COPD 환자 운동 예측을 위한 불균형 데이터 처리 기법의 효율성에 관한 연구

진현석[1], 조세현[2], 최자윤[2], 김경백[1]
[1]전남대학교 인공지능융합학과
[2]전남대학교 간호학과
ggyo003@jnu.ac.kr, sehyunstar@naver.com, choijy@jnu.ac.kr, kyungbaekkim@jnu.ac.kr

A Study on the Efficiency of Imbalanced Data Processing Techniques for Exercise Prediction in COPD Patients

Hyeonseok Jin[1], Sehyun Cho[2], Jayun Choi[2], Kyungbaek Kim[1]
[1]Dept. of Artificial Intelligence Convergence, Chonnam National University
[2]Dept. of Nursing, Chonnam National University

Abstract

COPD(Chronic Obstructive Pulmonary Disease)는 장기간에 걸쳐 기도가 좁아지는 폐질환으로, 규칙적 운동은 호흡을 용이하게 하고 증상을 개선할 수 있는 주요 자가관리 중재법 중 하나이다. 건강정보 데이터와 인공지능을 사용하여 규직적 운동 이행군과 불이행군을 선별하여 자가관리 취약 집단을 파악하는 것은 질병관리 측면에서 비용효과적인 전략이다. 하지만 많은 양의 데이터를 확보하기 어렵고, 규칙적 운동군과 그렇지 않은 환자의 비율이 상이하기 때문에 인공지능 모델의 전체적인 선별 능력을 향상시키기 어렵다는 한계가 있다. 이러한 한계를 극복하기 위해 본 연구에서는 국민건강영양조사 데이터를 사용하여 머신러닝 모델인 XGBoost와 딥러닝 모델인 MLP에 오버샘플링, 언더샘플링, 가중치 부여 등 불균형 데이터 처리 기법을 적용 후 성능을 비교하여 가장 효과적인 불균형 데이터 처리 기법을 제시한다.

1. 서론

최근 장기간에 걸쳐 기도가 좁아지는 폐질환인 COPD(Chronic Obstructive Pulmonary Disease) 환자의 비율이 증가하고 있다. 2015-2019년 국민건강영양조사에따르면, COPD 발병은 5년 평균 12.9%였고, 70세 이상 응답자의 약 1/3 정도에서 발생하여 인구 고령화에 따른 증가가 뚜렷한 질환이다[1].

규칙적인 운동은 COPD 환자의 증상을 개선하고 호흡을 용이하게 하는 주요 자가관리법이다. COPD 진료지침에서는 신체활동으로 인한 일반적인 이점과 심혈관계 질환에 미치는 이점을 고려하여 COPD 환자의 규칙적인 신체활동을 권장하고 있다. 또한 규칙적인 운동을 포함한 호흡재활은 증상을 완화시키고, 삶의 질을 향상시킬 수 있어 환자의 지속적인 참여를 권장하고 있다[2]. 이를 위해 건강정보 데이터와 인공지능을 사용하여 COPD 환자 중 규칙적 운동 이행군과 불이행군을 선별하여 자가관리 취약 집단을 확인함으로써 환자관리의 경제적인 측면에서 긍정적인 효과를 낼 수 있다[3-4]. 하지만 많은 양의 데이터를 확보하기 어렵고, 운동이 필요한 환자와 그렇지 않은 환자의 비율이 상이한 클래스 불균형 문제로 인해 인공지능 모델의 전체적인 선별 능력을 향상시키기 어렵다는 한계가 존재한다.

본 논문에서는 이러한 한계를 극복하기 위해 국민건강영양조사 데이터에 불균형 데이터 처리 기법을 적용하고 성능을 비교하여 가장 효과적인 불균형 데이터 처리 기법을 제시한다.

2. 불균형 데이터 처리 기법

대표적인 불균형 데이터 처리 기법에는 오버 샘플링, 언더 샘플링, 가중치 부여 기법이 있다. 각각의 불균형 데이터 처리 기법은 데이터의 특성과 목적에 따라 효과가 달라질 수 있다. 본 연구에서는 이러한 기법들을 COPD 환자 건강정보 데이터에 적용하여 운동 예측 모델의 성능을 향상시키는데 가장 효과적인 기법을 탐색한다.

2.1 오버샘플링

오버샘플링은 데이터 세트의 균형을 맞추기 위해 소수 클래스의 데이터를 인위적으로 증가시켜 균형

을 맞추는 기법이다. 대표적으로 소수 클래스 데이터 포인트 사이를 보간하여 새로운 데이터를 생성하는 SMOTE(Synthetic Minority Over-sampling Technique)와 같은 통계 기반의 오버샘플링 기법이 사용되며, 최근에는 Variational AutoEncoder(VAE)[5]를 기반으로 데이터의 분포를 학습하여 새로운 데이터를 생성하는 TVAE(Triplet-based Variational Auto Encoder)[6], GAN(Generative Adversarial Network)[7]을 기반으로 조건부 확률 분포를 학습하여 유사한 데이터를 생성하는 CTGAN(Conditional Tabular Generative Adversarial Network)[8] 등 딥러닝을 기반으로 한 오버 샘플링 기법이 사용되고 있다. 오버샘플링은 데이터의 손실 없이 소수 클래스 데이터를 증가시켜 더 복잡한 패턴을 학습할 수 있지만, 데이터의 양이 증가함으로 인해 계산 비용 및 과적합 위험이 증가한다는 단점이 있다.

2.2 언더샘플링

언더샘플링은 다수 클래스의 데이터를 줄여 균형을 맞추는 기법이다. 대표적으로 Tomek Links와 같이 다수 클래스와 소수 클래스 사이의 경계에 위치한 데이터 포인트를 제거하는 통계 기반 방법이 사용된다. 언더샘플링은 데이터를 줄여 계산 비용을 줄일 수 있다는 장점이 있지만, 중요한 정보가 제거될 수 있다는 단점이 있다.

2.3 가중치 부여

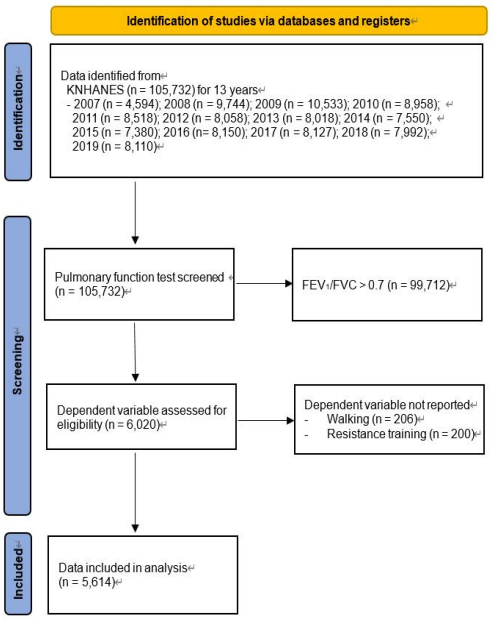
가중치 부여 방법은 소수 클래스의 데이터를 학습할 때 수식 1과 같이 손실 함수에 높은 가중치를 부여하여 불균형 데이터를 학습하는 기법이다. 가중치 부여 방법은 실제데이터의 분포를 변경하지 않아 오버샘플링 대비 과적합 위험이 상대적으로 낮다는 장점이 있지만, 모델 별로 가중치 부여 방식이 다르고 적절한 가중치를 설정하지 못하면 다수 클래스의 예측 성능이 저하될 수 있다는 단점이 있다.

$$weighted\ loss = \frac{1}{n}\sum_{i=1}^{n} w_j * \ell(y_i, \hat{y_i}) \quad (1)$$

3. 실험

3.1 데이터셋 생성 및 전처리

본 논문에서는 2007년도 부터 2019년도까지 수집된 국민건강영양조사 데이터[9]를 사용하였다. COPD 환자 운동 예측을 위해 40세 이상, COPD 진단을 받은 환자 데이터를 추출하였으며, 종속변수인 운동여부(BE5_1) 특징은 규칙적인 운동의 기준인 5일을 절단점으로 하여[10] 운동 이행군과 불이행 환자로 분류하였다. 자세한 프로세스는 그림 1과 같다. 또한 체계적문헌 고찰을 통해 운동 예측과 관련성이 있다고 판단된 53개의 독립변수 중 변수중요도 및 결측치 비율을 고려하여 최종적으로 34개의 독립변수를 갖는 5060건의 데이터를 추출하였으며, 8:2의 비율로 분할하여 각각 학습, 테스트에 사용하였다.



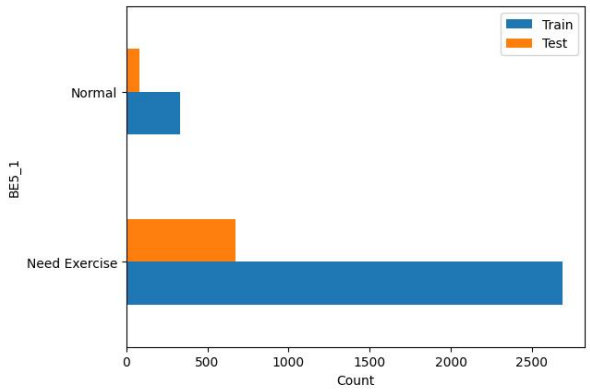
(그림 1) COPD 환자 데이터 추출 프로세스.

3.2 불균형 데이터 처리 기법 비교

Intel(R) Core(TM) i7-10700, NVIDIA GeForce RTX 3070, Scikit-Learn 1.4.0, XGBoost 2.0.3, Tensorflow 2.10 환경에서 COPD 환자 운동 예측 모델의 성능을 향상시키는데 가장 효과적인 불균형 데이터 처리 기법을 비교 분석하기 위해 오버샘플링(SMOTE, CTGAN), 언더 샘플링(Tomek Links), 가중치 부여 방법을 머신러닝 모델인 XGBoost(Extreme Gradient Boosting)[11], 딥러닝 모델인 MLP(Multi-layer Perceptron)에 적용하였으며, CTGAN은 SDV 라이브러리[12]를 사용하여 구현하였다. 평가는 그림 2와 같이 불균형한 데이터의 특징을 고려하여 수식 2로 계산되는 클래스별 F1 Score에 전체 데이터 개수를 클래스 개수로 나눈 값으로 가중치를 부여하는 Weighted F1 Score를 사용하였으며, 계산식은 수식 3과 같다.

$$F1Score = 2 * \frac{(\mathrm{Pr}eision * Recall)}{(\mathrm{Pr}eision + Recall)} \quad (2)$$

$$Weighted\ F1\ Score = \sum_{i=1}^{k} w_i * F1\ Score_i \quad (3)$$



(그림 2) COPD 환자 운동 클래스 분포.

실험 결과 표 1과 같이 가중치 부여 기법이 XGBoost, MLP를 대상으로 가장 높은 Weighted F1 Score를 달성하여 오버샘플링, 언더샘플링 기법 대비 데이터의 손실이 없고, 실제 데이터의 분포를 변경하지 않음으로 인해 소수클래스의 특징을 가장 효과적으로 반영할 수 있음을 확인하였다.

<표 1> 불균형 데이터 처리 기법 비교 결과

모델	적용 기법	Weighted-F1
XGBoost	미적용	0.84
	SMOTE	0.85
	CTGAN	0.84
	Tomek Links	0.84
	가중치 부여	**0.86**
MLP	미적용	0.82
	SMOTE	0.84
	CTGAN	0.83
	Tomek Links	0.84
	가중치 부여	**0.85**

4. 결론

본 논문에서는 COPD 환자 운동 예측 모델의 성능을 향상시키기 위해 대표적인 불균형 데이터 처리 기법인 오버샘플링, 언더샘플링, 가중치 부여 방식을 머신러닝 모델인 XGBoost, 딥러닝 모델인 MLP에 적용하여 성능을 비교하였으며, 가중치 부여 방식이 운동 예측 모델의 성능을 높이는데 가장 효과적인 처리 기법임을 제시하였다.

향후 연구에서는, 운동 예측 성능을 보다 고도화하기 위해 가중치 부여 방식을 적용한 모델의 구조를 개선하는 연구를 수행할 계획이다.

사사문구

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 인공지능융합혁신인재양성사업 연구 결과로 수행되었음(IITP-2023-RS-2023-00256629)

본 연구는 한국연구재단 연구과제로 수행되었습니다. (This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2022R1A2C1010364).)

참고문헌

[1] Kim, Sang Hyuk, et al. "Recent prevalence of and factors associated with chronic obstructive pulmonary disease in a rapidly aging society: Korea National Health and Nutrition Examination Survey 2015 - 2019." Journal of Korean Medical Science 38.14 (2023).

[2] 대한결핵 및 호흡기학회. "COPD 진료지침 2014 개정" (2014): 46-47.

[3] Spruit, Martijn A., et al. "Profiling of patients with COPD for adequate referral to exercise-based care: the Dutch model." Sports Medicine 50 (2020): 1421-1429.

[4] 이태헌, and 이남. "중증 COPD 환자에 대한 포괄적인 운동프로그램의 장기 효과-단일사례연구." 대한심장호흡물리치료학회지 8.2 (2020): 1-9.

[5] Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).

[6] Ishfaq, Haque, Assaf Hoogi, and Daniel Rubin. "TVAE: Triplet-based variational autoencoder using metric learning." arXiv preprint arXiv:1802.04403 (2018).

[7] Goodfellow, Ian, et al. "Generative adversarial nets." Advances in neural information processing systems 27 (2014).

[8] Xu, Lei, et al. "Modeling tabular data using conditional gan." Advances in neural information processing systems 32 (2019).

[9] 질병관리청 국민건강영양조사 원시자료, https://knhanes.kdca.go.kr/knhanes/sub03/sub03_02_05.do

[10] 질병관리청. (2013). 2012 Korea national health and nutrition examination survey results. 서울, 대한민국: 보건복지부. https://knhanes.cdc.go.kr/knhanes/index.do

[11] Chen, Tianqi, and Carlos Guestrin. "Xgboost: A scalable tree boosting system." Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.
[12] Patki, Neha, Roy Wedge, and Kalyan Veeram achaneni. "The synthetic data vault." 2016 IEEE i nternational conference on data science and advan ced analytics (DSAA). IEEE, 2016.