의사 레이블링을 통한 레이블이 없는 데이터 보완 연구

유민희, 유헌창
고려대학교 컴퓨터정보통신대학원

fbwbsgml@korea.ac.kr, yuhc@korea.ac.kr

Research on supplementing unlabeled data through pseudo-labeling.

Min-Hee Yoo, Heon-Chang Yu
Graduate School of Computer & Information Technology, Korea University

요 약

레이블링 작업은 데이터 분석 시 필요한 사전 작업중 하나이다. 모든 데이터들에 대해 레이블링 작업은 시간/인적 자원을 필요로 하기에, 해당 작업을 보완할 방법이 존재한다면 요구되는 리소스를 줄여 효율성을 크게 향상시킬 수 있다. 본 논문에서는 통신회사에서 적재된 데이터 셋에 대하여 레이블이 없는 데이터(Unlabeled-data)에 대해 의사 레이블링(Pseudo-labeling), SMOTE 를 통한 데이터 증강을 활용하여 기존에 활용되지 못한 데이터를 추가하여 모델에 학습시킨다. 실험을 통해 의사 레이블을 통한 모델 학습 방법이 기존 도메인 지식의 레이블 방법보다 효율적이고 성능이 우수함을 확인하였다.

1. 서론

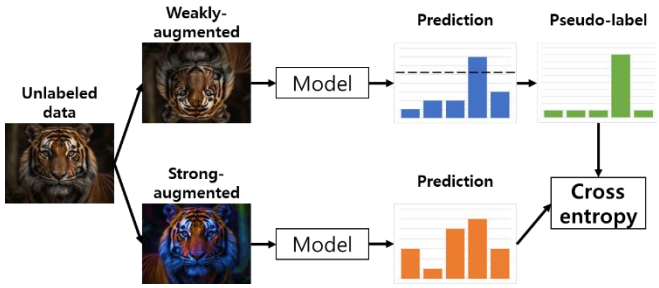
수집된 데이터에 대한 레이블링 작업은 데이터 분석에 있어서 필요한 사전단계 중 하나이다. 이를 위해선 많은 시간과 인적 자원이 필요하기에 다양한 분야에서 데이터 레이블링을 위한 인력 및 시간을 투자하고 있다. 이미지 분류의 경우 레이블이 없는 데이터를 활용하기 위해 레이블이 존재하는 데이터(labeled data)와 레이블이 없는 데이터를 모두 활용하는 준지도 학습(Semi-supervised learning) 방법이 활용되고 있으며, 관련 연구 역시 활발하다. 이를 준용하여 이미지가 아닌 데이터들을 대상으로 준지도 학습을 적용, 레이블이 없는 데이터에 대해 의사 레이블링을 진행하여 도메인 지식에 의존하여 분류한 방식에 데이터 관점의 분류기법을 추가하여 학습 데이터 확보 및 모델에 학습에 적용하여 성능을 향상시킬 뿐만 아니라, 모델 학습에 필요한 데이터 레이블링에 필요한 시간/인적 자원을 줄여 효율화 할 수 있는 방안에 대해 제안한다.

2. 관련연구

Kihyuk Sohn 이 제안한 Fix-match[1]는 레이블이 없는 데이터에 대한 의사 레이블링, 일관성 규제(consistency regularization) 및 이를 적절히 활용한 SSL(Semi Supervised Learning) 방법이다.

의사 레이블링 방법은, 우선 라벨이 있는 데이터로 모델을 학습시킨 후 라벨이 없는 데이터를 대상으로 예측을 진행한다. 이때 임계 값(Threshold)를 기준으로 데이터에 대해 의사 레이블링 작업을 진행한 후 이를 라벨이 있는 데이터 셋에 포함시켜 라벨이 없는 데이터에 대해 의사 레이블 부여가 가능하다.

이에 더해 일반화 작업을 위해 이미지에 두가지 증강 방법인 약한 증강(weak augmentation)/강한 증강(strong augmentation)을 적용하여, 약한 증강 데이터를 통해 의사 레이블을 부여하고, 강한 증강과의 cross entropy 를 최소화하는 방안으로 학습을 진행한다.



(그림 1) Fix-match

이와 관련하여 국내에선 이미지 데이터셋을 사용하여 클래스 불균형 상황에서의 의사 레이블연구[2] 역시 제안되고 있다.

3. 레이블이 없는 데이터의 모델 학습 활용 방안 연구

3.1 사용 데이터

학습에 사용된 Data set 은 통신 관련 회사에서 수집된 데이터로, <표 1>과 같은 분포를 이루고 있다.

구분 A 와 B 는 고객에게 제공 가능, 불가능을 의미하고 있다.

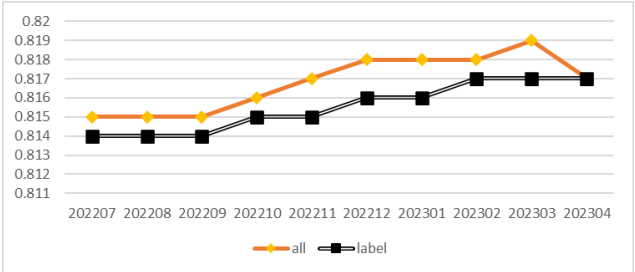
<표 1> 데이터 분포

구분	라벨	분포
A	Y	81.3%
B	N	2.9%
	P	1.4%
	D	0.3%
	etc	14.1%

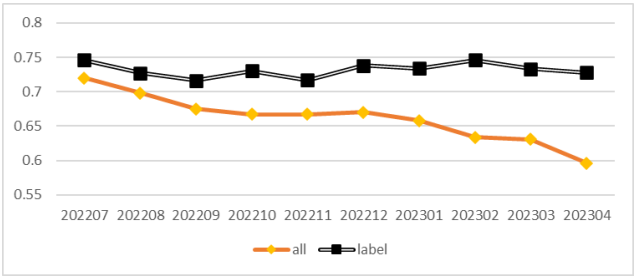
전체 데이터 중 14.1%가 레이블이 없는 데이터인 etc 로 분류되어 있으며, 구분으로 보았을 경우 B 의 비율이 매우 적게 분포되어 있는 data Set 으로, 본 실험의 목표는 구분 A/B 를 학습하는 binary classification 모델을 만드는 것이다. 사용한 Feature 개수는 116 개로 일자 별로 '건물의 종류', '건물내 통신에 활용되는 장비 존재 여부'와 같은 카테고리 형 변수 71 개와 '건물과 장비 간의 거리', '건물 밀집도'와 같은 연속형 변수 45 개로 구성되어 있다. 현업에서의 분석환경을 고려하여 불균형 데이터셋 연구에 제안되어지는 Boosting 계열 알고리즘인 LightGBM[3][4]을 사용하였다.

3.2 레이블 모델 학습

Etc 에는, 부정확한 데이터가 다수 포함되어 있으므로 기존엔 도메인 지식에 기반하여 라벨 N, P, D 만 활용하여 학습을 진행하였다. A/B 를 라벨로 학습한 모델을 'ALL'로 표기하였으며, B 중 라벨 N, P, D 만 활용하여 학습한 모델을 'label'로 표기하였다. 위 조건에 맞는 A와 B를 구분하는 LightGBM 모델을 학습, 현업에선 데이터가 매월 추가되기에 데이터 누적에 따른 추세를 보기 위해 월 별 추가 학습을 진행하였다.



(그림 2) all-label accuracy

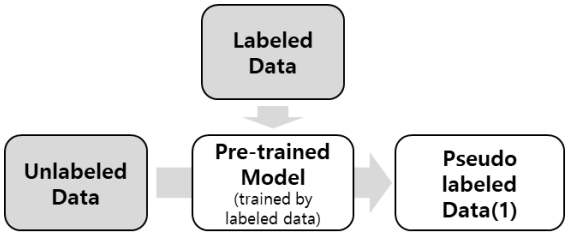


(그림 3) 구분 B 에 대한 all-label precision

Accuracy 를 비교하는 (그림 2)을 보면 학습 데이터는 구분 A 에 전체 데이터의 81%가 포함되어 있는 매우 불균형 하기에, accuracy 는 유의미한 차이를 보기 어렵다. 다만 precision 에서는 차이를 보이는데 (그림 3)를 보면, 매월 학습이 진행됨에 따라 모든 데이터를 넣은 'all'의 경우 최종월엔 0.13 정도의 차이를 보이는 것이 확인된다. 현업에서는 모형이 B 로 예측한 데이터에 대해 별도 관리와 같은 활동이 이어지기에 본 연구에서는 accuracy 보단 모형이 B 라 판단하였을 때 실제로 B 인 경우에 대한 정확도 지표인 Precision 을 중요 지표로 선정하였다.

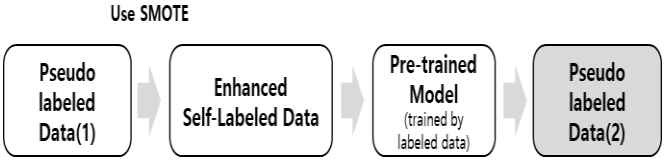
3.3 의사 레이블 모델 학습 설계

본 연구에서 이전 연구에서는 활용되지 못한 레이블 etc 를 의사 레이블링을 통해 모델학습에 포함하고자 한다. 학습을 진행하기 위해 etc 로 분류된 데이터들을 대상으로 레이블이 존재한 데이터로 학습된 pre-trained model 을 통해 pseudo-labeled Data(그림 4)을 생성한다. 여기서 레이블이 존재하는 데이터는, 11 개월 치 데이터를 사용하였으며, 레이블이 없는 데이터는 그 이후 10 개월치 데이터를 사용하였다.



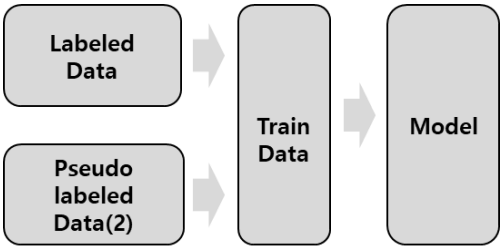
(그림 4) Pseudo-labeled Data

Augmented data 를 생성하기 위해 Pseudo-labeled Data 에 SMOTE[2]를 적용하였다(그림 5).



(그림 5) SMOTE 적용

이를 통해 Pseudo-labeled Data 를 생성하고, 기존 Labeled Data 에 추가하여 모델 학습을 위한 Train Data 를 생성, 모델학습을 진행하였다(그림 6).

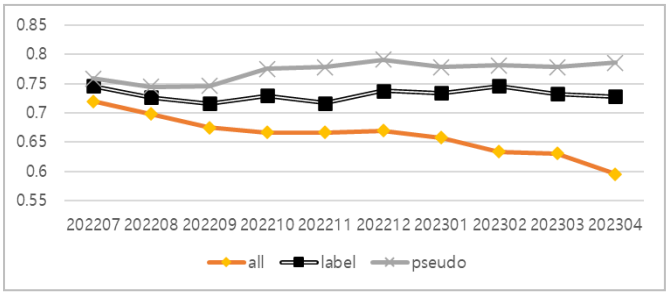


(그림 6) 최종 모델학습

매월 학습을 진행하며, 최초 학습시에는 레이블이 있는 데이터를 통해 Pre-trained Model 을 생성하며, 그 다음 학습시에는 전월 최종모델이 Pre-trained Model 을 대체하여 의사 레이블링을 누적하여 진행하도록 설계하였다.

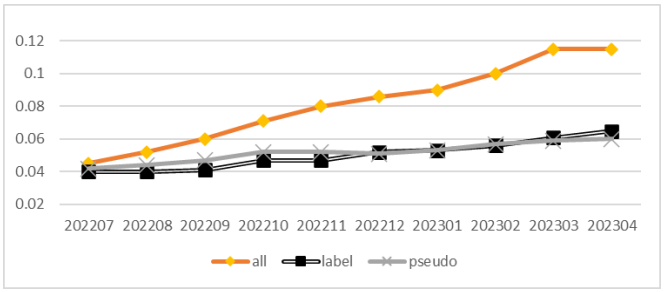
4. 실험 결과 및 분석

본 연구는 도메인 지식으로 분류된 라벨을 사용한 모델과 의사 레이블이 부여된 데이터를 추가한 모델 간의 성능을 비교하였다. 의사 레이블이 부여된 데이터를 활용한 모델을 'pseudo'로 표기하였으며 Pre-trained Model 생성을 위한 레이블이 존재하는 데이터는 최초 11 개월치 데이터를 사용, 그 이후 10 개월 데이터를 의사 레이블이 부여된 데이터로 추가하여 학습하는 형태로 설계 및 연구를 진행하였다.



(그림 7) all-label-pseudo precision

의사 레이블이 부여된 데이터를 활용하는 경우, precision(그림 7)에서 최종 월에 0.058 상승한 것을 확인하였다.



(그림 8) 구분 B 에 대한 all-label-pseudo recall

또한 Recall(그림 8)값은 차이가 없음을 확인하였다. 이를 통해 레이블이 없는 데이터에 대해 의사 레이블링을 통해 기존 레이블이 존재하는 데이터로 학습한 모델보다 B 로 예측하는 대상 수는 비슷하나, 중요 정확도 지표인 precision 이 향상된 모델 학습이 가능함을 확인하였다. 이를 통해 본 학습 방법을 현업에 적용시에 새로 적재되는 데이터에 대한 레이블링 작업을 줄일 뿐만 아니라, 기존 모델보다 정확한 대상을 예측할 수 있음을 확인하였다.

5. 결론

본 논문에서는 의사 레이블링을 통해 기존 도메인 지식에선 활용되지 못한 레이블이 없는 데이터에 대해 추가 분류 및 모델 학습을 진행하였다. 결과적으로 기존 레이블이 존재하는 데이터를 통해 학습된 모델보다 precision 은 향상하였으며, recall 은 동일 수준이 도출되었다. 이를 통해 초기 분류된 라벨이 있다면 추후 데이터들은 별도 인적자원이 투입된 작업 없이 기존 방법보다 우수하고 효율적으로 모델 학습이 가능함을 확인하였다.

향후 과제로는 본 모델을 운영하면서 정확도를 모니터링하고, 시간이 누적됨에 따라 레이블이 존재하는 데이터보다 레이블이 없는 데이터가 많아지는 상황 속에서 변동성에 대해 추가 연구할 계획이다.

참고문헌

[1] SOHN, Kihyuk, et al. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems, 2020, 33: 596-608.

[2] 배제언; 배성호. 클래스 불균형 준 지도 학습에서 의사 레이블의 통계적 특성을 이용한 불균형 손실. 한국정보과학회 학술발표논문집, 2022, 1625-1627.

[3] FERNÁNDEZ, Alberto, et al. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research, 2018, 61: 863-905.

[4] KE, Guolin, et al. Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 2017, 30.

[5] KONG, Jiawen, et al. Improving imbalanced classification by anomaly detection. In: International Conference on Parallel Problem Solving from Nature. Cham: Springer International Publishing, 2020. p. 512-523.