특징조합 교란자 균형을 통한 인과정규화된 로지스틱 회귀 개선

서석인 ○, 황형주, 양홍석, 김기응

한국과학기술원

{siseo, hjhwang}@ai.kaist.ac.kr, hongseok00@gmail.com, keeeung.kim@kaist.edu

Improving Causally-Regularized Logistic Regression via Confounder Balancing with Feature Combinations

Seokin Seo○, HyeongJoo Hwang, Hongseok Yang, Kee-Eung Kim

KAIST

요 약

안정적 학습(stable learning)은 훈련데이터와 테스트데이터 간의 분포변화(distribution shift)에 강건한 학습을 목표로 한다. 본 논문에서는 이진 분류문제에서 제시된 기존 안정적 학습 방법 중 하나인 인과정규화된 로지스틱 회귀(Causally-Regularized Logistic Regularization; CRLR)를 포함하는 더 일반적인 인과정규화 방법을 제시하고자 한다. 기존 알고리즘의 경우 각 특징(feature) 하나를 처방변수로 취급한 후 나머지 특징에 대해 교란자 균형(confounder balancing) 방법을 적용하여 샘플들의 가중치를 학습하였는데, 이를 확장하여 특징의 조합을 처방변수로 취급한 뒤 나머지 특징들의 균형을 맞추는 방법을 제안한다. 또한, 제안하는 방법이 기존 방법보다 효과적임을 간단한 실험을 통해 보인다.

1. 서 론

대다수의 기계학습 알고리즘들은 훈련 데이터셋과 테스트 데이터셋이 동일한 분포로부터 독립적으로 샘플되었다는 가정 하에서 작동하고 있다. 하지만 기계학습 알고리즘을 실세계 문제에 적용하는 경우, 데이터 수집과정은 항상 이러한 가정을 만족하지는 않는다. 구체적인 예로, 실세계에서 수집된 이미지 데이터를 기반으로 분류(classification) 문제에 기계학습 알고리즘을 적용하는 경우, 분류하고자 하는 레이블과 연관된 인과적 특징(causal feature) 뿐만 아니라, 레이블과 직접적인 관련은 없지만 데이터 수집 과정에서의 편향(bias)으로 인해 특정 특징이 수집된 샘플들에 많이 포함되어 있는 경우 이러한 특징을 보고 분류기(classifier)가 학습될 수 있는 위험이 있다. (예:, 강아지 이미지 분류 데이터 수집 시 강아지와 초록색 들판이 같이 있는 이미지만 항상 같이 수집하는 경우)

이러한 편향을 데이터 선택 편향(data selection bias)이라고 부르는데, 이러한 편향이 존재하는 경우, 편향으로 인해 자주 드러났던 특징은 테스트시에는 드러나지 않는 경우가 많으므로, (위 예시에서 들판이 없는 강아지 이미지의 경우) 분포변화(distributional shift)가 일어나, 일반화 성능(generalization performance)이 크게 떨어질 수 있다. 이러한 면에서 미루어 볼 때, 데이터 선택 편향에 강건한 모델 학습 알고리즘 개발은 중요하며, 관련 연구분야로 안정적 학습(stable learning) 혹은 분포외 일반화(out-of-distribution generalization)가 있다.

기존 데이터 수집 편향에 안정적 학습 연구로 인과정규화된 로지스틱 회귀 (Causally-Regularized Logistic Regularization; CRLR) [1] 방법이 있다. CRLR은 이진 분류(binary classification) 문제의 접근법인 로지스틱 회귀(logistic regression) 학습 시에, 인과추론(causal inference) 분야에서 사용되는 교란자 균형(confounder balancing) 방법을 적용하여 각 샘플의 가중치(weight)를 학습하고, 이 가중치가 적용된 이진교차엔트로피 손실(weighted binary cross entropy loss) 최소화를 통해 안정적인 로지스틱 회귀모델을 학습하는 방법을 제시하였다.

이 논문*에서는 기존 방법인 CRLR을 확장하여, 처방변수의 조합을 균형의 기준으로 도입하는 방법을 제안한다. 이 방법은 기존 방법보다 많은 경우의 수를 고려하여 교란자 균형을 이루도록 샘플 가중치를 학습하며, 이를 통해 더 데이터 선택 편향에 안정적인 학습방법을 도모한다. 또한, 데이터 선택 편향이 존재하는 간단한 상황에서 기존 방법들과 비교 실험을 진행하여, 제안하는 방법이 효과적인지 보이고자 한다.

*이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. 2022-0-00311, 일상적 물건들의 다접촉 로봇 조작을 위한 목적지향 강화학습 기술 개발, No. 2020-0-00940, 안전한 강화학습 원천 기술 개발 및 자연어 처리에의 응용)

2. 배 경

2.1. 인과추론 및 관측연구 (observational study)

인과추론(causal inference)은 변수 간의 인과관계(causal relationship)에 따른 영향을 연구 분야이다. 인과추론이 문제시하는 상황은 대표적으로 잠재결과 프레임워크(potential outcomes framework)[2]라고 부르는 방식으로 표현되는데, 이 프레임워크는 다음과 같은 변수들을 포함한다.

- **처방 (Treatment; T)**: 원인에 해당하는 변수
  (예: 투약여부, 절단, …)
- **결과 (Outcome; Y)**: 결과에 해당하는 변수
  (예: 혈압변화, 사망여부, …)
- **교란자 (Confounder; X)**: 데이터 수집 시 처리와 결과 모두에 영향을 주는 변수
  (예: 환자의 증상, 체온, 혈압, …)

여기서 인과효과 식별(causal effect identification)의 목표는 교란자의 영향을 받아 배정된 처방에 따른 결과를 예측하는 것이 아니라, 특정 처방을 임의로 배정했을 시의 결과를 추정할 수 있도록 통계적 추정치로 식별하는 것이다. 이를 위해서는 교란자와 상관없이 무작위로 처방을 하여 결과를 관측하는 무작위 비교연구(randomized controlled trial; RCT)을 통해 데이터 수집하는 것이 필요하나, 이는 비용과 윤리적인 측면 등에서 현실적이지 않은 경우가 많다.

관측연구(observational study)[3]는 이미 수집된 데이터를 바탕으로 이러한 인과효과를 식별 및 추정하는 방법을 탐구한다. 이러한 방법 중 하나로 교란자 분포(confounder distribution)를 RCT를 통해 수집된 것처럼 교란자변수가 처방변수와 독립적으로 분포되도록 하는 샘플의 가중치를 찾는 교란자 균형(confounder balancing) 방법이 있다. 직접적인(direct) 교란자 균형 방법에서는 가중치를 찾기 위해서 처방변수가 주어졌을 때 교란자 조건부 확률분포 간의 모멘트를 일치시키는 방법을 사용한다.

2.2. 인과정규화된 로지스틱 회귀 (Causally-Regularized Logistic Regularization; CRLR)

이진특징 $\mathbf{X} \in \{0,1\}^{n\times p}$ 로부터 이진 레이블 $\mathrm{Y} \in \{0,1\}^{n}$ 을 예측하는 이진분류(binary classification) 문제를 고려한다. CRLR은 특징들 중 하나를 처방변수로, 나머지 특징들을 교란자변수로, 레이블을 결과변수로 취급한 뒤 처방변수의 값과 상관없이 교란자변수가 분포되도록 균형화하는 샘플들의 가중치를 찾는다. 즉, X_j를 j번째 특징에 해당하는 확률변수, X_{-j}를 j번째 특징을 제외한 나머지 특징들의 확률변수라고 하면, $\mathbb{E}[\mathrm{X}_{-j}|\mathrm{X}_j = 1] = \mathbb{E}[\mathrm{X}_{-j}|\mathrm{X}_j = 0]$ 를 만족하도록 아래와 같은 손실함수를 통해 가중치 $\mathrm{W} \in \mathbb{R}^n$를 학습할 수 있다.

$$L_j(\mathrm{X};\mathrm{W}) = \left\| \frac{\sum_{i:X_{ij}=1} W_i \mathbf{X}_{i,-j}}{\sum_{i:X_{ij}=1} W_i} - \frac{\sum_{i:X_{ij}=0} W_i \mathbf{X}_{i,-j}}{\sum_{i:X_{ij}=0} W_i} \right\|$$

이때, 모든 특징들에 대해 최소화하는 가중치를 찾기 위해서, CRLR에서는 다음과 같은 목적식을 최소화하는 가중치를 학습하는 방법을 제안하였다.

$$L_{CRLR}(X;W) = \sum_{j=1}^{p} L_j(\mathrm{X};\mathrm{W})$$

학습된 가중치를 이용하여, 다음과 같이 가중치가 적용된 이진교차엔트로피 손실함수(weighted binary cross entropy loss function) 최소화를 통해 로지스틱 회귀모델을 학습한다.

$$\mathrm{J}(\mathrm{X},\mathrm{Y};\mathrm{W},\beta) = \sum_{i=1}^{n} W_i(Y_i \log\sigma(X_i\beta) + (1-Y_i)\log(1-\sigma(X_i\beta)))$$

3. 제안 방법

CRLR에서는 특징들 중 하나만을 처방변수로 취급하여 나머지 특징들의 1차 모멘트인 평균을 일치시키는 방법을 사용하였다. 본 연구에서는 이 방법을 보다 일반적으로 확장하여 특징 하나뿐만 아니라, 특징들의 조합을 처방변수로 취급한 후, 특정 조합이 주어졌을 때 나머지 조합들의 분포를 일치하는 방법인 CRLR+를 제안한다.

먼저 특징의 인덱스 $\mathrm{P} = \{1,\dots,\mathrm{p}\}$의 멱집합 (power set), 즉 모든 가능한 조합의 집합을 $\mathbb{C} = P(S)$ 로 정의한다. 이때 각 조합 $C \in \mathbb{C} \setminus \{\emptyset\}$에 해당하는 열을 포함하는 특징과 포함하지 않는 특징을 각각 $\mathrm{X}_{i,C}$, $\mathrm{X}_{i,-C}$로 정의하면, 다음과 같이 조합 C에 대한 교란자 균형 손실함수를 정의할 수 있다.

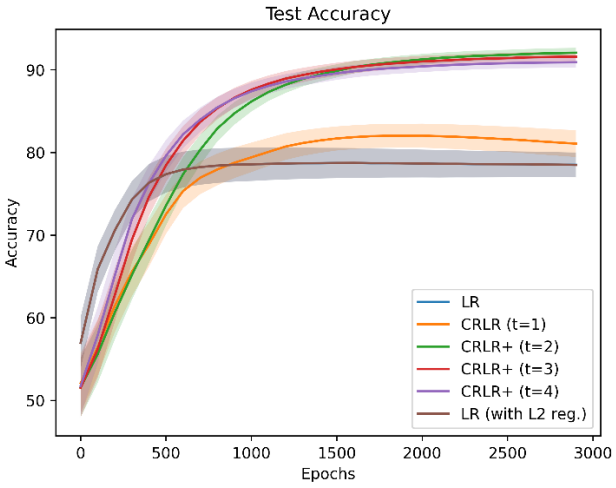
$$L_C(\mathrm{X};\mathrm{W}) = \left\| \frac{\sum_{i:X_{ij}=1,\forall j\in C} W_i \mathbf{X}_{i,-C}}{\sum_{i:X_{ij}=1,\forall j\in C} W_i} - \frac{\sum_{i:X_{ij}=0,\exists j\in C} W_i \mathbf{X}_{i,-C}}{\sum_{i:X_{ij}=0,\exists j\in C} W_i} \right\|$$

이때, 멱집합에 대한 손실함수를 고려하는 것은 계산적으로 비효율적일 수 있으므로, 전체집합 $\mathbb{C}$의 부분집합을 고려하는 방법을 제안한다.

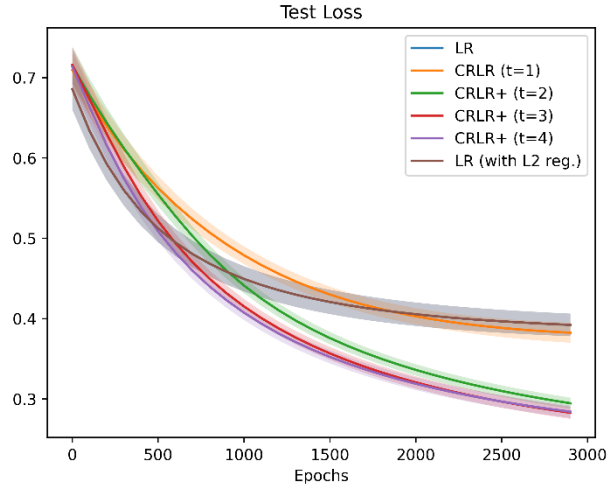
이를 위해, t보다 작거나 같은 관계수(cardinality)를 갖는 공집합이 아닌 집합을 원소로 갖는 멱집합의 부분집합 $\mathbb{C}_t = P_{\le t}(S) \setminus \{\emptyset\}$ 고려하면, t를 조절하여 계산복잡도와 일반화 성능의 정도를 조절할 수 있을 것으로 기대된다. 이렇게 정의된 $\mathbb{C}_\mathrm{t}$에 대해 가중치를 학습하는 손실함수를 아래와 같이 정의한다.

$$L_{\mathbb{C}_\mathrm{t}}(\mathbf{X};W) = \sum_{C\in\mathbb{C}_\mathrm{t}} L_C(\mathbf{X};\mathrm{W})$$

이때, t=1일때는 $L_{CRLR}(X;W)$ 과 일치하므로, 위 손실함수는 CRLR의 가중치 학습 손실함수보다 일반적인 손실함수이다. CRLR에서와 마찬가지로, CRLR+는 제시한 손실함수를 사용하여 가중치 W를 학습한 후, 이진교차엔트로피 손실함수 $\mathrm{J}(\mathbf{X},\mathrm{Y};\mathrm{W},\beta)$ 를 최소화하여 안정적 로지스틱 회귀모델을 학습한다.



[그림 1] 테스트 데이터에 대한 정확도



[그림 2] 테스트 데이터에 대한 교차엔트로피 손실

4. 실험

4.1 실험 환경 및 방법

이 논문에서 제안하는 방법이 풀고자 하는 문제는 훈련시와 테스트시의 분포변화(distributional shift)가 일어나는 경우이다. 이러한 현상이 일어나는 상황에서 실험을 위해 간단한 이진분류 데이터셋을 생성한 후, 훈련 시에 레이블에 강하게 연관되어 있지만, 테스트 시에는 연관성이 떨어지는 변수들을 추가하여, 이 상황에서 제안하는 알고리즘이 테스트 성능을 높이는 데 효과적인지 검증하였다.

이진분류 데이터셋을 생성하기 위해, 먼저 2개의 클래스에 해당하는 2차원 가우시안 분포로부터 5000개의 샘플들을 수집하고, 2:1 비율로 학습 데이터와 테스트 데이터로 나누었다. 이때 3개의 방해변수(nuisance variable)를 추가하였으며, 각 특징변수는 다음과 같다.

- X1, X2: 가우시안 분포로부터 샘플하는 변수
- X3: X1, X2를 선형조합하여 얻어낼 수 있는 불필요한 (redundant) 정보를 담은 변수
- X4, X5: 학습시에는 0.99 확률로 레이블과 연관된 값(1또는 -1)을 가지나, 테스트시에는 0.5 확률로 레이블과 같은 값을 가지는 변수들로, X5는 X4와 절대값은 같고 반대부호를 가지는 변수

제안하는 알고리즘인 CRLR+가 효과적임을 보이기 위해, 비교 알고리즘으로 로지스틱 회귀와, L2 정규화기법을 적용한 로지스틱 회귀, 그리고 CRLR과 함께 실험을 하였다. 또한, 최대조합개수인 t를 조절해가며 테스트 데이터에서의 성능이 어떻게 변화하는지 관측하였다. 가중치 기반의 방법들에서 가중치를 학습하기 위해서는 이진특징이 필요한데, 이때 각 특징을 특징의 평균보다 큰지를 비교하여 이진특징을 생성하였으며, 가중치와 모델을 번갈아가며 업데이트하였다.

4.2 실험 결과

그림 1에서 관측할 수 있듯이, 방해변수가 있는 상황에서는 간단한 이진분류 문제라도 로지스틱 회귀 알고리즘은 테스트시에 정확도 80%를 넘기지 못하는 것을 볼 수 있었다. CRLR은 로지스틱 회귀 알고리즘보다 좋은 성능을 보였으나 최대조합개수가 2 이상일 때 테스트시 더 높은 성능을 보였다. 이는 손실을 보여주는 그림 2에서도 뚜렷하게 비교되는데, 최대조합개수가 1일때는 로지스틱 회귀에 비해 교차엔트로피 손실이 유의미하게 낮아지는 것을 확인할 수 없었으나, 2 이상일때는 확연히 낮아지는 것을 관측할 수 있었다. 또한 t가 일정수준(t=2)보다 높아지면 수렴하는 테스트 성능은 비슷한 정도임을 관측할 수 있었다.

5. 결론

이 논문에서는 특징 하나씩만을 처방변수로 취급하여 교란자 균형 방법을 적용하던 기존 방법을 일반화하여 특징들의 조합을 처방변수로 취급한 뒤 교란자 균형방법을 적용하는 방법을 제안하였다. 또한 실험을 통해 제안하는 방법이 분포변화가 뚜렷한 상황에서 효과적일 수 있음을 보였다.

참고문헌

[1] Zheyan Shen, Peng Cui, Kun Kuang, et al. “Causally regularized learning with agnostic data selection bias”, 2018 ACM Multimedia Conference on Multimedia Conference. ACM, 2018: 411-419.

[2] Pearl, Judea. “Causality.” Cambridge university press, 2009.

[3] Rosenbaum, Paul R. “Observational study.” Encyclopedia of statistics in behavioral science (2005).