동적 가중치 부여 다중 비용 함수를 이용한 범주 불균형 데이터 분류

신창욱 [1O], 권오욱 [2], 차정원 [1]

창원대학교 [1], 한국전자통신연구원 [2]

papower1@changwon.ac.kr, ohwoog@etri.re.kr, jcha@changwon.ac.kr

Category Imbalance Classification using Dynamic Weighted Multiple Cost Functions

Chang-Uk Shin[1O], Oh-Woog Kwon[2], Jeong-Won Cha[1]

Changwon National University[1], Electronics and Telecommunications Research Institute[2]

요 약

본 연구에서는 범주 불균형 텍스트 분류를 위한 새로운 학습 방법을 제안한다. 비용 민감 학습 방법을 이용해 모델을 학습하고, 학습 중 나타난 경사 수치와 성능을 분석한다. 그리고 그 분석에 기반해, 둘 이상의 비용 함수를 조합하는 새로운 비용 민감 학습 방법을 제안한다. 제안 방법은 두 다른 특성을 가진 비용 함수를 조합하여 높은 성능을 달성할 수 있었다. 기존 인공신경망 학습에는 단 하나의 비용 함수를 사용하였다. 본 연구의 결과로부터, 둘 이상의 비용 함수를 이용해 모델을 학습하여 높은 성능을 달성할 수 있음을 보인다.

1. 서 론

범주 불균형 데이터셋을 이용해 분류 모델을 학습하기 위한 연구가 지속되고 있다. 높은 성능의 분류 모델을 학습하기 위해서는 대용량 고품질의 분류 데이터셋이 필요하다. 그러나, 대부분의 자연 발생 데이터에는 범주 편향이 존재한다. 그리고 그런 범주 불균형 데이터셋을 이용해 모델을 학습하게 되면 학습된 모델에 편향이 그대로 전이되어, 모델의 성능이 하락하게 될 가능성이 있다. 따라서, 주어진 데이터셋을 샘플링하여 범주 불균형이 해소된 데이터셋을 작성하거나, 혹은 편향을 완화할 수 있는 학습 방법 등을 이용해 인공신경망 분류 모델을 학습한다.

인공신경망의 경우, 비용 함수를 변형하여 범주 불균형 분류를 수행할 수 있다. 저빈도 클래스에 높은 가중치를 부여함으로써 고빈도 클래스에 편향되었던 모델의 편향을 완화하는 것이다. 빈도 이외에도, 학습 중 클래스별 정밀도 등을 이용해 비용 민감 학습을 수행한 사례도 있었다[1]. 이러한 방법들은 비용 민감 학습(cost sensitive learning)이라고 불린다.

본 연구에서는 대화 내 발화의 감정과 화행을 분류하는 문제에 비용 민감 학습 방법을 적용하고, 분석한다. 그리고 그 분석에 기반해 새로운 비용 민감 학습 방법을 제안한다.

2. 제안 방법

본 논문에서는 두 가지 비용 민감 학습 방법을 이용해 분류 모델을 학습하고, 그 학습 특성에 기반해 설계된 새로운 학습 방법을 제안한다.

두 가지 비용 민감 학습 방법 중 첫 번째 학습 방법은 드리클레 분포(Dirichlet distribution)를 이용한 동적 가중치 부여 방법이고, [2]에서 제안되었다. 기존 학습 코퍼스 내 발생 비율 등으로 설정된 가중치를 부여하는 대신, 드리클레 분포에서 가중치를 추출해 사용함으로써 성능을 개선하였다. 동적 가중치 부여 방법은 학습 중 큰 경사도가 고르지 않게 발생하였고, 이 특성이 학습 속도 개선에 이용될 수 있으리라 판단하였다.

Fuzzy Label은 [3]에서 영감을 받아 본 연구에서 제안하는 학습 전략이다. 정답 지표에 특정 수치를 부여하고, 정답 이외의 지표는 학습 코퍼스 내 발생한 빈도에 기반해 수치를 부여한다. 기존 정답 이외의 지표에 0을 부여하는 것에 비해 안정된 학습을 수행하고, 또한 높은 성능을 달성할 수 있으리라 기대하였다.

본 연구에서 수행된 실험을 통해, 두 개의 각기 다른 비용 민감 학습 전략들이 서로 다른 특성을 가지고 있음을 알 수 있었다. 이 두 특성이 가진 장점을 취한다면, 더욱 개선된 모델을 작성할 수 있을 것이라 판단하였다. 따라서 위 기술한 두 비용 민감 학습 방법을 결합한 새로운 비용 민감 학습 방법을 제안한다. 두 비용 민감 학습 방법의 결합 방법을 변경해가며 실험을 수행하였다.

3. 실 험

3.1. 실험 설정

본 연구에서 제안하는 학습 방법의 유용성을 확인하기 위해 범주 불균형 데이터를 대상으로 실험하였다. 본 연구에서 실험에 사용한 데이터셋은 DailyDialog 대화 데이터셋[4]이다. 최초 영어로 공개된 데이터셋을 직접 번역하고, 중복 제거 등의 처리를 수행하였다. DailyDialog 데이터셋으로 인공신경망 모델을 학습하기에는 데이터셋의 양이 부족하다고 판단하였다. 따라서, 이를 보강하기 위한 방법으로 [5]에서 제안한 디노이징 메커니즘을 적용하여 데이터셋을 증축하였다. 그리고 마지막 턴의 감정과 화행을 추론하는 문제이기에, 3턴 이상의 대화는 마지막 대화를 제거해가면서 증축을 수행하였다.

마지막으로, 감정과 화행 중 더욱 편향이 심한 감정 분류에서, 최고빈도로 발생하는 '감정 없음' 지표의 샘플을 임의 제거함으로써 그 편향을 일부 완화하였다. 이는 데이터셋 중 약 80%를 차지하는 '감정 없음' 지표로 인해 모든 입력에 대해 '감정 없음'을 출력하는 현상이 발생하였고, 이를 해결하기 위해 적용한 것이다. 최종 실험에 사용한 데이터셋의 통계량을 [표 1, 2, 3]에 정리하였다.

표 1 최종 실험에 사용된 데이터셋의 통계량

구분	수량	단위
대화	43,914	대화
발화	318,837	발화
평균 발화 길이	7.03	단어/발화
평균 대화 길이	7.26	발화/대화

표 2 최종 실험에 사용된 데이터셋의 감정 지표 분포

구분	수량	비율(%)
감정 없음	21,957	50.00
행복	16,843	38.35
놀람	2,138	4.87
슬픔	1,254	2.86
화남	1,177	2.68
혐오	336	0.77
두려움	209	0.48

표 3 최종 실험에 사용된 데이터셋의 화행 지표 분포

구분	수량	비율(%)
inform	22,969	52.30
question	9,308	21.20
commisive	5,855	13.33
directive	5,782	13.17

분류를 위한 인공신경망 구조는 [2]에서 제안된 구조를 채용한다. [6]의 구조를 기반으로 하고, 멀티 태스크 학습과 비용 함수를 수정하였다.

본 논문에서는 제안하는 비용 함수의 효용성을 보이기 위해, 동적 비용 민감 학습 방법, fuzzy label 그리고 그 변형 형태를 실험하였다.

동적 가중치 부여는 학습 데이터셋 내의 출현 비율을 이용하였다. 즉, 학습 데이터셋에 고빈도로 나타난 클래스의 샘플에는 적은 가중치를, 저빈도로 나타난 클래스의 샘플에는 높은 가중치를 부여한다. 따라서, [식 1]과 같이 연산한다. 식에서 y는 원 핫 표현(one-hot representation)으로 주어진다. 따라서, 가중치 벡터 중 해당 샘플의 정답에 위치한 값만이 실제 비용에 영향을 미친다.

$$\mathcal{L} = \text{cross\_entropy}(w\tilde{y}, y) \quad (1)$$

Fuzzy label은 모델의 추론 결과에 곱연산되는 동적 가중치와 달리, 정답 지표로써 부여된다. 따라서 항상 단 하나의 값이 1이고 나머지가 0으로 부여되는 원 핫 표현이 아니라, 모든 클래스에 값을 부여하는 것이다. Fuzzy label의 값 또한 학습 데이터셋 내의 비율을 이용하며, 다음과 같이 계산된다.

매 샘플의 정답에 해당하는 클래스에는 고정된 수치 0.7을 부여한다. 그리고 0.3을 정답이 아닌 클래스에 나누어 부여한다. 이를테면 카테고리의 수가 3이고, 첫 번째 클래스가 정답이라면, [0.7, x, y]로 정답을 부여한다. 그리고 두 번째 클래스와 세 번째 클래스의 학습 코퍼스 내 비율이 2:1이라면, 최종 [0.7, 0.2, 0.1]이 된다.

제안하는 비용 결합 방법을 구현하기 위해 여러 방법을 실험하며 검증하였다. 첫 번째 방법은 두 학습 방법을 동등하게 고려하도록 설정한 방법이다. 항상 두 비용에 0.5를 곱한 값이 최종 비용이 되도록 설정한다[식 2].

$$\mathcal{L}_{\text{total}} = 0.5\mathcal{L}_{\text{dirichlet}} + 0.5\mathcal{L}_{\text{fuzzy}} \quad (2)$$

두 번째 제안 방법은 모델로 하여금 두 학습 방법 중 적절한 것을 선택할 수 있도록 학습하는 방법이다. 학습 중 역전파 알고리즘을 통해, [식 3]의 α가 적절히 업데이트될 수 있도록 하였다.

$$\mathcal{L}_{\text{total}} = \alpha\mathcal{L}_{\text{dirichlet}} + (1-\alpha)\mathcal{L}_{\text{fuzzy}} \quad (3)$$

세 번째 방법은 두 번째 방법과 같이 두 비용 함수의 비율을 모델이 학습하도록 하고, 학습된 비율을 드리클레 분포에 통과해 나온 값을 가중치로 사용하는 방법이다[식 4]. 즉, 모델이 학습한 비용 함수 결정 비율에는 오류가 있다고 가정하고, 여기에 드리클레 분포로 다시 한번 노이즈를 추가하여 일반화 성능을 개선하고자 하는 것이다[2].

$$\mathcal{L}_{\text{total}} = d(\alpha)\mathcal{L}_{\text{dirichlet}} + (1-d(\alpha))\mathcal{L}_{\text{fuzzy}} \quad (4)$$

3.2. 실험 결과 및 분석

표 4 실험 성능

구분	비용 민감 학습 방법	화행 분류 (MAP)	감정 분류 (MAP)
1	-	62.42	43.60
2	동적 가중치 부여	64.65	44.46
3	Fuzzy Label	63.47	47.16
4	결합 (1:1로 고려)	65.44	48.36
5	결합 (비율 학습)	**66.44**	48.55
6	결합 (비율 학습 + 드리클레)	63.73	**54.44**

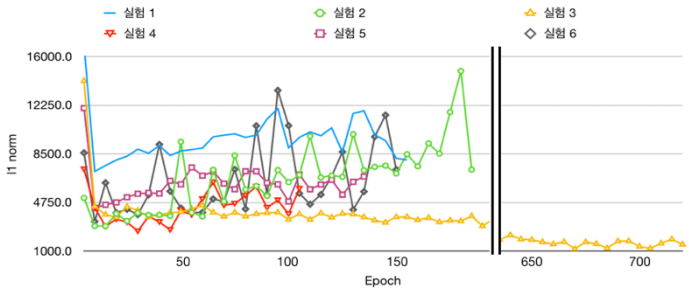

그림 1 실험 중 발생한 경사도

위 [표 4]는 동적 가중치 부여 방법과 제안하는 fuzzy label, 그리고 그 둘을 함께 사용한 제안 비용 함수 결합 방법의 성능을 정리한 표이다. 그리고 [그림 1]은 학습 중 발생된 경사도의 l1-norm을 그린 것이다.

[그림 1]에서, 동적 가중치 부여 방법(실험 2)은 학습 초반부를 제외하고는 높은 경사도로 모델의 가중치를 매번 크게 변경시킨다. 큰 경사도는 인공신경망 모델을 빠르게 극값에 도달할 수 있도록 도와주지만, 최적의 지점을 지나칠 가능성이 있다.

반면, fuzzy label은 경사도가 작고 고르게 발생하며, 비교적 오래 학습이 진행된 것을 알 수 있다. 이는 기존 경사도와 비용은 정답으로 원 핫 표현이 주어지고, 그 중 하나의 값만이 경사도 생성에 영향을 주는데 반해, 제안 fuzzy label은 항상 모든 클래스가 경사도 생성에 기여하기 때문이라고 분석된다. Fuzzy label의 성능이 더 높은 것으로 미루어 보았을 때, 이렇게 고르고 작은 경사도로 학습을 진행하는 것이 주어진 환경에 더 적합하였다고 볼 수 있다.

제안하는 비용 결합 방법은 위 두 학습 방법을 함께 사용한 것으로, 그 특성이 각각 반영되어, 학습 시간과 경사도 norm이 각각 두 방법의 중간에 위치할 것이라 예상하였다. 그러나, 경사도 norm의 값이 두 방법의 사이에 위치한 것과 달리, 학습 시간은 상당히 줄었고, 또한 성능도 개선되었다. 이는 제안 방법이 최초 임의의 모델 가중치 시작 위치에서 더욱 가까운 극값을 더욱 잘 찾아내었음을 의미한다.

4. 결론 및 향후 연구

본 연구에서는 새로운 범주 불균형 학습 방법과 여러 비용 함수를 결합하는 방법을 제안하였다. 기존의 연구에서는 하나의 비용 함수를 이용해 인공신경망 모델을 학습시켜 왔다. 본 연구에서는 두 개의 비용 함수를 이용해 하나의 모델을 학습하는 새로운 방식을 탐구하였다. 설정된 실험에서 제안 방법을 사용하였을 때 단일 비용 함수 사용 대비 감정 9.98%p, 화행 1.79%p의 성능 향상을 달성할 수 있었다.

본 연구에서 제안한 방법으로 수행된 실험의 결과는 적게 학습될수록 화행 분류의 성능이 높고, 반대로 길게 학습될수록 감정 분류의 성능이 높은 경향을 보였다. 즉, 이 두 문제가 각기 다른 시점에 최적화되는 것으로 분석된다. 따라서, 이러한 적합 정도의 차이에 따라 차별적으로 학습하거나, 혹은 최적으로 학습이 완료된 일부 문제에 과적합하지 않으면서도 아직 학습되지 않은 문제를 학습할 수 있는 방법에 대해 연구가 필요할 것으로 보인다.

Acknowledgement

이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임(2019-0-0004, 준지도학습형 언어지능 원천기술 및 이에 기반한 외국인 지원용 한국어 튜터링 서비스 개발).

참 고 문 헌

[1] Yue Geng and Xinyu Luo, “Cost-Sensitive Convolution based Neural Networks for Imbalanced Time-Series Classification”, arXiv, 2018.

[2] 신창욱, 차정원, “범주 불균형 분류 문제를 위한 동적 비용 민감 학습”, 한국 컴퓨터 종합학술대회 논문집, 2019.

[3] David Warde-Farley and Ian Goodfellow, “Adversarial Perturbations of Deep Neural Networks”, Pertubations, Optimization, and Statistics, pp. 311, 2016.

[4] Yanran Li et al., “DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset”, arXiv, 2017.

[5] 김태형 외, “디노이징 메커니즘을 통한 한국어 대화 모델 정규화”, 정보과학회논문지, 45권, 6호, pp. 572-581, 2018.

[6] Chang-Uk Shin and Jeong-Won Cha, “End-to-End Task Dependent Recurrent Entity Network for Goal-oriented Dialog Learning”, Computer Speech & Language, vol. 53, pp. 12-24, 2019.