



기계학습 공정성 관련 연구 동향

Trends of Research Into Fairness in Machine Learning

강송희 Kang, Songhee | 선임연구원 Senior Researcher, SPRI | dellabee@spri.kr

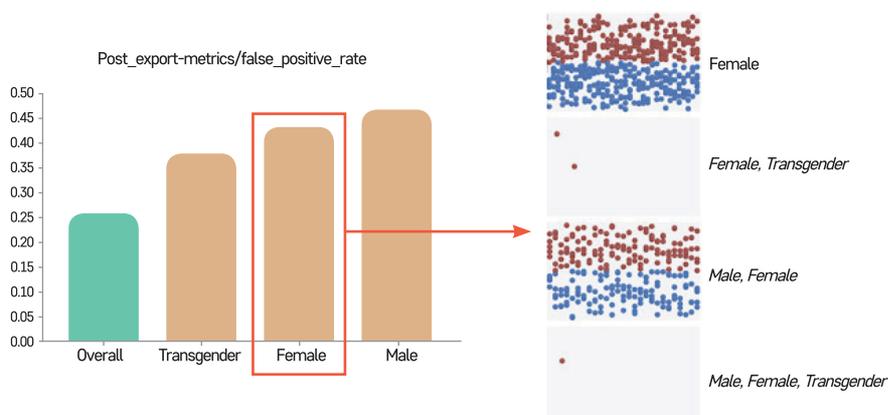
프로퍼블리카(ProPublica)의 2016년 기사 이후 ‘편향된 기계(Biased Machine)’에 대한 경각심이 커지고 있다. 인간사회에서 공정성이란 분배적, 절차적, 상호작용적 관점을 모두 포함한다. 그러나 기계학습에서 활용할 수 있도록 이러한 개념을 수학적으로 정의하기란 어렵다. 공정성에 대한 단일한 최상의 정의가 존재하지 않으며, 분배 관점에서 구분된 일부 통계적 공정성의 개념들조차 동시에 만족되기 어렵다는 제약이 있다. 최근 범용기술로 각광받고 있는 기계학습의 공정성에 대한 연구는 사회적 공정성의 개념과 격차를 메우는 방향으로 더 많은 연구와 투자가 이루어져야 할 것이다.

Since the 2016 ProPublica report "Machine Bias", there has been a growing awareness of biased machine. In human society, fairness encompasses distributive, procedural and interactive perspectives. However, it is difficult to mathematically define what fairness is for use in machine learning. There is no single best definition of fairness and even some of the concepts of statistical fairness from the distribution point of view are difficult to be satisfied at the same time. We will have to encourage more research and investments in fairness of machine learning, which has recently come into the spotlight as a general-purpose technology in a way that narrows the gap with the concept of social fairness.

👁 들어가는 말

최근 글로벌 소프트웨어 기업들은 '공정한 기계학습'에 대해 자체적으로 연구개발 투자를 단행하고 있다. 구글이 2019년 11월에 소개한 공정성지표 서비스(베타)에는 텐서플로(Tensorflow) 모델분석, 공정성지표, 텐서플로 데이터 검증, WIT(What If Tool)¹ 등을 포함하고 있으며 깃허브(Github)에 공개되었다. 이 중에서도 공정성지표는 텐서플로 확장(TensorFlow Extended, TFX) 컴포넌트에 포함되어 있으며, 평가도구를 통해 다섯 가지의 공정성지표를 확인하고 관련 데이터를 검증 및 확인해볼 수도 있다.

■ [그림 1] 구글 공정성지표(베타) 서비스를 활용하여 특정 그룹의 위양성률 데이터를 확인하는 모습



※ 자료 : 구글시블로그(2019.12.), <https://ai.googleblog.com/2019/12/fairness-indicators-scalable.html>

마이크로소프트 역시, 분류용 기계학습 모형의 편향을 다양한 공정성의 정의에 맞추어 줄일 수 있는 도구를 개발하여 깃허브에 공개하였다. IBM Research Trusted AI 또한 AI Fairness 360 오픈소스 툴킷을 공개하고, 70개 이상의 공정성지표와 10개의 최신 편향완화 알고리즘을 포함하였다고 발표했다.

위 사례들에서 알 수 있듯이 공정성에 대한 지표와 개념, 원칙은 분야와 상황에 따라 달리 적용해야 하고, 산업계 차원의 합의가 이루어지지 않고 있다. 따라서 아직은 대부분의 기업들이 자율주행, 인력채용 등 사용자와 직접 맞닿아 있는, 공정성이 매우 중요한 분야에 기계학습의 적용이나 관련 연구개발을 위한 투자를 선뜻 하기가 어렵다. 머신러닝이 사업현장에 적용되고 실질적으로 수익을 내기 위해서는 기술 자체의 확보만이 중요한 것이 아니다.

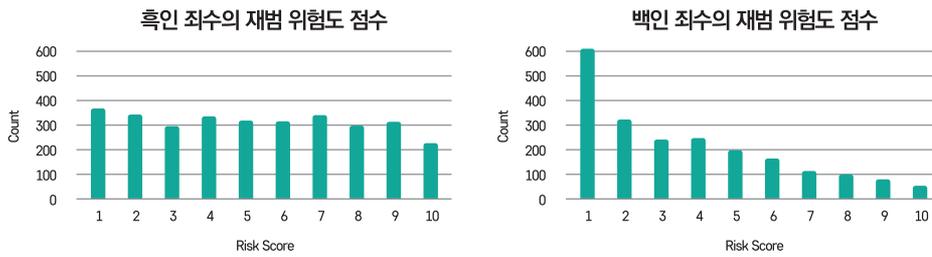
¹ 5가지 공정성 정의에 따라 각각 기계학습 모델의 최적화 전략을 결정하는 도구

📍 기계학습의 편향성에 대한 문제 제기

비영리 미디어 재단인 프로퍼블리카(ProPublica)는 2016년 Machine Bias²라는 기사를 발표했다. 이 기사는 미국 사법부가 사용하는 위험평가 소프트웨어(Risk Assessment Software)가 흑인 공동체에 대해 편향된 예측값을 도출한다고 주장했다. 여기서 사용된 위험평가 알고리즘인 COMPAS는 과거 유죄판결을 받은 사람들의 재범 가능성을 추정하기 위해 과거 데이터를 기반으로 예측값을 도출했는데, 2년 후 실제로 범죄를 다시 저지르지 않았음에도 불구하고 흑인이 백인보다 높은 위험도 판정을 받을 가능성이 거의 두 배에 이르렀고 백인은 흑인보다 낮은 위험도 판정을 받을 가능성이 훨씬 더 높았다. 전반적인 COMPAS 알고리즘의 정확도는 61% 정도였다.

■ [그림 2] COMPAS의 잘못된 예측결과

	백인	흑인
재범하지 않은 고위험군	23.5%	44.9%
재범을 저지른 저위험군	47.7%	28.0%



※ 자료 : ProPublica의 Broward County 데이터 분석자료(2016)

이와 같은 ProPublica 기사의 결론은 세간의 관심은 물론이고 기계학습 알고리즘의 공정성을 연구하는 연구자들로 하여금 폭발적 관심을 일으켰다. 연구자들은 공정한 결과물을 보장하기 위해 잘못된 편향을 내재한 모델의 위험을 완화시키는 방법을 개발하기 시작했다. 실제로 ICML 2018년 최우수논문상 5편 중 2편이 기계학습 공정성에 관한 논문이었고, arXiv에는 공정성과 관련된 논문들이 매주 수편씩 업로드 되는 중이다.

이후 기계학습 모형이 통계적으로 공정하다 하더라도 데이터에 의해 편향이 될 수 있을 뿐 아니라 산업계, 학계에서 아직 기계학습에 적용할 수 있는 통계적 공정성 개념에 대한 합의가 존재하지 않으며, 그러한 합의가 이루어졌더라도 사회에서 인식하는 공정성의 개념과 격차가 있다는 것을 설명한다.

² <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>

1. 입력 데이터로 인한 기계학습 모형의 편향

그렇다면 기계학습 모형이 편향을 갖게 되는 원인은 무엇인가? Barocas and Selbst(2016)은 기계학습 알고리즘의 편향을 발생시키는 원인을 훈련 데이터셋 내에 있는 축적된 인간의 편향에 기인한다고 보았다.

- ① 치우친 표본(Skewed Sample)

우연히 초기 편향이 발생하는 경우 시간이 지남에 따라 편향이 증폭된다.
- ② 오염된 사례(Tainted Example)

축적된 데이터에 존재하는 사람의 편견을 알고리즘에서 특별히 교정하지 않고 유지하는 경우 시스템 구조상 동일한 편향이 복제될 수밖에 없다.
- ③ 제한된 속성(Limited Feature)

특정 데이터의 속성(Feature)에 관련하여 소수 그룹에 대해서는 제한되거나 낮은 신뢰도의 정보만 수집될 수 있다.
- ④ 표본 크기의 불일치(Sample Size Disparity)

소수 그룹에서 제공되는 훈련 데이터가 대다수 그룹에서 제공되는 훈련 데이터보다 훨씬 적은 경우 소수 그룹을 정확히 모델링할 수 있는 가능성이 낮아진다.
- ⑤ 대리 변수의 존재(Proxy)

기계학습 훈련에 공정성 측면에서 민감한 데이터 속성(Feature: 인종, 성별 등)을 사용하지 않더라도 이를 대리하는 다른 속성(Feature: 이웃 등)이 항상 존재할 수 있어, 이러한 속성이 포함되어 있으면 편향이 계속 발생한다.

[표 1] 데이터로 인한 기계학습 편향의 유형

구분	원인	예시
관측되지 않은 성능 차이	치우친 표본	초기 범죄율이 높은 곳으로 더 많은 경찰관을 파견하는 경향이 있고, 그러한 지역에서 범죄율에 대한 기록이 더 높아질 확률이 높음
	오염된 사례	구글 뉴스 기사에서 남성-프로그래머의 관계는 여성-주부와의 관계와 매우 유사한 것으로 밝혀짐(Bolukbasi et al., 2016)
관측된 성능 차이가 존재하는 표본의 복제	제한된 속성	-
	표본 크기의 불일치	-
예측 결과의 불일치	대리 변수의 존재	인종, 성별과 같은 민감한 속성에 대해 이웃과 같은 속성이 대리 변수로서 작용하여 예측 결과가 편향될 수 있음

※ 자료 : <https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb>

2. 기계학습 모형이 이해하는 수학적 공정성의 한계

기계학습 모형이 이해할 수 있는 공정성은 수학적으로 정의되어 정량평가가 가능해야 한다. 이에 대한 공정성의 정의는 무려 20여 가지³로 분류되며, 대부분 분배적 관점에서 통계적 공정성에 초점을 두고 있다. 즉, 통계적 공정성에 대한 단일한, 혹은 최상의 정의에 대한 합의가 존재하지 않는 것이 현실이다.

■ [표 2] 통계적 공정성에 대한 수학적 정의 중 일부

구분	정의	수학적 정의
예측결과 기반	그룹 공정성 (통계적 동등성, 동등한 승인율)	그룹별로 긍정적 예측값을 할당받을 확률이 동일
	조건부 통계적 동등성	특정 데이터 속성(요소)을 통제했을 경우 그룹 별로 긍정적 예측값을 할당받을 확률이 동일
예측·실제결과 기반	예측적 동등성 결과 동등성	그룹별로 긍정적 예측값의 비율이 실제로 동일해야 함
	위양성율(Type I Error/ False Positive Error Rate) 균형	그룹별로 위양성 예측값을 할당받을 확률이 동일
	위음성율(Type II Error/ False Negative Error Rate) 균형	그룹별로 위음성 예측값을 할당받을 확률이 동일
	동등확률	그룹별로 실제 값 기반 진양성율(TPR, True Positive Rate)과 위양성율(FPR, False Positive Rate)은 동일
	조건부 사용 정확도 동등성	그룹별로 예측 값 기반 양성예측도(PPV, Positive Predictive Value)와 음성예측도(NPV, Negative Predictive Value)가 동일
	전체 정확도 동등성	그룹별로 전체적인 예측 정확도(진양성: True Positive, 진음성: True Negative)가 동일
	치료 동등성	그룹별로 위양성(False Positive)과 위음성(False Negative)의 비율이 동일

※ 자료 : FairWare'18. 2018.5.29., 스웨덴, Verma, S., & Rubin, J. (2018, May)에서 재발췌

이러한 통계적 공정성에는 현재 기준으로 [표 3]과 같은 한계가 있다. 우선 개념정의에 한계가 있을 뿐 아니라 모순적 공정성 개념이 공존하고, 예측의 정확도와 공정성은 서로 절충 관계이며, 지금까지 기계학습 분야에서 논의되어 온 수학적 공정성의 개념들이 분배적 공정성만을 다룬다는 것이다.

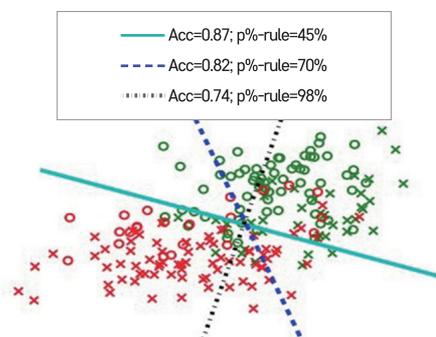
³ Verma, S., & Rubin, J. (2018, May), Fairness definitions explained. In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare) (pp. 1~7). IEEE, Gajane, Pratik, and Mykola Pechenizkiy. (2017) "On formalizing fairness in prediction with machine learning." arXiv preprint arXiv:1710.03184

[표 3] 통계적 공정성의 한계

구분	예시
개념 정의의 한계	단일한 최상의 수학적 개념 정의가 되지 않음
모순적 공정성 개념의 공존	통계적 공정성(그룹 공정성, 예측적 패리티, 동등확률)의 요건들이 동시에 만족되기 어렵다는 제약 조건
예측의 정확도와 절충 관계	공정성을 높이면 정확도가 떨어지는 경향
분배적 공정성에 종속	통계적 공정성은 분배적 관점에서 정의되고 있으며, 절차적 공정성이 갖춰지면 높아질 수 있는 가능성이 있으나, 일반적으로 통계적 공정성이 있어도 절차적 공정성이 있다고 볼 수는 없음

우선, 공정성의 개념을 정의하는 것 자체가 어렵다. ①(개념 정의의 한계) 기계학습 모형이 이해할 수 있는 단일한 최상의 수학적 개념이 정의가 되지 않은 상태다. 나아가, 세부 분류의 ②(모순적 공정성 개념의 공존) 통계적 공정성의 요건들이 동시에 만족되기 어렵다는 제약도 존재한다. 그룹 공정성과 예측적 동등성, 그룹 공정성과 동등확률, 예측적 동등성과 동등확률은 각각 동시에 만족될 수 없는 요건이다.(Chouldechova, 2016; Kleinberg et al., 2016) 모순적 공정성 개념의 공존 부분은 실제로 모든 공정성에 관한 수학적 정의를 한 번에 만족시키는 것이 어렵기 때문에 어느 하나의 공정성을 만족시키려다 보면 전체 알고리즘의 정확도가 떨어지는 것은 피하기 어렵다. 왜냐하면 각각의 공정성에 대한 수학적 정의들이 모델의 정확도와도 연관이 되기 때문이다. 물론 이러한 제약 조건이 정확도에 미치는 영향은 데이터 집합의 특성, 사용된 공정성의 정의 및 알고리즘에 따라 달라진다. 하지만 일반적으로 공정성을 높이면 정확도가 떨어지는 경향이 있기 때문에 실제로는 절충을 고려하지 않을 수 없다. 일례로 Zafar et al.(2017)의 연구에 따르면 선형분류 문제에 대해 그룹 공정성을 25%p 높일 때 정확도가 5% 떨어졌으며, 그룹 공정성을 28%p 높일 때에는 정확도가 8%떨어졌다.

[그림 3] 선형 분류 문제에서 그룹 공정성과 정확도 간 절충 관계

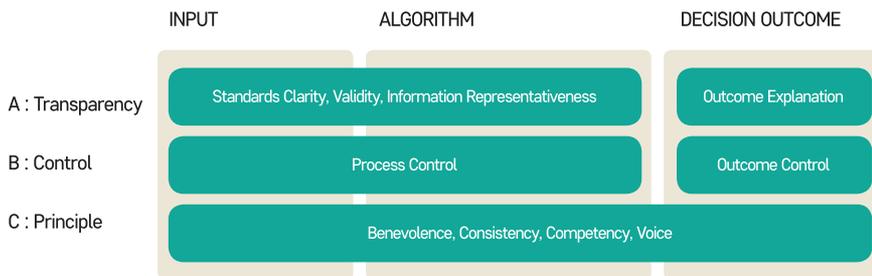


※ 자료 : Zafar et al.(2017)

이는 곧 ③(결과의 정확도와 절충적 관계) 목표하는 공정성 수준, 정확도 수준이 모델의 세부 구현에 의해 조절 가능하고 서로 절충이 필요한 관계라는 뜻이다.

공정성을 높이는 데 도움이 된다고 주장하는 알고리즘들은 대부분 전처리나 후처리 시에, 혹은 훈련 시에 제약조건을 최적화함으로써 적용된다. 알고리즘 기반의 의사결정에서 통계적 공정성을 달성하기 위한 절차적 공정성에 대한 연구(Lee et al., 2019)도 이뤄지고 있는데, 이는 투명성과 통제라는 요소를 전반적인 프로세스에서 관리하는 방식에 대한 틀을 제공하고 있다. 이를 종합해 보았을 때 ④(분배적 관점에 종속) 통계적 공정성은 분배적 관점에서 정의되고 있으며, 절차적 공정성이 갖춰지면 높아질 수 있는 가능성이 있으나 통계적 공정성을 달성했다고 해서 절차적 공정성을 갖췄다고 볼 수는 없다.

■ [그림 4] 알고리즘 기반 의사결정에서 절차적 공정성의 틀



※ 자료 : Lee et al.(2019), Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation CSCW2019

한편, 실무적으로 기계학습의 편향을 판단, 혹은 측정하는 데 핵심이 되는 성별, 인종과 같은 민감한 데이터 속성을 중심으로 한 최근 연구동향을 살펴보면, 훈련 시에 민감한 속성에 대한 정보를 명시적으로 사용하고 있다. 하지만 이러한 정보는 실제로는 제공되지 않을 수도 있다. 일부 통계적인 기법을 통해서 이러한 속성에 대한 정보 없이도 공정성을 개선하는 방법에 대한 연구도 존재하기는 하나(Hashimoto et al., 2018), 전반적인 정확도를 높이면서 관련기준을 명시적으로 고려하는 경우 오히려 공정성이 높아질 수 있다는 것이 최근 대부분 연구들의 주장이다. 결국 연구들에서 볼 때, 민감한 데이터 속성에 대한 편향을 제거하기 위해 해당 속성을 명시적으로 고려하는 방법을 취하고 있다는 아이러니가 발생할 수 있다.

3. 사회적 공정성과 기계학습 공정성의 개념적 차이

기계학습의 공정성에 관한 개념 정의는 인간 사회에서 흔히 말하는 공정성과는 괴리가 있다. 먼저 사회과학 분야에서 논의해 왔던 공정성의 유형은 세 가지로 분류될 수 있다. 분배적 공정성, 절차적 공정성, 상호작용 공정성이 그것이다. 분배적 공정성은 보상이나 칭찬과 같은 조직의 자원분배에 대한 구성원의 공정성 인식을 의미한다.⁴ 이와 같은 분배적 공정성에 대한 인식은 결과가 동일하게 적용되었다고 인식할 때 높아진다. 한편, 절차적 공정성은 업무 맥락 속 판단과 의사결정 절차의 공정성을 일컫는다.⁵ 이러한 절차적 공정성은 일관성, 편견 억제성, 정확성, 정정 가능성, 대표성, 윤리성 등을 만족할 때 높아진다. 마지막으로 상호작용 공정성은 신중하고 예의 있게 의사결정에 대한 설명을 하거나 정보를 전달할 때 개인이 받는 대우를 일컫는다.⁶ 이는 절차를 수행하거나 성과를 결정할 때 개인이 공손하고 정중하게 대우받거나, 그에게 가는 정보가 얼마나 시의성, 구체성, 진실성을 지니는가에 따라 높아질 수 있다.

이에 반해 기계학습 분야에서 논의되고 있는 공정성은 수학적으로 정의가 가능한 것으로, 위 세 가지 유형 중 분배적 공정성에 속하는 조금 더 협소한 개념이다. 따라서 기계학습 알고리즘이 어떠한 일부 통계적 공정성을 만족하는 모형을 사용하였을 뿐 아니라 데이터의 편향을 잘 통제하였다 하더라도, 사람들이 인식하는 '사회적 공정성'과는 개념적 격차가 존재하는 것이다.

시사점

한국 사회는 인공지능을 4차 산업혁명 시대 새로운 성장 동력이 되는 범용기술로 인지하고 있다. 기계학습을 포함한 인공지능의 파급력은 개인, 기업, 산업 및 사회에 이르기까지 광범위할 것이라 예측된다. 따라서 새로운 시대의 국가 경쟁력 제고를 위해 국제사회에서 이 기술과 관련하여 어떻게 포지셔닝 할지, 어떤 도메인에 실질적으로 투자할지에 대한 구체적인 국가전략 수립도 중요하다. 나아가 전략의 실현가능성을 높이기 위해 예산권 등, 실질적인 권한이 있는 범부처 조직을 설치하고 공정성에 대한 실무적인 논의를 포함한 장기적인 정책기획과 실행, 연구개발 투자를 단행할 필요도 있다. 특히, 기계학습의 공정성에 대한 연구는 사회적 공정성의 개념과 격차를 메우는 방향으로 더 많은 투자가 이루어져야 한다. 공정성에 대한 개념적 논의는 한 국가에서 독자적으로 추진할 수 없기 때문에 국제적인 공조가 필요하며, 정부기관뿐 아니라 기업과 학계를 포함하는 연구 공동체, 다양한 이익집단의 목소리가 한데 모아져야 할 것이다.

⁴ Homans, G. C. (1961), *Social Behavior: Its Elementary Forms*. New York: Harcourt, Brace, & Jovannovich, Inc.
⁵ Leventhal, G. S. (1980), *What Should Be Done with Equity Theory?* In Gergen, K. G. & Greenberg, M. S. & Willis, R. H. (eds.), *Social Exchange: Advances in Theory & Research*, NY: Plenum Press.
⁶ Bies, R. J., & Moag, J. F. (1986), *Interactional justice: Communication criteria of fairness*. In R.J. Lewicki, B. H. Sheppard, & M. H. Bazerman (Eds.), *Research on negotiations in organizations* (Vol. 1, pp. 43-55). Greenwich, CT: JAI Press.