# 범주 불균형 분류 문제를 위한 동적 비용 민감 학습

신창욱 ○, 차정원
창원대학교
papower1@changwon.ac.kr, jcha@changwon.ac.kr

# Dynamic Cost Sensitive Learning for Imbalanced Text Classification

Chang-Uk Shin, Jeong-Won Cha
Changwon National University

**요 약**

본 연구에서는 범주 불균형 분류 문제를 해결하기 위한 새로운 학습 방법을 제안한다. 범주 편향을 완화하기 위해 학습 데이터셋 내 범주의 비율에 기반하는 드리클레(Dirichlet) 분포를 이용해 동적으로 가중치를 부여하였다. 대화 내 발화의 감정과 화행을 분류하는 문제에 제안하는 비용 민감 학습 방법을 적용하였을 때, Macro Average Precision(MAP) 기준 화행 약 2.2%p, 감정 약 0.9%p 가량의 성능 향상을 달성할 수 있었다. 본 연구에서 수행한 실험 결과를 통해, 제안하는 비용 민감 학습 방법이 범주 불균형 데이터셋의 학습에 효과적임을 확인하였다.

## 1. 서 론

자연언어처리에서 분류 문제는 가장 중요한 문제 중의 하나이다. 분류 데이터셋을 이용해 텍스트 분류 모델을 작성할 때, 해당 데이터셋 내의 범주 분포는 학습된 분류 모델의 분류 성능에 큰 영향을 미친다. 그리고, 많은 자연 발생 데이터는 범주 불균형 문제를 가지고 있다. 이에, 그러한 범주 불균형 데이터셋을 적절히 모델링하기 위한 연구들이 진행되어 왔다.

범주 불균형 데이터셋을 학습하기 위한 연구들은 크게 두 가지 접근 방법으로 구분할 수 있다. 첫번째 접근 방법은 데이터 샘플링이라 불리우는 방법이다. 데이터 샘플링은 데이터셋 내 고빈도 샘플을 일부 제거하거나(언더 샘플링, under sampling) 저빈도 샘플을 복제하는 방식(과표본화, over sampling)으로 데이터셋 내 범주 불균형을 완화한다[1].

두번째 방법은 데이터셋의 범주 불균형 정도를 손실함수(loss function)에 이용하여 편향을 완화하는 방식이다. 이 방법은 비용 민감 학습(cost sensitive learning)이라고 불린다. 인공신경망의 경우, 손실함수에 그 비용을 개입시켜 비용 민감 분류기(cost sensitive classifier)를 구현할 수 있다. 비용을 설정하는 방법 등에 관해 많은 연구가 있어 왔다[2, 3].

이외에도 언더샘플링과 과표본화를 함께 사용하거나, 저빈도 샘플을 변형해 새로운 샘플을 생성하는 등의 접근으로 효과를 본 연구들이 있었다[1].

그러나, 데이터 샘플링 중 과표본화 방법은 실제 1회 발생한 샘플을 수차례 모델에 반복 학습시키게 됨으로써 모델에 또다른 편향을 야기할 수 있다. 반대로 언더 샘플링은 학습 데이터셋에 발생한 일부 샘플을 학습에 사용하지 않는 것이므로, 해당 샘플에 나타난 유용한 정보가 소실될 가능성이 있다. 마지막으로 '비용 민감 학습'의 경우, 기존에는 각각의 비용(cost)을 전문가가 직접 설정하거나, 데이터셋 내 클래스의 발생 빈도로 설정하는 등의 방식을 취해 왔다[4]. 그러나, 고정 비용을 사용하는 방법이 오히려 성능을 저하시키기도 한다.

범주 불균형 데이터셋을 특별한 처리 없이 일반적인 분류 기법으로 학습하게 되면, 고빈도 범주(major class)로 편향되는 현상이 나타난다. 본 논문에서는 이러한 편향을 완화하여 결국 범주 불균형 데이터셋을 효율적으로 학습할 수 있는 방법에 대해서 탐구한다.

## 2. 제안 방법

기존의 비용 민감 학습 방법들은 학습 코퍼스 내 각 클래스의 발생 빈도 등에 기반하여 고정된 가중치를 부여하였다. 그러나 설정한 구조와 데이터셋에 고정된 가중치를 부여하였을 때는 오히려 성능이 하락하였다[표 5]. 본 연구에서는 그러한 성능의 하락이 고정된 가중치로부터 발생한 또다른 편향에 의한 것이라 가정하였다. 그리고 그러한 편향을 감쇄하기 위하여 매 학습 스텝마다 드리클레(Dirichlet) 분포에서 클래스 가중치를 추출하여 사용한다.

이러한 접근은 이전의 일반화(generalization) 성능 개선을 위한 연구들로부터 영감을 받았다[5, 6]. [5]에서는 매 순전파마다 임의의 노드들을 제거하여 노드들 사이의 상호 순응(co-adapting)을 저지하고, 결국 모델의 과적합을 방지하고자 하였다. [6]에서는 매 역전파마다 임의의 노이즈를 역전파의 경사(gradient)에

더하여 과적합을 방지하도록 하였다.

이 연구들의 공통점은 모델의 학습에 임의의 변화를 추가하는 것이다. 그리고 그렇게 추가된 임의의 변화가 모델의 일반화 성능을 높일 수 있다고 주장한다. 본 연구에서 제안하는 비용 민감 학습 방법 또한 위의 연구들과 같은 접근을 취하고 있다. 고정된 가중치를 분류기에 부여하는 대신, 드리클레 분포에서 가중치를 추출해 부여함으로써, 사전 지식인 학습 코퍼스 내 클래스의 발생 빈도를 고려하면서도 모델의 편향을 방지하여 최종 성능을 향상시킬 수 있다.

## 3. 실험

### 3.1. 실험 설정

실험에 사용한 데이터셋은 [7]에서 제안한 DailyDialog 대화 데이터셋이다. 최초 영어 대화 데이터셋이 공개되었고, 이를 직접 번역하여 작성한 한국어 대화 데이터셋으로 실험을 수행한다. 정제와 번역을 수행한 후의 데이터셋의 통계는 [표 1]과 같다.

표 1 번역 DailyDialog 대화 데이터셋의 통계량

| 구분 | 수량 | 단위 |
|---|---|---|
| 대화 | 11,160 | 대화 |
| 발화 | 87,495 | 발화 |
| 평균 발화 길이 | 7.41 | 단어/발화 |
| 평균 대화 길이 | 7.84 | 발화/대화 |

실험에 사용한 데이터셋은 양이 적으며 분류 클래스 간 편향 또한 상당히 강한 편이라 판단하였다. 이에, [8]에서 제안한 디노이징 메커니즘을 적용해 데이터셋을 증축하였다. 또한, 3턴 이상의 대화는 마지막 턴을 제거하면서 다시 증축하였다. 마지막으로, 가장 많은 빈도로 발생하여 강한 편향을 야기하는 '감정 없음' 지표에 해당하는 대화를 임의 제거하여 학습 데이터셋 내 편향을 완화하였다. 최종 모든 전처리가 완료된 데이터셋의 통계량은 [표 2, 3, 4]와 같다.

실험에 사용한 인공신경망 기반모델은 [9]에서 제안한 구조이다. 본 논문에서 설정한 실험은 대화 내 감정과 화행의 추론을 함께 수행하는 멀티태스크 분류이므로 short-term memory module의 slot을 감정 지표 4개과 화행 지표 7개로 설정하였다. 그러나, 각 지표들의 분산 표현은 입력과 출력 어디에도 나타나지 않으므로 지표들의 분산 표현이 학습되지 않는다. 따라서, 최종 추론 모듈은 지표의 분산 표현을 추론하는 회귀 계층(regression layer)와 분류를 수행하는 분류 계층(classification layer)으로 구성하였다. 이렇게 함으로써 각 지표의 분산 표현을 업데이트되도록 할 수 있다.

표 2 최종 실험에 사용된 데이터셋의 대화 통계량

| 구분 | 수량 | 단위 |
|---|---|---|
| 대화 | 43,914 | 대화 |
| 발화 | 318,837 | 발화 |
| 평균 대화 길이 | 7.26 | 발화/대화 |

표 3 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 4 최종 실험에 사용된 데이터셋의 화행 지표 분포

| | | |
|---|---|---|
| inform | 22,969 | 52.30% |
| question | 9,308 | 21.20% |
| commisive | 5,855 | 13.33% |
| directive | 5,782 | 13.17% |

Short-term memory module의 slot vector들은 연결되어 각 태스크의 분류 모듈에 입력된다. 그리고 두 태스크에 동시에 입력되는 dummy slot RNN을 두어 두 태스크 간의 공유 정보가 관리될 수 있도록 하였다. 따라서 감정 추론 모듈에는 4개의 slot vector와 하나의 dummy slot vector가 연결되어 입력되고, 화행 추론 모듈에는 7개의 slot vector와 하나의 dummy slot vector가 연결되어 입력된다.

추론 모듈의 분류 계층은 두번째 계층의 활성 함수(activation function)가 softmax로 부여된 다층 퍼셉트론(multi-layer perceptron, MLP)을 이용한다. 이 MLP에 제안하는 비용 민감 학습 전략을 적용하여 실험을 수행한다.

실험으로, 어떠한 비용 민감 학습 방법도 적용하지 않은 실험(실험 1), 학습 코퍼스 내 클래스의 발생 비율을 가중치로써 부여한 실험(실험 2), 그리고 제안하는 드리클레 분포에서 추출한 벡터를 가중치로써 부여한 실험(실험 3)을 수행한다.

### 3.2. 실험 결과 및 분석

[표 5]에 본 연구에서 수행한 실험의 성능을 실었다. 수행된 모든 실험은 3.1. 실험 설정에서 기술한 구조를 따르고, 최종 분류 모듈에 제안하는 비용 민감 학습 전략의 적용 유무에 따라 실험을 구분하였다. 성능 척도로는 고빈도 범주만을 추론하여도 높은 성능을 달성할 수 있는 정확도(accuracy) 대신에 각 범주를 동등하게 고려하는 Macro Average Precision(MAP)을 사용한다.

표 5 실험 성능.
모든 성능은 3회 반복 실험의 평균치이다.
(* : 본 논문에서 사용한 데이터셋과 구조에 [9]에서 제안한 다단계 분류 모듈을 적용해 수행한 실험)

| 구분 | 비용 민감 학습 전략 | 화행 분류 MAP | 감정 분류 MAP |
|---|---|---|---|
| [9]* | 다단계 분류 | 62.25 | 40.46 |
| 실험 1 | - | 62.42 | 43.60 |
| 실험 2 | 정적 가중치 | 59.39 | 26.27 |
| 실험 3 | 동적 가중치 | **64.65** | **44.46** |

표 6 실험 3의 감정 분류 혼동 행렬

| 추론 / 정답 | 감정 없음 | 행복 | 놀람 | 슬픔 | 화남 | 혐오 | 두려움 |
|---|---|---|---|---|---|---|---|
| 감정없음 | **1,375** | 104 | 11 | 4 | 8 | 0 | 0 |
| 행복 | 524 | **624** | 6 | 2 | 1 | 0 | 0 |
| 놀람 | 29 | 5 | **31** | 2 | 1 | 0 | 0 |
| 슬픔 | 116 | 8 | 0 | **56** | 0 | 0 | 0 |
| 화남 | 65 | 5 | 4 | 1 | **5** | 0 | 0 |
| 혐오 | 6 | 1 | 2 | 0 | 0 | 0 | 0 |
| 두려움 | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| 정밀도 (%) | 64.89 | 83.53 | 57.41 | 86.15 | 33.33 | 0.00 | 0.00 |

[표 6]을 보면, 제안하는 비용 민감 학습 방법을 적용하였음에도, 대부분의 추론이 고빈도 범주로 치우쳐 있음을 알 수 있다. 이러한 상황에서 '화남' 클래스의 경우 총 80개의 샘플 중 5개만을 맞추었다. 그러나 모델이 '화남' 클래스를 추론한 샘플이 15개에 불과하기 때문에, '화남'의 정밀도는 5/15인 33%에 달한다. 또한, 이 33%는 최종 지표인 MAP에 4.7%p(33%/7) 가량 기여한다. 이러한 결과는 저빈도 클래스의 샘플의 수가 현저하게 적은 범주 불균형 데이터셋에 대해서는 MAP가 적절하지 않을 수 있음을 시사한다.

## 4. 결론 및 향후 연구

본 논문에서는 범주 불균형 문제를 해결하기 위한 비용 민감 학습 방법으로써 동적 가중치 부여 방법을 제안한다. 드리클레 분포를 이용하여 학습 단계마다 가중치를 동적으로 부여할 수 있도록 하였다.

기존 비용 민감 학습 방법들은 데이터셋 등으로부터 수집된 통계 혹은 전문가의 통찰에 기반해 클래스 가중치를 설정하고, 이를 학습에 적용하는 방식을 취하고 있다. 그러나, 본 논문에서 설정한 실험에서는 기존의 비용 민감 학습 방법이 오히려 더 좋지 못한 성능을 내었다. 본 논문에서는 그러한 현상이 고정된 가중치로부터 발생하는 또 다른 편향이라 가정하였다. 따라서, 고정된 가중치 대신에 드리클레 분포에서 가중치를 샘플링하여 적용함으로써 이를 개선하였다. 설정된 실험에서 제안 방법은 MAP 기준 화행 약 2.2%p, 감정 약 0.9%p 가량의 성능 향상을 달성하였다.

본 논문에서 제안하는 동적 비용 민감 학습 방법은 학습 데이터셋의 클래스 분포만을 이용하고 있다. 실제 인공신경망 기반 분류 모델 설계에서 손실에 영향을 미치리라 판단되는 요소는 학습 데이터셋 내 클래스 분포 이외에도 데이터 샘플들의 벡터 공간 내 응집도나 극단치 등 여러 요소가 존재한다. 학습 데이터셋 내 클래스 분포 이외의 이러한 정보들을 이용하여 비용 민감 학습 방법을 개선할 수 있으리라 예상된다.

## Acknowledgement

이 논문은 2019년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2017R1D1A1B03033534).

## 참 고 문 헌

[1] Nitesh V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique", Journal of Artificial Intelligence Research, 2002.

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

[3] Salman H. Khan et al., "Cost-Sensitive Learning of Deep Feature Representations From Imbalanced Data", IEEE Transactions on Neural Networks and Learning Systems, 29, 8, 2018.

[4] Charles X. Ling and Victor S. Sheng, "Cost-Sensitive Learning and the Class Imbalance Problem", Encyclopedia of Machine Learning, 2008.

[5] Nitish Srivastava et al., "Dropout: A Simple Way to Prevent Neural Networks from Overfitting", Journal of Machine Learning Research, 15, 2014.

[6] Arvind Neelakantan et al., "Adding Gradient Noise Improves Learning for Very Deep Networks", arXiv, 2015.

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

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

[9] 신창욱, 차정원, "대화에서 멀티태스크 학습을 이용한 감정 및 화행 분류", 제30회 한글 및 한국어 정보처리 학술대회 논문집, 2018.