**Good Counterfactuals and Where to Find Them: A Case-Based Technique for Generating Counterfactuals for Explainable AI (XAI) /** [**https://arxiv.org/ftp/arxiv/papers/2005/2005.13997.pdf**](https://arxiv.org/ftp/arxiv/papers/2005/2005.13997.pdf)

**0. Counterfactual Explanation – Promise, Problems & Prospects**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Counterfactual Promise:**  Counterfactual (**발생하지 않았지만 발생할 수 있었던** 사건) thinking은 **철학, 심리학, 계산, 법률 등의 분야에서 나름의 역할**을 하고 있으며 관련된 논문도 많이 있다.  **Counterfactual Problems:**   |  |  | | --- | --- | | Prolixity | 현재 많은 XAI 시스템은 random perturbation and search 등의 방법을 이용하여 counterfactual case를 만든다. 즉 **많은 서로 다른 counterfactual case가 주어진 test case를 고려**하여 만들어진다.   * Prolixity는 **예측을 반대로 뒤집는 test case**의 **feature에 최소한의 변화**를 주는 방법을 적용하여 최소화된다. * 따라서 다음과 같은 Loss function을 이용한다. * 여기서 는 **test case에 대한 벡터**이고 는 **counterfactual vector**이다. 는 학습 모델의 f에서의 **flapped(desired) prediction**이다. 또한 는 **balance weight**이다. | | Sparcity | 좋은 **explanatory counterfactual은 sparse할 필요**가 있다. 즉 counterfactual의 **feature 중 최소한의 것만 수정**되어야 한다. | | Plausibility | 생성된 counterfactual은 **주어진 도메인의 valid data point가 아니거나, 어렵거나 불가능한 feature change를 suggest**할 수 있다는 것이다.   * 많은 연구자들은 다음을 제안한다.   + feature를 잠그는 것, 즉 **feature의 변화를 금지**하는 것   + 사용자가 **feature weight에 대한 입력을 제안할 수 있게** 하는 것   Counterfactual은 문제 도메인에 대한 실제 경험이므로, XAI 시스템은 **training data를 이용하여 적당한 counterfactual을 찾아야** 한다. |   **CBR’s Prospect for Counterfactuals:**   * 많은 XAI 기술에서는 random perturbation을 사용하는데, 이런 기술은 **case-based 접근을 이용하여 이익**을 얻을 수 있는 **sparcity 및 plausibility problem**이 생길 수 있다. * CBR은 factual case를 이용하여 예측에 대해 성공적으로 설명했고, **sparse하고 plausible한 counterfactual case**를 생성할 수 있다. 그러나 실제로 사용할 때는 **case-base가 실제로 좋은 counterfactual을 포함하는지 결정**해야 한다.   + 좋은 counterfactual은 **query case와 2 feature 넘게 다르지 않은 NUN으로 정의** |

**1. Experiment 1. Plotting Counterfactual Potential**

|  |
| --- |
| 여기서는 **20개의 classic dataset의 counterfactual potential**을 계산하였는데, 그 결과는 다음 표와 같다.   * Good counterfactual인 **1-diff와 2-diff는 드물고**, 따라서 **case-based approach는 좋은 방법이 아니다.** |

**2. A Case-Based Technique for Good Counterfactuals**

|  |
| --- |
| **Overview:**  counterfactual들은 **plausibility**를 보장하기 위하여 주어진 **training data에서 현실에 기반을 둔 타당한 데이터들**이어야 한다.   * 따라서 “**good counterfactual의 패턴을 재사용하는 형태의 새로운 case-based counterfactual-XAI 기술”**을 만들었다.   + 이 패턴들은 **case-base에 이미 존재**한다.   + 새로운 **target problem**과 그 문제에 대한 **solution**에 대해 설명할 수 있는, **유사한 counterfactual**을 생성한다.   + 새로운 counterfactual을 생성할 때, **이미 존재하는 good counterfactual**들은 **어떤 feature를 적용해야 하는지에 대한 힌트**를 제공한다.   + **Explanatory competence**라는 개념에 의존한다.   **Explanatory Competence:**  Predictive competence 또는 simply competence의 개념과 같이 **어떤 case-base에도 적용**할 수 있는 개념이다. Explanatory competence의 basic unit은 **explains(c, c’)**로, **어떤 케이스 c는 c’라는 문제의 해결 방법을 설명하기 위해 사용될 수 있다**는 것을 의미한다.  Case-base c의 **explanatory competence**는 **(1)**과 같은 coverage set으로 나타낼 수 있으며, **degree of explanatory competence**는 **(2)**로 나타낼 수 있다. |

**[NEXT PAGE]**

**[PREVIOUS PAGE]**

|  |
| --- |
| **Leveraging Counterfactual Cases for Explanation:**  우리는 **대부분의 case-base가 query가 좋은 counterfactual과 연관되는 예시를 제공**한다고 예상할 수 있다.   * **(case, corresponding good counterfactual)을 XC의 explanation case라고 한다.** * 예를 들어 Abalone dataset에서는 좋은 counterfactual은 드물지만 **similarity tolerance를 0.02로 설정하면 20% 이상의 case가 좋은 counterfactual과 연관**될 수 있다. * 어떤 주어진 case-base에 대해서도, counterfactual이 생성되는 동안 사용할, 이 설명에 대응되는 case-base를 만들 수 있다. * 아래 수식에서 정의에 다르면 **explanation case는 서로 대칭**이다. **(각 case는 상대방 case의 query 또는 counterfactual이다.)**     위 그림은 **case C**와 **corresponding XC: xc(x,x’), xc(y,y’), xc(z,z’)**에 대한 2-class case-base를 보여준다.   * 머신러닝 모델에 의해 분류된, 설명되어야 하는 **2개의 query p와 q**에 대해, **class** **C에는 이들에 대한 좋은 counterfactual이 없다**고 가정한다. 즉 **새로운 좋은 counterfactual**을 찾아야 한다.   **A Case-Based Approach to Generating Good Counterfactuals**  여기서는 좋은 counterfactual을 생성하기 위한, **가까운 explanation case에 대한 retrieving, reusing, revising**을 이용한, **classical case-based reasoning** 접근 방법을 제안한다. |

**[NEXT PAGE]**

**[PREVIOUS PAGE]**

|  |
| --- |
| 1. **query가 p와 가장 비슷한 XC case**를 정의한다. **(Figure 1의 xc(x,x’)에 해당)**   * xc(x,x’)는 x’라는 좋은 counterfactual을 가지고 있고 p는 x와 유사하므로, **x’는 새로운 counterfactual p’에 대한, p를 설명하기 위한 적절한 basis**이다. * x와 x’의 class를 구분하게 하는, **x, x’ 사이의 difference feature**들은 p’를 생성하는 데 결정적인 역할을 한다.   2. xc(x, x’)의 각 match feature에 대하여 **p**에 있는 이 값들을 새로운 counterfactual **p’**에 복사한다. 비슷하게 xc(x, x’)의 각 difference feature에 대해서 **x’**에 있는 이 값들을 **p’**에 복사한다.   * 이때 **p’는 p와 x’의 feature value의 조합**으로, feature의 차이가 최대 2개이므로 **candidate good counterfactual**이다.   3. p’가 good counterfactual이 되게 하기 위하여 **p와는 서로 다른 class**에 있어야 한다.   * p’의 class는 머신러닝 모델에 의해 결정되고, **p와 class가 다르다면** p를 설명할 수 있는 good counterfactual이 된다.   4. p’가 retrieval/reuse된 이후에 **target query와 class가 서로 같을 수 있다.**   * q로부터 xc(y, y’)를 재사용하여 생성된 **새로운 counterfactual q’**가, difference feature의 차이가 class를 바꿀 만큼 충분하지 않아서 **q와 서로 같은 class**일 수 있다. **(Figure 1a)**   5. q’가 **valid counterfactual이 아니기** 때문에 **difference feature의 값에 대한 revise를 위해 adaptation step**을 수행할 수 있다.   * q에 대한 **feature difference 증가 없이 match feature를 변경할 수 없다**. * q’에 대한 difference feature를 다양한 방법으로 revise할 수 있다.   + q와의 **거리를 증가시키기 위한 perturbation**   + q가 y’와 서로 같은 class가 될 때까지 **q의 ordered nearest neighbours에 대하여 iterate**하는 것     - 이때 각 nearest neighbour의 difference feature의 값들은 **새로운 candidate q’’**를 생성하게 한다. 이때 **q’’의 class가 q와 달라지면 adaptation이 성공적으로 종료**된다.     - 어떤 neighbour도 class change를 유도하지 않으면 adaptation은 실패한다.     - Figure 1(a)에서는 neighbor **nn의 difference feature value**가 q’’를 생성하는 데 사용되었고, 이때 **class change**가 발생하여 **q’’는 q의 good counterfactua**l로 판명되었다. |

**3. Experiment 2: Evaluating Explanation Competence**

|  |
| --- |
| **Method: Data & Procedure**  각 dataset은 **서로 다른 복잡도**(class의 개수, feature 및 training example)의 classification task를 의미한다.  핵심적인 evaluation metric은 **(a) good counterfactual (explanatory competence)과 연관되는 target/query case의 비율, (b) 새로 생성된 good counterfactual과 target/query 사이의 거리**이다.    **Results & Discussion: Explanatory Competence**  각 dataset에 대해 Explanatory competence (설명되는 query의 비율)은 아래 그림에 나타나 있다. **(dataset 즉 baseline, retrieval과 adaption을 이용하여 생성된 synthetic counterfactual)**     * 이 결과는 explanatory competence가 **case-based counterfactual 접근 방법에 의해 얼마나 좋아질 수 있는지**를 나타낸다. |