연관된 특성을 조작하여 결정경계를 조사하는 반사실 설명

김희동[O1], 이성환[1]
[1]고려대학교 인공지능학과
hd_kim@korea.ac.kr, sw.lee@korea.ac.kr

Manipulating Related Features to Probe the Decision Boundary for Counterfactual Explanation

Hee-Dong Kim[O1], Seong-Whan Lee[1]
[1]Department of Artificial Intelligence, Korea University
hd_kim@korea.ac.kr, sw.lee@korea.ac.kr

요 약

고성능의 근래 인공지능 모델은 퍼셉트론에 기반한 신경망 구조로 높은 성능을 달성하고 있지만 신경망 구조의 내재적 특성인 블랙박스성 때문에 예측에 대한 해석이 어렵다는 문제가 있다. 이 문제 해결을 위해 모델의 결정을 변화시킬 새로운 입력 예시를 보여주며 결정경계를 설명하는 반사실 설명 기술이 연구되고 있으며, 모델의 투명성을 높여주어 결정이 중요한 분야에 인공지능 모델의 활용을 증대시키고 있다. 하지만 만들어진 반사실 설명은 종종 현실 데이터와 동떨어진 예시를 만들어내는 경향이 있는데, 그 이유는 지엽적인 특성만 변경시킬 경우 다른 특성과 조화되지 않아 인간이 보기에 이질감이 느껴지는 예시를 만들어내기 때문이다. 본 연구는 현실 문제에 직접 활용될 수 있도록 이질감이 없는 현실 데이터와 유사한 예시를 만들어내기 위해 목표 특성과 연관된 특성을 함께 조작하여 자연스러운 반사실 예시 생성을 목표로 한다. 이러한 반사실 설명 예시는 모델의 결정이 중요한 분야에서 결정경계를 설명함으로써 모델 활용도를 증대시키는데 일조한다.

1. 서 론[1]

근래 인공지능 모델은 특정 태스크에서 인간의 능력을 뛰어넘을 정도로 높은 성능을 보이며 빠르게 발전하고 있다. 하지만 인공지능 모델은 결정이 중요한 금융, 의료, 법률 분야에서 특히 활용도가 낮은 경향이 있는데, 그 이유는 잘못된 결정에 대한 책임소재가 불분명하기 때문이다. 이 책임소재를 설명하기 위해 설명 가능한 인공지능(XAI: eXplainable Artificial Intelligence) 분야가 활발하게 연구되고 있으며, 반사실 설명(Counterfactual explanation)이라는 모델의 결정을 바꾸는 예시를 통한 결정경계 설명 연구가 활발하게 진행되고 있다[1]. 이러한 예시는 대출심사 등 사용자에게 모델의 결정을 바꿀 수 있는 실질적인 목표 예시를 제공해 줄 수 있기 때문에 실용적인 연구분야이다. 하지만, 기존 연구에서는 반사실 설명이 현실 데이터와 동떨어진 경우가 많아서 사용자에게 혼란을 주는 설명을 만들어 낸다는 문제가 있다[2]. 이 문제를 해결하기 조작하는 특성의 개수를 줄이거나[3] 특성의 조작량을 줄인 설명을 만들어 내었지만[4], 정형 데이터에서만 효과를 거두고 있고, 이미지와 같은 비정형 데이터 에서는 효과적이지 않다는 문제가 남아있다[5].

본 연구에서는 이미지 데이터에서도 반사실 설명이 더 실용적으로 사용될 수 있도록 현실 데이터와 유사한 예시 생성으로 목표로 하는데, 이러한 예시를 현실세계에 있을법한 예시(Plausible explanation)라고 불린다. 연구의 방법론은 생성모델[6]을 기반으로 잠재공간(Latent space)에서 특정 클래스의 데이터 분포를 구하고, 그 분포를 통해 입력을 어떻게, 얼마나 수정해야 하는지 수정방향과 목표치를 제시한다. 이 과정에서 더 현실세계에 있을법한 예시로 보여지기 위해 연관된 특성을 함께 변경하는데, 이를 위해 적대적 생성 신경망(Generative Adversarial Network)의 잠재공간에서 목표 클래스로의 방향벡터(Direction Vector)에 클래스의 중심(Center-of-Target)을 사영시켜 목표 클래스의 특성을 가장 많이 반영하는 지점(Projection Point)을 계산한다. 최종적으로 만들어진 반사실 설명이 분류기의 결정경계를 넘었는지 검사하여 설명의 타당성까지 검증을 완료한다.

[1] 이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00984, 플러그앤플레이 방식으로 설명가능성을 제공하는 인공지능 기술 개발 및 인공지능 시스템에 대한 설명 제공 검증)

결론적으로, 입력 샘플에서 클래스를 변경시키는 경로중에서 대상의 중심과 가장 가까운 지점을 찾는 것이다.

본 연구는 이미지와 같은 비정형 데이터 영역에서도 효율적으로 동작하도록 고안되었으며, 이미지 데이터에서도 동작하는 IM1 점수[7]와 타당성 점수로 평가하였다. IM1 점수에서 고성능을 달성할수록 본 연구의 목표인 현실세계에 있을법한 예시로 평가된다. 본 연구를 통해 반사실 설명이 현실세계에 있을법한 예시를 생성하기 위해 연관된 특성 조작의 가치를 보여주고자 한다.

2. 방법론

본 연구는 결정을 바꿀 예시를 생성할 때 연관된 특성을 최대한 많이 반영시키는 것을 목표로 한다. 이를 위해 잠재공간에서 클래스의 결정경계를 탐색하기 위해 입력에서부터 분류기의 결정을 효율적으로 바꿀 수 있게 해줄 수정방향과 수정 정도를 구한다. 이때 수정 정도는 입력과 근접한 잠재 벡터를 구하기 위해 고려해야 하는 요소이다. 잠재공간에서 근접한 지점은 생성 결과도 비슷하다는 특성을 이용해서 이 문제를 잠재공간에서의 최소자승법(Least-Squares Solution) 으로 변환하여 해결한다. 최소자승법 계산을 위해 타겟의 중심과 방향벡터를 구해야 한다. 방법론의 개요는 그림 1에 묘사되어 있으며, 2.1, 2.2, 2.3절에서 자세히 다루도록 한다.

2.1. 타겟의 중심 계산 알고리즘

타겟의 중심(CT)은 생성모델의 잠재공간에서 타겟 클래스의 연관된 특성을 가장 많이 가지고 있는 지점이다. 이 지점 계산을 위해 잠재공간의 각 지점이 생성모델을 통과했을 때 타겟 클래스로 분류되는지 검사하여 잠재공간의 지점별 레이블을 확보한다. 이후 잠재공간에서 타겟 클래스로 분류될 지점들을 모두 모아 세트를 구성한 후 요소별 분포를 구해낸다. 이 요소별 분포에서 확률값이 가장 높은 지점이 타겟의 중심이다. 다음에 확률값이 가장 높은 지점을 구하기 위해 가우시안 혼합 모델(Gaussian Mixture Model)을 활용하여 각 요소별 확률밀도함수를 구한다. 다음에 확률값이 가장 높은 지점인 극대값을 구하기 위해 확률밀도함수를 활용하여 각 요소별 극대값을 구한다. 마지막으로, 각 요소별 극대값으로 이루어진 지점을 타겟의 중심으로 지정한다.

2.2. 방향벡터 계산

방향벡터는 그림 1에서와 같이 입력 지점으로부터 타겟 클래스로 변경하기 위한 수정방향이다. 이 수정방향은 클래스의 중심들을 이어서 만들어진다. 예를 들어, 클래스 c로부터 클래스 c'으로 변경하는 경우, 방향벡터는 다음 수식처럼 클래스 중심의 차로 계산된다.

$$n = CT_{c'} - CT_{c}.$$

구해진 방향벡터 방향으로 입력 벡터를 이동시키는데, 생성모델의 잠재공간은 표준 정규분포를 따르기 때문에 $\frac{CT_{c'}-CT_{c}}{\|CT_{c'}-CT_{c}\|_2}$ 로 정규화 하여 사용한다.

2.3. 사영 연산

마지막 단계로 사영 연산으로 사영 포인트를 구한다. 그림 1과 같이 타겟의 중심을 입력 지점을 지나고있는 방향벡터로 사영하여 사영 연산을 한다. 이 사영 지점은 타겟의 중심으로부터 가장 거리가 짧은 지점을 계산하기 위함으로, 입력의 특성을 보존하면서도, 분류기의 예측 결과를 바꾸며, 타겟 클래스의 연관된 특성을 가장 많이 보존하는 지점이다. 사영 연산으로 반사실 설명을 만들어낼 잠재벡터($z_{c'}$)은 다음과 같이 계산된다.

$$z_{c'} = z_c + \lambda n, \quad \lambda = n \cdot \left(CT_{c'} - z_c\right).$$

위 수식에서 λ는 사영 연산을 하기 위해 방향벡터와 입력벡터와 대상의 중심을 잇는 벡터의 내적 연산으로 구해진다. 그리고, λ는 방향벡터를 따라 조작할 명시적인 조작량이 된다.

3. 실험

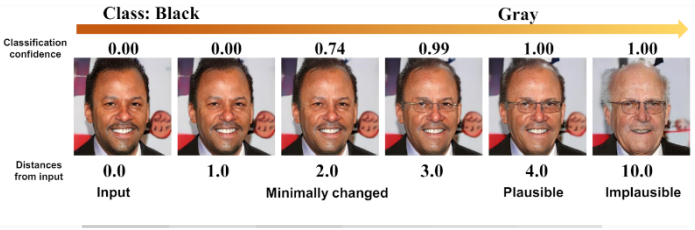


그림 2는 이미지 데이터에서 검정색 머리에서 회색 머리로 클래스를 변경하는 과정이다. 최소수정(Minimally changed)는 잠재공간에서 z_c와 $z_{c'}$ 사이에서 보간을 통해 추출을 하고, 분류기의 결정경계를 넘은 직후의 데이터를 나타낸다. 그리고 $z_{c'}$ 로부터 생성된 이미지는 현실세계에 가장 있음직한 적정 수정된 반사실 설명이다. 하지만, 우측 그림 2의 우측 예시는 $z_{c'}$ 을 넘어 외삽하게 되어 과도한 수정을 하게 되면 오히려 현실세계에 있음직하지 않은 예시(Implausible)이다. 과도한 수정은 잠재공간에서 입력과 크게 동떨어진 예시가 만들어져 정체성을 잃게 만든다.

따라서, 최소수정과 현실세계에 있음직한 반사실 설명 두가지가 분류기의 결정경계를 효과적으로 설명하는 반사실 설명이다. 게다가, 회색 머리로 변경할 때 연관된 특성인 나이, 안경을 함께 조작해 줌으로써 더 현실세계이 있음직한 예시로 만들어 준다. 이러한 연관된 특성을 가지는 이유는 대상의 중심으로부터 사영을 통한 지점은 대상의 중심으로부터 가장 가깝기 때문에 연관된 특성을 많이 포함하기 때문이다.

그림 3은 숫자 데이터에서 최소수정과 적정 수정을 나타낸다. 파란색 점선 상자는 특성이 줄어든 부분을 나타내며, 주황색 점선 상자는 특성이 강조된 부분을 나타낸다. 각 부분은 타겟 클래스로 분류되기 위해 수정되어야 할 부분을 나타내며, 적절히 조작된 반사실 예시는 분류기가 결정경계를 결정할 때 어떤 부분을 얼마나 집중하고 있는지 예측하게 해준다. 마찬가지로, 최소수정은 입력에 비해 클래스의 예측 결과를 바꾸기 위한 최소 수정량이며, 현실세계에 있음직한 예시는 목표 클래스의 특성을 더 많이 가진 설명을 만들어 낸다.

4. 결론 및 향후 연구

이 논문에서는 연관된 특성을 함께 조작하여 분류기의 결정경계를 넘으면서도 현실 세계에 있음직한(Plausible) 반사실 설명을 만들어낸다. 연관된 특성을 최대한 많이 포함시키기 위해 생성모델의 잠재공간에서 대상의 중심을 정의하고, 입력 지점으로부터 클래스를 변경하기 위한 방향벡터로 사영 하여 최소 거리의 지점을 찾아내어 연관된 특성을 가장 많이 포함시키도록 유도하는 것이다. 사영 연산으로 인해 명시적으로 정해진 수정량은 과도한 수정을 자연스럽게 예방하게 되고, 현실세계에 있음직하지 않을 예시가 만들어지는 것을 방지한다. 따라서, 본 논문을 통해 연관된 특성의 중요성을 강조하고, 실용적인 반사실 설명을 만들어낼 수 있다.

참 고 문 헌

[1] S. Wachter, B. Mittelstadt, and C. Russel, “Counterfactual explanations without opening the black box: Automated decisions and the GDPR,” Harv. JL & Tech. 31 (2017): 841.

[2] Eoin M. Kenny and Mark T. Kean, “On generating plausible counterfactual and semi-factual explanations for deep learning,” Proceedings of the AAAI Conference on Artificial Intelligence (2021) Vol. 35. No. 13.

[3] H.-G. Jung, S.-H. Kang, H.-D. Kim, D.-O. Won, and S.-W. Lee, “Counterfactual explanation based on gradual construction for deep networks,” Pattern Recognition 132 (2022): 108958.T.

[4] Laugel, M.-J. Lesot, C.Marsala, X. Renard, M. Detyniecki "Comparison-based Inverse Classi-fication for Interpretability in Machine Learning,” 17th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2018) p.100-111.

[5] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, S. Lee, “Counterfactual visual explanations,” Proceedings of the International Conference on Machine Learning (pp. 2376-2384). PMLR.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial networks,” Proceedings of ACM 63.11 (2020): 139-144.

[7] A. V. Looveren and J. Klaise, “Interpretable counter-factual explanations guided by prototypes,” Proceedings of the ECML PKDD (2021) pp. 650-665.