기계학습 프로젝트 제안서

하이퍼 파라미터 튜닝을 통한 유방암 진단 모델 성능 개선

과목명	기계학습
지도교수	윤일동
학과	컴퓨터공학부
학번	202204495
이름	홍창희
제출일	2024.12.10

프로젝트 개요

유방암은 전 세계적으로 중요한 건강 문제 중 하나로, 조기 진단은 환자의 생존율을 크게 향상시킬 수 있습니다. 본 프로젝트는 머신러닝 기법을 활용하여 유방암 진단 모델을 설계하고, 하이퍼파라미터 튜닝을 통해 모델 성능을 최적화하는 것을 목표로 합니다.

이를 위해 UCI Breast Cancer Wisconsin 데이터셋을 활용하여 로지스틱 회귀(Logistic Regression), 랜덤포레스트(Random Forest), 신경망 모델(Neural Network)을 적용하고, 각 기법의 하이퍼 파라미터 튜닝 적용 전후의 성능을 비교 분석합니다.

데이터셋 설명

사용할 데이터는 Kaggle에서 제공하는 Breast Cancer Wisconsin (Diagnostic) Data Set 입니디.

https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data/data

이 데이터는 유방암 진단 문제를 해결하기 위해 설계된 데이터셋으로, 유방암 종양의 세포 핵의 물리적 특성을 분석하여 이를 기반으로 양성과 악성을 분류하는 이진 분류하는데 사용됩니다. 이 데이터는 테이블 형태의 csv파일로 제공됩니다. 각 feature는 세포 핵의 10가지 물리적 특성에 대한 각각의 평균(mean), 표준오차(se), 최대값(worst)으로 구성되어 있어 총 30개의 컬럼을 갖습니다. 레이블은 양성(B)과 음성(M) 2개의 값을 갖습니다.

세포 핵 특성의 목록은 아래와 같습니다.

- radius: 중심에서 둘레의 점까지 거리의 평균
- texture: 회색조 값의 표준 편차.
- perimeter: 둘레
- area: 면적
- smoothness: 반지름 길이의 국소적 변화
- compactness: 둘레^2/ 면적 - 1.0
- concavity: 윤곽에서 오목한 정도
- concave points: 윤곽에서 오목한 점의 개수
- symmetry: 대칭성
- fractal dimension: 프랙탈 차원

프로젝트 목표

1. 다양한 머신러닝 모델 적용
    - 로지스틱 회귀, 랜덤포레스트, 신경망 모델을 적용하여 유방암 진단 문제를 해결한다.
2. 하이퍼파라미터 튜닝
    - 로지스틱 회귀, 랜덤포레스트 모델, 신경망 모델에서 Random Search를 활용하여 각 모델의 성능을 최적화한다.
3. 모델 성능 비교
    - 정확도, 정밀도, 재현율, F1-score 등 다양한 지표로 각 모델의 하이퍼파라미터 튜닝 전후 성능을 비교한다.
4. 최적의 모델 선정
    - 성능과 실행 효율성을 고려하여 유방암 진단에 가장 적합한 모델을 제안한다.

분석 및 예측 과정

1. 데이터 전처리
    - 결측치 처리 – 컬럼의 값 중 결측치가 있다면 해당 컬럼의 중앙값으로 대체
    - 불필요한 컬럼 제거 – 데이터 중 ID 컬럼은 학습에 불필요하므로 제거함
    - 라벨 인코딩 – 이진 분류를 위해 진단 결과(M, B)를 숫자형 라벨(M=1, B=0)로 변환
    - 특성과 라벨 분리 – 데이터셋을 특성(X)와 라벨(y)로 분리
    - 정규화 – 모델 학습 효율성을 높이기 위해 StandardScaler를 사용해 모든 특성을 정규화
    - 훈련/테스트 데이터 분리 – 데이터 중 80%를 학습 데이터로, 20%를 테스트 데이터로 분리
2. 모델 학습 및 튜닝
    - 로지스틱 회귀
        - 규제강도(C)와 최적화 알고리즘(solver)를 Random Search로 최적화
    - 랜덤포레스트
        - 결정 트리의 개수(n_estimators), 최대 깊이(max_depth), 최소 샘플 분할(min_samples_split)을 Random Search로 최적화
    - 신경망 모델
        - Dropout 레이어에 적용되는 비율(rate) 파라미터를 Random Search로 최적화
        - Early Stopping의 patience 파라미터를 Random Search로 최적화
3. 성능 평가
    - 평가 지표 – 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1-score
    - 평가 방법:
        - 테스트 데이터를 활용하여 각 모델의 예측 성능을 평가
        - 하이퍼파라미터 튜닝 전후 성능 비교

활용 기술 및 도구

- 프로그래밍 언어: Python
- 라이브러리:
    - 데이터 처리: pandas, Numpy
    - 모델 학습: scikit-learn, TensorFlow, Keras
    - 시각화: matplotlib, seaborn

기대 효과

1. 머신러닝 기법을 활용하여 유방암 진단 모델의 신뢰도와 정확도를 향상.
2. 각 모델의 성능을 비교하여 데이터 특성에 가장 적합한 알고리즘 제안.
3. 하이퍼파라미터 튜닝의 효과를 분석하여 머신러닝 모델의 일발화 성능을 높임.