ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
https://www.jiisonline.org
J Intell Inform Syst 2022 December: 28(4): 179~190 http://dx.doi.org/10.13088/jiis.2022.28.4.179

검증데이터 기반의 차별화된 이상데이터 처리를 통한 데이터 불균형 해소 방법

황철현
한양여자대학교 빅데이터과
(chhwang@hywoman.ac.kr)

데이터 불균형은 한 분류의 데이터 수가 다른 분류에 비해 지나치게 크거나 작은 현상을 의미하며. 이로 인해 분류 알고리즘을 활용하는 기계학습에서 성능을 저하시키는 주요 요인으로 제기되고 있다. 데이터 불균형 문제 해결을 위해서 소수 분포 데이터를 증폭하는 다양한 오버 샘플링(Over Sampling) 방법들이 제안되고 있다. 이 가운데 SMOTE는 가장 대표적인 방법으로 소수 분포 데이터의 증폭 효과를 극대화하기 위해 데이터에 포함된 잡음을 제거(SMOTE-IPF)하거나, 경계선만을 강화(Borderline SMOTE) 시키는 다양한 방법들이 출현하였다. 이 논문은 소수분류 데이터를 증폭하는 전통적인 SMOTE 방법에서 이상데이터(Anomaly Data)에 대한 처리방법개선을 통해 궁극적으로 분류성능을 높이는 방법을 제안한다. 제안 방법은 실험을 통해 기존 방법에 비해 상대적으로 높은 분류성능을 일관성 있게 제시하였다.

주제어 : 데이터 불균형, 데이터 증폭, 이상데이터, Borderline SMOTE

논문접수일 : 2022년 10월 16일 논문수정일 : 2022년 11월 20일 게재확정일 : 2022년 11월 25일
원고유형 : Regular Track 교신저자 : 황철현

1. 서론

데이터 불균형은 한 분류의 데이터 수가 다른 분류의 데이터 수에 비해 압도적으로 많거나 적은 경우를 말하며 데이터의 왜곡된 특성으로 인해 발생한다. 또한 이러한 문제는 기계학습 과정에서 분류 예측성능에 악영향을 미치는 것으로 알려져 있다(Ali, 2019).

데이터 불균형은 단지 교육 분야만의 문제가 아니라 사기 탐지, 위험예측, 사건·사고 예측 등과 같은 금융, 의료, 자연재해 관리 등의 다양한 분야에서 흔히 관측되는 현상이다(Wu, 2003; Ghorbani & Ghousi, 2020). 현실세계에서는 데이터가 희박하게 나타나는 소수분류에 더 많은 의미를 부여하고 집중적인 노력을 수행하지만, 불행히도 기계학습에서는 다수 분류에 데이터가 편향되어 소수 분류의 예측력이 떨어지는 문제가 있다.

예를 들어 최근 주목받고 있는 대학의 학생성과관리영역에서 학습부진 학생에 대한 조기예측 프로그램을 살펴보자. 학습부진 학생은 전체학생 가운데 항상 소수에 속하고, 높은 가치를 가지지만, 기계학습의 훈련과정에서 다수 분류인 전체 학생에 편향 학습되어 부진학생에 대한 식별 정확도가 현저하게 떨어지는 문제를 가진다.

이러한 데이터 불균형 문제를 해결하기 위한 노력 가운데 SMOTE(Synthetic Minority Over sampling Technique)는 가장 많이 활용되는 표준적인 방법이다. SMOTE는 소수 분류데이터 사이에 임의의

데이터를 생성(Over Sampling)하는 방법으로 데이터 전처리 방법 가운데 가장 영향력이 높은 방법이다. 특히 기존의 Random Over Sampling에 비해 과 적합(Over Fitting)을 상당 부분 감소시킨다는 장점이 있다(Krawczyk, 2016). 앞서 제시한 장점에도 불구하고 SMOTE는 샘플링 과정에서 소수 분류에 포함된 잡음까지 증폭시키거나, 데이터 이동(data shift) 문제에 취약한 단점이 있다(Cheng et al., 2019; Fernandez., 2018).

SMOTE-IPF와 Borderline SMOTE는 앞서 제시한 SMOTE의 단점을 보완한 방법이다. 우선 Borderline SMOTE는 소수 분류 데이터를 '안전, 위험, 경계' 구역의 데이터로 분류한 뒤 경계구역에만 샘플링 함으로써 분류 간 경계선을 강화시킨다. 이러한 방법들은 훈련데이터의 복잡성을 감소시킴으로써 훈련의 효율성을 증가시켜 분류 예측 의 정확도를 향상 시키는 효과를 기대할 수 있다(Han et al., 2017; Saez et al., 2015).

앞서 제시된 두 가지 방법은 데이터 불균형 해소를 위한 증폭 방법 가운데에서 가장 많이 활용되고 있음에도 불구하고 이상데이터에 대해서는 다루고 있지 않다. 이로 인해 이상데이터는 데이터 전처리 과정에서 대부분 제거된다.

반면, 최근에는 이상 데이터에 대한 중요성이 강조되면서, 멀티모달(Multi Modal)이나 적대적 생성모델(GAN)을 활용하여 이상상황을 탐지하는 연구가 늘어나고 있으며 검출 정확도 역시 증가하고 있다. 이와 같은 상황에서 이상데이터에 대해 어떻게 활용할 것인지에 대한 연구가 필요하다(Choi & Kim, 2019; Lee & Kim, 2022; Shin et al., 2021).

본 논문은 앞서 제시된 두 가지의 방법과는 달리 이상데이터를 제거하지 않고 검증데이터와 결합하는 방법을 제안한다. 또한 이상데이터 가운데 향후 활용 가능성이 높은 데이터를 식별하여 1차 증폭하고 기존의 SMOTE를 통해 2차 증폭하는 방법을 통해 성능을 향상 시킬 수 있는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 2장에서는 데이터 샘플링과 노이즈 필터링에 관한 기존 연구를 제시하고, 3장에서는 이상데이터를 활용한 노이즈 필터링과 샘플링 방법을 제안한다. 4장에서는 제안 방법의 성능측정을 위한 실험과정과 결과를 도출하고, 5장에서는 결론을 제시한다.

2. 관련 연구

2.1 데이터 불균형 관련 연구

분류 불균형 학습(CIL, Classification Imbalance Learning)이라고도 불리는 데이터 불균형은 지난 20년 동안 기계학습을 적용하는 사기탐지와 네트워크 침입탐지, 질병진단 등 다양한 분야에서 관찰되어 왔었고, 이를 극복하기 위한 관련 연구가 지속되어 왔다(Ali et al., 2019; Cheng et al, 2019).

데이터 불균형은 분류 간 데이터 분포차이로 인해 기계 학습모델이 다수 분류(majority class)에 편향되어 학습됨으로써 소수 분류(minority class)를 무시하는 문제이다. 이러한 데이터 불균형을 해소하기 위한 기존의 방법은 크게 훈련데이터를 직접 수정하는 샘플링을 수행하거나, 편향을 완화하고 불균형에 적응하도록 학습 알고리즘을 수정하는 방법, 그리고 앞의 두 가지 방법의 장점을 결합한 hybrid 방법 등이 있다(Krawczyk, 2016).

이 가운데에서도 데이터 불균형을 해결하기 위한 가장 기본적인 접근방법은 다수분류 데이터를 제거하는 언더샘플링(under sampling)과, 소수

분류 데이터를 생성하여 분류 간 데이터 분포를 일치시키는 오버샘플링(over sampling)방법이 있다. 하지만 언더샘플링은 학습데이터 부족 문제와 분류에 유용한 패턴이 손실되는 문제로 인해 활용이 제한되어, 오버샘플링이 주로 활용되고 있다(Fernandez., 2018).

2.2 SMOTE 관련 연구

SMOTE는 부족한 소수분류의 데이터를 생성하는 샘플링 과정에서 무작위로 데이터를 추출하거나 단순히 복제하는 기존의 방법에서 과적합(Over Fitting)이 발생되는 문제를 지적하고 이를 개선하고자 개발되었다.

2002년 Chawla 등에 의해 제안된 SMOTE 방법은 이후 과적합을 회피하면서도 분류 간 유사한 분포를 가지는 능력이 검증되어 많은 기계학습 모델에서 사용되었으며 사실상의 표준으로 인정받고 있다(Krawczyk, 2016; Chawla et al., 2002).

SMOTE는 소수분류의 데이터와 이웃데이터 사이에 무작위 보간(random interpolation)을 활용하여 샘플링을 수행한다. 다음 수식(1)은 SMOTE에서 보간에 의해 데이터를 생성하는 방법을 설명하고 있다.

$$X_{\neq w} = X_i + \lambda \times (X_{zi} - X_i), \lambda \in [0,1] \qquad (1)$$

Gazzah & Amara는 소수 분류의 데이터를 생성하는 sampling 과정에서보다 현실적인 데이터를 생성하는 방법을 제안하였다. 이러한 제안 방법들은 데이터의 가중치, 기하학적 모형, 다중 보간, 가우스 분포(Gaussian Distribution)와 같은 학습 데이터의 분포를 활용한다(Gazzah & Amara, 2008).

2.3 SMOTE 잡음 데이터 관련 연구

SMOTE는 잡음 증폭(noise amplification)이라는 문제를 안고 있다. 잡음 증폭이란 학습데이터 내부의 소수분류데이터에 잡음 데이터가 있는 경우, 잡음까지 증폭되어 성능이 저하되는 것을 말한다.

분류 알고리즘은 데이터 불균형보다 잡음에 더 예민하게 반응하는 것으로 알려져 있어, 자칫 데이터 불균형 해소로 인한 분류 알고리즘의 성능 이득보다 잡음 증폭으로 인한 손실이 더 클 수 있다.

소수분류 내에 포함된 잡음 데이터가 성능에 악영향을 미치는 방향으로 데이터가 증폭되는 것을 피하기 위한 첫 번째 방법은 SMOTE-IPF이다(Saez et al., 2015). 이 방법은 IPF(Iterative Partitioning Filter)라는 반복 앙상블 기반의 잡음 필터를 통해 잡음을 감지 및 제거하고 보다 규칙적인 분류 경계선을 구축하기 위해 사용된다.

또 다른 방법은 HAN 등에 의해 제안된 Borderline SMOTE로서 소수 분류 데이터를 '안전, 위험, 경계' 데이터로 분류한 뒤 안전·위험 데이터 주위에는 데이터를 증폭하지 않고, 경계 데이터 주위에만 집중적으로 데이터를 증폭하는 방법을 활용한다(Han et al., 2005).

잡음 데이터를 의미하는 위험데이터 구역의 데이터 주위에는 데이터를 증폭하지 않기 때문에 잡음 데이터 증폭을 회피할 수 있다.

앞서 과정에서 우리가 얻을 수 있는 이점은 훈련 과정에서 발생되는 복잡한 경계선을 단순화시킬 수 있다는 것이다. 이를 통해 궁극적으로는 분류 알고리즘의 과적합 발생 가능성을 저하시키고 안정적으로 시험과정의 예측 성능을 보장할 수 있다.

2.4 이상데이터 검출 관련 연구

이상데이터를 탐지하는 전통적인 방법으로는 K-Means, PCA, DBSCAN과 같은 비지도 학습을 주로 활용되었다.

최근에는 딥러닝을 기반으로 이상데이터를 검출하는 Deep AD의 출현으로 인해 장비의 적정 유지보수 시점을 예측하거나, 각종 시계열 데이터에서 이상데이터를 검출하기 위해 활용되고 있다(Serradilla et al., 2021; Nguyen et al., 2021). 이 논문은 이상데이터를 검출하는 것이 아니라 검출된 이상데이터에 대한 처리 방법을 개선하는 것을 목적으로 하기 때문에 이상데이터의 검출과정에 대해서는 다루지 않는다.

3. 이상데이터를 활용한 샘플링 방법

본 논문에서는 학습데이터에서 이상데이터를 검출하여 훈련과 검증데이터세트의 각 이웃데이터 사이에 데이터를 샘플링하는 방법을 제안한다. 이상 또는 잡음 데이터를 제거하거나 증폭과정에서 고립시켰던 기존의 방법과 달리 이상데이터를 각 상황에 맞게 증폭하여 활용하는 방법이다.

3.1 제안 방법의 의도

본 논문에서는 이상데이터를 단순히 제거 대상으로만 취급하지 말고 활용 가능성을 찾아보고자 하는 노력의 일환이다. 일반적으로 다수 데이터 보다 상대적으로 소수이기도 하지만 절대적인 수 측면에서도 소수이기 때문에 데이터 수와 다양성이 부족하다.

또한 이상데이터는 품질 문제를 포함하고 있는 잡음 데이터이기도 하지만 새로운 데이터 패턴의 출현으로 해석될 수 있다. 이러한 새로운 패턴의 데이터는 제거 대상이 아닌 적극적으로 증폭해야 하는 대상이다. 본 논문에서는 검출된 이상데이터가 새로운 데이터 패턴의 출현을 암시하는 데이터임을 판별하는 기준으로 검증데이터 세트를 활용하였다. 훈련데이터 보다 검증데이터에서 데이터의 이상 정도가 감소하였다면 증폭 대상으로 삼고자 한다.

추가적으로 본 논문은 기존의 데이터 증폭 방법인 SMOTE를 대체하는 방법이 아니라 SMOTE를 실행하기 이전에 이상데이터를 1차 증폭함으로써 2차 데이터 증폭 과정인 SMOTE의 효율성을 높이고자 하는 방법이다.

3.2 이상데이터 검출 및 구분 방법

본 논문에서 이상데이터는 훈련데이터세트 내부에 존재하는 소수분류 데이터 가운데 다른 데이터들과 떨어져 위치한 데이터로서 이웃데이터 간 밀도(density)를 기준으로 이상데이터를 추출한다.

밀도 기반의 이상데이터를 추출하는 가장 많이 사용하는 방법으로 DBSCAN을 들 수 있으며 본 논문에서도 이를 사용한다. DBSCAN은 수식 (2)와 같이 epsilon 거리 내 적정 수 이상의 데이터가 존재하는지를 기준으로 이상데이터를 검출하는 공간 군집 알고리즘이다(Serradilla et al., 2021).

$$eps(p) = q \in D | dist(p,q) \leq eps \quad (2)$$
$$|eps(q)| \geq Minpts$$

검출된 이상데이터는 훈련데이터 내 이웃 데이터 간 거리(이하 '훈련거리'라 칭함)와 검증데이터 간 거리(이하 '검증거리'라 칭함)를 산출하

여 이상데이터를 상세하게 구분한다. 만약 훈련거리가 검증거리보다 작은 경우에는 데이터 샘플링 환경이 훈련데이터세트에 적합하다고 판정하고, 반대로 검증거리가 훈련거리보다 작은 경우에는 검증거리가 소수분류의 샘플링 환경에 더 적합하다고 판정한다.

수식(3)은 훈련과 검증거리를 활용하여 샘플링에 적합한 데이터 세트를 도출하는 방법을 제시하였다.

$$Anomaly = \begin{cases} dist(N_{train}) \geq dist(N_{valid}) \; F_{train} \\ dist(N_{train}) < dist(N_{valid}) \; F_{valid} \end{cases} \quad (3)$$

** F_{train} : Friendly Train Data, F_{valid} : Friendly Valid Data*

다음 표1은 앞서 제시한 이상데이터 검출과 구분방법을 pseudo code로 구현하여 제시하였다.

〈표 1〉 이상데이터 검출 pseudo code

```
train, validation, test = split_data(Input)
anomaly_list = dbscan(train)
for   i=0, length(anomaly_list), i++ :
        if   (knn_dist(N_train, anomaly_list[i])≥
                knn_dist(N_valid, anomaly_list[i])) :
           then : F_train.append(anom[i])
           else : F_valid.append(anom[i])
```

3.3 이상데이터 처리 방법

앞 절에서 구분된 결과를 활용하여 이 절에서는 이상데이터와 이웃데이터의 사이에 무작위 데이터를 생성함으로써 이상데이터의 특이성을 감소시키고, 경계선을 명확하게 생성하도록 한다.

다음 수식(4)와 (5)는 이상데이터인 Ftrain(xi), Fvalid(xi)와 ngh(train(xi)), ngh(valid(xi)) 사이에 무작위 함수 u를 활용하여 데이터를 생성한다. 즉 훈련거리가 더 가까운 이상데이터(Ftrain (xi))는 훈련데이터 세트에서 이웃데이터를 선정(ngh(train(xi)))하고, 검증거리가 더 가까운 이상데이터(Fvalid(xi))는 검증데이터세트에서 이웃데이터를 선정 (ngh(valid(xi)))한다.

$$Synth_{train} = u \bullet (ngh(train(x_i)) - F_{train}(x_i)) \quad (4)$$
$$Synth_{valid} = u \bullet (ngh(valid(x_i)) - F_{valid}(x_i))$$
$$Synth = Synth_{train} + Synth_{valid}$$

다음 표2는 a와 b 사이의 무작위 데이터를 추출하는 함수인 synth(a,b)를 포함하여 이상데이터에 대한 샘플링을 훈련데이터와 검증데이터 각각의 집합에서 추출하는 방법을 pseudo code로 제시하였다.

〈표 2〉 데이터 샘플링 pseudo code

```
Synth_Data, Synth_train, Synth_valid = []
for   i, length (Ftrain(X_i)),i++) :
   for j, length (ngh_train(X_i)), j++) :
        Synth_k = Synth(ngh_train(X_i), Ftrain(X_i))
        Synth_train = Synth_train + Synth_k
for   i, length (Ftrain(X_i)),i++) :
   for j, length (ngh_valid(X_i)), j++) :
        Synth_k = Synth(ngh_train(X_i), Fvalid(X_i))
        Synth_valid = Synth_valid + Synth_k
Synth_Data = Synth_train + Synth_valid
```

4. 실험 및 결과 해석

4.1 실험 데이터 세트

제안된 이상데이터의 처리방법에 대한 효과검증을 위해 우리는 공개된 학생성과데이터를 사용하였다. 이 데이터는 포르투칼 Alentejo 지역의 2개 중등공립학교에서 수집된 357명 학생들의 수학성적 데이터이다(Cortez & Silva, 2008).

다음 표3은 실험에서 사용된 39개의 독립변수와 3개의 종속변수 설명과 데이터의 유형을 제시하였다.

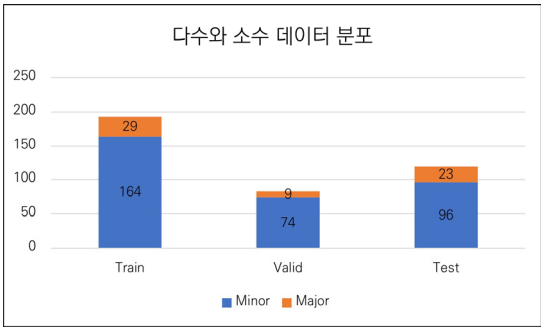
4.2 데이터 전처리

실험 데이터를 기계학습을 위해 필요한 데이터로 전환하기 위해 필요한 전처리 방법은 다음과 같다.

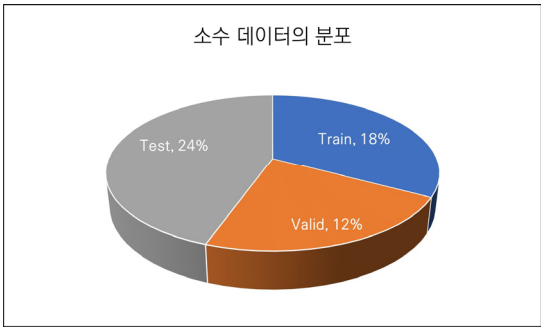
- 성적변수인 G1~G3의 3개 변수 가운데 최종 성적인 G3를 종속변수로 선정하였다.
- 학생성과에 대한 이진분류를 위해 G3의 값이 하위 5%인 6이하의 값을 가진 학생에 대해서는 '1' 값을 갖고, 7이상의 값을 가진 학생에 대해서는 '0'값을 갖는 'G3YN' 변수를 파생변수로 추가 생성하였다.
- 이진(binary) 특성을 가진 14개 변수와, 명목(nominal) 변수 특성을 가진 4개 속성에 대해 one-hot encoding을 수행하여 총 39개의 독립 변수가 확정되었다.
- 추출된 39개 독립변수에 대해 Scale을 일치시키는 Z-Score Standardization을 수행하였다.
- VIF(variance inflation factors)를 통해 다중공선성(multi colinearity)을 확인하였으나, VIF가 10을 넘지 않아 그대로 활용하였다.
- 제안 방법의 성능 검증을 위해 분류 알고리즘에서 사용할 훈련(train)용 193개, 검증(validation)용 83개, 시험(test)용 119개로 분리 구축하였다.
- 각 학습 데이터에서 소수분류는 훈련 29개(18%), 검증 9개(12%), 시험 23개(24%)를 차지한다. 단 실험 데이터의 분포는 실험과정에서 random state별로 실험 데이터가 새롭게 생성되기 때문에 차이가 발생한다. <그림 1>와 <그림 2>는 이를 도식화하여 제시하고 있다.

〈표 3〉 실험 데이터 구성

Name	Description	Type
school	student's school	binary
sex	student's sex	binary
age	student's age	numeric
address	student's home address type	binary
famsize	family size	binary
Pstatus	parent's cohabitation status	binary
Medu	mother's education	numeric
Fedu	father's education	numeric
Mjob	mother's job	nominal
Fjob	father's job	nominal
reason	reason to choose this school	nominal
guardian	student's guardian	nominal
taveltime	home to school travel time	numeric
studytime	weekly study time	numeric
failures	number of past class failures	numeric
schoolsup	extra educational support	binary
famsup	family educational support	binary
paid	extra paid class(course subject)	binary
activities	extra-curricular activities	binary
nursery	attended nursery school	binary
higher	wants to take higher education	binary
internet	internet access at home	binary
romantic	with a romantic relationship	binary
famrel	quality of family relationship	binary
freetime	free time after school	numeric
goout	going out with friends	numeric
Dalc	workday alcohol consumption	numeric
Walc	weekend alcohol consumption	numeric
health	current health status	numeric
absences	number of school absences	numeric
G1	first period grade1	numeric
G2	second period grade1	numeric
G3	final period grade1	numeric



〈그림 1〉 다수와 소수데이터의 분포



〈그림 2〉 학습데이터 종류별 소수데이터의 분포

4.3 실험 설계

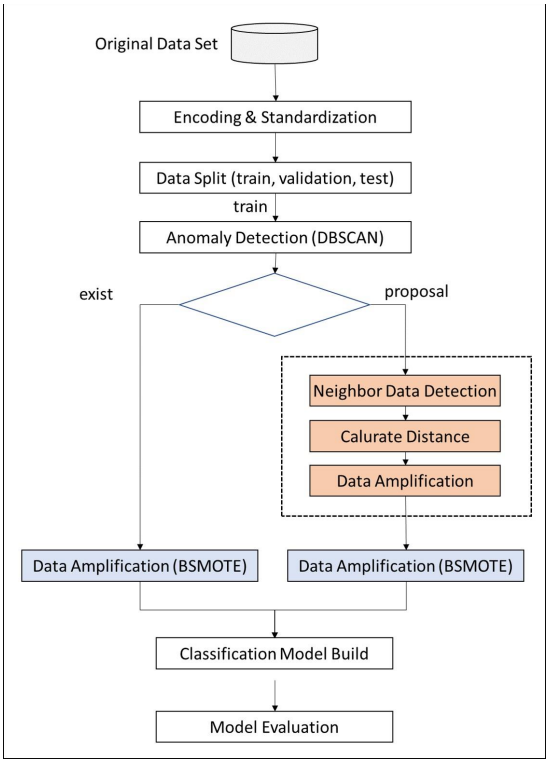
제안 방법이 분류 알고리즘의 성능을 향상시킬 수 있는지를 검증하기 위해 다음과 같은 환경을 구축하여 실험을 수행하였다.

- 다양한 분류 알고리즘 환경에서 수행된 결과를 수집하기 위해 XGBoost, Random Forest, LightGBM, CAT Boost, Decision Tree 등 총 5개 알고리즘을 활용하여 실험하였다.
- Random 값을 변화시켜 각 알고리즘 마다 100회의 데이터세트(train/valid/test)를 구축하였다.
- 각 알고리즘별로 데이터 세트에 대한 분류 모델을 구축할 때 마다 random search 방법을 활용하여 최적 parameter를 탐색하고 적용하였다.
- 제안방법에 따라 사전 증폭된 입력 데이터를 생성한 후 재차 Borderline SMOTE를 통해 불균형을 완전히 제거한 데이터 세트를 생성하였다.
- 제안 방법은 증폭 대상이 되는 이상데이터가 작거나 없을 경우 불균형을 완전히 제거하지 못하므로 Borderline SMOTE를 통해 불균형을 완전히 제거한다.
- 데이터를 증폭하지 않은 입력데이터(1종)를 Borderline을 통해 불균형을 제거한 데이터를 생성하였다.
- (case-1) Original Data → Proposal → BSMOTE, (case-2) Original Data → BSMOTE 두 사례에서 제안된 방법을 통해 사전 증폭된 데이터가 기존 BSMOTE 증폭방법을 통한 분류 방법의 성능에 미치는 영향을 파악할 수 있다.
- 분류 모델의 성능 지표는 ROC-AUC로 최적화한 후 측정 결과를 제안 방법과 기존 방법에 대해 비교한다.
- 제안 방법으로 인한 ROC-AUC값의 변화와 성능지표의 Top 3 분석을 통해 효과를 측정한다.

다음 그림 3은 실험을 위한 기능과 기능의 전체적인 흐름을 제시하였다. 두 방법 모두 기계학습 입력 전에 Borderline SMOTE 과정을 통해 데이터를 증폭하였기 때문에 두 방법의 데이터 모두 불균형 문제는 없는 상태이다. 따라서 제안 방법에 따른 성능효과를 기존 방법과 비교하기 용이하다.

4.4 실험 결과

실험에 적용된 분류 알고리즘에 대해 제안 방법이 미치는 영향은 표 4과 같다. 100회 반복을 통해 성능을 산출 한 결과, 각 실험에서 과반이 넘게 (55.8% ~ 74.0%) 기존 방법보다 높은 성능을 보인다.



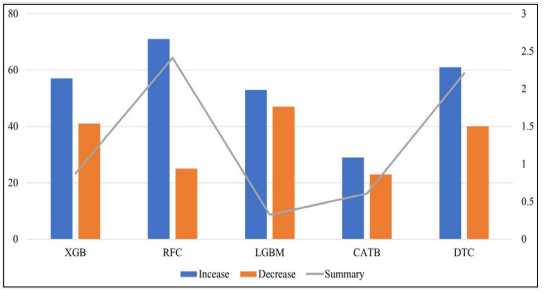
〈그림 3〉 제안 방법의 절차도

〈표 4〉 제안 방법의 효과 검증

Algorithms	Increase		Decrease		summary
	count	%	count	%	
XGBoost	57	58.2	41	40.6	(+)0.872
Random Forest	71	74.0	25	24.8	(+)2.412
Light GBM	53	53.0	47	46.5	(+)0.321
CatBoost	29	55.8	23	22.8	(+)0.607
Decision Tree	61	60.4	40	39.6	(+)2.204

제안 방법의 효과는 표 3에서 요약하여 제시하였다. 표 3은 증가분과 감소분을 상쇄시킨 값으로 제안 방법의 영향을 직관적으로 알 수 있다.

표 3에 의하면 각 알고리즘의 summary 값은 0.607에서 2.412까지 모두 증가('+')된 것으로 나타난다. 이와 같이 제안 방법은 실험 환경 하에서 분류 알고리즘의 성능을 향상시키고 있다는 것을 알 수 있다. 그림 2는 이를 도식화하여 제시하고 있다.



〈그림 4〉 분류 알고리즘별 성능 비교

다수의 실험 사례가 존재하는 기계학습사례에서 특히 중요한 것은 top 1 결과로써 가장 높은 성능이 도출되는 절차를 말한다. 이는 기계학습 구축과정에서 가장 높은 성능을 보이는 1개 모델만이 개발과정에서 실제로 구축되기 때문이다.

실험 결과에서 top 3를 분석하면 다음 표 5와 같다. top 3의 15개 사례 가운데에서 제안 방법을 통한 사례가 9개로 60%를 차지한다. 특히 총 5개 가운데 4개 알고리즘에서 제안 방법이 top 1 지표를 산출함으로써 제안 방법이 동일한 알고리즘에서 가장 높은 성능을 나타내는 확률이 80%에 달한다는 실험 결과를 도출하였다.

5. 결론

본 논문에서는 데이터 불균형으로 인한 분류 예측 성능이 떨어지는 데이터 환경에서 이상치를 활용하여 데이터를 증폭하는 방법을 제안하였다.

〈표 5〉 Top 3 성능 분석 결과

Algorithms	top 1		top 2		top 3	
	value	method	value	method	value	method
XGBoost	0.749	Legacy	0.745	Prop	0.740	Prop
Random Forest	0.646	Prop	0.621	Prop	0.610	Prop
Light GBM	0.786	Prop	0.768	Legacy	0.764	Legacy
CatBoost	0.703	Prop	0.681	Legacy	0.677	Legacy
Decision Tree	0.758	Prop	0.737	Legacy	0.728	Prop

이 방법은 전처리 과정에서 활용되기 때문에 기존 SMOTE와 같이 사용될 수 있다. 따라서 제안 방법에 따라 데이터를 1차 증폭하고, 증폭된 결과를 BSMOTE를 통해 증폭함으로써 최종 기계학습 데이터를 확정한다. 이를 통해 안정적인 학습데이터 증폭을 보장 받을 수 있다.

향후 연구 분야로는 데이터 증폭 패턴과 분류 알고리즘의 성능 변화의 관계를 규명하고, 데이터가 부족한 운영초기 환경에서도 안정적으로 분류성능을 유지할 수 있는 방법에 대한 연구가 필요하다.

참고문헌(References)

Ali, H., Salleh, M. N. M., Saedudin, R., Hussain, K., & Mushtaq, M. F. (2019). Imbalance class problems in data mining: a review. Indonesian Journal of Electrical Engineering and Computer Science, 14(3), 1560-1571.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research, 16, 321-357.

Cheng, K., Zhang, C., Yu, H., Yang, X., Zou, H., & Gao, S. (2019). Grouped SMOTE with noise filtering mechanism for classifying imbalanced data. IEEE Access, 7, 170668-170681.

Choi, N., & Kim, W. (2019). Anomaly Detection for User Action with Generative Adversarial Networks. Journal of Intelligence and Information Systems, 25(3), 43-62.

Cortez, P., & Silva, A. M. G. (2008). Using data mining to predict secondary school student performance.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. in: Proceedings of Second International Conference on Knowledge Discovery and Data Mining, Portland, OR, 226 – 231.

Fernández, A., Garcia, S., Herrera, F., & Chawla, N. V. (2018). SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research, 61, 863-905.

Gazzah, S., & Amara, N. E. B. (2008, September). New oversampling approaches based on polynomial fitting for imbalanced data sets. In 2008 the eighth iapr international workshop on document analysis systems (pp. 677-684). IEEE.

Ghorbani, R., & Ghousi, R. (2020). Comparing different resampling methods in predicting students' performance using machine learning techniques. IEEE Access, 8, 67899-67911.

Han, H., Wang, W. Y., & Mao, B. H. (2005, August). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In International conference on intelligent computing (pp. 878-887). Springer, Berlin, Heidelberg.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. Progress in Artificial Intelligence, 5(4), 221-232.

Lee, D., & Kim, N. (2022). Anomaly Detection Methodology Based on Multimodal Deep Learning. Journal of Intelligence and Information Systems, 28(2), 101-125.

Nguyen, H. D., Tran, K. P., Thomassey, S., & Hamad, M. (2021). Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management. International Journal of Information Management, 57, 102282.

Sáez, J. A., Luengo, J., Stefanowski, J., & Herrera, F. (2015). SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. Information Sciences, 291, 184-203.

Serradilla, O., Zugasti, E., Ramirez de Okariz, J., Rodriguez, J., & Zurutuza, U. (2021). Adaptable and explainable predictive maintenance: Semi-supervised deep learning for anomaly detection and diagnosis in press machine data. Applied Sciences, 11(16), 7376.

Shin, B., Lee, J., Han, S., & Park, C.-S. (2021). A Study of Anomaly Detection for ICT Infrastructure using Conditional Multimodal Autoencoder. Journal of Intelligence and Information Systems, 27(3), 57-73.

Wu, G., & Chang, E. Y. (2003, August). Class-boundary alignment for imbalanced dataset learning. In ICML 2003 workshop on learning from imbalanced data sets II, Washington, DC (pp. 49-56).

Abstract

Resolving data imbalance through differentiated anomaly data processing based on verification data

Chulhyun Hwang*

Data imbalance refers to a phenomenon in which the number of data in one category is too large or too small compared to another category. Due to this, it has been raised as a major factor that deteriorates performance in machine learning that utilizes classification algorithms. In order to solve the data imbalance problem, various ovrsampling methods for amplifying prime number distribution data have been proposed. Among them, SMOTE is the most representative method. In order to maximize the amplification effect of minority distribution data, various methods have emerged that remove noise included in data (SMOTE-IPF) or enhance only border lines (Borderline SMOTE). This paper proposes a method to ultimately improve classification performance by improving the processing method for anomaly data in the traditional SMOTE method that amplifies minority classification data. The proposed method consistently presented relatively high classification performance compared to the existing methods through experiments.

Key Words : Data Imbalance, Data Amplification, Anomaly Data, Borderline SMOTE

Received : October 16, 2022 Revised : November 20, 2022 Accepted : November 25, 2022
Corresponding Author : Chulhyun Hwang

* Corresponding Author: Chulhyun Hwang
Dept of Big Data, Hanyang Woman's University
200, Salgoji-gil, Seongdong-gu, Seoul, 04763, Korea
Tel: +82-2-2290-2201, Fax: +82-2-2290-2201, E-mail: chhwang@hywoman.ac.kr

저 자 소 개

황철현

현재 한양여자대학교 빅데이터과 교수로 재직 중이다. 배재대학교에서 컴퓨터공학을 전공하여 공학박사를 취득하였다. 주요 관심 분야는 Machine Learning, Tabular Data Deep Learning 등이다. 주요 논문을 IJECE, IJEECS, JICCE, JIPS, 한국정보통신학회지 등에 게재하였다.