데이터 불균형 기법의 부작용 완화를 위한 어텐션 기반 앙상블

박요한°[1], 최용석[1], Wencke Liermann[2], 이공주[1]
충남대학교
[1]{happy115012,yseokchoi,kjoolee}@cnu.ac.kr
[2]wliermann@o.cnu.ac.kr

Attention-Based Ensemble for Mitigating Side Effects of Data Imbalance Method

Yo-Han Park°, Yong-Seok Choi, Wencke Liermann, Kong Joo Lee
Chungnam National University

요 약

일반적으로 딥러닝 모델은 모든 라벨에 데이터 수가 균형을 이룰 때 가장 좋은 성능을 보인다. 그러나 현실에서는 특정 라벨에 대한 데이터가 부족한 경우가 많으며 이로 인해 불균형 데이터 문제가 발생한다. 이에 대한 해결책으로 오버샘플링과 가중치 손실과 같은 데이터 불균형 기법이 연구되었지만 이러한 기법들은 데이터가 적은 라벨의 성능을 개선하는 동시에 데이터가 많은 라벨의 성능을 저하시키는 부작용을 가지고 있다. 본 논문에서는 이 문제를 완화시키고자 어텐션 기반의 앙상블 기법을 제안한다. 어텐션 기반의 앙상블은 데이터 불균형 기법을 적용한 모델과 적용하지 않은 모델의 출력 값을 가중 평균하여 최종 예측을 수행한다. 이때 가중치는 어텐션 메커니즘을 통해 동적으로 조절된다. 그로므로 어텐션 기반의 앙상블 모델은 입력 데이터 특성에 따라 가중치를 조절할 수가 있다. 실험은 에세이 자동 평가 데이터를 대상으로 수행하였다. 실험 결과로는 제안한 모델이 데이터 불균형 기법의 부작용을 완화하고 성능이 개선되었다.

주제어: 에세이 자동 평가, 데이터 불균형, 가중치 손실, 앙상블, 어텐션

1. 서론

에세이 평가는 학생들의 논리적/비판적 사고와 글 작성 능력을 평가하기 위한 작업 중 하나이다[1]. 그러나 수작업으로 모든 에세이를 평가하는 것은 많은 시간 소요와 노력이 소모된다. 또한, 평가자의 주관적 판단으로 인해 같은 평가 기준이 있더라도 평가 결과가 다를 수 있다. 따라서 이러한 문제를 해결하기 위해 딥러닝 모델을 활용한 자동 에세이 평가에 관한 연구가 이루어 지고 있다[2, 3].

일반적으로 딥러닝 모델은 모든 라벨에 데이터 수가 균형을 이룰 때 모델이 각 라벨을 골고루 학습할 수 있다. 이는 모델의 성능을 향상시키는데 도움이 된다[4]. 그러나 에세이 자동 평가 데이터는 수집하기 어렵고 수집한 데이터들은 특정 점수 범위에 집중되어 있다. 그림 1은 AIHUB의 에세이 글 평가 데이터의 라벨별 분포도이다. 10 ~ 19점의 에세이는 전체의 10%를 차지하는데 비해 26 ~ 28점의 에세이는 약 41%를 차지하고 있다. 에세이 데이터를 그대로 활용할 경우 모델은 자연스럽게 높은 점수를 부여하는 경향을 가질 것이다. 이는 모델이 다른 범위의 점수(낮은 점수)를 예측하는 것이 어려울 수 있다.

데이터 불균형은 어떤 라벨이 다른 라벨보다 훨씬 많은 데이터를 가지거나 적은 데이터를 가지는 상황을 의미하며 이는 현실에서 빈번하게 발생하는 문제이다. 본 논문에서는 데이터가 적은 라벨은 TAIL이라고 하며, 빈도가 많은 라벨을 HEAD로 정의한다. 딥러닝 모델은 데이터가 불균형하다면 HEAD를 더 잘 학습하려는 경향이 있다. 최근에는 상대적으로 성능이 낮은 TAIL 데이터를 대상으로 성능을 높이려는 연구들이 이루

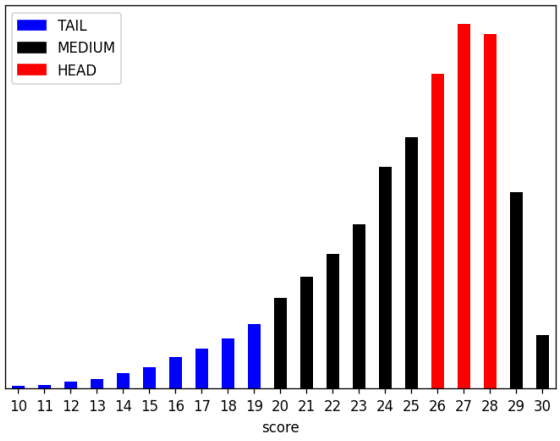

그림 1. 에세이 자동 평가에서 라벨별 데이터 분포

어지고 있다[5, 6, 7]. 데이터 불균형을 해결하기 위한 대표적인 기법은 가중치 손실이다. 가중치 손실은 각 라벨에 가중치를 할당하여 모델 학습에서 손실 값을 조정하는 기법이다. 일단적으로 가중치는 라벨별 빈도수에 반비례하도록 설정한다. 즉, HEAD의 가중치는 작아미여 TAIL의 가중치는 크게 조정이 될 수 있다. 본 논문에서는 에세이 자동 평가 모델을 학습할 때 가중치 손실 기법을 적용하여 데이터 불균형을 완화하고자 한다.

데이터 불균형을 완화하는 기법은 TAIL에 대한 성능을 개선하는 반면에 HEAD에 대한 성능을 저하시킬수 있는 문제가 있다[8]. 이를 보완하기 위해 본 논문에서는 어텐션 기반의 앙

상블을 제안한다. 앙상블은 다수의 모델 예측 결과를 결합하는 기법으로 각 모델의 예측 결과를 가중 평균하여 최종 예측을 수행한다. 기존의 앙상블 기법과 달리 우리는 각 모델에서 예측한 점수와 어텐션을 통해서 가중치가 동적으로 계산된다. 제안된 어텐션 기반의 앙상블은 데이터 불균형 기법으로 개선된 TAIL의 성능을 유지하면서 HEAD의 성능 저하를 줄이는 효과를 가진다. 이를 통해 전체적인 성능을 개선하여 데이터 불균형 문제를 해결한다.

2. 관련 연구

데이터 불균형 문제를 해결하기 위한 기법은 대표적으로 샘플링 기법과 가중치 손실 기법이 있다. 샘플링 기법은 데이터 셋의 구성을 조정하는 방식으로 HEAD과 TAIL 간의 데이터 분포를 균형화하는 데 활용된다. 샘플링 기법 중 하나인 오버샘플링은 TAIL의 데이터를 복제하거나 합성하는 기법이다. 즉, TAIL의 데이터를 늘림으로써 모델의 해당 라벨을 더 잘 학습할 수 있도록 하는 기법이다. [6]은 모델 학습 과정에서 중복된 데이터를 여러 번 사용함으로써 라벨별 학습량의 균형을 맞추는 기법이다. 그러나 TAIL의 데이터를 반복적으로 학습하기 때문에 과적합문제가 발생할 수 있다. [9]는 TAIL 데이터의 과적합 문제를 피하기 위해 데이터를 합성하여 생성하는 기법을 제안하였다.

가중치 손실은 각 라벨의 손실 함수에 가중치를 할당함으로써 모델의 학습을 조정하는 것이다. 즉, TAIL은 높은 가중치를 얻는 반면에 HEAD는 낮은 가중치를 부여하여 데이터 불균형을 완화시킨다. 가중치 손실의 기법 중 하나는 라벨의 빈도를 반영한 역빈도 가중치이다. 역빈도 가중치는 라벨 빈도의 역수를 가중치로 사용하여 TAIL의 가중치를 크게 만드는 방식이다. 또 다른 기법은 Focal loss이다[7]. Focal loss는 학습이 진행됨에 따라 어려운 샘플에 대한 가중치를 높게하는 기법이다. 대체적으로 TAIL은 학습량이 적기 때문에 어려운 샘플로 분류되어 가중치가 높아지면서 데이터 불균형이 완화된다.

앙상블은 여러 개의 모델이 예측한 결과를 조합하여 전체적인 성능을 향상시키는 기법이다. 앙상블 기법은 크게 3가지로 나눌 수 있다. 첫번째는 투표(Voting) 방식으로 각 모델이 예측한 결과들 중에서 다수결 투표를 통해 최종 예측 값을 결정하기 기법이다[10]. 하지만 투표는 분류 문제에 적합하며 실수 값을 예측하는 회귀 문제에 적합하지 않다. 두번째는 가중화 평균으로 모델 별 가중치를 계산하고 각 예측 결과에 대한 가중 평균으로 최종 예측하는 기법이다[11]. 세번째는 스태킹(Stacking)으로 여러 다른 모델의 예측 결과를 기반으로 메타 모델을 학습시켜 최종 예측을 수행하는 기법이다[12]. 이때 메타 모델은 각 모델의 예측 값을 입력으로 사용한다.

본 논문에서 에세이 자동 평가 모델은 데이터 불균형을 완화시키기 위해 역빈도 가중치를 적용하여 학습한다. 또한 데이터 불균형 기법으로 인한 HEAD의 성능 저하를 보완하기 위해 새로운 어텐션 기반의 앙상블을 제안한다.

3. 제안 기법

3.1 모델

그림 2는 에세이 자동 평가 모델의 구조도이다. 각 에세이는 N개의 문장으로 구성되어 있으며 ($E = \{sent_1, sent_2, ..., sent_N\}$), 각 문장은 M개의 단어로 구성되어 있다($sent_n = \{w_1, w_2, ..., w_M\}$).

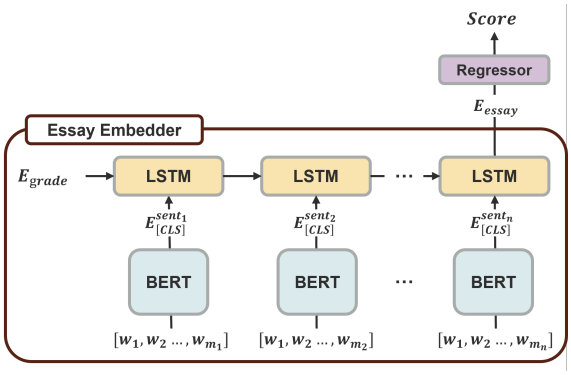

그림 2. 에세이 자동 평가 모델 구조

각 문장의 임베딩은 에세이에서 각 문장의 단어를 BERT[13]에 입력으로 넣어 주어 [CLS] 토큰의 임베딩을 사용한다 ($E^{sent_n}_{[CLS]}$). 각 문장의 임베딩은 단방향 LSTM[14] 모델에 입력으로 주어져 에세이의 시퀀스 정보를 고려한 에세이 임베딩 (E_{essay})을 추출한다. 이때, 에세이 작성자의 학년 정보를 임베딩(E_{grade})하여 LSTM의 초기 히든 벡터로 사용한다. 이를 통해 모델은 에세이 작성자의 특성을 고려할 수 있다. 마지막으로 추출된 에세이 임베딩은 선형 계층을 통해 에세이 점수를 예측한다.

3.2 데이터 불균형

본 논문에서는 데이터 불균형을 완화하기 위해 가중치 손실 기법을 도입한다. 이 기법은 라벨마다 다른 가중치를 손실 값에 곱하는 기법이다. 수식 1은 가중치 손실을 적용한 최종 손실 함수(L_w)이다. B는 배치 사이즈를 나타내며, s_i과 L_i는 각각 i번째 데이터에 대한 라벨과 손실 값을 나타낸다.

$$L_w = \frac{1}{B}\Sigma_{i=1}^{B} w_{s_i} L_i \quad (1)$$

각 라벨의 가중치(w_{s_i})는 역빈도의 루트값을 기반으로 계산된다2. 여기서 N_s는 라벨 s에 속하는 데이터의 빈도를 나타낸다.

$$w_s = \frac{\Sigma_{\hat{s} \subset S} \sqrt{N_{\hat{s}}}}{\sqrt{N_s}} \quad (2)$$

기존의 데이터 불균형 기법은 주로 분류 문제에 적용되었지만 에세이 자동 평가와 같은 회귀 문제에서는 실수값을 예측해야 하기 때문에 다른 특성을 고려해야 한다. 분류 문제에서는 라벨 간에 서로 다른 특성을 가지지만 회귀 문제에서는 인접한 라벨이 유사한 특성을 가질 가능성이 크다. 예를 들어 12점의 에세이는 25점에 비해 14점의 에세이와 유사한 특성을 가질 수 있다. 따라서 회귀 문제에서는 라벨간 유사성을 고려하는 것이 중요하다.

이러한 관점을 기반으로 [8]은 회귀 문제의 데이터 불균형을 해결하기 위해 라벨 분포 평탄화(Label Distribution Smooth, LDS)를 제안하였다. 라벨 분포 평탄화는 각 라벨 데이터 빈도를 인접한 라벨의 데이터 빈도를 반영하여 재조정하는 기법으로 각 라벨의 빈도수에 인접한 라벨의 데이터 빈도수를 추가하여 조정한다. 이를 위해 가우시안 필터를 적용하여 빈도를 추가하였으며 라벨 값의 차이가 커질 수록 반영 비율이 작아지도록 하였다. 본 논문에서는 회귀 문제의 특성을 고려한 가중치 손실을 적용하기 위해 라벨 분포 평탄화를 통해 데이터 빈도를 재조정하고 수식 2로 가중치를 다시 계산하였다. 이때 가우시안 필터의 크기는 5로 설정하였다.

3.3 앙상블

[8]에서 데이터 불균형를 완화하면 TAIL에 대한 성능은 개선되는 반면에 HEAD의 성능은 저하되는 문제를 발견하였다. 본 논문은 이러한 문제를 해결하고자 HEAD에서 좋은 성능을 보이는 기본 모델(Base)과 데이터 불균형 기법을 적용하여 TAIL에서 성능이 우수했던 모델(Expert, 전문가 모델)의 장점을 동시에 활용하는 가중 평균 기반의 앙상블을 사용한다. 가중 평균 기반의 앙상블 모델은 두 모델의 측정 값을 가중 평균하여 최종 에세이 점수를 예측한다(수식 3).

$$score_{ensemble} = w_{Base} \cdot score_{Base} + w_{Expert} \cdot score_{Expert} \quad (3)$$

스택킹(Stacking)은 앙상블 기법 중 하나로 여러 모델의 예측값을 입력으로 활용하여 메타 모델을 구축하고 학습하여 최종 예측을 수행하는 기법이다. 스택킹 방법은 각 기본 모델이 생성한 예측값을 사용해 메타 모델을 학습시킨다. 그러므로 메타 모델의 가중치는 추론 단계에서 입력 데이터와 관계없이 동일한 값을 가진다. 이런 특성으로 인해 스택킹 방법은 두 가지의 제한 사항을 가지고 있다. 첫째, 메타 모델의 가중치는 추론 단계에서 고정되어 있어 입력 데이터와 관계없이 동일한 가중치를 가지고 있다. 둘째, 스택킹에서는 각 모델의 예측 결과만을 활용하므로 입력 데이터와 모델 간의 특성을 반영할 수 없다.

본 논문에서는 스택킹의 한계를 극복하기 위해 어텐션 기반의 앙상블을 제안한다. 어텐션 기반의 앙상블은 각 모델의 예측 결과에 대한 가중 평균을 계산할 때, 입력 데이터에 따라 가중치를 동적으로 조절할 수 있다. 그림 3은 어텐션 기반의 앙상블 모델의 구조이다.

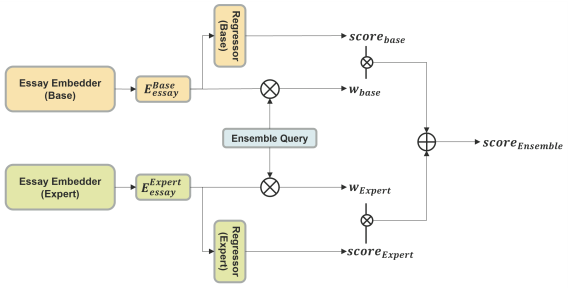

그림 3. 앙상블 모델 구조

가중치는 수식 4와 같이 계산된다. 먼저 앙상블 쿼리(E_Q)와 에세이 임베딩(E^{Δ}_{essay})간의 코사인 유사도를 계산한다. 앙상블 쿼리는 동적으로 모델의 가중치를 계산하기 위한 임베딩으로 학습되는 파라미터이다. 계산된 코사인 유사도 값을 softmax 함수에 입력하여 정규화한 뒤 가중치로 사용한다. Softmax 함수를 사용하면 가중치의 합이 항상 1이 되므로 가중치는 0과 1 사이 값을 가지며, 각 모델에 대한 상대적 기여도를 나타낸다.

$$w_{\Delta} = \frac{\exp\left(E^{\Delta}_{essay} \cdot E_Q\right)}{\exp\left(E^{Base}_{essay} \cdot E_Q\right) + \exp\left(E^{Expert}_{essay} \cdot E_Q\right)}, \quad \Delta \subset \{Base, Expert\} \quad (4)$$

4. 실험 및 결과

4.1 데이터

본 논문에서 실험에 사용된 데이터는 AIHUB[1]의 '에세이 글 평가'에서 논술형 에세이를 사용하였으며, 10점 미만의 에세이 148개를 제외하였다. 모델의 검증 데이터는 'Training' 파일의 10%을 임의로 샘플링하였으며, 평가 데이터는 'Validation" 파일을 사용하였다. 학습, 검증, 평가 데이터 수는 각각 16,341개, 4,087개, 3,265개이다.

데이터 불균형 기법을 적용하기 위해서는 라벨별로 빈도수가 필요하다. 그러나 에세이 점수는 실수 값을 가지기 때문에 1점 단위로 라벨을 정의하였으며 데이터 분포는 그림 1과 같다.

4.2 하이퍼 파라미터

표 1은 본 실험에서 사용된 에세이 자동 평가 모델의 하이퍼 파라미터이다. 실험에 사용된 BERT 모델은 어절 단위의 한국어 BERT[2]이며, 차원수는 768차원이다. 모델 학습 과정에서

[1]https://www.aihub.or.kr/
[2]https://aiopen.etri.re.kr/

BERT의 매개변수는 학습하지 않는다. LSTM은 3개의 층으로 구성되어 있으며, 각 층의 차원 수는 각각 256, 128, 64이다. Regressor는 총 3개의 선형 계층으로 구성되어 있으며, 차원 수는 각각 64, 16, 8, 1이다. 최적화 함수는 AdamW를 사용하였으며, 학습률은 0.005으로 설정하였다. 학습률 스케쥴러는 코사인을 사용하였으며, 전체 학습 과정에서 10번의 주기를 가지도록 설정하였다.

표 1. 하이퍼 파라미터

Hyper-parameters		Value
Hidden Dimension Size	BERT	768
	LSTM	256,128,64
	Regressor	64,16,8,1
Learning Rate	BERT	0
	Others	0.005
Optimizer		AdamW
Learning Rate Scheduler		Cosine
Dropout Rate		0.1

4.3 평가 방법

에세이 자동 평가 모델은 성능 평가 지표로서 카파(kappa) 계수를 평가 지표로 사용하였다. 카파는 모델의 예측 값과 실제 점수 간의 일치 정도를 측정하며 -1에서 1사이의 값을 가진다.

본 논문에서는 데이터 불균형 기법의 성능을 관찰하기 위해 라벨별 성능을 비교하였다. TAIL은 10 ∼ 19의 점수를 사용하였으며 전체 데이터의 10%의 비율을 차지한다. HEAD의 데이터는 26 ∼ 28점의 데이터로 선정하였으며 전체 데이터의 41%를 차지한다. 라벨에 대한 평가 지표로는 평균 제곱근 오차(Root Mean Squared Error, RMSE)를 사용하였다. 평균 제곱근 오차는 모델이 예측한 결과와 정답 간의 차이를 측정하는 지표로 낮을수록 좋은 성능을 가리킨다.

4.4 실험 결과

표 2은 데이터 불균형 기법에 따른 실험 결과이다. 먼저, 라벨 간 불균형한 데이터를 사용하여 학습한 모델(Base)은 0.578의 카파 계수로 나타났다. 그러나 데이터 불균형 완화 기법으로 역빈도 기법을 적용한 모델(Inverse)은 0.579로 개선이 되었다. 이것은 역빈도를 활용한 가중치 손실 기법이 모델의 성능을 약간 향상시킨 것을 보여준다.

추가로 라벨 분포 평탄화를 적용한 모델(LDS)은 0.625이다. 이것은 인접한 라벨 간의 유사성을 고려한 라벨 분포 평탄화를 적용함으로써 회귀 문제에서 데이터 불균형을 효과적으로 보완하고 모델의 성능을 개선한 것으로 볼 수 있다. 특히, TAIL에 대한 평균 제곱근 오차가 6.429에서 5.78로 줄어들었다는 점으로 보아 TAIL에 대한 예측이 개선된 것을 보여준다. 그러나 데이터 불균형 기법을 적용하였을 때 TAIL의 성능이 개선된 반면에 HEAD에 대한 성능은 크게 떨어졌다.

마지막으로 라벨 분포 평탄화 모델을 전문가 모델로서 베이스 모델과 앙상블(Ensemble)할 때 카파 계수는 0.633으로 가장 좋은 성능을 보였다. TAIL의 경우 평균 제곱근 오차가 6.075로 전문가 모델에 비해서는 성능이 떨어지지만 베이스 모델보다는 좋은 성능을 보이고 있다. 반면, HEAD에서는 베이스 모델보다 떨어지지만 앙상블 모델이 전문가 모델에 비해 오차가 작은 것을 볼 수 있다. 이러한 결과는 양쪽 라벨에 대해서 중간 성능을 보이고 있지만 성능 하락 폭이 상대적으로 적기 때문에 TOTAL에서는 가장 좋은 성능을 보여주었다. 이는 두 모델의 장점을 결합하여 성능이 개선된 것으로 볼 수 있다.

표 2. 데이터 불균형 기법에 따른 에세이 자동 평가 결과

Model		Base	Inverse	LDS	Ensemble
Kappa(↑)		0.578	0.579	0.625	**0.633**
RMSE (↓)	TOTAL	2.916	2.947	2.934	**2.861**
	HEAD	**1.405**	1.579	1.580	1.408
	MEDIUM	**3.040**	3.524	3.449	3.095
	TAIL	6.429	6.048	**5.780**	6.075

본 논문은 제안한 어텐션 기반의 앙상블 모델의 성능을 평가하기 위해 두 가지 다른 앙상블 기법과 비교 실험을 진행하였다. 첫 번째는 평균(Average) 앙상블 기법으로, 두 모델의 예측 결과를 동일하게 0.5의 가중치로 평균을 취하는 기법이다. 두 번째는 스택킹(Stacking) 앙상블 기법으로, 두 모델의 예측 결과를 입력으로 하는 메타 모델을 학습하는 기법이다. 실험에서는 메타 모델을 2개의 선형 계층로 구성하였다.

그림 4는 각 앙상블 기법에 따른 가중치의 차이를 보여주고 있다. 가로 축은 데이터의 정답 라벨을 나타내며, 세로 축은 베이스 모델과 전문가 모델 간의 가중치 차이를 나타낸다. 스택킹 모델의 가중치는 두 선형 계층의 매개변수를 내적하여 계산하였다.

평균(Average)과 스택킹(Stacking) 앙상블은 입력 데이터와 상관없이 가중치가 고정되어 있다. 특히 스택킹의 가중치 차이는 약 0.45로 베이스 모델의 가중치가 더 높은 것을 볼 수 있다. 이는 메타 모델을 학습할 때에도 HEAD에 집중되어 베이스 모델의 가중치를 높게 학습한 것으로 보인다. 반면 어텐션 기반의 앙상블은 가중치 값이 동적으로 조절되고 있다. TAIL에서는 전문가 모델의 가중치가 큰 값을 가지며, HEAD에서는 베이스 모델의 가중치를 크게 설정하고 있다. 이것은 데이터 불균형 기법의 부작용를 해결하기 위해 동적으로 최적의 가중

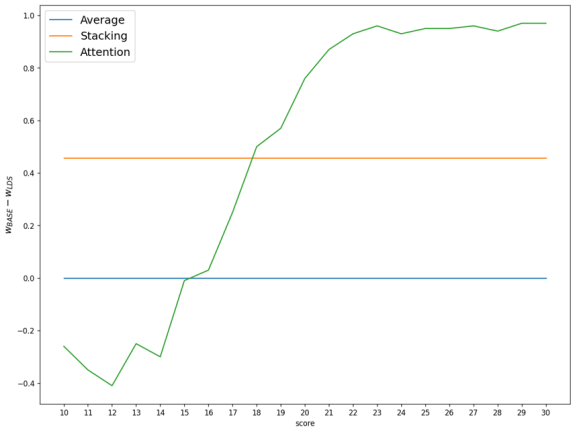

그림 4. 앙상블 기법 별 가중치 차이

표 3. 앙상블 기법에 따른 에세이 자동 평가 결과

Ensemble		Average	Stacking	Attention
Kappa(↑)		0.611	0.593	**0.633**
RMSE (↓)	TOTAL	2.887	2.916	2.861
	HEAD	1.490	1.323	1.408
	MEDIUM	3.245	2.957	3.095
	TAIL	5.981	6.389	6.075

치를 선택한 것으로 볼 수 있다.

표 3은 세 가지 앙상블 기법에 대한 비교 결과이다. 이 실험에서 어텐션 기반의 앙상블이 가장 좋은 성능을 보였으며 스택킹의 카파 계수는 가장 낮았다. 스택킹은 베이스 모델이 더 크기 때문에 HEAD의 성능이 가장 좋은 것을 볼 수 있다. 그러나 전문가 모델의 가중치가 값이 낮아 TAIL의 경우 평균보다 성능이 떨어지고 있다. 어텐션 기반의 앙상블은 어텐션을 통해 입력 데이터마다 최적의 가중치를 선택함으로써 TAIL과 HEAD에 대한 예측을 개선한 것으로 볼 수 있다.

5. 결론

본 논문에서는 에세이 데이터를 대상으로 데이터 불균형 문제에 대한 연구를 수행했다. 이를 해결하기 위해 가중치 손실 기법과 라벨 분포 평탄화 기법을 결합하여 데이터 불균형을 개선하고자 하였다. 실험 결과로는 가중치 손실을 적용하였을 때 전체 성능이 개선되었으며, 라벨 분포 평탄화를 적용하였을 때 가장 좋은 성능을 보였다.

그러나 데이터 불균형 완화 기법은 HEAD에 대한 성능 저하를 야기시킨다. 이를 완화시키기 위해 앙상블 기법을 적용하였다. 앙상블 중 하나인 스택킹은 고정된 가중치를 사용하여 데이터에 따라 가중치를 동적으로 조절하지 못하는 한계가 있다. 이에 본 논문에서는 어텐션 기반의 앙상블을 제안하여 입력 데이터에 따라 동적으로 가중치를 조절하도록 하였다. 실험 결과 동적인 가중치를 사용하는 기존의 앙상블에 비해 어텐션 기반의 앙상블이 우수한 성능을 보여주었다.

감사의 글

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2023-00241142).

참고문헌

[1] M. A. Hussein, H. Hassan, and M. Nassef, "Automated language essay scoring systems: A literature review," *PeerJ Computer Science*, Vol. 5, p. e208, 2019.

[2] F. Nadeem, H. Nguyen, Y. Liu, and M. Ostendorf, "Automated essay scoring with discourse-aware neural models," *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications*, pp. 484–493, 2019.

[3] M. Uto, Y. Xie, and M. Ueno, "Neural automated essay scoring incorporating handcrafted features," *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6077–6088, 2020.

[4] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, Vol. 6, No. 1, pp. 1–54, 2019.

[5] S. García and F. Herrera, "Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy," *Evolutionary computation*, Vol. 17, No. 3, pp. 275–306, 2009.

[6] E. Burnaev, P. Erofeev, and A. Papanov, "Influence of resampling on accuracy of imbalanced classification," *Eighth international conference on machine vision (ICMV 2015)*, Vol. 9875, pp. 423–427, 2015.

[7] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

[8] Y. Yang, K. Zha, Y. Chen, H. Wang, and D. Katabi, "Delving into deep imbalanced regression," *International Conference on Machine Learning*, pp. 11 842–11 851, 2021.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, Vol. 16, pp. 321–357, 2002.

[10] J. A. Morgan-Benita, C. E. Galván-Tejada, M. Cruz, J. I. Galván-Tejada, H. Gamboa-Rosales, J. G. Arceo-Olague, H. Luna-García, and J. M. Celaya-Padilla, "Hard voting ensemble approach for the detection of type 2 diabetes in mexican population with non-glucose related features," *Healthcare*, Vol. 10, No. 8, p. 1362, 2022.

[11] J. Briskilal and C. Subalalitha, "An ensemble model for classifying idioms and literal texts using bert and roberta," *Information Processing & Management*, Vol. 59, No. 1, p. 102756, 2022.

[12] N. Habbat, H. Nouri, H. Anoun, and L. Hassouni, "Sentiment analysis of imbalanced datasets using bert and ensemble stacking for deep learning," *Engineering Applications of Artificial Intelligence*, Vol. 126, p. 106999, 2023.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Jun. 2019. [Online]. Available: https://aclanthology.org/N19-1423

[14] A. Graves and A. Graves, "Long short-term memory," *Supervised sequence labelling with recurrent neural networks*, pp. 37–45, 2012.