 기준이 확실한 데이터라고 할 수 있고, 분류 경계선 근처의 데이터일수록 구분이 모호한 데이터일 확률이 높다. 해당 데이터를 검출하기 위해서는 일반화 모델(early-stop)의 학습시 확인된 Loss(추론값과 정답의 오차) 값의 평균값을 기준을 사용한다. 분류가 확실한 대상이라면 Loss 값이 작다고 할 수 있고, 분류가 불확실한 대상일수록 Loss 값이 크다고 할 수 있다. 역시나 기하학점 관점에서는 Loss 값이 큰 대상일수록 구분이 모호한 영역인 경계값 주변에 위치할 가능성이 높고, Loss값이 작은 확실한 분류 대상인 경우 경계값에서 먼 데이터로 간주된다.



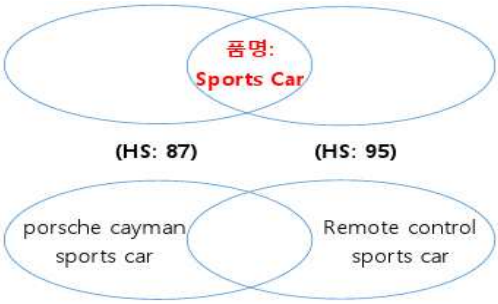
(그림 3) 노이즈 검출 2단계

2단계 노이즈 검출 알고리즘을 통해, 노이즈 데이터를 검출할 수 있다면, 노이즈 데이터를 학습에서 제외하는 방향으로 데이터를 정제한 뒤, 모델 학습이 가능하다. 학습 데이터 외에도 테스트 데이터도 동일하게 노이즈를 제거하고 모델의 성능평가를 수행하였다.

3. 자연어분류 태스크 적용(HS분류)

딥러닝 기법을 활용한 노이즈 검출 알고리즘을 자연어 분류 태스크에 적용하였다. HS코드 분류는 수출입 무역신고 데이터에서 상품명을 기준으로 HS코드를 분류하는 문제이다. 실제 업무에서는 상품명(거래품명) 외에, 관세품명, 모델명, 규격, 성분 등 다양한 특징정보를 활용해 관세사가 HS코드 분류업무를 수행한다. 그러나 실제 유통되는 데이터에는 전체 데이터를 확보하기는 쉽지 않다. 가장 흔하게는 상품명(거래품명)이 무역, 물류 업무 공통으로 유통되는 데이터이고, 이 거래품명을 기준으로 검증작업을 수행하였다.

품명 데이터를 HS코드 분류함에 있어서, 품명에 HS코드를 분류하기에 특징이 부족한 데이터인 경우가 많다. (그림 4)의 예와 같이 단순히 Sports Car라고 하면 해당 HS 분류가 HS87(일반 자동차)인지, HS95(장난감)의 데이터인지 특징이 부족하다고 할 수 있다. 품명이 Sports Car인 경우, 그림 4와 같이 좀 더 구체적인 정보를 제공할 때, 구분할 수 있는 특징점이 검출되고, HS코드를 분류하기가 쉬워진다.



(그림 4) 품명-HS코드 분류 문제

품명-HS코드 분류 데이터에도 노이즈(분류하기에 특징이 모호한) 데이터가 존재하고, 제안하는 노이즈 검출 알고리즘을 활용해, 노이즈 데이터를 정제한 뒤, 데이터 정제 전후의 성능평가 결과를 비교하였다. HS코드는 HS 6단위까지는 국제 공통코드이며, HS 8,9,10,12 단위 등 각 국가마다 세율의 기준이 되는 HS코드 단위가 세분화된다. 본 연구에서는 한국 HS 10단위 중, HS 2단위를 학습 및 추론에 사용하였다. HS 2단위의 카테고리는 총 96가지이며, 01에서부터 97까지 사용된다.(77은 유보), 아래의 학습 결과는 품명 Text에 대해서 라벨링된 HS 2단위 96개의 카테고리를 분류하도록 모델을 학습하였다.

<표 1> 품명-HS코드 분류모델 성능평가

모델	Accuracy	Precision	Recall	전체 데이터 건수
노이즈 정제 전 학습모델	73%	72%	72%	170 만건
노이즈 정제 후 학습모델	89%	90%	90%	117 만건

<표 1>의 결과는 전체 데이터 170만건의 데이터에서, 노이즈 데이터 정제(제거)후 결과를 나타낸다. 전체 데이터중 53만건의 데이터가 노이즈 데이터로 검출되었고, 테스트셋은 전체 데이터의 10%를 사용한 결과이다. 노이즈 데이터 검출을 통해 31%(53만건)의 데이터가 제외되었고, 해당 데이터를 제외 후 분류모델의 정확도가 16% 개선되는 결과를 확인하였다. 추출된 노이즈 데이터를 샘플링한 결과는 <표 2>와 같다. 아래의 데이터를 보면 해당 HS2 단위를 분류하기에 text에 그 특징점이 부족하다고 할 수 있다.

<표 2> 품명-HS코드 분류 노이즈 데이터 샘플

HS2단위	구분	노이즈 품명	정상/오류 유형
HS22 (주류)	정상	coca cola white peach	(확실한 정답)
HS95 (장난감)	정상	video eletronic game parts	(확실한 정답)
HS84 (기계류)	정상	floor polishing machine	(확실한 정답)
HS22 (주류)	노이즈	bitburger drive	오류 CASE 1 (불확실한 정답)
HS87 (일반차량)	노이즈	sports car	오류 CASE 2 (불확실한 노이즈)
HS01 (산 동물)	노이즈	mouse for lab	오류 CASE 3 (확실한 노이즈)

학습시 사용한 딥러닝 모델은 CNN을 자연어 문장에 적용한 선행연구 [5]을 참고하여 적용하였다. HS코드 분류에 사용되는 품명(거래품명)은 명사 위주의 짧은 단어로 구성되는 경우가 많은데, 짧은 자연어 문장에는 CNN을 적용하는 것이 LSTM 계열보다 성능이 우수했다.

4. 이미지 태스크 적용(Mnist손글씨)

제안한 노이즈 검출 알고리즘이 이미지 분류 태스크에 잘 동작하는지를 확인하기 위해, Minst 손글씨 데이터에도 적용해 보았다. 이미지 분류 태스크에도 분류 특징점이 부족하거나, 라벨링 오류 데이터가

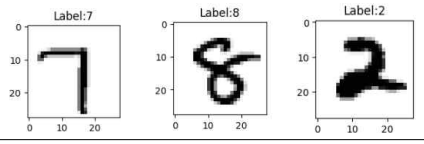
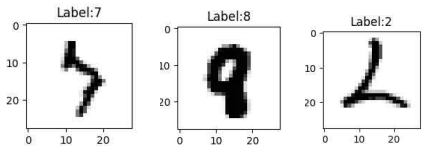
존재할 수 있다. 따라서 데이터 정제는 중요한 요소이며, Mnist 데이터셋에도 사람이 구분하기에 모호한 데이터들이 다수 포함되어 있다. Mnist 손글씨 데이터는 학습데이터 6만건, 테스트 데이터 1만건으로 구성되어 있다. 제안한 노이즈검출 알고리즘을 통해 학습 및 테스트 데이터에서 노이즈 데이터를 정제(제외)후 모델을 학습하고 평가한 결과는 <표 3>과 같다.

<표 3> Mnist 손글씨 데이터 분류모델 성능평가

모델	Accuracy	학습 건수	테스트 건수
노이즈 정제 전 학습 모델	99.16%	60,000	10,000
노이즈 정제 후 학습 모델	99.70%	59,706	99,917

<표 3>과 같이 노이즈 검출 알고리즘을 통해, 학습데이터에서 294건의 데이터 검출되었고, 테스트 데이터에서 83건의 데이터가 검출되었다. 표 4에서 검출된 노이즈 데이터를 제거 후 학습 및 평가를 수행한 결과이다. Mnist 손글씨 데이터에서 SOTA accuracy 99.87에 근접한 결과를 확인하였다. 아래 <표 4>는 테스트 데이터에서 검출한 노이즈 데이터 샘플 중 일부이다.

<표 4> Mnist 데이터 정상 및 노이즈 데이터 샘플

구분	Mnist 손글씨 테스트 데이터셋 샘플
정상	
노이즈	

5. 결론

자연어, 이미지 분류모델 개발에 있어서, 데이터의 노이즈 정제를 통해 성능이 개선 가능함을 확인하였다. 데이터의 노이즈 정제 방법으로 분류 경계값 주변의 노이즈 검출 방법을 제안하였고, 해당 알고리즘은 자연어, 이미지 분류 태스크에 종속적이지 않고, 범용적으로 활용 가능함을 검증하였다. 본 논문에서 제안한 방법으로 개발된 분류모델이 제시된 데이터셋 내에서 보다 우수한 성능으로 동작함을 확인하였지만, 실제 real 데이터에서도 성능이 더 우수함을 확인해야하는 과정이 남아있다. 또한 real 데이터에는 정제(제외)된 데이터가 존재하고, 해당 노이즈 데이터가 입력되었을 때 성능이 어떻게 변화하는지에 대한 검토도 필요하다. 향후에는 본 논문에서 실험한 태스크 외에 오디오 노이즈 제거를 통한 성능 개선 문제와, 좀 더 나아가 LLM(초거대언어모델)에서 환각답변 문제 해결을 위한 데이터 정제방법으로 연구를 확장할 계획이다.

참고문헌

[1] Lee, Eunji, et al. "Classification of goods using text descriptions with sentences retrieval," arXiv preprint arXiv:2111.01663, 2021.

[2] Chen, Xi, Stefano Bromuri, and Marko Van Eekelen. "Neural machine translation for harmonized system codes prediction." Proceedings of the 2021 6th International Conference on Machine Learning Technologies. 2021.

[3] Ding, Liya, ZhenZhen Fan, and DongLiang Chen, "Auto-categorization of HS code using background net approach," Procedia Computer Science 60: 1462-1471, 2015.

[4] Luppes, Jeffrey, Arjen P. de Vries, and Faegheh Hasibi. "Classifying short text for the harmonized system with convolutional neural networks." Radboud University, 2019.

[5] Kim, Yoon. "Convolutional neural networks for sentence classification. arXiv 2014." arXiv preprint arXiv:1408.5882, 2019.

[6] Rolnick, David, et al. "Deep learning is robust to massive label noise." arXiv preprint arXiv:1705.10694, 2017.

[7] Song, Hwanjun, et al. "Learning from noisy labels with deep neural networks: A survey," IEEE transactions on neural networks and learning systems, 2022.

[8] Smart, Brandon, and Gustavo Carneiro. "Bootstrapping the relationship between images and their clean and noisy labels," Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023.