DAEM-ERC: 대화에서의 감정 인식을 위한 데이터 증강과 앙상블 기반 모델

유소영, 이한별, 조예지, 권나현, 김정현

세종대학교

{yooso0731, gks312, yezy0390}@sju.ac.kr happyday5707@naver.com, j.kim@sejong.ac.kr

DAEM-ERC: Data Augmentation and Ensemble-based Model for Emotion Recognition in Conversation

Soyoung Yoo, Hanbyul Lee, Yeji Cho, Nahyeon Kwon, Junghyun Kim
Sejoung Univ.

요 약

대화에서의 감정 인식(Emotion Recognition in Conversations, ERC)은 사람과 컴퓨터 간의 상호작용을 위한 핵심 기술이다. 감정 인식을 위해 사용되는 데이터에는 오디오, 비디오, 텍스트 등이 있고, 이러한 데이터들로부터 얻은 정서적 정보를 결합하여 멀티모달 감정 인식을 구현할 수 있다. 본 연구에서는 발화 텍스트, 오디오, 생체 신호 데이터를 사용하여 각 데이터에 특화된 개별 분류기를 생성한 뒤 Weighted soft voting 앙상블을 통해 최종 감정 분류를 진행하는 멀티모달 감정 인식 모델을 제안한다. 또한 우리는 각 데이터 특성을 고려한 증강 기법을 사용하여 심각한 클래스 불균형 문제를 완화했다. 결과적으로 우리의 제안 모델은 소수 클래스 감정 분류에 강점을 가지며 Weighted f1 스코어 0.91을 달성했다.

1. 서 론

최근 인공지능 기술의 발전에 따라, 점점 더 사람과 유사한 인공지능을 만들기 위한 연구들이 진행되고 있다. 그 중 대화에서의 감정 인식(ERC)은 사람과 컴퓨터 간의 자연스러운 대화 및 상호작용을 구현하기 위한 핵심 기술이다[1]. 사람의 대화는 언어적인 요소뿐만 아니라 표정, 목소리 변화와 같은 비언어적인 특성을 포함한다. 이에 따라 최근에는 여러 가지 유형의 데이터를 함께 처리하는 멀티모달 기반 감정 인식 연구가 활발히 진행되고 있다. 이러한 멀티모달 기반 감정 인식에는 주로 텍스트와 오디오, 이미지 데이터가 사용되어 왔다[2].

생체 신호는 사람의 감정 상태에 따라 변하고, 웨어러블 기기나 특정한 센서를 통해 측정 및 수집될 수 있다. 이러한 생체 신호 데이터를 감정 인식에 사용하면 발화자의 감정 상태를 더 정확하게 인식할 수 있다[3]. 따라서 우리는 텍스트, 오디오 데이터와 함께 생체 신호 데이터를 사용한 멀티모달 감정 인식 연구를 진행하고자 한다.

본 연구에서 사용한 데이터 셋은 특정 클래스의 데이터 개수가 전체의 약 85% 이상을 차지하는 클래스 불균형 문제가 있다. 이러한 극심한 클래스 불균형은 다수(majority) 클래스에 편향된 학습을 하는 문제를 일으키므로 이를 해결할 적절한 방법이 필요하다. 우리는 소수(minority) 클래스에 해당하는 데이터가 부족한 근본적인 문제를 해결하기 위해 데이터 증강 기법 중 오버샘플링을 고려하였다. 이때, 텍스트와 오디오, 생체 신호 데이터는 모두 다른 특성을 가지므로 데이터 타입 별로 다른 오버샘플링 기법을 적용하였다.

본 연구에서는 텍스트, 오디오, 생체 신호 데이터를 모두 활용하여 감정을 분류하는 모델을 제안한다. 텍스트, 오디오, 생체 신호 데이터는 동일한 발화에 대해 상이한 특징 정보를 가지고 있으므로 각 특성을 고려한 개별 분류기를 설계하고 예측 결과의 신뢰도를 높이기 위하여 개별 분류기를 앙상블로 결합한다.

실험 결과에서 제안된 개별 모델은 학습 데이터를 증강했을 때 증강하지 않았을 때보다 클래스의 데이터 수가 매우 작은 소수 클래스 감정 분류에 강점을 보였다. 앙상블을 통해 개별 분류기보다 결합된 분류기의 성능이 향상되었음을 확인하였다. 또한 앙상블로 결합 시 weight를 조절하여 전체 성능을 유지하면서 개별 클래스에 대한 분류 성능을 더욱 개선하였다.

2. 사용 데이터 셋 및 전처리

2.1 사용한 데이터 셋

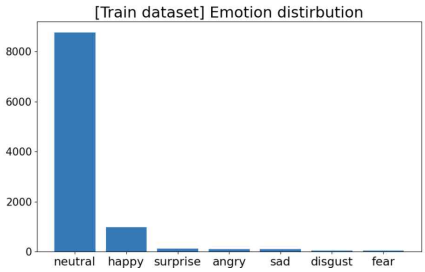
본 논문에서 우리는 한국어 멀티모달 감정 데이터셋 2020 (KEMDy2020)[4]을 사용한다. 이는 6개 주제에 대한 일반인 80명의 자유 발화 데이터 셋으로, 텍스트, 오디오, 센서 데이터로 구성된다. 이 데이터 셋은 40개의 세션을 포함하며, 각 세션은 6개의 주제에 대한 대화로 이뤄진다. 이때 특정 주제에 대한 개별 발화들을 세그먼트라고 한다.

데이터 셋의 감정은 7개(기쁨, 놀람, 분노, 중립, 혐오, 공포, 슬픔)의 클래스로 이루어지며, 세그먼트별로 10명의 평가자가 평가한 결과 중 많이 선택된 클래스로 분류된다. 클래스는 두 개 이상의 값을 가지는 경우(멀티 클래스)도 있다.

2.2 데이터 전처리

텍스트, 오디오, 센서 데이터에 대한 개별 전처리를 수행하기 전, 먼저 결측 세그먼트와 멀티클래스 세그먼트를 제거했다. 이후 40개의 세션에 대하여 80%는 학습, 20%는 테스트 데이터 셋으로 분할했다.

그림 1에서 볼 수 있듯이, 본 연구에서 사용한 데이터는 심각한 클래스 불균형 문제가 존재한다. 각 데이터 셋의 약 85% 이상은 “중립(Neutral)” 클래스에 해당한다. 따라서 우리는 텍스트, 오디오, 센서 데이터의 개별 특성에 맞는 오버 샘플링 방법으로 증강하여 불균형 문제를 개선했다. 이때, “중립”과 “기쁨”의 데이터 개수에 맞춰 각각 데이터를 증강하여 실험을 진행했고, 성능 비교를 통해 “기쁨” 클래스의 데이터 개수를 증강 기준값으로 선택하였다.



감정	데이터 개수
Neutral	8750
Happy	977
Surprise	129
Angry	105
Sad	97
Disgust	45
Fear	37

그림 1. 학습 데이터 셋의 클래스 분포.

2.2.1 텍스트 데이터

텍스트 데이터는 원본 데이터에 포함되어있는 불필요한 태그를 제거하고 Mecab 형태소 분석기를 사용하여 토큰화를 진행했다.

자연어 처리 분야에서 데이터 증강은 하나의 단어로도 문장의 의미가 달라질 수 있기 때문에 성능 향상에 유의미한 영향을 줄 것으로 예상되는 텍스트 데이터 증강 기법을 선별하여 사용했다[5]. 실험에 사용한 기법은 특정 단어를 동의어로 대체하는 방법인 Synonym Replacement(SR), 임의의 단어를 추가하는 방법인 Random Insertion(RI), 문장 내 임의의 두 단어의 위치를 바꾸는 방법인 Random Swap(RS) 이다. 위 3가지 기법을 증강할 클래스에 랜덤하게 적용했다.

이후, 증강된 데이터에 대해서 한국어로 사전 학습된 “FastText”로 각 단어마다 크기가 300인 임베딩 벡터를 계산하고 최대 토큰 길이는 전체 문장 토큰 길이 중 3사분위에 해당하는 값인 21으로 지정했다. 해당 값보다 문장 길이가 긴 경우에는 자르고, 이보다 짧은 경우에는 제로패딩을 진행했다.

2.2.3 오디오 데이터

오디오 데이터는 0.6초부터 35초까지 모두 다른 발화 시간으로 구성되어 있기 때문에 전체 데이터 발화 시간 분포에서 3사분위 값인 8초로 최대 길이를 설정했다. 이에 따라 8초보다 긴 경우에는 자르고, 짧으면 앞과 뒤에 랜덤하게 침묵을 추가했다.

본 연구에서는 음의 높낮이, 말하기 속도와 같이 사람마다 상이한 음성적인 특징들을 반영하기 위해 오디오 측면 데이터 증강 기법을 선택하여 데이터를 증강했다. 그 중 랜덤 노이즈 추가, 피치 조정, 시간 축 이동, 스트레칭을 사용했다.

마지막으로 Python의 librosa 라이브러리를 사용하여 오디오 데이터의 고유한 특징인 Mel-Frequency Cepstral Coefficient(MFCC)를 추출했다. 이때 오디오 데이터는 샘플링 레이트(sampling rate)를 16kHz로 설정하여 디지털 값으로 변환했다. 그런 다음 특징 추출의 개수는 40, 윈도우 길이는 1024, 그리고 홉(hop) 크기는 512로 설정하여 MFCC 특성을 추출했다.

2.2.4 센서 데이터

센서 데이터는 Electrodermal Activity(피부 전도성, 이하 EDA)와 피부 온도 데이터로 이루어져 있다. 센서 데이터는 각 발화마다 발화 시간이 다르므로 발화 별로 수집한 데이터 길이에 차이가 있다. 이를 해결하기 위해 각 데이터의 최대 길이인 141에 맞춰 제로패딩을 추가했다.

이후 데이터 불균형을 해결하기 위해서 Synthetic Minority Over-sampling Technique(SMOTE)를 사용한다[6]. 이 증강 기법은 낮은 비율로 존재하는 클래스의 데이터를 최근접 이웃 알고리즘[7]을 활용하여 새롭게 생성하는 방법으로, 과적합 발생 가능성이 적다는 장점이 있다.

3. 모델링

본 연구에서 제안하는 모델의 구조는 그림 2와 같다.

3.1 텍스트 데이터 분류기

텍스트 감정 분류는 필터로 문장을 스캔하여 문맥적 의미를 파악하는 “TextCNN”[8]의 구조에 영감을 받아 설계한 분류기를 사용한다. 모델은 각각 4x300, 5x300 컨볼루션 계층과 두 개의 Fully-connected 계층으로 이루어져 있다. 모델에 단어 임베딩 벡터가 입력되면 위 두 개의 컨볼루션 연산이 병렬로 수행되며 각각의 피처맵이 생성된다. 생성된 피처맵은 모두 연결된 후, Fully-connected 계층을 통과하며 감정 분류를 진행한다.

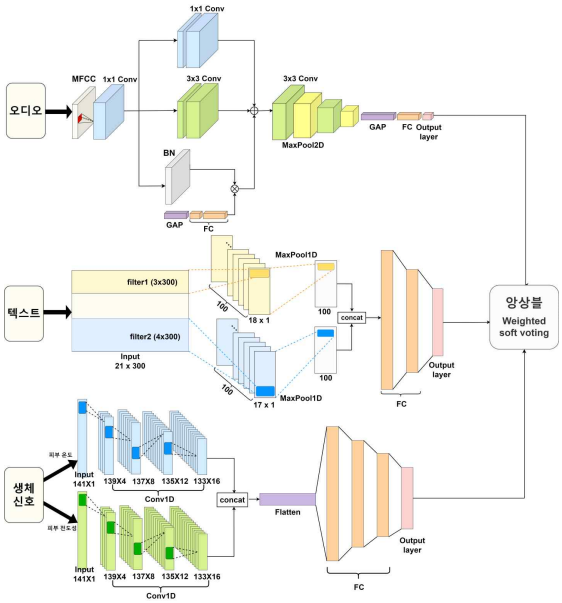


그림 2. 제안 모델의 구조도.

3.2 오디오 데이터 분류기

오디오 데이터는 우선 하나의 1x1 컨볼루션 계층을 통해 저차원의 여러 특징들이 추출된다. 그런 다음 병렬로 구성된 세 개의 특징 추출 블록을 통해 고차원의 정보들을 다양하게 학습한다. 이때, 각 블록은 1x1 컨볼루션 블록, 3x3 컨볼루션 블록, SE-Net[9] 블록으로 구성된다. 세 개의 피처 맵을 모두 더한 후 3x3 컨볼루션과 맥스 풀링 계층을 통과하면서 피처 맵의 크기는 점점 작아진다. 이 과정을 통해 중요한 고차원 특성들이 추출되며, 마지막으로 Global average pooling 계층과 두 개의 Fully-connected 계층을 거쳐 감정 분류를 진행한다.

3.3 센서 데이터 분류기

센서 데이터 모델은 EDA와 피부 온도 데이터의 특징을 각각 추출하고 이를 결합하여 동시에 학습한다. 모델은 2개의 컨볼루션 블록이 병렬로 이루어져 있어 EDA와 피부 온도 데이터에 대한 피쳐맵을 각각 추출한다. 이때, 각 블록은 4개의 1x3 컨볼루션 계층으로 구성된다. 이후 추출된 피처맵이 결합되고 4개의 Fully-connected 계층을 통해 감정 분류를 진행한다.

3.4 앙상블

본 논문에서는 텍스트, 오디오, 센서 데이터를 결합하여 분류하는 멀티모달 기반 감정 인식을 위해 개별 분류기의 결과를 앙상블 한다. 이때 개별 클래스에 대한 분류 성능을 더욱 개선하기 위해 더 다양한 감정에 대한 분류를 달성한 모델에 높은 가중치를 부여하는 Weighted soft voting을 사용한다.

4. 모델 실험 및 성능 비교

텍스트, 오디오, 센서 데이터에 대한 개별 분류기는 모두 Cross-entropy를 손실 함수로 하고 Adam 알고리즘을 통해 최적화됐다. 모델의 성능은 Weighted f1 스코어와 Macro f1 스코어로 평가된다. 이때, Weighted f1 스코어는 클래스별 샘플 수에 따라 가중치를 부여하여 계산된 값이고 Macro f1 스코어는 클래스별 f1 스코어의 평균이다.

4.1 데이터 증강 기준에 따른 성능 비교

모델	증강 기준	F1 score		Time (h:m:s)
		Macro	Weighted	
텍스트	**기쁨**	**0.33**	**0.86**	**0:02:17**
	중립	0.26	0.73	0:12:44
오디오	**기쁨**	**0.27**	**0.90**	**1:34:13**
	중립	0.22	0.86	3:29:24
센서	**기쁨**	**0.14**	**0.85**	**0:02:42**
	중립	0.13	0.43	0:13:31

표 1. 증강 기준에 따른 개별 분류기의 성능과 학습 시간.

2.2에서 언급한 대로, 표 1에서 우리는 “중립”과 “기쁨” 클래스를 기준으로 각각 데이터를 증강하여 개별 분류기의 성능 및 학습 소요 시간을 확인했다. 증강하기 전 학습 데이터 셋의 크기는 10140행이었고, “중립”에 맞춘 경우는 61250행, “기쁨”에 맞춘 경우는 14612행이었다. 이에 따라 “중립” 데이터 개수에 맞춰 증강한 데이터를 사용한 경우, “기쁨”에 맞춘 경우보다 학습에 최대 약 6배가량 더 긴 시간이 소요됐다. 또한 모든 개별 분류기의 f1 스코어는 “기쁨”에 맞춘 경우, “중립”에 맞춘 경우와 비교해 더 높았다. 이러한 실험 결과를 통해 우리는 "기쁨"에 맞춰 증강한 데이터 셋을 사용했다.

4.2 데이터 증강 여부에 따른 성능 비교

모델	증강 여부	F1 score								
		분노	혐오	공포	기쁨	중립	슬픔	놀람	Macro	Weighted
텍스트	X	0.00	0.00	0.50	0.14	0.93	0.00	0.00	0.22	0.91
	O	**0.09**	**0.29**	**0.47**	**0.18**	**0.92**	**0.11**	**0.22**	**0.33**	**0.86**
오디오	X	0.00	0.00	0.00	0.36	0.85	0.00	0.00	0.17	0.71
	O	**0.07**	**0.12**	**0.20**	**0.50**	**0.94**	**0.00**	**0.14**	**0.27**	**0.90**
센서	X	0.00	0.00	0.00	0.00	0.94	0.00	0.00	0.13	0.94
	O	**0.05**	**0.00**	**0.00**	**0.04**	**0.91**	**0.00**	**0.00**	**0.14**	**0.85**

표 2. 데이터 증강 여부에 따른 개별 분류기 성능 비교.

표 2는 텍스트, 음성, 센서 데이터에 대한 개별 분류기에서 데이터 증강 여부에 따른 감정 클래스별 f1 스코어와 Macro f1 스코어(macro), Weighted f1 스코어(weighted)이다.

먼저 텍스트 모델의 경우, 증강 후 “공포”과 “중립”을 제외한 5개의 클래스에서 f1 스코어가 증강 전 데이터를 사용했을 때와 비교하여 향상됐다. 오디오 모델의 경우, 모든 감정 클래스의 f1 스코어가 향상됐으며, 마지막으로 센서 모델에서는 “분노”, “기쁨” 클래스의 f1 스코어가 향상됐다. 결과적으로 실험에 사용된 개별 분류기는 증강 전에 비해 모두 더 다양한 클래스를 분류해 클래스별 f1 스코어와 Macro f1 스코어가 함께 향상되었다.

4.3 앙상블 모델에 따른 성능 비교

앙상블 기법	F1 score								
	분노	혐오	공포	기쁨	중립	슬픔	놀람	Macro	Weighted
Soft voting	0.00	0.00	0.29	0.15	0.94	0.00	0.13	0.22	0.93
Weighted soft voting	**0.04**	**0.21**	**0.50**	**0.19**	**0.93**	**0.11**	**0.19**	**0.31**	**0.91**

표 3. 앙상블 기법에 따른 성능 비교.

표 3에서 우리는 Weighted soft voting을 이용하여 텍스트, 오디오, 센서 데이터의 특성을 종합한 제안 모델의 성능을 Soft voting과 비교했다. 이때, Weighted soft voting은 각 모델의 성능에 따라 예측한 클래스에 가중치를 곱했다. 가중치는 여러 실험을 통해 확인한 최적의 값으로 설정하여, 텍스트 모델에는 0.5, 오디오 모델에는 0.3, 센서 모델에는 0.2 비율의 가중치를 사용했다.

그 결과, 두 앙상블 기법에 대한 Weighted f1 스코어는 비슷했지만, 클래스 간 성능 비교를 위한 Macro f1 스코어는 Weighted soft voting이 더 높았다. 이는 개별 분류기에 따른 적절한 가중치를 곱하여 감정을 분류하는 우리의 제안 모델이 soft voting 앙상블 기법이 적용된 모델보다 소수 클래스를 더 잘 분류한다는 것을 의미한다.

5. 결론

본 논문에서는 효과적인 멀티모달 감정 인식을 위한 예측 모델을 제안한다. 제안 모델은 텍스트, 오디오, 센서 데이터에 따른 증강 과정과 개별 분류기 및 이를 결합한 앙상블로 구성된다. 이때 데이터 증강은 "기쁨" 클래스에 맞춰 진행했으며 이를 통해 실험에 사용한 데이터 셋의 극심한 클래스 불균형 문제를 완화했다. 이후 증강 여부에 따른 개별 분류기의 Weighted f1 스코어와 Macro f1 스코어를 비교한 결과, 증강하지 않은 데이터를 사용한 경우 Weighted f1 스코어는 높았지만, 단순히 다수 클래스에만 치중해서 예측하는 경향을 보였다. 반면 증강한 데이터를 사용한 경우 보다 다양한 감정 클래스를 분류하여 Macro f1 스코어가 더 높았다.

이후 증강한 데이터로 학습한 개별 분류기의 결과를 Weighted soft voting를 사용해 앙상블 했고 Soft voting과의 성능 비교를 진행했다. 실험 결과, Soft voting보다 Weighted soft voting을 통한 앙상블 방법의 Macro f1 스코어가 더 높았고, 이는 소수 클래스에서의 분류 성능이 더 좋다는 것을 의미한다.

본 연구에서 설계된 개별 분류기는 각각 텍스트, 오디오, 센서 데이터를 사용하는 향후 연구에서 데이터의 특징을 학습하는 데 도움을 줄 수 있다. 또한 우리의 제안 모델은 심각한 클래스 불균형 문제가 존재하는 멀티모달 감정 인식 과제에서 유용하게 활용될 수 있을 것이다.

참고문헌

[1] Chudasama et al. "M2FNet: multi-modal fusion network for emotion recognition in conversation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

[2] Xing, Songlong, Sijie Mai, and Haifeng Hu. "Adapted dynamic memory network for emotion recognition in conversation." IEEE Transactions on Affective Computing 13.3, 1426-1439, 2020.

[3] Haag, Andreas, et al. "Emotion recognition using bio-sensors: First steps towards an automatic system." Affective Dialogue Systems: Tutorial and Research Workshop, ADS 2004, Kloster Irsee, Germany, June 14-16, 2004. Proceedings. Springer Berlin Heidelberg, 2004.

[4] K. J. Noh and H. Jeong, “KEMDy20,” https://nanum.etri.re.kr/share/kjnoh/KEMDy20?lang=ko_KR

[5] Wei et al., "Eda: Easy data augmentation techniques for boosting performance on text classification tasks." arXiv preprint arXiv:1901.11196 2019.

[6] Fernández et al. "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary." Journal of artificial intelligence research 61, 863-905, 2018.

[7] Pandey, Amit, and Achin Jain. "Comparative analysis of KNN algorithm using various normalization techniques." International Journal of Computer Network and Information Security 11.11, 36, 2017.

[8] Chen et al., "Convolutional neural network for sentence classification," MS thesis. University of Waterloo, 2015.

[9] Hu, Jie et al., "Squeeze-and-excitation networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.