LLM 모델을 활용한 화합물 특성 예측에서 지속적인 데이터 불균형 문제 해결

Do Yeon Kim.[1], Yong Oh Lee, Ph.D.[2,3]
[1]*Dept. of Computer Engineering, Hongik University, Seoul*
[2]*Dept. of Industrial and Data Engineering, Hongik University, Seoul, yongoh.lee@hongik.ac.kr*
[3]*Hongik University Bio-Health Convergence Research Center*

Ⅰ. Background

- 화학 화합물의 특성 예측 및 생성에서 **SMILES**는 분자 구조를 효율적으로 인코딩하는 방식으로, 최근 **LLM(Large Language Model)**을 활용하는 방식으로 발전하고 있음.
- 기존의 descriptor 기반 머신러닝 모델이나 descriptor-free Transformer 모델 모두 데이터셋 자체의 **데이터 불균형 문제로 인해 소수 클래스의 예측 성능이 낮아지는 경향**이 존재함.
- Llama 기반의 생성 모델인 **LlaMol**을 활용하여 데이터 불균형 문제 해결 방향을 제시하고자 함.

Ⅱ. Method

- **Benchmark Dataset 및 불균형도**

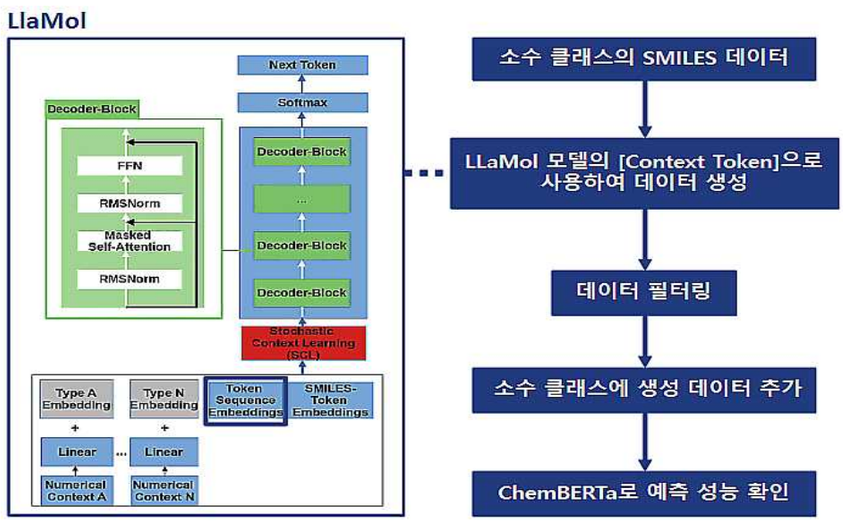
  Tox21(5343 SMILES, 불균형도 11:1), **HIV**(41127, 28:1), **Clintox**(1480, 15:1), **BBBP**(2039, 3:1)

Nc1ccc(C(=O)N2CCC(=O)c3ccc(Cl)cc3-2)cc1

BrC1=CC=CC=C1C(=O)C2=CC=CC3=C2NC(=O)C3

[그림 1] SMILES 데이터 예시

- **Oversampling을 위한 생성 모델: LlaMol**
  - **Llama**기반의 SMILES 생성 모델
  - **[Context Token]**으로 특정 SMILES 사용 가능



[그림 2] LlaMol 모델 구조 및 실험 진행 순서

- **Oversampling 데이터 필터링**

  1. **Validity**(화학적으로 유효한 분자의 비율) = 1
  2. **Novelty**(학습 데이터에 포함되지 않은 분자의 비율) = 1
  3. 생성된 데이터 간의 **중복 제거**

- 사용한 **분류 모델**: **ChemBERTa**

  - SMILES 형식으로 인코딩된 화학 화합물 구조를 학습하여 특성을 예측하는 **RoBERTa** 기반 모델

Ⅲ. Results

[표 1] Tox21 데이터셋에서 구간에 따른 성능 비교

(OS: OverSampling)

구분	Base	OS 1	OS 2	OS 3	OS 4	OS 5
불균형도	11:1	9:1	6:1	5:1	4:1	3:1
정확도 (or ROC)	0.729	0.935	0.942	0.943	0.954	0.947
민감도☆	0.516	0.706	0.639	0.718	0.797	0.790
특이도	0.980	0.707	0.980	0.965	0.967	0.959

[표 2] Clintox 데이터셋에서 구간에 따른 성능 비교

(OS: OverSampling)

구분	Base	OS 1	OS 2	OS 3	OS 4	OS 5
불균형도	15:1	10:1	7:1	6:1	5:1	4:1
정확도 (or ROC)	0.551	0.673	0.731	0.722	0.809	0.788
민감도	1.000	0.125	0.524	0.462	0.548	0.556
특이도☆	0.000	0.467	0.934	0.978	0.912	0.883

[표 3] HIV 데이터셋에서 구간에 따른 성능 비교

(OS: OverSampling)

구분	Base	OS 1	OS 2	OS 3	OS 4	OS 5
불균형도	28:1	23:1	20:1	18:1	16:1	15:1
정확도 (or ROC)	0.795	0.960	0.977	0.969	0.951	0.934
민감도☆	0.358	0.373	0.442	0.662	0.713	0.729
특이도	0.993	0.985	0.991	0.986	0.988	0.985

[표 4] BBBP 데이터셋에서 구간에 따른 성능 비교

(OS: OverSampling)

구분	Base	OS 1	OS 2	OS 3	OS 4	OS 5
불균형도	3:1	2.9:1	2.7:1	2.5:1	2.3:1	2.1:1
정확도 (or ROC)	0.922	0.941	0.936	0.945	0.947	0.949
민감도	1.000	0.315	0.612	0.685	0.704	0.712
특이도☆	0.000	0.524	0.894	0.881	0.780	0.747

- 모든 데이터셋에 대해 생성 데이터를 추가하였을 때 **정확도**가 **향상**
- **민감도**가 **대체로 증가**→ 생성 데이터를 추가하여 소수 클래스의 데이터의 예측 성능 향상
- **특이도**의 경우, 모든 데이터셋에서 공통적으로 증가하다 감소하는 지점 존재(구간 2 또는 3)→ 추가하는 생성 데이터 개수의 조정 필요

Ⅳ. Conclusion

- **ChemBERTa**를 사용한 화합물 특성 예측에서 **데이터 불균형** 문제가 여전히 존재함을 확인하였으며, 이러한 문제를 **LlaMol을 활용한 SMILES 합성으로 데이터 증강**을 통해 해결 가능함을 보임.
- 향후 데이터 불균형을 해결하기 위해 생성 모델의 효율성을 극대화하여 데이터를 증강하는 방식을 연구할 예정임.

※ 본 연구는 '바이오헬스 혁신융합대학 정책연구'와 '홍익대학교 학술진흥연구비'의 지원을 받아서 수행하였습니다.

대한의료정보학회 The Korean Society of Medical Informatics
홍익대학교 HONGIK UNIVERSITY
Artificial Intelligence Applications Lab