JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지 Vol. 26, No. 11: 1608~1614, Nov. 2022

이상 데이터를 활용한 성과부진학생의 조기예측성능 향상

황철현*

Improvement of early prediction performance of under-performing students using anomaly data

Chul-Hyun Hwang*

*Assistant Professor, Department of Big Data, Hanyang Women's University, Seoul, 04763 Korea

요 약

최근 학생 수 감소로 인한 대학 간 경쟁이 심화되면서 성과부진학생을 조기에 예측하고, 중도이탈을 예방하기 위해 다양한 노력을 기울이는 것은 대학의 필수 업무로 인식되고 있다. 이를 위해서는 학생의 성과를 정밀하게 예측하는 우수한 성능의 모델이 필수적이다. 본 논문은 성과부진학생을 식별하기 위한 분류 예측 모델에서 이상 데이터를 제거하거나 증폭을 통해 예측 성능을 향상시키는 방법에 대해 제안한다. 기존 이상데이터 처리방법은 주로 데이터를 삭제하거나 무시하는데 집중되었지만 이 논문에서는 잡음과 변화지표를 구분하는 기준을 제시하고, 데이터를 삭제하거나 증폭함으로써 예측 모델의 성능을 높이는데 기여한다. 제안 방법의 검증을 위해 공개된 학습 성과 데이터를 활용한 실험에서 기존 방법에 비해 제안방법이 분류 성능을 향상시킬 수 있는 다수의 사례를 발견할 수 있었다.

ABSTRACT

As competition between universities intensifies due to the recent decrease in the number of students, it is recognized as an essential task of universities to predict students who are underperforming at an early stage and to make various efforts to prevent dropouts. For this, a high-performance model that accurately predicts student performance is essential. This paper proposes a method to improve prediction performance by removing or amplifying abnormal data in a classification prediction model for identifying underperforming students. Existing anomaly data processing methods have mainly focused on deleting or ignoring data, but this paper presents a criterion to distinguish noise from change indicators, and contributes to improving the performance of predictive models by deleting or amplifying data. In an experiment using open learning performance data for verification of the proposed method, we found a number of cases in which the proposed method can improve classification performance compared to the existing method.

키워드 : 학생성공, 학생성과예측, 데이터 증폭, 이상 데이터, 데이터 불균형

Keywords : Student Sucess, Student Performance Prediction, Data Amplification, Anomaly Data, Data Imbalance

Received 5 September 2022, **Revised** 22 September 2022, **Accepted** 30 September 2022

*** Corresponding Author** Chul-Hyun Hwang(E-mail:chhwang@kbu.ac.kr, Tel:+82-31-570-9606)
Assistant Professor, Department of Big Data, Hanyang Women's University, Seoul, 04763 Korea

Open Access http://doi.org/10.6109/jkiice.2022.26.11.1608 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

국내 대학의 입학 가능 인원이 2040년에는 28만 명까지 감소하면서 신입생 유치를 위한 대학 간 경쟁이 날로 심해지고 있다. 또한 이미 입학한 재학생의 경우에도 학습 부진과 타 대학으로의 재입학 등의 사유로 학업을 중도 포기하는 비율이 지속적으로 증가하고 있다[1].

실제로 국내 4년제 대학의 최근 3년간 중도탈락 학생 비율이 4.6%로 사상 최고치를 지속적으로 기록하고 있다. 문제의 심각성을 반영하여 정부(교육부)는 2021년 대학기본역량진단 평가에서부터 신입생 충원율과 함께 재학률에 대한 평가 비중을 대폭 학대하도록 하였다[2].

대학 학위교육에서 학습부진 등의 사유로 인한 중도탈락 문제는 단지 국내 문제만은 아니고 선진국 대부분이 겪는 문제이다. 미국의 경우, 전체 580개의 4년제 공립대학중 정시 졸업률이 50% 이상인 대학이 50여개 밖에 되지 않는 등 이미 매우 낮은 수준을 보이고 있다[3].

이러한 문제를 극복하고자 선진국을 중심으로 학생상담과 개인화된 교육프로그램을 연계하는 등 종합적인 대책을 마련하기 위해 노력하고 있다. 정시졸업 가능성이 낮은 학생들을 조기에 식별하는 예측모델은 성과부진학생 관리의 핵심 프로세스이다[4, 5].

학습성과 부진학생을 예측하기 위해 사용되는 기계학습 방법은 주로 분류 알고리즘이 많이 활용되어 왔다. 최근에 앙상블(Ensemble)이나 부스팅(Boosting)과 같은 통계기반 기계학습 알고리즘들이 많이 활용되고 있다. 또한 딥러닝에 대한 꾸준한 관심을 바탕으로 표 형태 데이터 기반 딥러닝(Tabular Deep Learning)에 대한 다양한 방법들이 제안되고 있다[6].

하지만 학습성과 부진학생 예측은 데이터 불균형으로 인해 예측성능이 저하되는 구조적인 문제를 가지고 있다. 데이터 불균형은 한 분류의 데이터 수가 다른 분류의 데이터 수에 압도적으로 많거나 적은 경우를 말하는데, 부진학생의 수가 전체 학생에서 항상 소수로 발생되기 때문에 데이터 불균형은 피할 수 없는 문제이다.

본 논문에서는 데이터 불균형을 해소하기 위해 우선 이상 데이터를 추출한 후, 이상데이터에서 삭제와 증폭 데이터로 구분하여 삭제와 증폭을 수행하는 1차 over-sampling을 수행한다. 1차 over sampling이 완료된 데이터를 활용하여 데이터 불균형을 해결하기 위한 SMOTE (Borderline SMOTE)를 최종적으로 수행한다.

본 논문의 구성은 다음과 같다. 2장에서는 학생성과 예측과 데이터 증폭 관련 기존 연구를 살펴보고, 3장에서는 논문의 제안한 방법인 이상데이터에 대한 검출과 증폭방법에 대해 살펴본다. 4장에서는 제안방법 효과를 검증할 수 있도록 구성된 실험환경과 측정방법에 대해 제시하고, 마지막으로 5장에서는 결론을 도출한다.

Ⅱ. 관련 연구

2.1. 학생성과예측 관련 연구

학생의 성공을 예측하거나 부진학생을 조기에 식별하는 연구는 다양한 교육 데이터 마이닝(EDM)과 기계학습 방법의 발전을 통해 성장해 왔다[5].

초기 연구에서는 학생·부모 학력과 직업, 신분, 장애 등의 사회 인구학적인 특징과 함께 교육 과정의 특징을 결합한 데이터를 주로 활용하였다. 그리고 성공여부를 조기에 예측하기 위해 Navie bayes, Decision Tree, MLP 등의 분류 알고리즘이 주로 활용되었다[6].

최근에는 초기의 알고리즘과 더불어 RandomForest, Adaboost, GBM과 같은 다양한 분류 알고리즘을 활용하여 성공 요인을 발견하거나 예측 성능을 더욱 발전시키기 위한 다양한 연구가 진행되었다[6~7].

또한 대용량의 교육 데이터를 처리하기 위해 빅데이터 플랫폼을 활용하는 방법, 성과 예측 분야에 딥러닝(Deep Learning) 기술을 적용하는 방법, 과거의 연구결과에 대한 체계적인 문헌 검토를 수행하려는 연구가 진행되었다[6]. 이 같은 활동은 학생성과예측이 실험실에서 벗어나 점차 상용화되고 서비스되어가고 있음을 알려준다.

2.2. 데이터 불균형 관련 연구

분류 불균형 학습(CIL, Classification Imbalance Learning)이라 불리는 데이터 불균형 문제는 학습 데이터에 포함된 데이터 분포 차이로 발생된다. 이는 기계학습 모델이 다수 분류(major class)에 편향되어 훈련됨으로써, 소수 분류(minority class)를 무시하게 되어 발생되는 문제이다.

데이터 불균형을 해결하기 위해 가장 많이 사용하는 방법은 SMOTE이다. SMOTE는 무작위나 단순 복제로 데이터가 증폭되는 과적합 문제를 극복하기 위해 2002

년 chawla 등에 의해 제안되었으며, 이후 많은 기계학습에서 사용되면서 사실상의 표준으로 간주되고 있다[8].

SMOTE는 소수 데이터 주변에 존재하는 k개의 인접 데이터 가운데 하나를 임의로 선정한 후 합성(Synthetic) 방법을 통해 가상의 데이터를 생성하는 방법이다[9]. 인접한 데이터를 찾기 위해서는 KNN 알고리즘을 활용하고 합성을 위해서는 보간(interpolation)을 활용한다. SMOTE의 데이터 생성 방법은 다음 수식 (1)과 같다.

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

하지만 SMOTE는 데이터를 증폭하면서 소수 분류에 잡음 데이터를 포함하고 있다면 잡음 데이터까지 증폭해버리는 문제점을 안고 있다. 특히 분류 알고리즘에서는 예측성능이 데이터 불균형보다 잡음 데이터에 더 민감하게 반응하기 때문에 잡음까지 증폭되는 문제는 매우 중요하다.

SMOTE-IPF와 Borderline SMOTE는 SMOTE가 잡음까지 증폭시키는 문제점을 해결하기 위해 제안되었다. 우선 S´aez 등이 제안한 잡음필터링 방법인 SMOTE-IPF는 IPF라는 반복 앙상블 기반의 잡음필터를 통해 잡음 문제를 해결하고자 한다 [10]. 이와 함께 HAN 등에 의해 제안된 Borderline SMOTE는 소수 분류 데이터를 '안전·위험·경계'의 3가지 유형으로 데이터를 분류한 뒤 경계 데이터만 증폭함으로써 잡음 데이터가 증폭되는 문제점을 피하도록 한다[11]. 본 논문에서는 SMOTE를 활용한 데이터 증폭 과정에서 잡음데이터를 처리하기 위한 다른 접근 방법을 제시한다.

2.3. 이상치 탐지 관련 연구

이상치를 탐지하기 위한 방법으로는 크게 지도 방식(Supervised Method)과 반지도 방식(Semi-Supervised Method), 비지도 방식(Unsupervised Method)으로 나눌 수 있다.

지도에 의한 이상치 탐지 방법은 검출 정확도가 높은 장점이 있지만 학습을 위한 이상치 샘플을 확보하기 위해 시간과 비용이 많이 드는 단점이 있다. 특히 정상과 이상치의 데이터 분포에 많은 차이가 있는 경우에 활용하기 힘들다.

반지도 방식은 이상과 이상치의 분포 차이가 매우 큰 경우, 정상 데이터를 둘러싼 discriminative boundary를 설정하고, boundary를 최대한 좁혀 boundary 밖의 데이터를 모두 비정상으로 간주하는 방법으로 one-class classification이라고도 부른다. one-class classification에서 deep learning을 활용하는 deep AD는 특별한 라벨 없이 활용 가능한 이점으로 인해 다양한 산업에서 활용되고 있으며[12], 특히 예측유지보수나 의료영상의 이상 탐지와 같이 고전적인 알고리즘으로 힘든 문제 영역에서도 높은 성능을 내고 있다[13][14].

비지도 방식은 차원 축소 알고리즘이 주로 활용되는데 차원 축소와 복원 과정을 통해 이상치를 검출한다. PCA(Principal Component Analysis) 알고리즘을 주로 활용되었으며 최근에는 neural network 기반의 Autoencoder에 대한 연구가 많이 이뤄지고 있다. 최근에는 LSTM과 결합하여 시계열 데이터에서 이상 데이터를 탐지하는 영역에서도 활용되고 있다[15][16].

Ⅲ. 이상 데이터 처리방법

3.1. 이상 데이터의 검출 및 분류 방법

학습부진학생에 대한 조기 예측에서 이상 데이터라고 함은 어떤 학생의 데이터가 다른 학생의 데이터와 현저하게 차이가 나는 경우를 말한다. 이는 밀도 또는 거리와 관련이 있다. 따라서 이상 데이터(anomaly data)를 검출하기 위한 알고리즘으로 밀도 기반의 공간 군집 알고리즘인 DBSCAN(Density-based spatial clustering of applications with noise)을 활용한다.

DBSCAN은 각 데이터의 epsilon 거리 내 적중 수 이상의 데이터가 존재하는지를 기준으로 이상데이터를 검출하는 공간 군집 알고리즘이다[16]. 다음 수식(2)는 DBSCAN에서 이상데이터를 검출하는 방법을 제시한다.

$$\begin{aligned} & eps(p) = q \in D | dist(p,q) \leq eps \\ & |eps(q)| \geq Minpts \end{aligned} \quad (2)$$

이상데이터가 검출되면 '순수잡음'과 '변화지표'의 2가지 분류로 구분한다.

첫째, 순수잡음이란 발생·수집과정에서 외부 잡음이 포함되었거나, 다시는 재현되지 않을 드문 사례를 말한다. 이 데이터는 성능에 악영향을 주므로, 훈련 데이터에서 삭제하거나 증폭과정에서 배제한다.

둘째, 변화지표란 소수 분류의 데이터 부족으로 인해 정상 데이터가 필요한 최소한의 데이터를 확보하지 못

한 경우이다. 이런 데이터는 미래의 데이터 발생 징후를 담고 있어 높은 우선순위로 증폭해야 한다.

본 논문에서는 순수잡음과 변화지표 데이터를 구분하기 위해 훈련데이터와 검증데이터 세트에서 각각의 이웃데이터를 산출한 후, 만약 검증데이터에서 이웃 데이터 간 거리가 가깝다면 변화지표로 분류한다.

다음 수식 (3)은 이상데이터에서 잡음과 변화지표 데이터를 구분하는 방법을 제시하였고, 표 1에서는 pseudo code를 제시하였다.

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

** PN : Pure Noise, CI : Change Indicator*

Table. 1 Pseudocode for anomaly data detection

```
1 : train, validation, test = split(Input)
2 : anom = dbscan(train)
3 : for  i, length(anom) :
        if ( dist(N_train(anom[i])) ≥ dist(N_valid(anom[i])) :
          then : Noise.append(anom[i])
          else : ChgIdx.append(anom[i])
```

3.2. 이상 데이터의 처리 방법

순수잡음 데이터에 대한 처리방법은 훈련데이터세트에서 데이터를 삭제하도록 한다. 이를 통해 다른 데이터에 미치는 영향을 최소화할 수 있다. 데이터가 부족한 소수 분류의 데이터를 삭제하는 것은 매우 위험한 결정이지만 실제 잡음 데이터는 성능에 미치는 악영향이 더 크기 때문에 삭제하도록 한다. 변화지표 데이터는 발생 가능성이 높은 데이터이기 때문에 주변의 이웃 데이터와 사이에 데이터를 생성하는 증폭과정을 거친다.

수식 (4)는 이상 데이터 주위에 이웃 데이터를 기반으로 새로운 데이터를 생성하는 방법을 제시한다. p'_i는 이상데이터로 검출된 데이터이며 r은 0과 1 사이의 값을 임의로 추출한 값이다. $dist_{ij}$는 이상 데이터와 가장 근접한 이웃 데이터 사이의 거리를 나타낸다. 이상 데이터와 근접한 부진학생 데이터 사이에 증폭할 데이터를 생성한다.

$$synthetic_{ij} = p'_i + r \times \frac{dist_{ij}}{2} \quad (4)$$

다음 표2는 순수잡음과 변화지표로 구분하는 방법을 pseudo code로 제시하였다. 훈련이웃거리(Train Distance)와 검증이웃거리(Valid Distance)를 우선 산정하여 비교한 후 검증이웃거리가 더 가까울 경우에 변화지표로 선정한다. 검증이웃거리가 더 가깝다는 것은 현재보다 미래에 더 중요한 데이터로 취급되기 때문에 이를 반영한 것이다.

Table. 2 Pseudocode for anomaly data classification

```
1 : for  i, length(ChgIdx) :
2 :   for  j, length(N_train(anom[i,j]) :
3:      if ( dist(N_train(anom[i])) ≥ dist(N_valid(anom[i])) :
4:         then : Noise.append(anom[i])
5:         else : ChgIdx.append(anom[i])
```

Ⅳ. 실험 및 결과해석

4.1. 실험 데이터 세트

제안된 이상데이터 처리방법의 효과를 검증하기 위한 실험에서 UCI Machine Learning Repository에 공개된 학습 성과 데이터를 활용한다.

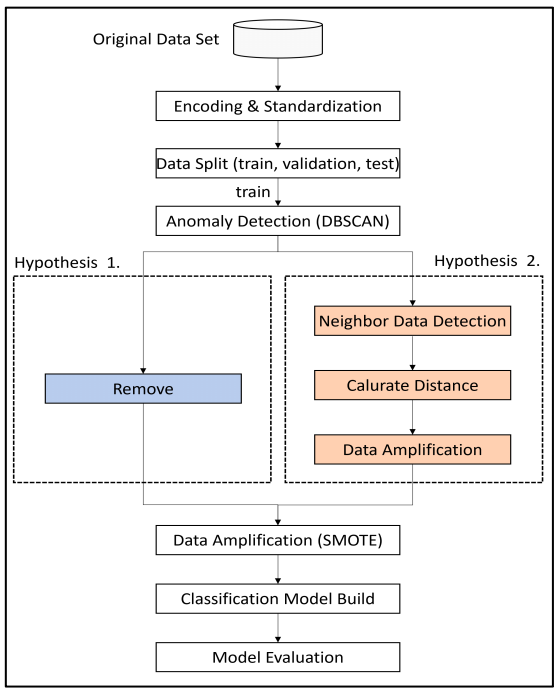

Fig. 1 Anomaly data processing flow

이 데이터는 포르투칼 Alentejo 지역의 2개 중등공립 학교에서 수집된 Student Performance Data Set이다 [17]. 실험 데이터에는 성별, 나이, 부모의 직업, 교육 수준, 통학거리, 주당 공부시간 등 30개의 Background 변수와 함께 3개(G1,G2,G3) 성적변수가 포함되어 있으며, 논문에서는 Background 변수를 독립변수로, 최초 성적을 나타내는 G3변수를 종속변수로 활용하였다.

4.2. 데이터 전처리

실험데이터를 부진학생 예측모델에서 사용하기 위해 다음과 같은 절차로 전처리 과정을 수행하였다.

- G1 값이 하위 5%인 0~6 사이는 '1', 나머지는 '0'값을 가지는 G1YN 변수 추가
- 이진(binary)과 명목(nominal) 변수 특성을 가진 26개 속성에 대해 one-hot encoding을 수행
- 39개 독립변수에 대해 Scale을 일치시키는 Z-Score Standardization을 수행
- VIF(variance inflation factors)를 산출하여 변수간 다중 공선성(multicolinearity)을 확인
- 훈련(train)용 174개, 검증(validation)용 75개, 시험(test)용 108개로 구분하여 구축

4.3. 실험 설계

제안된 이상데이터 처리방법이 분류예측 모델에 미치는 영향을 살펴보기 위해 다음의 환경에서 실험을 수행한다.

- 제한된 실험환경에서 실험 횟수를 증가시키기 위해 Random_Stats를 0~101까지 증가 시키면서 실험
- XG Boost, Random Forest, Light GBM, Cat Boost, Decision Tree, Neural Network의 총 6가지 분류 알고리즘을 활용
- 각 분류 알고리즘의 성능 최적화를 위해 Random Search Optimization과, k=3인 교차 검증을 수행
- 기존 데이터에 Borderline SMOTE 증폭방법을 적용한 분류 성능과 제안 방법에 의한 분류 성능을 상호 비교
- 성능지표는 상세지표를 종합한 ROC-AUC를 활용

4.4. 실험 결과

제안 방법의 효과를 규명하기 위해 각 알고리즘이 제안 방법을 적용하기 전·후로 구분하여 ROC-AUC를 산출하여 보면 다음 표 3과 같다. Random Fores, LightGBM, Decision Tree는 ROC-AUC가 상승되었으며, 나머지 XGBoost, CatBoostm Neural Net은 ROC-AUC가 감소되었다.

표 4에서는 대응표본 t-검정(Matched Sample T-Test)을 활용하여 ROC-AUC값의 통계적 유의성을 살펴보았다.

Table. 3 Descriptive statistics of ROC-AUC value

Algorithms	Before		After	
	mean	std	mean	std
XGBoost	0.619	0.053	0.617	0.053
Random Forest	0.510	0.021	0.521	0.029
Light GBM	0.618	0.062	0.625	0.061
CatBoost	0.522	0.047	0.510	0.030
Decision Treee	0.574	0.066	0.589	0.067
Neural Net	0.558	0.049	0.538	0.061

Table. 4 Matched sample t-test result

Algorithms	t-test			
	statistics	p-value	mean(x1)	mean(x2)
XGBoost	0.179	0.861	0.562	0.559
Random Forest	-1.993	0.072	0.538	0.590
Light GBM	-2.031	0.067	0.551	0.582
CatBoost	-0.341	0.740	0.574	0.582
Decision Tree	-0.867	0.404	0.569	0.590
Neural Net	-0.512	0.619	0.592	0.597

random forest, light GBM, Decision Tree의 3가지 알고리즘에서 $p < 0.5$ 이하로 통계적으로 유의하게 나타났다. 이 제안 방법을 적용하기 전·후로 비교하면 ROC-AUC가 약 2%에서 5%까지 상승한 것을 알 수 있다. 따라서 제안 방법은 제시한 3가지 알고리즘으로 구현될 경우 유의미한 효과를 볼 수 있다는 것을 알 수 있다.

다음 표 5와 그림 2은 앞서 통계적으로 유의하다고 판정된 3가지 알고리즘에 대해 각각 100번의 실험과정에서 산출된 성능지표(ROC-AUC)를 기존 SMOTE 방법에 의한 성능지표와 비교한 자료를 제시하였다.

Table. 5 Performance change by algorism

Algorithms	Positive		Negative	
	Count	Rate	Count	Rate
Random Forest	65	68%	31	32%
Light GBM	59	56%	41	41%
Decision Tree	55	55%	45	45%

통계적으로 유의한 수준의 변화를 보이는 3가지 알고리즘의 경우 전체 실험 사례 가운데 50% 수준 이상으로 성능 향상이 관찰되었다. 표 3에서 보듯이 성능 증가(55~68)와 감소 사례(32-45)의 비율은 뚜렷한 차이를 보이고 있다.

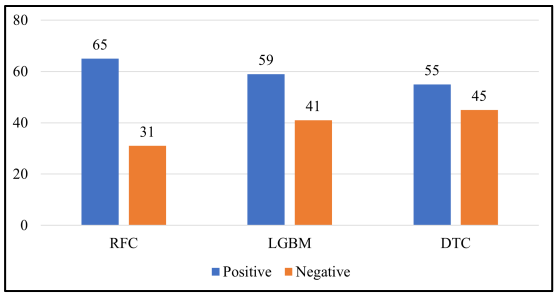

Fig. 2 Performance change by algorithm

다음 표 4는 통계적으로 유의한 3가지 알고리즘의 최고 성능을 제시하였다. 알고리즘별로 총 200회 (기존:100회, 제안:100회) 실험에서 3가지 알고리즘 모두 개선 방법에서 최고 성능이 산출되었다.

Table. 6 Best Indicator by reference

Algorithms	Best Indicator	Reference
Random Forest	0.623762	Proposal
Light GBM	0.803211	Proposal
Decision Tree	0.754982	Proposal

Ⅴ. 결 론

논문에서는 학생성공의 주요 척도인 학습 성과예측에서 부진학생예측 정확도를 높이기 위해 이상데이터를 활용하는 방법을 제안하였다. 제안 방법은 이상데이터로 추출된 데이터를 순수잡음, 변화지표의 2가지 유형으로 구분한 후에 순수잡음은 훈련데이터에서 제거하고, 변화지표는 이웃 데이터를 증폭하는 방법을 제안하였다.

제안 방법의 효과를 검증하기 위해 공개된 학생성과 데이터를 기반으로 총 6개의 분류 알고리즘을 활용한 실험을 수행하였고, 이 가운데 통계적으로 유의한 3개의 분류 알고리즘 사례를 도출하였다.

통계적으로 유의한 3개 알고리즘별로 각각 구현된 100회의 실험사례에서 가장 높은 지표를 제안방법으로 구현하여 달성하였고, 기존의 방법에 비해 지표가 상승된 비율이 50% 넘는 등 다양한 학습상황에서 성능을 향상 시킬 수 있음을 확인하였다.

본 논문은 위의 연구결과를 바탕으로 학생의 학습성과 이력데이터가 기계학습 모델에 입력되기 전에 이상데이터를 검출하고, 잡음분리과정을 통해 증폭 데이터를 선정하는 기준을 제시하였다. 이를 통해 부진학생예측에 대한 정확도를 향상시키는데 기여하였다.

References

[1] D. W. Youn, “The present and future of university restructuring (Focused on quota policy),” *Korea Higher Education Research Institute*, pp. 1-54, Dec. 2021.

[2] Ministry of Education. 2021 Basic Competency Assessment for Universities Basic plan [Internet]. Available: https://www.moe.go.kr/boardCnts/view.do?boardID=294&boardSeq=78253&lev=0&searchType=null&statusYN=W&page=1&s=moe&m=020402&opType=N.

[3] Complete College America, Four-year myth: Making college more affordable [Internet]. Available: http://completecollege.org/wp-content/ uploads/2014/11/ 4-Year- Myth.pdf, 2014.

[4] H. Lakkaraju, E. Aguiar, C. Shan, D. Miller, N. Bhanpuri, R. Ghani, and K. L. Addison, “A Machine Learning Framework to Identify Students at Risk of Adverse Academic Outcome,” in *Proceedings of the 21st ACM SIGKDD, International Conference on Knowledge Discovery and Data*, Sydney, Australia, pp. 1909-1918, 2015.

[5] B. Albreiki, N. Zaki, and H. Alashwal, “A Systematic Literature Review of Student’ Performance Prediction Using Machine Learning Techniques,” *Education Science*, vol. 11, no. 9, pp. 1-27, Sep. 2020.

[6] E. Alyahyan and D. Düstegör, “Predicting academic success in higher education: Literature review and best practices,” *International Journal of Educational Technology in Higher Education*, vol. 17, no. 3, Feb. 2020.

[7] W. Xing and D. Du, “Dropout Prediction in MOOCs: Using Deep Learning for Personalized Intervention,” *Journal of Educational Computing Research*, vol. 57, no. 3, pp. 547 - 570, Mar. 2019.

[8] B. Krawczyk, “Learning from imbalanced data: open challenges and future directions,” *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221-232, Apr. 2016.

[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, no. 1, pp. 321–357, Dec. 2019.

[10] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, “SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering,” *Information Sciences*, vol. 291, no. 10, pp. 184-203, Jan. 2015.

[11] H. Han, W. Wang, and B. Mao, “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning,” in *Proceedings of International Conference on Intelligent Computing*, Berlin, Heidelberg, pp. 878-887, 2017.

[12] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Smek, M. Kloft, T. G. Dietterich, and K. -R. Muller, "A Unifying Review of Deep and Shallow Anomaly Detection," in *Proceedings of the IEEE*, vol. 109, no. 5, pp. 756-795, May 2021.

[13] O. Serradilla, E. Zugasti, J. Ramirez de Okariz, J. Rodriguez, and U. Zurutuza, “Adaptable and Explainable Predictive Maintenance: Semi-Supervised Deep Learning for Anomaly Detection and Diagnosis in Press Machine Data,” *Applied Sciences*, vol. 11, no. 16, pp. 73-76, Aug. 2021.

[14] H. Zhao, Y. Li, N. He, K. Ma, L. Fang, H. Li, and Y. Zheng, "Anomaly Detection for Medical Images Using Self-Supervised and Translation-Consistent Features," *IEEE Transactions on Medical Imaging*, vol. 40, no. 12, pp. 3641-3651, Dec. 2021.

[15] H. D. Nguyen, K. P. Tran, S. Thomassey, and M. Hamad, "Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management," *International Journal of Information Management*, vol. 27, pp. 102282, Apr. 2021.

[16] M. Easter, H. P. Kriegel, J. Sander, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in large Spatial Databases with Noise," in *KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Portland: OR, USA, pp. 226-231, 1996.

[17] P. Cortez and A. Silva. “Using Data Mining to Predict Secondary School Student Performance,” in *Proceedings of 5th Future Business Technology Conference (FUBUTEC 2008)*, Portugal, pp. 5-12, 2008.

황철현(Chul-Hyun Hwang)

1991년 금오공과대학교 전자공학과(공학사)
1995년 경남대학교 컴퓨터공학과(공학석사)
2015년 배재대학교 컴퓨터공학과(공학박사)
2022~현재 한양여자대학교 빅데이터과 조교수
※관심분야 : 빅데이터, 인공지능