차원 축소 기법을 활용한 의료 데이터 클러스터링 정확도 향상

이연우 [1], 이연지 [2], 이일구 [3]
[1] 성신여자대학교 융합보안공학과 학부생
[2] 성신여자대학교 융합보안공학과 박사과정
[3] 성신여자대학교 융합보안공학과, 미래융합기술공학과 교수

20221124@sungshin.ac.kr, cselab.lyj@gmail.com, iglee@sungshin.ac.kr

Enhancing Clustering Accuracy of Medical Data Utilizing Dimension Reduction Technique

Yeon-Woo Lee[1], Yeon-Ji Lee[1], Il-Gu Lee[2]
[1]Dept. of Convergence Security Engineering, Sungshin Women's University
[2]Dept. of Convergence Security Engineering, Convergence Technology Engineering Sungshin Women's University

요 약

AI 의 분류 성능 저하를 개선하기 위해 under sampling 이나 경계 데이터를 제거하는 등 다양한 연구가 진행되어 왔지만, 여전히 데이터 경계의 복잡한 분포로 인한 근본적인 문제는 해결되지 않고 있다. 이러한 문제를 해결하기 위해 본 연구는 PCA 기반의 차원 축소 방법을 제안한다. 실험 결과에 따르면 제안한 방법은 클러스터 간 분리도 향상과 분류 성능을 원본 데이터를 이용한 종래의 방법 대비 10% 개선할 수 있었다.

1. 서론

AI 의 분류 성능 저하는 클래스 불균형과 클래스 중첩으로 인한 모호한 경계 등 다양한 요인에 의해 발생한다[1,2]. 클래스 불균형 문제를 해결하기 위한 종래 기법으로는 under sampling[3]과 Tomek-link 쌍 제거 기법 [4] 등이 있다. 이들 모두 종래 대비 개선된 분류 정확도를 달성하고 있으나, 의료 데이터 환경에서는 구분이 어려운 경계에 있는 데이터셋이 주요한 데이터이기 때문에 이와 같은 방식은 적합하지 않다. 따라서 본 연구에서는 차원 축소 기법을 통해 클래스 간 거리를 넓히는 방법을 제안한다. 이를 통해 클래스 경계를 명확히 하여 분류 성능 향상을 기대할 수 있다.

2. PCA 기반의 차원 축소 방법

본 연구에서는 AI 학습에 사용되는 고차원의 데이터를 저차원으로 축소하여 분류 성능에 미치는 영향을 분석하였다. 그림 1 은 제안 방법의 동작 흐름을 나타낸 것이다. 데이터셋을 불러온 후, 문자열의 범주형 데이터를 정수형으로 바꾸는 라벨 인코딩을 수행하였으며, 분류 결과값인 타겟 라벨 컬럼 diagnosis 를 분리한다. 이후 모든 데이터를 정규화하여 차원 축소에 적합한 형태로 데이터를 전처리한다. 처리된 데이터는 각각 PCA, UMAP, t-SNE 차원 축소 기법으로 차원 축소를 진행한다. 이후 원본 데이터와 차원 축소 기법을 적용한 각각의 데이터셋으로 학습 및 분류를 수행한다.

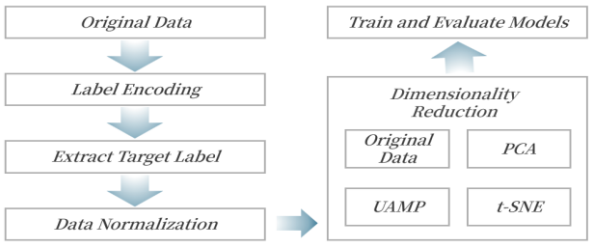

그림 1. 제안 방법의 동작 흐름

3. 성능 평가 및 분석

표 1 은 실험 환경 및 평가 지표를 정리한 것이다. 본 연구에서는 breast cancer 데이터셋[5]을 사용하며, 라벨 인코딩 및 데이터 정규화로 전처리된 데이터는 데이터 손실을 최소화하고 분류 성능을 안정적으로 유지하기 위해 30 차원에서 5 차원으로 축소한 후, 5 차원에서 2 차원으로 점진적인 차원 축소를 진행하였다. 이때 사용한 차원 축소 모델은 PCA, UMAP, t-SNE 세 가지이며, 원본 데이터와 차원 축소를 수행한 각각의 데이터셋에 대해 Random Forest 와 Decision Tree 모델로 학습 및 분류를 수행하였다. 그림 2 는 차원축소 기법과 원본 데이터셋의 라벨을 클러스터링 한 결과이다.

실험 요소	내용
데이터셋	Kaggle breast cancer
차원 축소 방법	PCA, UMAP, t-SNE
분류기	Decision Tree, Random Forest
성능 평가 지표	Accuracy, Confusion Matrix

표 1. 실험 환경 및 평가 지표

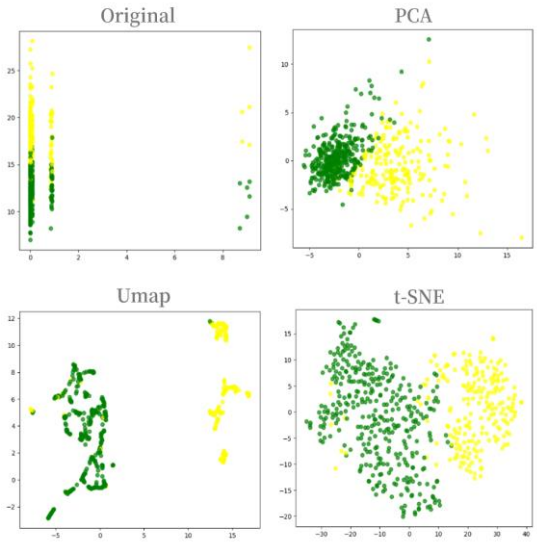

그림 2. 데이터 분포 및 라벨 기반 클러스터링 결과

그림 2 를 통해 차원 축소를 하지 않은 원본 데이터셋은 각 클래스의 데이터가 섞여 있어 명확하게 분리되지 않지만, 차원 축소 방법을 적용한 데이터셋은 명확하게 클러스터링되었으며, 그룹간 거리도 넓어짐을 확인할 수 있다. 그 중에서도 UAMP 기법을 활용한 클러스터가 가장 명확하게 구분되었으며, PCA 와 t-SNE 는 UMAP 보다는 분류 성능이 나빴지만 원본 데이터 보다 분류 성능이 개선되었다.

표 2 는 원본 데이터와 차원 축소를 거친 각각의 데이터셋의 분류 정확도를 비교한 결과이다. 모든 조건에서 차원 축소한 데이터셋의 학습 결과가 더 높게 확인되며, 가장 성능이 좋은 것은 PCA 기법임을 확인하였다.

Dimension Reduction Scheme	Results	
	Random Forest	Decision Tree
Original	0.9211	0.8684
PCA	**0.9825**	**0.9561**
UMAP	0.9386	0.9298
t-SNE	0.9474	0.9298

표 2. 분류 정확도 비교

그림 3 은 원본 데이터와 PCA 차원 축소를 적용했을 때 confusion matrix 를 시각화 및 비교한 것이다. 이 실험 결과에 따르면, PCA 를 적용했을 때 정확도가 향상되었으며, 라벨이 0 일 때의 오분류가 감소했다.

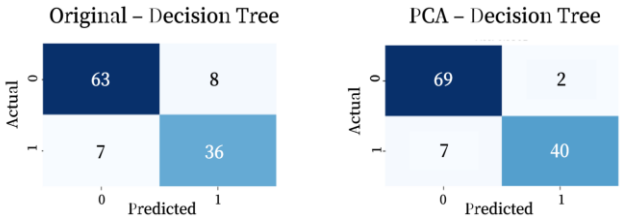

그림 3. confusion matrix 시각화

4. 결론

본 연구에서는 차원 축소 방법이 분류 성능에 미치는 영향을 비교 분석하였다. 실험 결과에 따르면 차원 축소를 적용한 데이터에서 전반적으로 더 높은 분류 정확도를 보였으며, 데이터 분포 및 클러스터링 결과에서도 라벨 기반의 클러스터가 명확히 형성되었다. 이를 통해 차원 축소 방법이 고차원 데이터의 정보 손실을 최소화하면서도 분류 성능을 향상시킬 수 있음을 확인하였다. 향후 연구에서는 클러스터간 거리가 넓어질수록 분류 정확도 또한 비례하여 향상될 수 있도록 개선된 차원 축소 방법을 개발하고 평가할 계획이다.

ACKNOWLEDGMENT

본 논문은 2024 년도 산업통상자원부 및 한국산업기술진흥원의 산업혁신인재성장지원사업 (RS-2024-00415520)과 과학기술정보통신부 및 정보통신기획평가원의 ICT 혁신인재 4.0 사업의 연구결과로 수행되었음 (No. IITP-2022-RS-2022-00156310)

참고문헌

[1] Le Wang, Meng Han, Xiaojuan Li, Ni Zhang, Huanhuan Cheng, Review of Classification Methods on Unbalanced Data Sets, IEEE Access, vol.9, pp.78504-78524, 2021

[2] Seo-Woo Choi, Myeong-Jin Han, Yeon-Ji Lee, Il-Gu Lee, "Classification of Malware Families Using Hybrid Datasets," Journal of the Korea Institute of Information Security & Cryptology (JKIISC), Vol.33, No.6, pp. 1067-1076, Dec. 2023 (KCI)

[3] Bartosz Krawczyk, Colin Bellinger, Roberto Corizzo, Nathalie Japkowicz, Undersampling with Support Vectors for Multi-Class Imbalanced Data Classification, IEEE International Conference on Big Data, 2021, pp. 3043–3052

[4] Debashree Devi, Saroj kr. Biswas, Biswajit Purkayastha, Redundancy-driven modified Tomek-link based undersampling: A solution to class imbalance, Pattern Recognition Letters pp.1-10, 2016

[5] Yasser H, 2021, “Breast Cancer Dataset’, Kaggle, Restrieved from http://www.kagle.com/datasets/yasserh/breast-cancer-dataset