Re-evaluationg the impact of data cleaning on ML classification

(부제) Anaysis of Data cleaning effect based on Noise Level in Datasets

분산 클라우드 컴퓨팅 – 팀 프로젝트 1차발표

데이터사이언스학과 황선진
자동차공학과 김진우
데이터사이언스학과 야오와말

목차

1 프로젝트 배경

- 프로젝트 관련 논문
  - CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks (2021, ICDE)
    - (제안) ML 분류 모델 성능에 대한 데이터 정제 영향을 체계적·실험적 분석
    - (의의) 데이터 정제가 ML 분류 모델 성능 향상에 도움 : 향후 데이터 정제 연구의 기반 마련

✓ 오류 유형(5가지)

- Missing Values(결측치)
- Outliers (이상치)
- Mislabels(라벨링 오류)
- Inconsistencie(일관성 오류)
- Duplicates(중복값)

데이터 정제 →

✓ 분류 모델(7가지)

- Linear Regression
- Discision Tree
- Random Forest
- Adaboost
- XGBoost
- KNN (k- Nearest Neighbors)
- Naive Bayes

1 프로젝트 배경

- 프로젝트 관련 논문 : 한계점
  - 데이터 정제 효과가 데이터셋에 의존(depend on Dataset) 함을 파악하였으나
    데이터셋의 오류 유형만 언급했을 뿐, 구체적인 오류 데이터 비율·관련 추가 실험이 존재하지 않음

✓ 논문 발췌 : Overall Observations summary

TABLE 16. Summary of Empirical Findings for Single Error Types

Error Type	Impact on ML	Does the impact depend on			
		Datasets	Scenarios	Cleaning Algos	ML Algorithms
Duplicates	Varying (Mostly S & N)		No	Yes	No
Inconsistencies	Varying (Mostly S)		No	N.A.	No
Missing Values	Varying (Mostly P & S)	Yes	No	Yes	No
Mislabels	Varying (Mostly P & S)		Yes	N.A.	No (except Boosting)
Outliers	Varying (Mostly S)		No	Yes	No (except KNN)

Strong Dependency on Dataset.
the cleaning impact depends on datasets — while two datasets may contain errors of the same type, the distributions of those errors can be vastly different. Therefore, practitioners should never make arbitrary cleaning decisions dealing with dirty data in ML classification tasks.

✓ 논문 발췌 : Datasets

We collected 14 real-world datasets with varying error types and error rates

TABLE 3. Dataset and Error Types

Datasets	Error Types				
	Inconsistencies	Duplicates	Missing Values	Outliers	Mislabels
Citation		x			
EEG				x	x
Marketing			x		x
Movie	x	x			
Company	x				
Restaurant	x	x			
Sensor				x	
Titanic			x		x
Credit			x	x	
University	x				
USCensus			x		x
Airbnb		x	x	x	
BabyProduct			x		
Clothing					x

※ 단, 오류 유형이 아닌 오류 비율에 대한 구체적인 언급은 없음

1 프로젝트 배경

- 문제의식 : 아이디어 제안
  - (문제의식) 해당 논문에서 Error rates(오류 데이터 비율)과 관련된 실험이 존재하지 않음
  - (아이디어 제안) Error rates(오류비율) 수준에 따른 데이터 정제효과 분석

아이디어 예시 : 결측값이 데이터의 10%일때/30%일때 ML모델 성능 관련 데이터 정제 효과 비교·분석

(결측값 10% : 데이터 정제후 ML모델 성능)

	A	B	C	D	E
1	ID	Age	Experience	Income	ZIP Code
2	1	25	1	49	91107
3	2	45	19	34	90089
4	3	39	15	11	94720
5	4	35	9	100	94112
6	5	35	8		91330
7	6	37	13		92121
8	7	53	27	72	91711
9	8	50	24	22	93943
10	9	35	10	81	90089
11	10	34	9	180	93023
12	11	65	39	105	94710
13	12	29	5	45	90277
14	13	48	23	114	93106
15	14	59	32	40	94920

⟷

(결측값 30% : 데이터 정제후 ML모델 성능)

	A	B	C	D	E
1	ID	Age	Experience	Income	ZIP Code
2	1	25	1	49	91107
3	2	45	19	34	90089
4	3	39	15	11	94720
5	4	35	9	100	94112
6	5	35	8		91330
7	6	37	13		92121
8	7	53	27	72	91711
9	8	50	24	22	93943
10	9	35	10		90089
11	10	34	9		93023
12	11	65	39		94710
13	12	29	5	45	90277
14	13	48	23		93106
15	14	59	32	40	94920

2 프로젝트 방법론

- 데이터 정제 – ML 모델성능 평가 프레임워크 구성 및 실험
  - (참고) CleanML 논문의 공개 코드를 활용 : https://github.com/chu-data-lab/CleanML
  - 오류유형 중 Missing Values(결측치), Outliers(이상치), Duplicates(중복값) 선택하여 실험
    - (선택 이유) 데이터 분석 시 자주 접하는 오류유형이며, 다양한 오류 탐지 및 정제 방법 적용 가능
      * 다른 오류 유형(Mislabels(라벨링 오류), Inconsistencie(일관성 오류))의 경우 탐지 및 정제 방법이 제한적
  - 논문의 실험 데이터셋 중 Airbnb 데이터셋 선택 : 선택한 3가지 오류유형 모두 포함

✓ 논문 발췌 : 오류유형에 따른 탐지 및 정제 방법

TABLE 2. Automatic Cleaning Methods

Error Type	Detection Method	Repair Method
Missing Values	Empty Entries	Deletion
		Mean_Mode, Mean_Dummy Median_Mode, Median_Dummy Mode_Mode, Mode_Dummy
		HoloClean
Outliers	SD	Mean, Median, Mode
	IQR	
	IF	HoloClean
Duplicates	Key Collision	Deletion
	ZeroER	
Inconsistencies	OpenRefine	Merge
Mislabels	cleanlab	cleanlab

✓ 논문 발췌 : Datasets and error types

TABLE 3. Dataset and Error Types

Datasets	Error Types				
	Inconsistencies	Duplicates	Missing Values	Outliers	Mislabels
Citation		x			
EEG				x	x
Marketing			x		x
Movie	x	x			
Company	x				
Restaurant	x	x			
Sensor				x	
Titanic			x		x
Credit			x	x	
University	x				
USCensus			x		x
Airbnb		x	x	x	
BabyProduct			x		
Clothing					x

2 프로젝트 방법론

- 데이터 정제 – ML 모델성능 평가 프레임워크 구성 및 실험
  - 프레임워크 프로세스 : ❶ - ❸ 반복 수행
  - (결과 분석) 데이터 정제 전/후 성능 지표 값 비교·분석

❶ 데이터셋 오류비율 조정 →

- Missing Values(결측치)
- Outliers (이상치)
- Duplicates(중복값)

❷ 데이터 정제 수행 →

- Mean
- Median
- Mode
- Deletion
- HoloClean
- Cleanlab

❸ ML 모델 성능 평가 및 평가지표 값 저장

- Linear Regression
- Discision Tree
- Random Forest
- Adaboost
- XGBoost
- KNN (k- Nearest Neighbors)
- Naive Bayes

3 프로젝트 기대효과

- 데이터 오류 비율에 따른 데이터정제 방법 선택에 인사이트 제공
  - CleanML 논문의 한계점인 오류 비율 관련 연구 부족 : 관련 실험 프로젝트 수행
  - 오류 비율 관련 실험 프로젝트를 통해 CleanML 논문 실험결과·시사점 재확인
  - 데이터 오류 비율은 쉽게 파악 가능하기 때문에 실제 데이터 분석 및 ML 모델 구현에 도움

- 오류비율 외에도 CleanML 논문에서 다루지 못한 회귀모델, 딥러닝 모델, 데이터 증강 활용 등 추가 연구 가능

  * CleanML 논문은 분류모델 및 머신러닝 모델에 초점을 맞추고 연구를 수행

감사합니다