약물 특성 임베딩에 따른 약물 부작용 유사 레이블 분류

함유진, 황재성, 김문규, 박동현*
세종대학교
yjham@sju.ac.kr, jshwang@sju.ac.kr, mgkim@sju.ac.kr, parkdh@sejong.ac.kr

Classification of drug side effects pseudo-label according to drug property embedding

Yujin Ham, Jaesung Hwang, Mungyu Kim, Donghyeon Park
Sejong University

요 약

인공지능을 활용해서 약물의 성분과 효능만으로 해당 약물의 이상 반응을 확인할 수 있다면 사람들이 약을 복용하는 데 있어서 이상 반응에 대한 장벽을 해결하는 데 도움을 줄 수 있을 것이다. 이에 따라 본 논문에서는 의약품의 성분과 효능의 정보를 담은 임베딩을 입력 값으로 이상 반응 수도 레이블을 분류해서 같은 이상 반응 수도 레이블을 갖는 기존 의약품의 이상 반응을 참고하여 이상 반응을 확인할 것을 제안한다. 그 결과 의약품의 성분과 효능에 따라 의약품 부작용 수도 레이블을 일정 수준 분류할 수 있음을 확인했다.

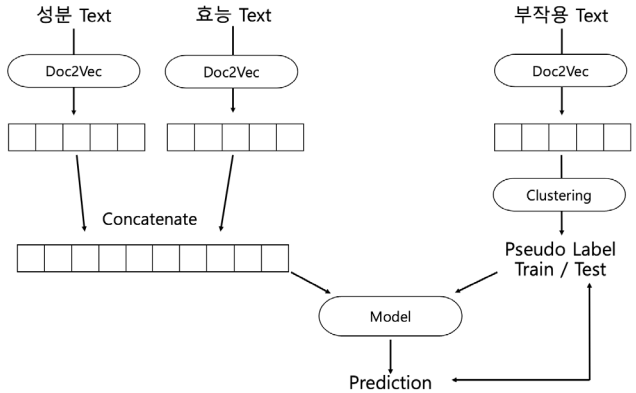
I 서론

보건의료빅데이터개방시스템[1]에 따르면 매달 약 2천만 명의 사람들이 39억 개의 약을 먹는다. 많은 사람이 많은 약을 먹는 만큼 약의 이상 반응도 적지 않게 발생할 것이다. 하지만, 이상 반응이 발생했을 때, 복용한 많은 약 중 어느 약에 의한 결과인지 고민에 빠지고 전문가의 도움 없이 개인의 힘으로 알아내기는 어렵다. 인공지능을 활용해 약물 이상 반응을 예측할 수 있다면 사람들의 약물 이상 반응 고민을 해결하는 데 도움을 줄 수 있을 것이다. 인공지능을 활용한 의약품 이상 반응 확인의 한 방법으로 본 논문에서는 비슷한 특성을 지니는 의약품은 비슷한 부작용을 나타낼 것이란 가정하에 특성이 비슷한 기존 약품의 부작용을 참고해 새로운 약품의 부작용을 확인할 것을 제안한다. 따라서 의약품의 성분과 효능 정보를 담은 임베딩 입력 값을 넣고 이상 반응 수도 레이블을 분류해서 의약품의 특성에 따라 비슷한 부작용을 갖는지 확인한다. 실험을 위해 공공데이터포털[2][3]의 데이터세트를 이용했다. 여러 기계 학습 모델 중 성능이 좋았던 모델 3개와 2개의 딥러닝 모델을 채택하여 실험하였다. 그 결과 이상 반응 수도 레이블 분류에서 테스트 정확도 0.852(Class2), 0.720(Class3), 0.670(Class5), 0.615(Class10)의 성능을 각각 보였다. 이는 약물 성분과 효능 정보만 단편적으로 주어졌을 때, 개인이 스스로 어느 약이 어떤 이상 반응을 일으켰는지 확인하고 상황을 해결하는 데 도움을 줄 수 있을 것이다.

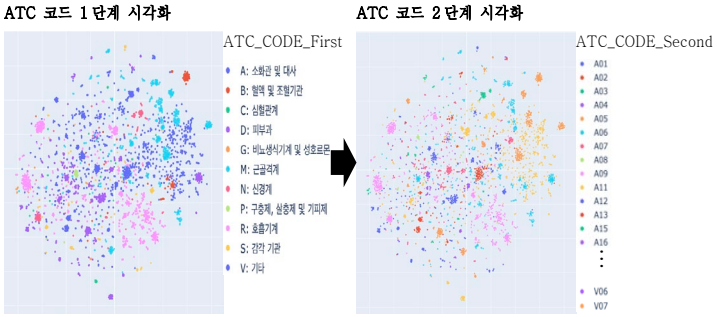
II 본론

II.1 연구방법 및 정의

Doc2Vec[4]을 사용해 성분·효능 임베딩을 만들어 입력값으로 사용한다. 또한, Doc2Vec[4]을 사용해 만든 부작용 임베딩에 대해 K-Means[5] 방법을 이용해 각 의약품의 이상 반응 수도 레이블을 만들어 출력 값으로 사용한다. 마지막으로, 약품의 성분·효능 임베딩을 입력 값으로 넣어 의약품의 이상 반응 수도 레이블을 출력 값으로 분류하도록 여러 기계학습 모델과 딥러닝 모델을 훈련한다. 전반적인 연구 방법은 [그림 1]과 같다.



[그림 1] 연구 방법 도식도



[그림 2] 의약품 성분·효능 임베딩 시각화 결과, ATC 코드 2단계는 1단계를 다시 주요 치료적 그룹으로 눈 결과이며 본 논문에서는 55개가 존재한다.

II.2 데이터셋

공공데이터포털[2][3]에서 "식품의약품안전처_의약품

*박동현: Corresponding Author (parkdh@sejong.ac.kr)

		Class2		Class3		Class5		Class10	
		Train	Test	Train	Test	Train	Test	Train	Test
ML	Random Forest	0.755	0.697	0.710	0.655	0.662	0.567	0.555	0.466
	CatBoost	0 .792	0 .738	0.715	0.670	0.662	0.595	0.564	0.464
	KNeighbors	**0.883**	**0.852**	**0.743**	**0.720**	**0.707**	**0.670**	**0.667**	**0.615**
DL	Convolutional Neural Network	0.818	0.748	0.737	0.668	0.544	0.487	0.470	0.411
	Multi-layer Perceptron	0.810	0.771	0.732	0.687	0.625	0.572	0.535	0.477

[표 1] 성분·효능 임베딩에 따른 이상반응 수도 레이블 분류 성능 평가

제품 허가정보"[2], "식품의약품안전처_의약품개요정보(e약은요)"[3] 데이터세트를 활용신청 후 승인받아서 사용했다. 전처리 완료 후 최종적으로 사용한 데이터는 3,994 개의 행과 11 개의 열로 이루어져 있다.

II.3 Doc2Vec을 이용한 의약품 성분·효능 임베딩

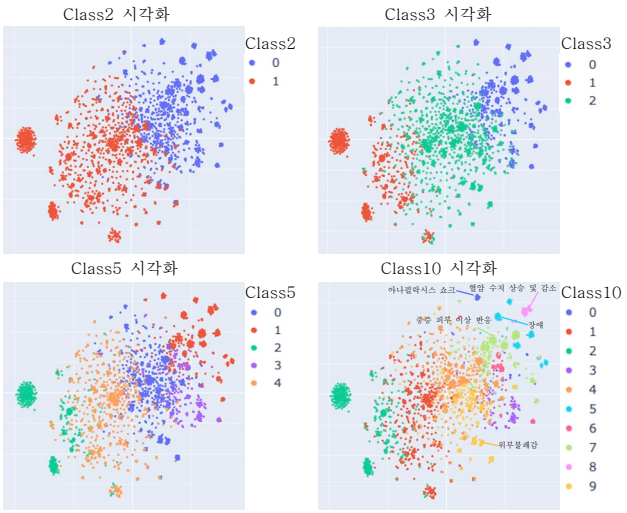
의약품의 성분과 효능을 각각 임베딩 하고, 성분 임베딩과 효능 임베딩을 이어 붙여 600차원의 성분·효능 임베딩을 만들어 분류 모델의 입력 값으로 사용했다. 성분·효능 임베딩이 의약품의 특성을 잘 대표하는지 확인하기 위해 t-SNE[6] 방법을 이용해 성분·효능 임베딩을 2차원으로 차원을 축소해 시각화 했다. 의약품마다 ATC Code[7](1단계, 2단계)에 따라 색을 다르게 표현하였다. [그림 2]를 보면 같은 색상의 객체끼리 잘 군집화 된 것을 확인할 수 있다. 따라서 성분·효능 임베딩이 의약품의 특성을 잘 반영했다고 할 수 있다. 다만 일부 색상의 경우 군집화 되면서 펴져 있는 것을 확인할 수 있다. 이는 하나의 의약품이 여러 개의 ATC Code[7]를 가질 수 있기 때문이라고 판단된다.

II.4 Doc2Vec을 이용한 이상반응 Pseudo Label 생성

비슷한 부작용을 가진 의약품끼리 클래스를 만들어서 약품 부작용 수도 레이블을 생성해 출력 값으로 사용하려 한다. [그림 3]을 통해서 이상 반응에 따라 수도 레이블이 잘 생성되었음을 확인했다. 자연어로 서술된 약품의 부작용을 Doc2Vec[4] 방법을 이용해 300차원의 부작용 임베딩을 만들고 K-Means[5] 방법을 이용해서 비슷한 부작용 임베딩끼리 클래스를 만들었다. 이때 클래스의 개수를 2, 3, 5, 10으로 선택하여 'Class2', 'Class3', 'Class5', 'Class10' 총 4개의 이상 반응 수도 레이블을 생성하였다. 클래스가 잘 생성되었는지 확인하기 위해 정성평가를 진행하였다. 클래스의 개수가 2, 3, 5일 때는 부작용 문장의 길이(즉, 부작용의 개수)가 클래스별로 확연하게 차이 났다. 클래스의 개수가 10개인 경우 0, 5, 7, 8, 9 클래스는 순서대로 아나필락시스 쇼크, 장애, 중증 피부 이상 반응, 혈압 수치 상승 및 감소, 위부 불쾌감 관련 증상으로 클래스에 따른 비교적 뚜렷한 이상 반응의 종류를 알 수 있었다. 나머지 클래스 중 일부 클래스는 다른 클래스와 부작용이 비슷했고, 일부 클래스는 군집화 된 데이터의 양이 많았다. 이 점을 보완하고 나머지 클래스의 이상 반응 종류도 뚜렷이 알기 위해서는 더 세분화된 클래스로 나누는 등의 시도를 해야 할 것으로 보인다.

II.5 모델 및 실험 결과

분류에 자주 사용되는 기계학습 모델들을 사용해보았다. 여러 모델 중에서 KNeighbors[5]와 트리 기반의 모델의 성능이 가장 우수했다. 따라서 높은 성능을 보인 3개의 기계 학습 분류 모델 KNeighbors[5], CatBoost[8], Random Forest[9]와 간단한 딥러닝 모델 Multilayer Perceptron[10], Convolutional Neural Network[11]를 채택하여 사용하였다. 훈련 정확도를 손해 보더라도



[그림 3] 의약품 부작용 임베딩 클래스 시각화 결과

과적합 되지 않도록 모델을 훈련시켰다. 클래스를 더 세밀하게 나눌수록 정확도가 떨어지는 특징이 있었다. 각 모델의 성능은 [표 1]과 같다. 테스트 정확도 0.852(Class2), 0.720(Class3), 0.670(Class5), 0.615(Class10)의 성능을 각각 보였다. 이는 약물의 성분과 효능 정보만으로 어느 정도의 약물 이상 반응이 있을지 미리 판단하는 데 도움이 될 것이라 기대한다. 또한, 딥러닝 모델의 성능이 좋을 것이라는 예상과 달리 KNeighbors[5] 모델의 성능이 가장 좋았다. 이는 수도 레이블을 만들 때 사용한 K-Means[5] 방법과 KNeighbors[5] 방법이 이웃 객체를 통해서 분석하는 방법이라는 공통점을 가지고 있기 때문일 것이다.

III 결론

본 논문에서는 의약품 성분·효능 임베딩에 따라 이상 반응 수도 레이블을 분류하는 실험을 진행했다. KNeighbors[5]를 사용하여 테스트 정확도 0.852(Class2), 0.720(Class3), 0.670(Class5), 0.615(Class10)의 성능을 보였다. 향후 연구에서는 의약품의 비슷한 부작용을 지닌 클래스 레이블을 분류하는 것이 아닌 의약품의 특성에 따라 구체적으로 어떤 이상 반응이 발생할 수 있는지 자연어를 생성할 수 있다면 의약품 이상 반응 확인에 더 도움이 될 것으로 예상한다.

IV Acknowledgement

이 논문은 2022년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(NRF-2022R1F1A1069639)

참 고 문 헌

[1] "의약품 사용 통계." 보건의료빅데이터개방시스템, n.d., opendata.hira.or.kr/op/opc/olapMsupInfo.do. 2022년 08월 24일 접속.

[2] "식품의약품안전처_의약품 제품 허가정보." 공공데이터포털, n.d., www.data.go.kr/tcs/dss/selectApiDataDetailView.do?publicDataPk=15095677. 2022년 08월 24일 접속.

[3] "식품의약품안전처_의약품개요정보(e약은요)." 공공데이터포털, n.d., www.data.go.kr/tcs/dss/selectApiDataDetailView.do?publicDataPk=15075057. 2022년 08월 24일 접속.

[4] Le, Quoc, and Tomas Mikolov. "Distributed representations of sentences and documents." International conference on machine learning. PMLR, 2014.

[5] Pedregosa, Fabian, et al. "Scikit-learn: Machine learning in Python." the Journal of machine Learning research 12 (2011): 2825-2830.

[6] Van der Maaten, Laurens, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.11 (2008).

[7] "Anatomical Therapeutic Chemical (ATC) Classification." World Health Organization, n.d., www.who.int/tools/atc-ddd-toolkit/atc-classification. 2022년 08월 24일 접속.

[8] Prokhorenkova, Liudmila, et al. "CatBoost: unbiased boosting with categorical features." Advances in neural information processing systems 31 (2018).

[9] Breiman, Leo. "Random forests." Machine learning 45.1 (2001): 5-32.

[10] Noriega, Leonardo. "Multilayer perceptron tutorial." School of Computing. Staffordshire University (2005).

[11] Simard, Patrice Y., David Steinkraus, and John C. Platt. "Best practices for convolutional neural networks applied to visual document analysis." Icdar. Vol. 3. No. 2003. 2003.