

석사학위논문

소프트웨어 결함 예측에서 트리 기반 앙상블 학습의  
하이퍼파라미터 최적화 기법 비교

Comparison of hyperparameter optimization techniques of tree-based ensemble learning in software  
defect prediction

이 유 진

한양대학교 대학원

2025년 2월

석사학위논문

소프트웨어 결함 예측에서 트리 기반 앙상블 학습의  
하이퍼파라미터 최적화 기법 비교

Comparison of hyperparameter optimization techniques of tree-based ensemble learning in software  
defect prediction

지도교수 Scott Uk-Jin Lee

이 논문을 공학 석사학위논문으로 제출합니다.

2025년 2월

한양대학교 대학원


지능정보융합공학과

이 유 진

이 논문을 이유진의 석사학위 논문으로 인준함

2025년 2월

심 사 위 원 장 : 조 성 현

(인) 

심 사 위 원 : 이 석 복

(인) 

심 사 위 원 : Scott Uk-Jin Lee

(인) 

한양대학교 대학원

## 목차

제 1 장 서론 .....	1
1.1 연구 배경 및 목적 .....	1
1.2 연구 질문 및 목표 .....	2
1.3 논문의 구성 .....	3
제 2 장 관련연구 .....	1
2.1 소프트웨어 결함 예측에 사용되는 앙상블 모델 .....	5
2.2 하이퍼파라미터 최적화 기법을 적용한 앙상블 및 단일 모델 .....	6
제 3 장 이론적 배경 .....	8
3.1 트리 기반 앙상블 모델 .....	8
3.1.1 랜덤 포레스트 .....	9
3.1.2 엑스트라 트리 .....	10
3.1.4 XGBoost .....	12
3.1.5 LightGBM .....	13
3.1.6 CatBoost .....	14
3.1.7 AdaBoost .....	15
3.2 하이퍼파라미터 최적화 기법 .....	16
3.2.1 그리드서치 .....	17
3.2.2 랜덤서치 .....	18
3.2.3 베이지안 최적화 .....	18
3.2.4 유전 알고리즘 .....	19
제 4 장 하이퍼파라미터 최적화 기법 비교 실험 .....	21
4.1 데이터셋 설명 .....	21
4.1.1 사용된 나사 데이터셋의 특징 .....	21
4.1.2 데이터 전처리 과정 .....	22
4.2 실험 설계 및 평가 방법 .....	23
4.2.1 실험 설계 방식 .....	23
4.2.3 성능 평가 지표 .....	32

제 5 장 실험 결과 및 분석 .....	35
5.1 모델별 최적화 성능 분석 .....	35
5.2 최적화 알고리즘별 성능 비교 .....	41
5.2.1 최적화 알고리즘에 따른 모델 성능 비교 .....	41
5.2.2 최적화 알고리즘별 모델 간 일관성 평가 .....	43
5.3.1 모델과 최적화 알고리즘 조합이 성능에 미치는 상호작용 효과 분석 .....	46
5.3.2 최적의 모델-알고리즘 조합 도출 .....	47
제 6 장 결론 .....	50
참 고 문 헌 .....	53

#### <표 차례>

표 1 소프트웨어 결함 데이터 셋 설명 .....	22
표 2 트리 기반 앙상블 모델에 대한 기본 하이퍼파라미터 설정 .....	26
표 3 트리 기반 앙상블 모델 최적화를 위한 하이퍼파라미터 범위 .....	30
표 4 트리 기반 앙상블 모델별 최적화 기법 적용에 따른 성능 비교 .....	36
표 5 트리 기반 앙상블 모델에 대한 최적화 기법별 성능 비교 .....	41
표 6 최적의 모델-알고리즘 조합 도출 결과 .....	48

#### <그림 차례>

그림 1 훈련 데이터와 테스트 데이터의 분할의 예시 코드 .....	25
그림 2 랜덤 포레스트 모델의 초기화 예시 코드 .....	28
그림 3 정확도 계산식 .....	32
그림 4 정밀도 계산식 .....	32
그림 5 재현율 계산식 .....	33
그림 6 F1 스코어 계산식 .....	33
그림 7 최적화 방법별 모델 성능(F1 Score) .....	38
그림 8 최적화 방법별 모델 성능(Accuracy) .....	39
그림 9 최적화 방법에 의한 모델 간 평균 정확도 .....	43
그림 10 모델-최적화 알고리즘 상호 작용 그래프 .....	46

## 요약

소프트웨어 결함 예측은 소프트웨어 개발 과정에서 잠재적인 오류를 조기에 탐지하고 품질을 개선하며 유지보수 비용을 절감하는 데 중요한 역할을 한다. 특히, 정확하고 효율적인 결함 예측은 개발 리소스를 최적화하고 프로젝트의 성공 가능성을 높이는 데 기여한다. 본 연구는 소프트웨어 결함 예측 문제를 해결하기 위해 트리 기반 앙상블 모델(Random Forest, ExtraTrees, Gradient Boosting, XGBoost, LightGBM, CatBoost, AdaBoost)을 대상으로 네 가지 하이퍼파라미터 최적화 기법(그리드 서치, 랜덤 서치, 베이지안 최적화, 유전 알고리즘)이 모델 성능에 미치는 영향을 체계적으로 비교하였다. 각 최적화 기법은 다양한 소프트웨어 결함 데이터셋을 활용하여 평가되었으며, 성능 비교를 위해 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 스코어(F1 Score), AUC(Area Under the Curve)와 같은 다섯 가지 지표를 사용하였다.

연구 결과, 최적화 기법은 모델 성능 향상에 중요한 영향을 미치며, 특히 Genetic Algorithm은 Gradient Boosting, CatBoost, XGBoost와 같은 복잡한 구조의 모델에서 가장 우수한 성능을 나타냈다. 반면, Grid Search와 Random Search는 ExtraTrees, LightGBM과 같은 상대적으로 단순한 모델에서 안정적이고 일관된 성능을 보였다. 이러한 결과는 모델의 구조와 하이퍼파라미터 탐색 기법 간의 상호작용 효과가 존재함을 시사한다.

또한, 모델과 최적화 기법 간 상호작용을 분석한 결과, 특정 모델에서 특정 최적화 기법이 유의미한 성능 차이를 나타냄을 확인하였다. 예를 들어, Bayesian Optimization은 CatBoost와 Gradient Boosting에서 일관된 성능을 보였지만, RandomForest나 Adaboost에서는 상대적으로 낮은 성능을 기록하였다. 이러한 분석은 각 모델의 특성에 따라 최적화 기법을 선택하는 것이 성능 극대화에 필수적임을 보여준다.

본 연구는 소프트웨어 결함 예측에서 하이퍼파라미터 최적화 기법의 중요성을 실증적으로 평가하며, 모델 특성과 최적화 기법의 조합이 성능에 미치는 영향을 심층적으로 분석하였다. 이를 통해 소프트웨어 품질 보증 시스템에서 최적의 예측 성능을 달성하기 위한 구체적인 방향성을 제시하였다.