Fault Localization 모델 학습을 위한 체계적인 데이터셋 생성 기술

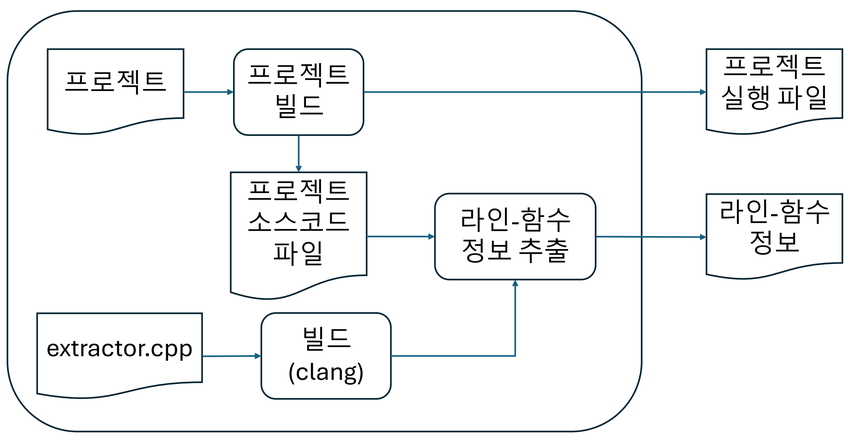
양희찬, 김문주 (한국과학기술원)

동기

- Fault Localization (FL) 기법들은 많이 존재한다 (SBFL, MBFL, 등). 하지만 소개 된 기법들은 특정 유형의 프로젝트에 좋은 효과를 보여주고 있으나, 모든 유형의 프로젝트에 좋은 성능을 보여주지 못하고 있다.
- 특정 유형에 국한 되는 문제를 해결하기 위해 **동적과 정적** 특징 정보를 가지고 **머신러닝 모델**을 학습 시켜 다양한 프로젝트에 적용할 수 있는 결함 위치 탐지 기법을 만드는 것을 목표한다.
- 따라서 체계적인 데이터셋 생성 기술은 FL 모델 학습과 이후 실험에 필요한 기술이 된다.

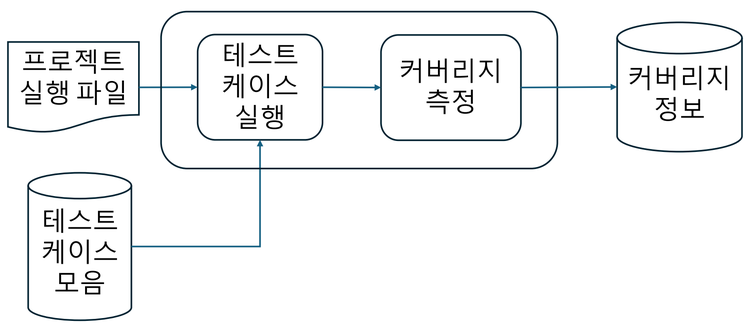
데이터셋 생성 과정 (SBFL 기준: 5개 단계)

1) 프로젝트 빌드 단계



- 빌드를 통해 프로젝트의 **실행 파일**을 생성한다
- 프로젝트의 소스 코드로부터 **라인-함수** 매핑 정보 추출한다
  - Clang Frontend API 활용

2) 테스트 케이스 실행 및 커버리지 정보 추출 단계

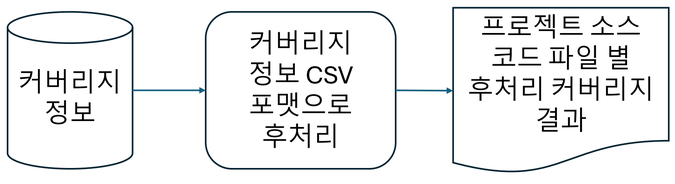


- 프로젝트의 실행 파일을 통해 각 테스트 케이스를 실행한다.
- Gcovr 활용 각 테스트 케이스의 **커버리지 정보**를 기록한다.
  - 라인, 함수, 파일 커버리지 결과
- 모든 테스트 케이스 실행 결과 **특징 (criteria) 정보** 계산한다.

프로젝트	기준	Buggy 항목 실행 후 pass 하는 TC 개수	Buggy 항목 실행 후 fail 하는 TC 개수	Buggy 항목 실행 하지 않는 TC 개수
Jsoncpp	파일	1,153	3	0
	함수	243	3	910
	라인	27	3	1,126

3) 커버리지 결과 CSV 포맷으로 후처리 단계

- Gcovr로 추출 한 테스트 케이스 별 **커버리지 결과**를 프로젝트 소스 코드 별로 CSV 포맷으로 **후처리**한다.
  - 각 라인 별로 테스트 케이스의 실행 여부를 기록한다

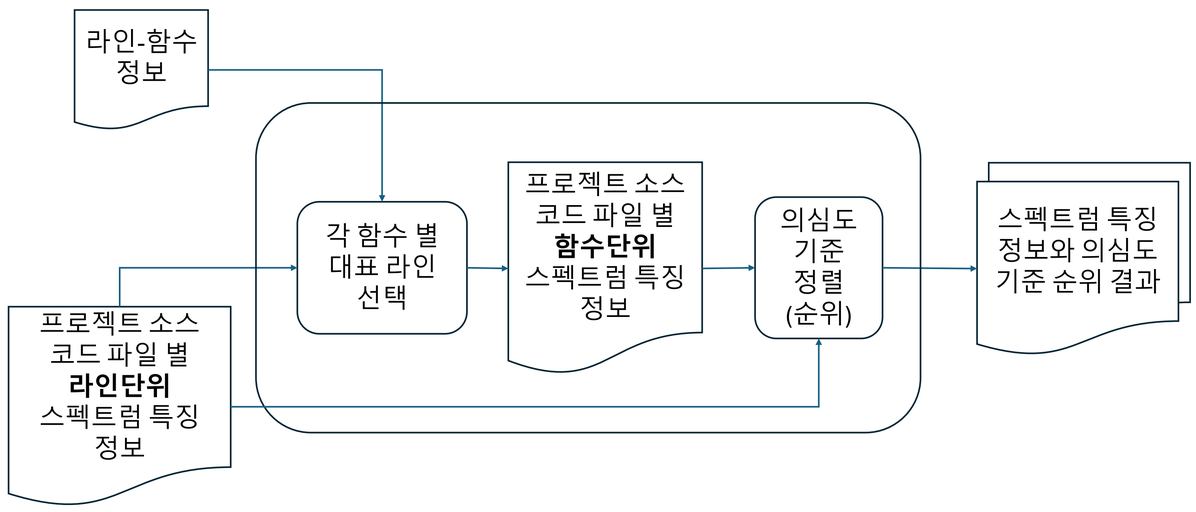


4) 스펙트럼 기반 특징 추출 단계

- 후처리 된 커버리지 정보로부터 **라인단위 스펙트럼 특징**을 계산하여 추출
  - 라인별 e_p, e_f, n_p, n_f 정보와 **의심도 점수** (SBFL Formula)

5) 의심도 순위 정렬 단계

- 라인단위 스펙트럼 특징 결과로부터 **라인단위**와 **함수단위**로 의심도 기준 정렬해서 순위를 매긴다.



OSS 프로젝트인 JsonCpp 적용 결과

- JsonCpp에 실제 bug 총 4개에 대해 적용
- 총 **1,156 테스트 케이스**에서 각각 버그 버전의 coincidentally passing TC 제외해서 데이터셋 생성
- 최종 산출물에 담긴 특징
  1. 버그 프로젝트, 파일명, 함수명, 라인번호
  2. SBFL의 e_p, e_f, n_p, n_f 정보
  3. SBFL Formula의 **의심도 점수**
  4. 의심도 기준 실제 순위 (rank)
  5. Bug 라인 여부
- 아래 데이터셋을 결함 위치 탐지 모델 학습 데이터로 사용

버그 프로젝트	파일명	함수명	라인 번호	e_p	e_f	n_p	n_f	Naish2 Susp.	Ochiai Susp.	Rank	Bug
Jsoncpp1	Json_reader.cpp	OurReader::decodeNumer()	1628	0	3	1,126	0	3.0	1.0	1	1
Jsoncpp1	Json_reader.cpp	OurReader::decodeNumer()	1618	6	3	1,120	0	2.99	0.57	2	0
…	…	…	…	…	…	…	…	…	…	…	…
Jsoncpp1	Json_value.cpp	Value::Value()	439	112	3	1,004	0	2.89	0.16	319	0

~

향후 연구 계획

- 여러 Real-World C/C++ Project에 해당 기술을 적용해서 생성된 **동적/정적 특징** 데이터셋으로 결함 위치 탐지 모델 학습
  - 동적: **SBFL**와 **MBFL**
  - 정적: **CCCC** 정적 분석 활용
- 학습 데이터에 따르는 **FL 실험** 결과 분석
  - 예) 동적 특징: 특정 테스트 케이스를 제외하여 생성한 데이터셋의 효과.
- FL 모델 추후 결정