집단 공정성을 고려하는 자기 지도 대조 학습

정채윤[2,1 O], 한성원[2,1], 김선동[1], 차미영[1,2]
기초과학연구원 데이터사이언스그룹[1], 한국과학기술원 전산학부[2]
lily9991@kaist.ac.kr, lion4151@kaist.ac.kr, sundong@ibs.re.kr, mcha@ibs.re.kr

Group-Wisely Fair Self-Supervised Contrastive Learning

Chaeyoon Jeong[2,1 O], Sungwon Han[2,1], Sundong Kim[1], Meeyoung Cha[1,2]
Data Science Group, Institute for Basic Science[1], School of Computing, KAIST[2]

요 약

이 연구에서는 자기 지도 대조 학습을 기반으로 편향을 해소하고 집단 공정성을 만족하는 데이터 표현을 생성하는 방법을 제시한다. 모델은 공정하게 데이터를 표현하기 위해 임베딩 공간상에서 보호 집단을 구분 불가능하도록 임베딩을 생성하여 이후 본 학습에서 보호 변수를 학습하지 못하도록 한다. 성별, 인종 등의 대표적인 보호 변수에 대해 실험을 진행한 결과, 우리의 모델은 4개의 데이터에서 큰 성능의 손실 없이 공정한 임베딩을 생성하는 결과를 보였다.

1. 서 론

기계학습(machine learning) 기술은 실제 사회의 데이터를 분석하고 현상을 이해하는 데 커다란 이바지를 하였다. 하지만 최근에는 기계 학습 모델들이 사회적 불평등과 편견을 답습하고 재생산한다는 문제점이 제기되고 있다. 학습 데이터에는 성별, 인종, 장애 여부, 종교 등의 보호 변수(protective attribute 또는 sensitive attribute)가 직, 간접적으로 포함되어 있다. 이러한 정보들은 모델이 개인이나 집단마다 다른 결과를 도출하도록 유도하고, 사회의 부당한 차별을 반영해 증폭시킬 위험성이 있다. 이에 따라, 기계학습 모델에서도 공정성(fairness)의 필요성이 대두되었다.

공정성이란 기계학습 모델이 데이터에 포함된 편향 또는 부정확성으로 인해 집단 또는 개인을 부당하게 차별하지 않도록 하는 개념이다[1]. 이 연구에서는 여러 공정성 지표 중 집단 공정성(group fairness)을 달성하는 것을 목표로 한다. 집단 공정성은 같은 보호 변수를 가지는 집단(i.e., 보호 집단 - protected group)이 보호 변수(e.g., 인종, 성별)와 무관하게 동등하게 대우받아야 한다는 개념으로, 기계학습 분야에서는 학습된 모델이 각 보호 집단마다 유사한 예측 결과 혹은 성능을 보여야 함을 뜻한다.

이 연구에서는 자기 지도 학습(self-supervised learning)에서 집단 공정성을 달성하기 위해 데이터의 편향이 제거된 공정한 임베딩(fair embedding)을 학습하는 모델을 제시한다. 이러한 임베딩은 어떠한 본 학습(downstream task)에 사용되어도 공평한 예측을 할 수 있도록 하는 것이 목적이다. 집단 공정성의 측면에 부합하는 임베딩을 생성하기 위해 이 연구에서는 각 집단이 임베딩 공간 내에서 구분 불가능하도록 임베딩을 학습한다. 즉, 각 보호 집단마다 임베딩이 거의 같은 분포를 하도록 임베딩을 생성하면, 이를 이용한 본 학습에서도 공정한 예측이 가능해진다.

이 논문은 대조 학습(contrastive learning)을 통해 위 문제를 해결한다. 대조 학습이란 유사한 샘플, 즉 양성 샘플(positive sample)은 임베딩 내의 거리를 가깝게 하지만, 그렇지 않은 샘플들, 즉 음성 샘플(negative sample)에 대해서는 멀어지도록 임베딩 공간을 학습하는 방법이다. 기존의 대조 학습에서는 자기 지도를 위해 유사한 속성을 가진 개체끼리 가까운 임베딩을 생성하는 것이 목적이기 때문에, 많은 경우에 주로 양성 샘플은 유사한 속성을 갖는 동일 집단 내에서 선택하고, 음성 샘플은 속성과 상관없이 동일 배치(batch) 내에서 선택하는 것이 보통이다. 그러나 공정한 임베딩을 생성하고자 한다면 유사성에 집중하기보다 보호 변수를 최대한 희석하는 것을 목표로 해야 한다. 이를 위해 우리는 대조 학습의 음성 샘플들을 같은 보호 집단 내로 한정하였고, 양성 샘플들은 두 샘플을 혼합하여 생성하였다. 그 결과, 네 종류의 데이터에서 제안한 임베딩을 예측 모델에 활용하였을 때 월등히 공정한 결과를 확인할 수 있었다.

2. 모델 설계

2.1 모델 개요

이 연구에서는 보호 변수를 공유하는 보호 집단을 분별할 수 없도록 임베딩을 생성하는 자기 주도 대조 학습 모델을 제시한다.

각 보호 집단 내의 인스턴스(instance)들이 닫힌

공간에서 서로 최대한 멀어진다면, 집단의 임베딩 분포는 균등 분포(uniform distribution)에 가까워진다. 만일 각 집단에 대해서 위 목표를 달성하게 된다면, 각 집단의 모든 인스턴스가 같은 균등 분포에 따라 흩뿌려지게 된다. 결국, 모든 그룹의 임베딩이 같은 분포를 하게 되고, 이 임베딩을 통해서는 각 보호 집단을 구분하지 못하게 된다.

2.2 손실함수(loss function)

대조 학습을 위해 InfoNCE 손실함수를 적용하였다[2]. 기준 샘플 x에 대한 한 개의 양성 샘플 x_t 와 $N-1$ 개의 음성 샘플로 이루어진 집합을 $X = \{x_1, \dots, x_N\}$라 하였을 때, 임베딩 모델 f에 대한 InfoNCE 손실 함수는 다음과 같다.

$$\mathcal{L}_N(x, X) = -\mathbb{E}_X\left[\log \frac{sim\big(f(x), f(x_t)\big)}{\sum_{x_j \in X} sim\Big(f(x), f\big(x_j\big)\Big)}\right]$$

위 수식에서 $sim(\cdot,\cdot)$ 은 두 임베딩의 유사도를 의미하는 함수이다. InfoNCE 손실함수의 값을 최소화하도록 훈련하면 기준 샘플과 양성 샘플 간의 임베딩 유사도는 높이는 동시에, 음성 샘플과의 임베딩 유사도는 낮추는 방식으로 임베딩이 학습된다.

2.3 양성 샘플

양성 샘플은 무작위로 두 개의 샘플을 혼합하여 생성한다. 특정 샘플 x 에서 보호 변수를 제외하고 무작위로 제거할 변수(feature)들을 선정하고, 그 변수를 역시 무작위로 선정한 다른 샘플 x'의 값으로 대체하여 양성 샘플 x_t를 생성한다. 무작위 변형을 하는 이유는 보호 변수와 연관이 있는 변수와의 과적합(overfitting)을 방지하기 위해서인데, 과적합 발생 시 모델이 결국 보호 변수를 학습하는 상황이 일어나기 때문이다. 상술한 과정으로 생성된 x_t 와 x 가 가까워지도록 학습한다. x_t 는 x 의 일부를 변형한 것이므로 서로의 임베딩 유사도를 높이는 방향으로 학습이 진행된다.

2.4 음성 샘플

동일 보호 집단의 임베딩이 최대한 멀어지도록 구성하면 임베딩의 분포가 균등 분포가 될 것이라고 가정하였다. 따라서, 음성 샘플은 동일 보호 집단 내에서, 즉 같은 보호 변수를 갖는 인스턴스들의 집단 내에서 추출한다.

2.5 모델의 구현

우리가 구현한 모델은 다층 퍼셉트론(multi-layer perceptron) 구조로, 임베딩을 생성하는 백본(backbone) 신경망에 3개의 층과 임베딩 공간 내에서 손실함수를 ㄴlayer)으로 이루어져 있다. 활성화 함수(activation function)로는 ReLU를 사용하였다. 각 배치(batch)의 크기는 128이며 총 200 에포크(epoch)의 학습을 거친다. 옵티마이저(optimizer)로는 Adam을 사용하였고 이때의 학습률(learning rate)은 1e-3, 가중치 감쇠 지수(weight decay factor)는 1e-6으로 설정하였다.

3. 실험 및 평가

모델의 성능을 평가하기 위해 각 데이터 셋에서 생성된 임베딩을 본 학습인 타겟 변수 예측에 적용하여 예측 성능을 평가하였다. 예측 모델로는 로지스틱 회귀(Logistic Regression)을 활용하였다.

3.1 데이터 정보

평가를 위하여 UCI Adult, UCI German credit, COMPAS, LSAC의 네 가지 데이터를 사용하였다. 각 데이터의 자세한 사항은 [표 1]에 기술되어 있다.

표 1. 실험에 사용된 데이터 종류

데이터	*본 학습*	*보호 변수*
UCI Adult	연봉이 50K를 넘는지 예측	성별, 인종
UCI German Credit	신용위험 예측	성별
COMPAS	재범위험 예측	성별, 인종
LSAC	법학교 입시 합격·불합격 예측	성별, 인종

3.2 평가 지표

Democratic parity distance[3]와 equalized odds[4]를 사용하여 임베딩의 집단 공정성을 평가하였고, 성능 지표로는 AUROC를 활용하였다. 일반적으로, 공정성 지표와 성능 지표 사이에는 trade-off가 있다. Demographic parity distance(이하 ΔDP)는 각 보호 집단 간 예측값의 차이이다. 보호 변수의 집합을 S라 할 때, ΔDP는 다음과 같다.

$$\Delta DP = \mathbb{E}_{s,s' \in S, s \neq s'}\left[\left|P\left(\hat{Y} = 1 \mid s\right) - P\left(\hat{Y} = 1 \mid s'\right)\right|\right]$$

Equalized Odds(이하 ΔEO)는 각 보호 집단 간 균등한 기회를 보장하는 지표이다. 보호 변수의 집합을 S라 할 때, ΔEO는 다음과 같다.

$$\Delta EO = \mathbb{E}_{s,s' \in S, s \neq s'}\left[\left|P\left(\hat{Y} = 1 | s, Y = y\right) - P\left(\hat{Y} = 1 | s', Y = y\right)\right|\right]$$

3.3 베이스라인과의 비교

성능 비교를 위한 베이스라인(baseline) 모델로 대조 학습을 활용한 최신 방법론인 SAINT를 채택하였다[5]. 이 방법은 음성 샘플을 같은 보호 그룹이 아닌 전체 데이터 셋에서 무작위로 선별한다. 즉, 자기 자신을 변형한 샘플을 제외한 모든 샘플과 멀어지도록

임베딩을 학습하고, 학습된 모델을 fine-tune 하는 과정을 거쳐 예측 성능을 끌어올렸다.

[표 2] ~ [표 4]는 각 데이터 및 보호 변수를 사용하여 분석한 결과를 나타낸다. 베이스라인 모델과 비교하였을 때 제안한 모델의 공정성 지표들이 크게 개선된 가운데, 비슷한 수준의 분류 성능을 보여주었다. 이 결과는 집단 공정성을 고려하더라도, 임베딩 모델의 분류 성능을 유지하면서 효과적으로 편향을 제거하여 공정한 임베딩을 구축할 수 있다는 점을 시사한다.

표 2. 각 데이터 및 보호 변수에 대한 ΔDP 수치

데이터	*Adult*		*Credit*	*COMPAS*		*LSAC*	
보호 변수	Gender	Race	Gender	Gender	Race	Gender	Race
Baseline	0.1865	0.0893	0.0291	0.0505	0.1206	0.0744	0.1277
Ours	**0.0571**	**0.0443**	**0.0043**	0.0518	0.1219	**0.0626**	**0.0985**

표 3. 각 데이터 및 보호 변수에 대한 ΔEO 수치

데이터	*Adult*		*Credit*	*COMPAS*		*LSAC*	
보호 변수	Gender	Race	Gender	Gender	Race	Gender	Race
Baseline	0.1906	0.1015	0.0970	0.1204	0.3263	0.0361	0.1181
Ours	**0.0692**	**0.0484**	**0.0284**	0.1175	**0.2988**	0.0376	**0.0730**

표 4. 각 데이터 및 보호 변수에 대한 AUROC 수치

데이터	*Adult*		*Credit*	*COMPAS*		*LSAC*	
보호 변수	Gender	Race	Gender	Gender	Race	Gender	Race
Baseline	**0.8988**	**0.8999**	0.7618	0.7333	0.7397	0.8576	0.8577
Ours	0.8432	0.8762	0.7595	0.7367	0.7411	0.8525	0.8397

3.4 하이퍼 파라미터 분석(hyper parameter analysis)

추가로, 손실함수를 변형하여 양성 샘플과의 유사성 및 음성 샘플과의 유사성 사이의 비중을 결정하는 하이퍼 파라미터 λ를 설정하였다.

$$\mathcal{L}_{N,\lambda}(x,X) = -\mathbb{E}_X\left[\lambda \cdot \log\left(sim\left(f(x),f(x_t)\right)\right) - \log\left(\sum_{x_j \in X} sim\left(f(x),f(x_j)\right)\right)\right]$$

하이퍼 파라미터의 값을 $\lambda = \{0.01, 0.1, 1\}$로 설정하고 각각의 경우에 대하여 성능 및 공정성을 측정하였는데, λ 값을 작게 할 경우 음성 샘플 간의 척력을 증가 시켜 보다 공정한 임베딩을 만드는 역할을 한다. [그림 1]은 UCI German credit 데이터 및 성별 변수에 대하여 3회씩 실험한 평균과 표준편차를 보여준다. λ 를 0.01까지 줄여도, 모델 성능(AUROC)은 큰 변화가 없는 반면, 공정성 지표인 ΔDP와 ΔEO가 큰 폭으로 감소하여 더욱 공정한 예측을 수행할 수 있었다.

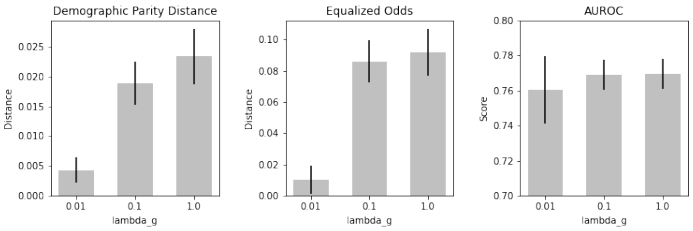

그림 1. λ의 값에 따른 각 지표의 결과

4. 결론

이 논문은 기계학습 모델이 데이터 내에 존재하는 보호 변수에 대한 편향에 구애받지 않고 공정하게 본 학습을 수행할 수 있도록 공정한 임베딩을 생산하는 자기 지도 대조 학습 방법을 제시한다. 공정한 임베딩을 형성하기 위해 대조할 음성 샘플을 동일 보호 집단 내에서 추출함으로써 보호 집단마다 임베딩 분포를 균등 분포에 가깝게 만들고 이에 따라 임베딩 공간상에서 각 보호 집단을 구분 불가능하도록 한다. 그 결과, 본 학습에서 분류 성능은 거의 유지하면서도 탁월하게 공정해진 수치를 확인하였다. 이러한 임베딩을 사용하여 사회의 여러 문제를 더욱 정당하고 공정하게 분석하고 예측할 수 있을 것이다. 추후 연구에서는 집단 공정성 외에 샘플 단위의 공정성 지표도 사용하여 더욱 공정한 임베딩 모델을 만들 수 있을 것으로 예상한다.

사사문구

이 논문은 기초과학연구원의 지원을 받아 수행된 연구임(IBS-R029-C2, IBS-R029-Y4).

참고문헌

[1] Oneto, L., Chiappa, S., “Fairness in machine learning”, Recent Trends in Learning From Data, Springer, 155-196, 2020

[2] Aaron van den Oord et al., “Representation learning with contrastive predictive coding”, arXiv preprint arXiv:1807.03748, 2018.

[3] Muhammad Bilal Zafar et al., “Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment”, Proceedings of the 26th International Conference on World Wide Web, 1171-1180, 2017.

[4] Moritz Hardt et al., “Equality of Opportunity in Supervised Learning”, 30th Conference on Neural Information Processing Systems, 3315-3323, 2016.

[5] Gowthami Somepalli et al., “SAINT: Improved Neural Networks for Tabular Data via Row Attention and Contrastive Pre-Training”, arXiv preprint arXiv:2106.01342, 2021.