Data Mining Project Report

**의료약물 정보 간의 연관성 분석**

생명공학과 202002134 오재은

바이오메디컬공학과 202104415 윤수진

**1. Motivation and contributions**

의약품 사용은 우리의 일상에서 불가피한 존재이지만, 부작용 또한 불가피하다. 특히 전세계적인 전염병인 COVID-19을 통해 많은 사람들이 부작용에 대해 많은 관심을 가지게 되는 계기가 되었다. 이러한 배경을 통해 의약품과 부작용 사이의 연관성을 객관적으로 이해하고 사람들이 보다 쉽게 이해할 수 있도록 프로젝트 주제를 선정하게 되었다.

따라서 우리는 ‘부작용’에 집중하여 다른 attribute(항목)과의 연관성을 탐색하고, 이를 기반으로 시각적으로 나타내어 직관적으로 이해하기 쉽도록 한다. 이 프로젝트를 통해서 의약품 사용자들이 부작용을 경험할 가능성을 이해하고 예측하는 데에 도움이 될 것이라 기대한다.

**2. Related work/methods**

1. Related work

기존에 존재하는 관련 연구들은 적당한 전처리 과정을 거친 후, 빈도수에 기반하여 시각화하는 연구들이 대부분이다. 특이점이 존재하는 연구는 therapeutic class와 habit forming 사이의 관계성을 찾는 연구이다.[[1]](#footnote-1) 단순히 빈도수에 기반한 것이 아니라, Side Effect에 집중하여 Analysis를 통해 연관성을 찾아 보는 것이 이 프로젝트와 기존 연구의 차이점이다.

2. Methods

우리 팀이 사용하고자 하는 dataset은 ‘250k Medicines Usage, Side Effects and Substitutes’라는 dataset을 kaggle에서 제공받아 사용하고자 한다. 이와 같은 dataset을 참고하여 여러 데이터 마이닝 기술로 유용한 정보를 추출하는 작업을 우리 팀의 프로젝트에 적용할 수 있다. 몇 가지 가능한 데이터 마이닝 기법들은 다음과 같다.

1) Clustering

sample들에 대한 지식없이 similarity에 근거하여 pattern space에 주어진 유한 개의 pattern들이 서로 가깝게 모여서 무리를 이루고 있는 pattern 집합을 cluster라고 하고 이러한 처리 과정을 Clustering이라고 한다. 우리가 가지고 있는 dataset을 활용하게 되면 의약품의 특성을 기반으로 비슷한 특성을 가진 의약품들을 그룹화할 수 있다. 예를 들면 유사한 효과를 가진 약물을 특성에 따라 그룹화하여 약물 대체품 식별을 용이하게 하거나 유사한 부작용이 있는 의약품을 그룹화하는 clustering이 있을 것이다.

2) Association Rule Learning, 연관규칙학습

의약품 사용과 부작용 사이의 연관성을 발견해 잠재적인 부작용을 예측하여 어떤 의약품을 사용하는 사람들이 특정 부작용을 경험할 가능성이 높은 지를 알 수 있다. 또는 Jaccard coefficient를 사용하여 약물의 활성 성분을 비교하고 유사한 구성의 대체제를 추천할 수도 있을 것이다.

3) Text Mining

Text Mining이란 비정형 텍스트를 정형화하여 의미 있는 패턴과 새로운 인사이트를 찾아내는 process이다. Text Mining을 활용하면 우리 팀의 dataset으로 의약품의 이름과 특성에 대한 정보를 분석하여 귀중한 정보와 키워드를 추출할 수 있다. 예를 들면 특정 부작용을 유발하는 의약품의 키워드를 추출하여 부작용 관련 정보를 도출할 수 있다.

**3. Approach and methodology**

이 Project는 의료 약물의 사용과 부작용 간의 연관성을 분석하는 것을 목표로 하고 있다.

첫 번째 가설로는 Attribute 중에서 Substitute와 Side Effect간의 연관성이 가장 클 것을 기대하고 다양한 기법을 사용했다.

스크린샷, 다채로움이(가) 표시된 사진

자동 생성된 설명

🡪 하지만 아래의 사진과 같이 어떠한 기준으로 Cluster가 5개로 나뉘었는지 아무런 연관성을 찾을 수 없었다. 또한, 그 Cluster들도 randomly하게 분포되어 있는 것을 확인했다.

데이터의 연관성을 찾기 위해 많이 쓰이는 Regression 기법으로 분석을 시도했다. 그리고 다양한 변수로 인한 non-linear한 data의 특징을 고려하여 Decision Tree(의사결정 나무), Random Forest(랜덤 포레스트) 기법을 사용하여 이 Data Set의 분석을 시도했다.

그리고 시각화를 통해서 의약품을 경험하고, 사용할 수 있는 많은 사람들이 직관적으로 결과를 이해할 수 있도록 한다.

**4. Processing/Results and Evaluation**

1) Dataset

우리가 사용하고자 하는 dataset은 ‘250k Medicines Usage, Side Effects and Substitutes’라는 csv 파일로, 전 세계적으로 사용 가능한 모든 제조업체의 248,000개 이상의 의료 약물에 대한 포괄적인 정보가 담겼고 248,218 rows와 58 columns로 구성되어 있다. 여기에는 약물의 name과 대체품, 부작용, 사용 용도, chemical class, habit forming, therapeutic class, action class가 있다. 이 데이터의 특징은 데이터의 수가 많고 모든 항목이 문자열로 이루어져 있으며 null 값이 많다.

‘부작용’에 집중하여 다른 attribute(항목)과의 연관성을 찾아 그 정보를 이용할 수 있다면, 의약품 연구에 유용한 정보가 될 수 있을 것이라고 생각했다. 그래서 우리는 여기서 약물의 side effect에 집중하여 side effect과 특정 attribute 사이에 연관성이 있다는 가정을 세우고 regression을 사용하여 데이터의 특징을 알아보고자 했다.

2) Data Preprocessing

원본 데이터는 null 값이 많고, 문자열이라는 특이점이 있다. 따라서 연관성을 찾기 위해 Data의전처리 과정이 매우 중요하다.

텍스트, 스크린샷, 폰트, 디자인이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 소프트웨어, 컴퓨터 아이콘이(가) 표시된 사진

자동 생성된 설명텍스트, 전자제품, 스크린샷, 소프트웨어이(가) 표시된 사진

자동 생성된 설명전처리는 각 attribute 마다 missing values의 비율이 45% 이상이면 해당 attribute를 제거하여 dimension을 축소하였고, 다음 단계에서는 가장 빈도가 높은 사용 용도 15개만 남기고 나머지 값은 제거하여 데이터 수를 줄였다. 그 다음 남은 null값을 완전히 없앤 후 label encoding을 하여 문자열을 숫자로 변환하였다.

🡪 원본 dataset과 전처리 후 남은 columns( id, name, substitute0, substitute1, substitute2, substitute3, substitute4, sideeffect0, sideeffect1, sideeffect2, sideeffect3, sideeffect4, use0, chemical class, habit forming, therapeutic class, action class) (전부 null 값 없음)

3) Regression analysis result

텍스트, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명1. Linear regression model (sideeffect0 기준)

처음에는 간단하게 9개의 attribute를 독립변수로, side effect0를 종속변수로 하는 선형회귀모델을 생성했는데 R square 값이 너무 낮아서 해당 모델로 독립변수가 종속변수를 잘 설명할 수 없다고 판단했다.

(독립변수를 선택 시, substitute0, substitute1, substitute2, substitute3, substitute4, use0, chemical class, therapeutic class, action class를 택하고 name, habit forming은 제외했다. name은 약물 이름의 개수가 너무 많았으며, habit forming은 값이 Yes 또는 No로 데이터가 구성되어 있는데, No의 값이 대부분이기에 독립변수로 쓰기에 적절하지 않다고 판단하여 제외했다.)

텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 도표, 다채로움이(가) 표시된 사진

자동 생성된 설명2. Decision tree

선형회귀 모델의 한계를 느낀 후 모델의 비선형성을 파악하고 어떤 변수가 예측에 중요도를 미치는지 확인하기 위해 decision tree와 random forest를 활용하였다. decision tree에서는 위의 그래프에 나와있는 9개의 독립변수를 바탕으로 종속변수인 side effect 0에서 4까지 각각 총 5개의 모델을 생성한 후, importance의 평균을 막대그래프로 나타낸 결과이다. 그리고 오른쪽 표는 feature importance를 나타낸 표이다. Chemical class, action class 순으로 높은 importance를 나타내었다.

텍스트, 스크린샷, 다채로움, 도표이(가) 표시된 사진

자동 생성된 설명텍스트, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명3. Random forest

Random forest에서도 마찬가지로 위의 그래프에 나와있는 9개의 독립변수를 바탕으로 종속변수인 side effect 0에서 4까지 각각 총 5개의 모델을 생성한 후, importance의 평균을 막대그래프로 나타낸 결과이다. 그리고 오른쪽 표는 feature importance를 나타낸 표이다. Chemical class, action class 순으로 높은 importance를 나타내었다.

4) Regression model Evaluation

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명Sideeffect0~4까지의 importance의 평균을 내기 전, Sideeffect0, sideeffect1, sideeffect2, sideeffect3, sideeffect4를 종속변수로 하는 5개의 Decision tree model과 random forest model의 MSE[[2]](#footnote-2)와 R2 square[[3]](#footnote-3)를 평가 지표로 활용했다.

모든 모델이 R2 square 값이 1에 가까운 값이 나오고 MSE의 값도 매우 낮은 값을 보이는 것을 모아 독립변수가 종속변수를 잘 설명할 수 있고 모델의 성능 또한 좋다고 판단했다.

5) Decoding Result

텍스트, 스크린샷, 그래프, 폰트이(가) 표시된 사진

자동 생성된 설명🡪 가장 빈도가 높은 ‘Chemical class’

부작용이 Chemical class와 어느 정도 연관성이 있음을 확인한 후, 그렇다면 chemical class마다 어떤 side effect와 관계가 있는지 알아보고자 했고, 코드 구현 후 왼쪽의 그래프와 같은 결과를 확인할 수 있었다.

이것을 다시 디코딩하면 특정 chemical class가 특정한 side effect 발생에 영향을 미치는지 알 수 있을 것이라 생각해 디코딩을 시도해 보았다.

텍스트, 폰트, 화이트, 대수학이(가) 표시된 사진

자동 생성된 설명(chemical class는 가장 빈도가 높은 상위 3개만, 각각의 chemical class 마다 가장 빈도가 높은 side effects는 0과 1만 일단 디코딩 시도함)

텍스트, 폰트, 화이트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

다시 숫자를 문자열로 변환해본 결과, 특정한 chemical class와 그에 해당하는 side effect가 무엇인지 알 수 있었다. 이처럼 연관성을 찾고 다시 디코딩으로 그 관계를 파악할 수 있다면, 그 기전을 파악할 때 유의미한 결과로 사용될 수 있을 것이라 기대한다.

6) Correlation matrix Heat Map Result & Discussion

스크린샷, 사각형, 직사각형, 텍스트이(가) 표시된 사진

자동 생성된 설명

위의 그림은 회귀모델을 바탕으로 chemical class와 action class 간에 서로 독립적인 관계인지 종속적인 관계인지 파악하기 위해 나타낸 히트맵이다.

‘Chemical Class’와 ‘Action Class’ 간에 서로 강한 상관관계가 있는 것을 알 수 있다. 이를 바탕으로 두 attribute가 서로 종속적인 관계가 있을 것이라고 판단했다. 그리고 앞선 regression 결과의 importance 또한 높지 않은 것으로 보아 다른 독립 변수도 연관성의 요인으로 고려해볼 필요가 있다고 판단했다. 또한, chemical class와 action class 사이의 연관성을 찾아보는 것도 의미 있는 작업이 될 수 있을 것이라고 생각했다.

**5. Conclusion and future work**

1) Conclusion

처음 시도에서 연관규칙학습, Clustering 등 다양한 기법으로 Substitute와 Side Effects 사이의 연관성을 찾았지만 실패했다. 따라서 Side Effect와 다른 Substitute간의 연관성을 찾기 위해 Regression기법을 통해 Data 분석을 시도했다. 결론적으로 ‘Chemical Class’와 ‘Action Class’가 약물의 부작용과 가장 높은 연관성을 가지는 것으로 나타났다.

다만, 한계점으로는 데이터의 많은 null 값과 문자열 데이터로 인해서 전처리 과정에 어려움이 있어 기법 사용에 제한이 많았다. 그리고 label을 인코딩 후, 다시 디코딩을 해야한다는 추가적인 과정이 소요되었다. 이렇게 Side Effects와 다른 Attribute들 간의 연관성을 찾았지만 더 유의미한 결과를 얻기 위해서는 다양한 독립 변수를 고려해야 하는 추가적인 분석이 필요할 것이다.

2) Future work

‘Chemical Class’와 ‘Action Class’, 그리고 ‘Side Effect’ Attribute들 간의 연관성을 찾을 수 있었다. Future work로는 더 다양한 기법을 사용하여 데이터 분석을 진행하여 부작용 예측 모델을 생성하고, 예측 모델의 성능을 향상시키고자 한다. 또한, 혹시 모를 다른 Attribute 간의 연관성을 탐색한다면 보다 정교한 예측 모델을 개발할 수 있을 것이다.

이 프로젝트에서 얻은 연관성을 기반으로 추가적인 연구를 통해 더 많은 연관성을 가진 데이터 베이스를 마련한다면, 연구 결과를 바탕으로 대체 의약품 개발 및 부작용 감소에 큰 도움이 될 수 있을 것으로 기대한다. 그리고 더 나아가서는 이후에 생산되는 대체 의약품 개발에 유의미한 인사이트를 제공할 수 있을 것으로 기대한다.

1. Pharma250k Medicines usage dataset

   <https://www.kaggle.com/code/khanhtang/pharma-250k-medicines-usage-dataset> [↑](#footnote-ref-1)
2. \* MSE: Mean Squared Error(평균 제곱 오차)의 약자로, 예측 값과 실제 값 간의 차이를 측정하는 데 사용되는 통계적 지표입니다. MSE는 회귀 분석이나 기계 학습 모델의 성능을 평가할 때 주로 사용됨 [↑](#footnote-ref-2)
3. \* R square: 독립변수가 종속변수를 얼마나 잘 설명하는지를 나타냄 (0에서 1사이의 값) [↑](#footnote-ref-3)