불균형 데이터의 이진 분류를 위한 앙상블 구성 방법

김영훈 [1], 이주홍 [2]
[1] 인하대학교 전기컴퓨터공학과 석사과정
[2] 인하대학교 전기컴퓨터공학과 교수

qorxjs@inha.edu, juhong@inha.ac.kr

Ensemble Composition Methods for Binary Classification of Imbalanced Data

Yeong-Hun Kim[1], Ju-Hing Lee[2]
[1]Dept. of Electrical and Computer Engineering, Inha University
[2]Dept. of Electrical and Computer Engineering, Inha University

요 약

불균형 데이터의 분류의 성능을 향상시키기 위한 앙상블 구성 방법에 관하여 연구한다. 앙상블의 성능은 앙상블을 구성한 기계학습 모델 간의 상호 다양성에 큰 영향을 받는다. 기존 방법에서는 앙상블에 속할 모델 간의 상호 다양성을 높이기 위해 Feature Engineering 을 사용하여 다양한 모델을 만들어 사용하였다. 그럼에도 생성된 모델 가운데 유사한 모델들이 존재하며 이는 상호 다양성을 낮추고 앙상블 성능을 저하시키는 문제를 가지고 있다. 불균형 데이터의 경우에는 유사 모델 판별을 위한 기존 다양성 지표가 다수 클래스에 편향된 수치를 산출하기 때문에 적합하지 않다. 본 논문에서는 기존 다양성 지표를 개선하고 가지치기 방안을 결합하여 유사 모델을 판별하고 상호 다양성이 높은 후보 모델들을 앙상블에 포함시키는 방법을 제안한다. 실험 결과로써 제안한 방법으로 구성된 앙상블이 불균형이 심한 데이터의 분류 성능을 향상시킴을 확인하였다.

1. 서론

앙상블은 두 개 이상의 모델을 결합하는 알고리즘이다. 여러 모델을 결합하여 적절한 방법으로 결과를 도출하는 방식으로 앙상블의 구성을 이루는 모델 간의 성능과 상호 다양성(Diversity)은 앙상블 성능에 영향을 미치는 주요 요소로 간주된다[1]. 앙상블을 구성하는 모델 간의 상호 다양성이 존재하지 않는다면 성능 개선을 기대할 수 없으며 과적합의 요인이 된다[2].

모델 간의 상호 다양성 수치를 산출하는 기존 다양성 지표는 각 모델의 예측 결과의 차이를 비교하는 방식이다[3]. 실험 데이터로 사용되는 불균형 데이터에 다양성 지표를 적용한다면 다수 클래스에 편향 된 수치가 산출 되는 문제가 있다. Pairwise 기반 다양성 지표는 후보 모델 개수의 증가에 따른 높은 계산 비용이 드는 문제를 가지고 있다.

이러한 문제로 다양성 지표를 사용하지 않고 모델 간의 상호 다양성을 높여 앙상블의 성능을 개선 시키기 위한 연구가 많이 진행되었다[4]. 단일 기계 학습 모델에서 Feature Engineering 인 Data Sampling 과 Feature Selection 기법을 적용하여 여러 학습 모델을 만들었다[5]. 또한 여러가지 기계학습 모델들을 사용하여 서로 이질적인 여러 학습 모델을 만들어 앙상블에 포함되는 모델들의 상호 다양성을 높였다[6].

앙상블에 포함될 후보 모델들이 많이 생성될수록 유사한 예측 결과를 산출하는 모델도 증가하는데 이는 앙상블 구성 시 모델 간의 상호 다양성을 낮추고 성능을 저하시키는 문제로 이어지며, 과적합의 요인이 된다.

모델 간의 다양성을 측정하는 기존 지표인 Disagreement (수식 1)은 불균형데이터를 사용하는 두 모델 간의 다양성 값이 다수 클래스에 편향되는 문제가 있다.

$$\text{Disagreement: } Dis_{i,k} = \frac{N^{01}+N^{10}}{N^{11}+N^{10}+N^{01}+N^{00}} \quad \text{(수식 1)}$$

<표 1> Relationship between a pair of classifiers

	D_k correct(1)	D_k correct(0)
D_i correct(1)	N^{11}	N^{10}
D_i wrong(0)	N^{01}	N^{00}

그러나 모델 간의 분류율 사이에 이율 배반성이 존재하기 때문에 모델 간의 측정된 다양성 결과가 일정 수준 이상으로 높아 진다 해도 앙상블의 성능이 계속해서 향상되지는 않는다[7].

본 논문에선 상호 다양성이 높은 모델들을 앙상블에 포함시키는 방법에 관하여 논한다. 이를 위해서 후보 모델의 개수를 줄이기 위해 유사한 모델들을 제거하여 남겨진 후보 모델들의 상호 다양성이 높아지게 만들어주는 모델 가지치기 방법을 제안하고, 기존의 다양성 지표의 단점을 해소하는 새로운 다양성 지표를 제안한다. 다수의 유사한 모델들이 제거된 후보 모델들 중에서 상호 다양성이 적절히 높은 모델들을 선별하는 유전자 알고리즘에 적용하여 앙상블을 구성하는 방법을 제안한다.

2. 본론

기존 다양성 지표 Disagreement 을 개선한 Class Disagreement (CD)는 (수식 2)와 같다. Major CD(수식 3)는 데이터의 다수를 차지하는 클래스(상환)만을 고려한 두 모델 간의 다양성 비율을 산출한다. Minor CD(수식 4)는 데이터의 소수를 차지하는 클래스(연체)만을 고려한 두 모델 간의 다양성 비율을 산출한다. 클래스별로 서로 다르게 분류한 결과를 비율로 변환하여 평균을 계산하는 방식으로 편향성 문제를 개선하였다.

$$CD = \frac{Major\ CD\ +\ Minor\ CD}{2} \quad \text{(수식 2)}$$

$$\text{Major CD} = \frac{B+C}{(A+B+C+D)} \quad \text{(수식 3)}$$

$$\text{Minor CD} = \frac{b+c}{(a+b+c+d)} \quad \text{(수식 4)}$$

<표 2> Major Class Confusion Matrix between two classifiers

	C_k correct(1)	C_k wrong(0)
C_i correct(1)	A	B
C_i wrong(0)	C	D

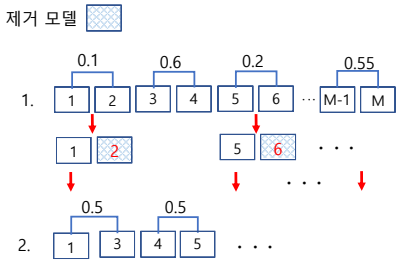
<표 3> Minor Class Confusion Matrix between two classifiers

	C_k correct(1)	C_k wrong(0)
C_i correct(1)	a	b
C_i wrong(0)	c	d

- A,B,a,b: 두 후보 모델의 예측 결과가 일치
- C,D,c,d: 두 후보 모델의 예측 결과가 불일치

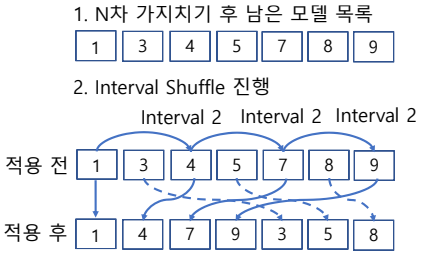
그러나, CD 와 같이 Pairwise 방식의 다양성 수치로는 모델의 성능을 확인할 수 없기 때문에 낮은 성능의 후보 모델을 배제할 수 없다. 따라서 먼저 낮은 성능의 모델들을 필터링하여 제거한 후에 유사한 모델들을 제거함으로써 후보 모델의 개수를 줄이는 가지치기 방법을 수행한다.

가지치기 방법은 후보 모델들을 나열하고 인접한 두 모델의 분류 유사성을 평가하여 유사한 두 모델 중 하나의 모델을 제거하는 방법이다. 모델의 유사성 평가방법은 CD 를 사용하여 상호 다양성을 계산하고 다양성 수치가 일정 수준 이하(본 논문은 0.1)면 유사 모델이라 판단한다.



(그림 1) 다양성 지표와 가지치기 방법을 결합하여 유사한 후보 모델을 제거하는 방법

이러한 가지치기를 반복 시행하기 위해서 Interval Shuffle 방법으로 모델 들의 순서를 변경한다. 이 가지치기 방법을 모델의 개수가 일정 수준 이하로 줄어들 때까지 반복적으로 수행하여 유사한 모델들을 제거한다.



(그림 2) 가지치기 후 인접한 모델을 변경하는 Interval Shuffle 진행 방법

남겨진 후보 모델 들에서 N 개의 앙상블을 구성할 때 각 앙상블이 서로 다른 모델들을 포함하도록 유전자 알고리즘을 사용한다. 초기 해 생성단계에서 랜덤하게 L 개의 모델을 조합하여 Population 을 생성한다. 적합도 평가는 Population 내 모든 집합 쌍을 조합하고 조합된 집합 간의 모델 교집합이 적은 상위 집합들을 선별한다. 교차연산에서 집합(부모)로부터 랜덤

으로 모델의 수를 절반씩 받아 조합하여 자식을 생성한다. 돌연변이 연산에서는 자식의 모델에서 무작위로 하나의 모델을 선택하고 랜덤하게 변경한다. 종료 조건은 앙상블마다 교집합을 이루는 모델의 수가 일정한 수 이하로 줄어들 때까지 반복적으로 수행하여 앙상블을 구성할 모델들을 조합한다.

3. 실험

본 논문에서 사용한 Data 는 2008 년부터 2016 년도 기간의 Lending Club Loan Default Data 로 연속형 변수 20 개 범주형 변수 7 개로 구성되어 있으며 396,030 개의 대출 관측치를 보유하고 있다. Data 전처리 과정에서 One-hot Encoding 을 적용하여 변수를 27 개에서 78 개로 변환시켰다.

학습 과정에서 Data 불균형 문제의 해결을 위해 Data Resampling 방법인 Random Under Sampling 을 사용하였다. 학습 모델은 DNN(Deep Neural Network)을 사용하였고 Feature Engineering 과정에서 Random Data Sampling 과 Random Feature Selection 을 적용하는 방법으로 3,000 개의 후보 모델을 생성하였다.

성능을 기준으로 모델을 제거할 때 사용하는 성능 지표는 신용 Data 의 특성상 분류 대상을 명확하게 분류할 필요가 있어 Recall 을 사용하였고, Positive Recall 과 Negative Recall 의 합이 하위 30% 이하인 모델을 제거하였다. 2,100 개의 후보 모델에서 CD 와 가지치기 방법을 사용하여 유사한 후보 모델을 제거해 적절한 상호 다양성을 충족하는 700 개의 후보 모델을 선별하였다. 이 후보 모델 들에서 앙상블을 구성할 때 모델을 선택하는 방법으로 유전자 알고리즘을 사용하여 20 개의 모델로 구성된 앙상블 2,000 개를 생성하였다.

결과 불균형 Data 인 점을 고려하여 Ensemble Vote 의 Threshold 을 8 로 설정하였다. 2,000 개의 앙상블 모델과 3,000 개의 후보 모델에서 랜덤 방식으로 동일하게 100 개의 모델을 추출한 후 평가지표 Recall 을 사용하여 100 개의 모델의 분류 성능을 평가하고 평균을 계산하였다. 제안하는 앙상블 모델의 평균이 DNN 모델 평균 대비 Positive Recall(상환)은 0.07% Negative Recall(연체)은 1.3% 증가하였다.

<표>4 제안 앙상블과 모델과 DNN 모델 100 개의 평균

Vote 6	*Positive Recall*	*Negative Recall*
Ensemble	0.850	0.751
DNN	0.843	0.738

기존 앙상블 방법인 Random Forest 와 XGBoost 결과와 비교하기 위해 Ensemble Vote 의 Threshold 을 11 로 조정해 Recall 수치를 맞추고 결과를 확인하였다. 위와 같은 방법으로 100 개의 모델을 추출하여 평균을 계산하였다. 제안하는 앙상블 모델의 평균이 비교 대비 Positive Recall(상환)은 대비 2% 증가하였고 Negative Recall (연체)은 동일한 수치를 보였다.

<표>5 제안 앙상블 모델의 100 개 평균과 RF,XGB 모형의 성능 결과

Vote 11	*Positive Recall*	*Negative Recall*
Ensemble	0.82	0.785
RF	0.80	0.78
XGB	0.80	0.79

4. 결론

본 논문에선 앙상블 성능 향상을 위해 후보 모델을 선별하고 앙상블을 구성하는 방법을 제안하였다. 기존 다양성 지표의 편향 문제를 개선한 지표와 가지치기 방법을 결합하여 적절한 상호 다양성을 보이는 후보 모델들을 선별했다. 유전자 알고리즘을 사용하여 앙상블을 구성할 모델들을 선택하여 결과를 확인하였다. 앙상블과 DNN 모델에서 동일하게 100 개의 모델을 추출하여 평균을 비교하였다. 또한, 기존 대표적인 앙상블 방법 RF 와 XGB 모델의 분류 결과를 Positive Recall, Negative Recall 의 수치를 산출하여 비교한 결과 두 비교에서 모두 제안한 모형의 예측율이 향상 되었음을 확인하였다.

참고문헌

[1] Bian, Shun, and Wenjia Wang. "On diversity and accuracy of homogeneous and heterogeneous ensembles." *International Journal of Hybrid Intelligent Systems* 4.2 (2007): 103-128.

[2] Bian, Yijun, and Huanhuan Chen. "When does diversity help generalization in classification ensembles?." *IEEE Transactions on Cybernetics* 52.9 (2021): 9059-9075.

[3] Kuncheva, Ludmila I., and Christopher J. Whitaker. "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy." *Machine learning* 51.2 (2003): 181.

[4] Wu, Yanzhao, et al. "Promoting high diversity ensemble learning with ensemblebench." *2020 IEEE Second International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 2020.

[5] 민성환. "부도예측을 위한 KNN 앙상블 모형의 동시 최적화." *지능정보연구* 22.1 (2016): 139-157..

[6] 박성욱, 김종찬, and 김도연. "앙상블 학습 알고리즘을 이용한 컨벌루션 신경망의 분류 성능 분석에 관한 연구." *멀티미디어학회논문지* 22.6 (2019): 665-675.

[7] Webb, Geoffrey I., and Zijian Zheng. "Multistrategy ensemble learning: Reducing error by combining ensemble learning techniques." *IEEE Transactions on Knowledge and Data Engineering* 16.8 (2004): 980-991.