<기존 classification, ranking, recommendation 연구에서의 fairness 정의 조사>

□ 기존 classification 연구에서의 fairness(공정성) 정의

o DL/ML 분야에서의 fairness는 개인이나 집단을 차별하지 않는 것을 의미한다.

o 특히 classification 연구에서 fairness는 classification 결과를 받는 대상에 의해 정의된다.

o On Formalizing Fairness in Prediction with Machine Learning, FATML 2018

- 이전 연구(reference.. 추가!)에서는 법으로 보호된 특징을 protected attributes라고 정의하며, 이에는 성별, 인종, 종교, 지역, 장애, 가족의 상태 등이 있을 수 있다.
- 기존 머신러닝 분야에서의 공정성을 정의는 다음과 같다
- 비자각에서 오는 공정성(fairness through unawareness): 예측 과정에서 protected attributes를 명백하게 사용하지 않으면 그 예측 프로그램은 비자각에서 오는 공정성을 가진다고 할 수 있다.
- 반사실적 공정성(counterfactually fairness): protected attribute 만 다르게 했을 때, 같은 결과물이 나온다면 이를 반사실적으로 공정하다고 한다.
- 집단적 공정성(group fairness): unprotected attribute를 가지는 두 부분 집합에 대해 그 예측값이 protected attribute가 될 확률이 매우 유사하다.
- 개인 공정성(individual fairness): 비슷한 개인은 그 분류 예측값도 유사하다.
- 선호되는 대우(preferred treatment): 그룹별로 예측 프로그램이 다른 경우에, 특정 그룹이 더 많은 편익을 얻는 경우에 선호되는 대우를 지닌다고 한다.
- 선호되는 영향력(preferred impact): 어떤 예측 프로그램이 다른 예측 프로그램에 비해 모든 그룹에 대해 적어도 하나 이상의 편익을 가질 때 선호되는 영향력을 지닌다고 한다.

- 추가로 머신러닝 분야에서 잘 고려하지 않았던 2가지의 fairness 개념을 제안함.
- 자원의 균등성(equality of resources): 특정 개인들의 의도적인 결정과 행동에서 나오는 결과일 때만, 사회적 혜택의 불균등한 분배가 공정하다. 즉, 각 개인이 자신들의 이익을 위해 의도적으로 나오는 야망과 선택에는 자원의 균등성이 적용될 수 있

으나, 개인들이 선택하지 않은 선천적으로 타고난 재능 같은 경우에는 고려하지 않아도 된다.

- 기능 가능성의 균등성(equality of capability of functioning): 존재하거나 행하는 상태에 대해 사회적으로 타고난 특징들은(나이, 성, 인종, 계층 등) 같은 기회에 있어서 동일하지 않은 힘을 가지도 있다고 해도 모든 사람들이 만족해야한다.

o Equality of opportunity in supervised learning, NIPS 2016

- 동등한 기회(equal opportunity): 실제값이 우세한 결과일 때, protected attributes가 있건 없건 간에 분류를 그 우세한 값으로 예측할 확률이 동등해야한다.

□ 기존 ranking 연구에서의 fairness(공정성) 정의

o classification과는 달리 ranking에서는 학습 데이터의 편향성, 유저 행동의 편향성(특정 종류의 결과에만 클릭하는 경향), 문서 자체의 편향성(such as different sections of resumes completed at different rates by men and women)의 특징이 있어 이를 위한 연구가 필요함.

o Fairness of exposure in rankings, KDD 2018.

- attention-based measure: 사용자로부터 클릭률 등의 proxy를 통해 얻을 수 있는 각 항목들의 attention 이나 노출이나 간섭 등을 통해 클릭될 수도 있는 잠재적 attention을 정량화하는 방법.
- Pm*n= probabilistic ranking of m items in n positions
- v_j = item과 상관없이 j 번째 position의 visibility(eyetracking등의 실험으로 얻음)
- u_i = item I의 관련도 relevance or utility (유저가 생각하는 이 항목 검색 쿼리와 연관 있는 정도-모든 유저의 평균값으로 정의, 단 실제상황에서는 예측값을 사용해야함)
- G_0 = majority/advantaged/unprotected group
- G_k = minority/disadvantaged/protected group (k>=1)
- $\text{Exposure}(G_k|P) = \frac{1}{|G_k|}\sum_{i:d_i \in G_k}\sum_{j=1}^{n} P_{i,j} v_j.$
- $\text{U}(G_k) = \frac{1}{|G_k|}\sum_{i:d_i \in G_k} u_i.$

- $$\frac{Exposure(G_0|P)}{U(G_0)} = \frac{Exposure(G_1|P)}{U(G_1)}$$

- fairness of exposure: 항목의 노출 위치와 사용자의 주의도를 충분히 고려했을 때, 어떤 항목이건 간에 '노출되는 양/해당 검색어와의 관련성'은 일정해야 한다.

o Equity of attention: Amortizing Individual Fairness in Rankings, SIGIR 2018.

- 이전 연구에서 집단 별로 공정성을 판별한 것과 달리, 이 논문은 개인의 수준에서 fairness를 연구함.
- 각 항목들에 사용자의 집중(attention)이 공평하게 분배될 수 있도록 하는 분할상환 공정성(amortized fairness)을 제안함.
- a_i = i번째 항목의 normalized attention [0,1]
- r_i = I번째 항목의 normalized relevance score [0,1].
- 각 항목들의 누적된 attention이 누적된 relevance와 비례하면 equity of amortized attention을 만족한다고 정의함(아래 수식).
- $$\frac{\sum_{l=1}^{m} a_{i1}^{l}}{\sum_{l=1}^{m} r_{i1}^{l}} = \frac{\sum_{l=1}^{m} a_{i2}^{l}}{\sum_{l=1}^{m} r_{i2}^{l}}, \ \forall u_{i1}, u_{i2}.$$

- equity of amortized attention을 만족하려면 A와 R의 분포가 동일해야하므로 아래 수식으로 (un)fairness 정의 가능.
- $$unfairness(\rho^1, ..., \rho^m) = \sum_{i=1}^{n} |A_i - R_i| = \sum_{i=1}^{n} \left| \sum_{j=1}^{m} a_i^j - \sum_{j=1}^{m} r_i^j \right|. \quad (1)$$

- 추후 필요하다면 위 정의를 이용한 ranking system 조사할 필요 있음.

o Measuring fairness in ranked outputs. SSDBM 2017.

- probability-based measure: 랜덤으로 ranking이 생성되었다고 했을 때의 순위에서의 예상 특성에서 편차를 측정하는 방법.
- 우선 비교 데이터로 protected group과 unprotected group의 항목들을 relevance가 작아지는 순으로 나열한 두 개의 리스트를 준비하여, 베르누이 시행을 따라 두 개의 그룹에서 골고루 n개의 항목을 순서를 가지고 뽑은 데이터를 준비한다.
- 베르누이 시행의 변수인 p와 j 번째까지의 protected group의 항목 수를 이용해, 3가지 측정 방법(rND, rKL, rRD)을 제안함.

- 만들어진 분포와 실제 분포를 비교해 보면 편차 유무를 확인할 수 있음.

o Fairness and transparency in ranking, CIKM DAB 2018

- 이 논문은 공정한 ranking 시스템이 지녀야 하는 세 가지 특성을 주장한다.
- 1. 각 그룹에 속하는 항목들이 충분히 존재해야한다.
- 2. individual fairness를 만족할 수 있도록, 비슷한 항목들에 대해서는 일관적인 처리를 해야한다.
- 3. 특별히 약자 집단(disadvantaged group/ protected group)의 항목에 대해서는 적절한 표현을 사용해야한다.

o FA*IR: A Fair Top-k Ranking Algorithm, CIKM 2017

- 기존 ranking system과 비교하여 큰 utility 손실 없이, utility를 최대화하면서 동시에 ranked group의 공정성을 만족하는 top-k ranking을 만드는 효과적인 알고리즘(FA*IR)을 제안.
- utility = ranking algorithm에 의해 계산되는 qualification(=relevance score)가 가장 높은 후보를 선택.
- selection utility = 모든 후보가 top-k에 포함된 것이 더 선호됨.
- ordering utility = 후보의 모든 쌍이 top-k에 포함된 것이 더 선호됨.
- 조건1. ranked group fairness: 랭킹 r은 protected group을 공정하게 represent해야함.
- 조건2. selection utility: r 은 가장 qualified된 후보를 포함해야한다.
- 조건3. ordering utility: r은 qualification의 내림차순 순으로 정렬되어야 한다.
- protected group 과 non-protected group 이 비슷한 모양의 qualification 분포를 가질 것이라고 가정하지 않음.
- 변수 p의 조정을 통해 fairness 와 utility 사이 trade-off가 가능.
- 추후 필요하다면 제안하는 ranking system 더 조사할 필요 있음.

□ 기존 recommender 연구에서의 fairness(공정성) 정의

o 대부분 collaborative filtering 중심.

o Multi-sided fairness for recommendation, FATML 2017.

- 추천 시스템에서 여러 가지 관점의 공정성에 대해 다룸.
- 여기서 소비자(consumer)는 추천을 받는 사람들이고, 제공자(provider)는 추천이 되

면 이득을 보는 사람들, 즉 추천되는 항목 관련 사람들이다.

- C-fairness: protected class의 소비자입장에서의 영향력을 크게 고려해야하는 추천 시스템에서의 공정성 정의.
- P-fairness: 제공자 입장에서의 공정성만 고려되어야 하는 공정성. (예: 시장 다양성과 독점 방지가 필요할 때 고려되어야함)
- CP-fairness: 소비자와 제공자 모두 보호 그룹에 속할 경우, 두 그룹 모두의 공정성을 고려해야 하는 경우의 공정성.

o Fairness Objectives for Collaborative Filtering, NIPS 2017.

- memory-based collaboative filtering을 쓰는 경우에 추천 받는 사람(subject)를 위한 group fairness 연구
- 총 5가지의 fairness metric을 이용하여, protected group의 예측 오차 값(prediction error)이 unprotected group의 예측 오차값과 비슷할 때 fair하다고 정의.
- 1. value unfairness: 그룹 별로 예측 오차값의 차의 평균.
- $$U_{\text{val}} = \frac{1}{n}\sum_{j=1}^{n}\left|\left(\mathrm{E}_g\left[y\right]_j - \mathrm{E}_g\left[r\right]_j\right) - \left(\mathrm{E}_{\neg g}\left[y\right]_j - \mathrm{E}_{\neg g}\left[r\right]_j\right)\right|$$
- 2. absolute unfairness: 그룹 별로 예측 오차값의 절댓값의 차의 평균.
- $$U_{\text{abs}} = \frac{1}{n}\sum_{j=1}^{n}\left|\left|\mathrm{E}_g\left[y\right]_j - \mathrm{E}_g\left[r\right]_j\right| - \left|\mathrm{E}_{\neg g}\left[y\right]_j - \mathrm{E}_{\neg g}\left[r\right]_j\right|\right| .$$
- 3. underestimation unfairness: 두 그룹 간 실제보다 적게 예측하는 경우의 불균형.
- $$U_{\text{under}} = \frac{1}{n}\sum_{i=1}^{n}\left|\max\{0, \mathrm{E}_g\left[r\right]_j - \mathrm{E}_g\left[y\right]_j\} - \max\{0, \mathrm{E}_{\neg g}\left[r\right]_j - \mathrm{E}_{\neg g}\left[y\right]_j\}\right| .$$
- 4. overestimation unfairness: 두 그룹 간 실제보다 크게 예측하는 경우의 불균형.
- $$U_{\text{over}} = \frac{1}{n}\sum_{j=1}^{n}\left|\max\{0, \mathrm{E}_g\left[y\right]_j - \mathrm{E}_g\left[r\right]_j\} - \max\{0, \mathrm{E}_{\neg g}\left[y\right]_j - \mathrm{E}_{\neg g}\left[r\right]_j\}\right| .$$
- 5. non-parity unfairness: 각 그룹의 평균 예측값의 차
- $$U_{\text{par}} = \left|\mathrm{E}_g\left[y\right] - \mathrm{E}_{\neg g}\left[y\right]\right| .$$
- 위의 unfairness를 기존 loss function에 더해서, collaborative filtering을 위한 matrix factorization 학습을 하면 더 공정한 결과를 얻을 수 있다.
- 가짜 데이터를 만들어서 실험 진행.

o Balanced Neighborhoods for Multi-sided Fairness in Recommendation, FAT 2018.

- 소비자 측면의 동등성 점수: $$E_c@k = \frac{\sum_{i \in U^+} \sum_{\rho \in P_i@k} \gamma(\rho)/|U^+|}{\sum_{i \in U^-} \sum_{\rho \in P_i@k} \gamma(\rho)/|U^-|}$$
- γ : ρ → {0,1} 영화가 보호 받는 장르이면 1.
- $P_i@k = \rho_1, \rho_2, \ldots, \rho_k$ 유저 i의 top-k 추천 항목들

o SLIM: Sparse Linear Methods for top-N Recommender system, ICDM 2011.

- $$\tilde{A} = AW,$$
- A = user-item 이미 구입하거나 점수 매긴 matrix
- W = aggregation coefficient sparse matrix(아래 최적화 방법을 통해 학습)
- $\tilde{A}$= user별 item 추천 점수 matrix
- 주어진 A를 이용해 아래 최적화 방법을 통해 W을 학습시키고, $\tilde{A}$에서 A를 제외한 항목들을 내림차순으로 N개 추출.
- $$\begin{aligned} \underset{W}{\text{minimize}} \quad & \frac{1}{2}\|A - AW\|_F^2 + \frac{\beta}{2}\|W\|_F^2 + \lambda\|W\|_1 \\ \text{subject to} \quad & W \geq 0 \\ & \text{diag}(W) = 0, \end{aligned}$$

o Calibrated Recommendation, RecSys 2018.

- 소비자 개인의 공정성을 고려하는 것이 아니라, 소비자의 다양한 관심사의 공정성을 고려.
- 추천되는 항목들의 장르 비율이 초기의 입력값으로 들어간 장르 비율과 같으면 공정하다(calibrated)라고 정의함.
- p = 유저 u가 장르 g의 영화에 대해 과거에 본 분포
- q = 유저 u가 추천받는 장르 g의 영화 리스트 분포
- $$p(g|u) = \frac{\sum_{i \in \mathcal{H}} w_{u,i} \cdot p(g|i)}{\sum_{i \in \mathcal{H}} w_{u,i}}, \quad q(g|u) = \frac{\sum_{i \in \mathcal{I}} w_{r(i)} \cdot p(g|i)}{\sum_{i \in \mathcal{I}} w_{r(i)}},$$
- KL-divergence를 이용해 calibration metric을 아래와 같이 표현할 수 있음.

- $$C_{\mathrm{KL}}(p,q) = \mathrm{KL}(p||\tilde{q}) = \sum_{g} p(g|u) \log \frac{p(g|u)}{\tilde{q}(g|u)},$$

o Bias Disparity in Recommendation Systems, CIKM DAB 2018

- 사용자의 입력 값에 있는 편향된 정보들을 추천 시스템이 더 증폭시켜 편향 불균형 상태가 되는 현상을 문제 삼음.
- 가상의 데이터를 이용
- 추천 시스템이 편향 불균형을 일으킬 수 있는 여러 조건과 장기적으로 미칠 수 있는 영향을 실험을 통해 알아냄.
- 또한 실제 데이터에서도 이런 편향 불균형 상태 관찰함.

- preference ratio = $$PR_S(G,C) = \frac{\sum_{u \in G} \sum_{i \in C} S(u,i)}{\sum_{u \in G} \sum_{i \in I} S(u,i)}$$
- where S(u,i) = 1 (유저 u 가 항목 I를 선택했을 경우), 0 (그렇지 않은 경우)

- bias disparity = $$BD(G,C) = \frac{B_R(G,C) - B_S(G,C)}{B_S(G,C)}$$

- where $$B_S(G,C) = \frac{PR_S(G,C)}{P(C)}$$,
- P(C) = |C|/m (전체에서 category C를 랜덤하게 고를 확률)
- UserKNN algorithm를 이용해 실험 진행: 주어진 유저와 jaccard similarity가 비슷한 k명의 유저를 골라 가장 utility 값이 높은 항목들 추천.

o Fairness in Package-to-group Recommendations, WWW 2017.

- 추천받는 대상이 여러명(그룹)인 경우에서의 공정성을 다룸
- 그룹의 모든 사람들이 충분한 항목 개수에 의해 만족한다면, 이 추천이 공정하다고 할 수 있으나 해당 문제는 NP-hard임.
- 이 논문에서는 이 문제를 coverage problem으로 모델링하여 효과적인 fair package를 찾는 greedy algorithm을 제시함.
- 다음과 같은 공정성의 두가지 측면을 봄.
- m-proportionality of fairness: 유저 u와 추천받는 아이템들(package) P에 대해, P에 유저 u가 좋아하는 항목들이 m개(m>=1) 이상 존재하면, P는 유저 u에 대해 m-proportional 하다.

- $$F_{\text{prop}}(G,P) = \frac{|G_p|}{|G|},$$
- envy-free: 유저 u 와 추천받는 아이템들 P, 그룹 G가 주어졌을 때, 유저 u의 어떤 항목 I에 대한 예측 rating 값이 그룹원들의 예측 rating의 상위 x%안에 들 때, 이를 envy-free라고 한다.
- m-envy-freeness: P 의 적어도 m개 이상의 항목들이 envy-free이면 P는 u에 대해 m-envy-freeness라고 한다.
- $$F_{\text{ef}}(G,P) = \frac{|G_{\text{ef}}|}{|G|},$$

o Fairness aware group recommendation with pareto-efficiency, RecSys 2017.

- 추천받는 대상이 여러명(그룹)인 경우에서의 공정성을 다룸
- 그룹의 모든 사람들이 충분한 항목 개수에 의해 만족한다면, 이 추천이 공정하다고 할 수 있으나 해당 문제는 NP-hard임.
- 따라서 이 논문은 이 문제를 multi-objective optimization 으로 해결하려고함.
- relevance: rel(u, i) ∈ [relmin , relmax]
- individual utility: 유저 u에 대해 항목 그룹 I(|i| = K)이 추천된다고 했을 때, U(u, I) : U ×I → [0, 1]. 다음과 같은 두 가지 semantic이 있을 수 있음.
- $(1)\ \text{Average: } U(u,I) = \frac{1}{K \times rel_{max}} \sum_{i \in I} rel(u,i)$
- $(2)\ \text{Proportionality: } U(u,I) = \frac{\sum_{i \in I} rel(u,i)}{\sum_{i \in I(u,K)} rel(u,i)}$
- 이 논문에서의 fairness는 같은 그룹 내의 유저가 얼마나 균등하게 만족했는지, 즉 유저 사이의 utility의 차이가 적음을 의미함.
- 아래와 같은 4가지 의미의 fairness가 있을 수 있음.

$$Least\ Misery:\ F_{LM}(g,I) = \min\{U(u,I), \forall u \in g\}$$
$$Variance:\ F_{Var}(g,I) = 1 - Var(\{U(u,I), \forall u \in g\})$$
$$Jain's\ Fairness:\ F_J(g,I) = \frac{(\sum_{u \in g} U(u,I))^2}{|U| \cdot \sum_{u \in g} U(u,I)^2}$$

- $$Min - Max\ Ratio:\ F_M(g,I) = \frac{\min\{U(u,I), \forall u \in g\}}{\max\{U(u,I), \forall u \in g\}}$$
- 또한 social welfare 도 고려하여, fairness와 social welfare 모두가 최대화되는 방

향의 추천시스템을 만들고자함.

- social welfare: 그룹 내의 모든 유저들의 utiliy의 평균.

o Top-N Group recommendations with fairness, SAC 2019.

- 추천받는 대상이 여러명(그룹)인 경우에서의 공정성을 다룸
- member utility: 유저 u의 top-k 추천 아이템을 유저 u의 ground truth 라고 할 때, 그룹의 추천 아이템 목록과 유저 u의 ground truth 유사도를 u의 member utility라고 한다.
- 그룹 g의 가장 작은 member utility를 fairness라고 정의하고, 이 값을 최대화하는 추천 시스템을 공정하다고 한다.

o Fairness in recommendation ranking through pairwise comparison, KDD 2019.

- 두 유저 그룹에 대해 이미 클릭된 항목이 추천될 가능성이 관련이 있지만 클릭되지 않은 항목의 추천 가능성보다 높아야 pairwise fairness를 가진다고 정의함.
- 아래 식으로 표현 가능.

$$P(c_{\mathbf{q}}(j,j')|y_{\mathbf{q},j} > y_{\mathbf{q},j'}, s_j = 0, z_{\mathbf{q},j} = \tilde{z})$$

$$= P(c_{\mathbf{q}}(j,j')|y_{\mathbf{q},j} > y_{\mathbf{q},j'}, s_j = 1, z_{\mathbf{q},j} = \tilde{z}), \quad \forall \tilde{z}.$$

- 세부적으로 pairwise fairness에서 같은 그룹 내에서 항목을 뽑는지, 혹은 서로 다른 그룹에서 뽑는지에 따라 intra-pairwise fairness와 inter-pairness로 정의 가능.