Journal of KIIT. Vol. 17, No. 12 pp. 21-28, Dec. 31, 2019. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2019.17.12.21

약물 관련 정보를 이용한 약물 부작용 예측

서수경*[1], 이태건*[2], 윤영미**

Prediction of Drug Side Effects Based on Drug-Related Information

Sukyung Seo*[1], Taekeon Lee*[2], and Youngmi Yoon**

이 논문은 2018년도 정부(미래창조과학부)의 재원으로 한국연구재단(No.2018R1A2B6006223)과 2019년도 가천대학교 미래혁신도전과제(2019-0331)의 지원을 받아 수행된 연구임

요 약

약물 부작용이란 질병의 예방, 진단 또는 치료에 사용된 약물로부터 발생한 유해하고 의도하지 않은 현상이다. 이러한 부작용은 환자를 죽음에 이르게 할 수 있으며, 약물 개발 실패의 주요 원인 중 하나이다, 따라서, 다양한 방법들이 부작용을 알아내기 위하여 시도되었다. 본 연구에서는 시스템스 바이올로지 접근법을 기반으로 기존 연구에서 주로 사용되었던 화학적 구조, 생물학적 정보 이외에도 다양한 표현형 정보를 사용하는 것에 주목하였다. 먼저, 5가지 적응증 데이터베이스, 화학적 구조, 타겟 유전자 정보를 수집하고 개별로 유사도를 계산하였다. 테이블은 하나의 약물-부작용에 대하여 앞서 생성된 유사도를 이용하여 생성되었고 다양한 기계학습 기법이 적용되었다. 결과는 AUC(Area Under the ROC Curve)값을 통해 확인하였다. 본 연구의 유의성은 비교 실험을 통하여 확인하였다.

Abstract

Side effects of drugs mean harmful and unintended effects resulting from drugs used to prevent, diagnose, or treat diseases. These side effects can lead to patients' death and are the main causes of drug developmental failures. Thus, various methods have been tried to identify side effects. These can be divided into biological and systems biology approaches. In this study, we use systems biology approach and focus on using various phenotypic information in addition to the chemical structure and target proteins. First, we collect datasets that are used in this study, and calculate similarities individually. Second, we generate a set of features using the similarities for each drug-side effect pair. Finally, we confirm the results by AUC(Area Under the ROC Curve), and showed the significance of this study through a comparison experiment.

Keywords
side effect prediction, indication, machine learning, systems biology

* 가천대학교 DataBio Lab 연구원
- ORCID[1]: http://orcid.org/0000-0001-7753-9213
- ORCID[2]: http://orcid.org/0000-0001-7001-3550
** 가천대학교 컴퓨터 공학과 교수(교신저자)
- ORCID: http://orcid.org/0000-0002-7420-4968

· Received: Nov. 16, 2019, Revised: Dec. 05, 2019, Accepted: Dec. 08, 2019
· Corresponding Author: Youngmi Yoon
Dept. of Computer Engineering, Gachon University, Korea,
Tel.: +82-31-750-4755, Email: ymyoon@gachon.ac.kr

Ⅰ. 서 론

세계보건기구는 약물 부작용을 질병의 예방, 진단 또는 치료에 사용된 약물로부터 발생한 유해하고 의도하지 않은 현상이라 정의하고 있다[1]. 이러한 예기치 못한 부작용은 환자를 죽음에 이르게 할 수 있을 뿐만 아니라 제약 업계에서 약물 개발 실패의 주요 원인 중 하나이다[2].

따라서, 약물의 부작용을 알아내는 것은 매우 중요한 일이며, 이를 위해 다양한 방법들이 시행되고 있다. 생물학적 실험 방법으로는 생화학 및 세포 분석법으로 임상 전에 화합물을 시험하는 방법이 가능하다[3]. 그러나 생물학적 실험을 통하여 가능한 약물-부작용 관계를 실험하는 것은 비용적, 시간적 측면에서 비효율적이다[4]. 그러므로 컴퓨터를 이용하여 가능한 후보 약물-부작용 관계를 실험적 방법보다 빠르게 도출해낼 수 있는 시스템스 바이올로지 접근법이 대두되었다.

기존 연구들은 주로 약물의 화학적 구조와 생물학적 정보인 타겟 유전자를 사용해 약물의 부작용을 예측하였다. Yamanishi는 앞서 언급된 두 가지 종류의 데이터를 통합하여 커널 회귀 모델을 구축하여 약물 부작용을 예측하였으며, 특히 생물학적 정보인 타겟 유전자가 예측 성능에 기여한다는 점을 보였다[5]. Liu는 표현형 데이터인 약물 적응증을 화학적, 생물학적 정보에 추가하였고, 이로 인하여 약물 적응증 정보가 성능 개선과 부작용 예측에 사용될 수 있음을 보였다[6].

본 연구에서는 기존 연구에서 주로 사용되었던 화학적 구조, 생물학적 정보인 타겟 유전자 이외에도 적응증 정보를 여러 데이터베이스로부터 수집하였다. 약물의 화학적 구조는 타니모토 계수를 이용하여, 5가지 적응증 정보와 타겟 유전자는 자카드 계수를 통하여 유사도를 산출하였다. 따라서 총 7가지 정보의 약물 유사도가 본 연구에서 개별 특성(Feature)로 이용되었다. 산출한 유사도는 각 약물-부작용 쌍에 값을 부여하는데 이용되었고 생성된 테이블에 5가지 기계학습 기법이 적용되었다. 따라서, 본 연구는 약물의 다양한 정보와 기계학습 기법을 이용하여 약물의 새로운 부작용을 예측하는 모델을 구축하였다.

본 연구에서는 모델 성능을 비교하기 위하여 AUC(Area Under the ROC Curve)를 이용하였다. 본 연구의방법이 약물 부작용을 예측함에 있어 보다 더 효율적임을 비교 실험을 통하여 보였다.

본 연구는 2절에서 연구와 관련된 주요 데이터베이스, 연구에서 사용된 기계학습 기법, 성능 평가를 위한 기법을 설명하였다. 3절에서는 본 연구의 전체적인 개요도 및 설명과 사용된 데이터의 수집 및 정제, 테이블 구축, 기계학습 기법 적용, 성능 평가 방법에 대하여 자세히 기술하였다. 4절에서는 본 연구의 성능 평가와 비교 연구 제시를 통하여 본 연구의 유의성을 입증하였다. 5절에서는 향후 연구 방향 가능성에 대하여 제시하였다.

Ⅱ. 관련 연구

2.1 관련 연구

본 연구는 기계학습을 기반으로, 약물의 적응증, 화학적 구조, 표적 유전자와 같은 정보를 이용하여 후보 약물 부작용을 찾는 모델을 구축하려 한다. 따라서 다음과 같은 관련 연구를 이용하여 본 연구에 필요한 데이터를 수집하고, 모델을 구축 및 평가하였다.

2.2 repoDB

본 연구는 재창출된 약물 적응증을 수집하기 위하여 repoDB를 이용하였다. 약물 재창출이란 승인된 약물을 이용하여 기존에 알려지지 않은 새로운 적응증을 찾아내는 연구로 질병을 치료하기 위한 새로운 약물의 개발 시간 및 비용을 감소시키는 것으로 알려져 있다. repoDB는 DrugCentral과 ClinicalTrials.gov로부터 재창출 된 약물을 모아놓은 웹 서비스이다. 본 연구는 repoDB에서 1519개의 약과 1229개의 질병에 대한 6677개의 약물-적응증 관계를 수집하였다[7][8].

2.3 XGBoost

XGBoost는 여러 개의 결정 트리를 이용하는 앙

상블(Ensemble)기법으로 분류와 회귀분석 모두에 사용되는 방법이다[9]. 앙상블 기법은 부스팅(Boosting) 기법과 배깅(Bagging)으로 나눌 수 있다. 그 중에서도 XGBoost는 약한 분류기 여러 개를 묶어 정확도를 향상시키는 부스팅 기법에 기반하였다. 이 기법은 실제값과 예측값의 차이를 줄이기 위하여 순차적으로 분류기를 개선함으로써 모델의 정확도를 높일 수 있으며, 더 이상 개선이 이루어지지 않을 때 멈추게 된다. 본 연구에서는 해당 기법을 약물의 부작용을 예측하기 위하여 사용하였다. 따라서 이 기법은 이미 관련이 있다고 알려진 약물 부작용 관계와 무작위로 추출된 약물 부작용 관계를 구별하기 위하여 사용되었다.

2.4 AUC 계산법

본 연구에서는 수행된 10 fold 교차 검증의 성능을 나타내기 위해 AUC(Area Under the ROC Curve)를 사용하였다. AUC를 계산하는 방법에는 AUC_{merge}와 AUC_{avg} 두 가지 방법이 있다[10]. AUC_{merge}는 모든 fold에서의 예측 결과를 정렬하여 단일 ROC곡선을 생성한 다음, 그 곡선의 아래의 면적을 구하는 방법이며, AUC_{avg}는 10 fold 각각에서 AUC를 구하고 이를 평균하여 산출하는 방법이다. 본 연구에서는 AUC_{merge}를 이용하여 하나의 세트 당 하나의 AUC를 구하였다.

III. 연구 방법

3.1 전체 개요도

본 연구의 전체 개요는 그림 1과 같다. 먼저, 연구에 사용될 데이터를 수집 및 정제하였다. 약물의 화학적 구조, 타겟 유전자, 적응증 데이터는 약물 유사도를 계산하는데 이용된다. 약물-부작용 정보는 기계학습을 위한 학습에 필요한 Positive와 Negative 약물-부작용 쌍을 생성하는데 사용된다.

Positive세트는 알려진 약물-유전자 정보를 이용하여 생성하고, Negative세트는 Positive세트에서 사용되지 않은 약물-유전자 쌍을 이용하여 생성된다.

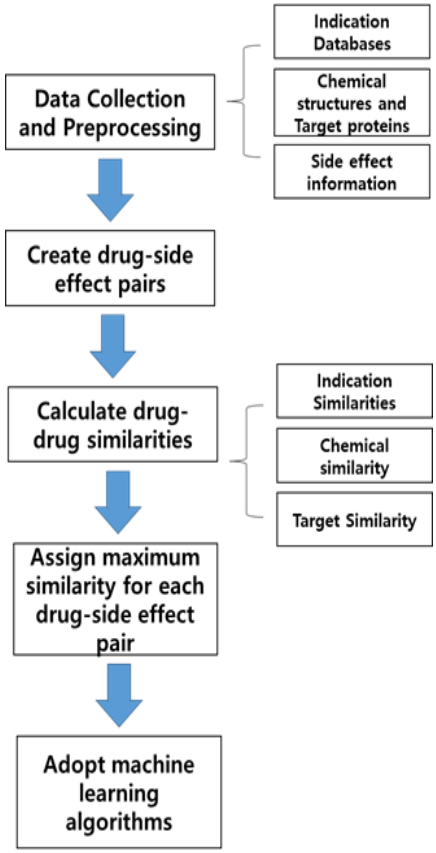


그림 1. 전체 개요
Fig. 1. System overview

위에서 수집된 약물 적응증, 화학적 구조, 타겟 유전자를 이용하여 약물-약물 유사도를 계산한다. 약물 적응증의 경우에는 적응증 데이터베이스별로 계산한다.

각 약물-부작용 쌍은 위에서 계산한 약물-약물 유사도를 이용하여 총 7가지 특성에 대한 값을 가지게 된다.

생성된 테이블에 다양한 기계학습 알고리즘(random forest, naive bayes, neural network, logistic regression, XGBoost)을 적용한다.

3.2 기존에 알려진 약물 적응증 수집

본 연구는 4개의 서로 다른 데이터베이스에서 알려진 약물-적응증 관계을 수집했으며, 각 적응증 데이터베이스는 단일 특성으로 사용되었다. 먼저, Therapeutic Target Database(TTD)는 문헌에서 선별된 적응증 정보를 수집하고 있는 데이터베이스이다[11].

표 1. 본 연구에서 사용된 약물-적응증 정보
Table 1. Experimental data sets

Type	Database name	Number of drugs	Number of indications	Number of drug-indication pairs
Repurposed drug indications	repoDB	1,519	1,229	6,677
Known drug indications	CTD	1,510	5,709	150,175
	PREDICT	590	304	1,912
	TTD	1,000	1,298	44,481
	Clinical Trials	772	7,883	40,638

각 질병에 대해 ICD-10(International Classification of Diseases) 코드를 Disease Ontology(DO)에서 제공하는 UMLS(Unified Medical Language System)에 매핑하여 1,000개의 약물과 1,298개의 질병에 대하여 44,481개의 약물-적응증 관계를 수집하였다[12]. Comparative Toxicogenomics database(CTD)에서는 1,510개의 약물과 5,709개의 질병에 대하여 150,175 약물-적응증 관계를 수집하였다[13]. Gottlieb의 PREDICT 논문에서는 제시된 590가지 약물과 304가지 질병에 대하여 1,912가지 약물-질병 관계를 수집하였다[14]. ClinicalTrials에서는 1단계에서 5단계까지 사용된 772개의 약물과 4,181개의 질병에 대하여 40,638개의 약물-적응증 관계를 수집했다[15]. 약물 이름은 DrugBank에서 제공한 약물 용어 파일을 이용하여 매핑하였으며, 질병 이름은 CTD에서 제공된 질병 어휘 파일을 사용하여 UMLS ID로 통일되었다.

3.3 약물의 화학적 구조, 표적 유전자 수집

화학적 구조는 약물의 화학적 특성을 나타내기 위하여 사용되었다. 본 연구는 PubChem과 DrugBank에서 1,878가지 약물에 대한 화학적 구조를 SMILES (Simplified Molecular-Input Line-Entry System) 형태로 수집하였다[15][16].

약물의 표적 유전자는 DrugBank에서 1,366가지 약물에 대하여 6,076개의 약물-표적 연관성을 수집하였다[17].

3.4 유사도 측정 및 세트 생성

본 연구에서는 위에서 언급한 약물의 적응증, 화학적 구조, 표적 유전자를 이용하여 약물의 유사를 계산하였다. 약물의 적응증과 표적 유전자의 경우 식 (1)과 같이 자카드 계수를 이용하여 유사도를 측정하였다.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

자카드 계수는 0에서 1 사이의 값을 가지며, 1에 가까울수록 유사도가 높은 것이다. 유사도를 측정하고자 하는 두 약물이 가지고 있는 약물 적응증 혹은 표적 유전자 전체 집합에 대하여 공통된 항목이 많을수록 높은 값을 갖는다.

화학적 구조를 이용한 약물의 유사도는 타니모토 계수를 통하여 측정되었다. 타니모도 계수란 식 (2)와 같이 교차 집합 대 조합 집합의 비율로 유사도를 계산하며 자카드 계수와 마찬가지로 0에서 1사이의 값을 가진다. N_a, N_b는 각각 약물 A, B의 속성의 수, N_c는 두 약물 사이의 겹치는 속성의 수를 나타낸다.

$$T(A,B) = \frac{N_c}{N_a + N_b - N_c} \quad (2)$$

알려진 최신 약물-부작용 관계는 SIDER 데이터베이스에서 107,878개를 수집하였다. 각 약물은 biodb.jp를 이용하여 PubChem 화합물 ID를 DrugBank ID로 통일하였다[18].

Positive 세트는 SIDER에서 수집한 알려진 약물-부작용 관계를 이용하여 생성되었으며 클래스 T (True)를 부여하였다. Negative세트는 Positive세트에서 사용된 약물과 부작용을 이용하여 전체 약물-부작용 쌍을 생성하고 Positive세트에서 사용된 쌍을 제외한 뒤 동수 추출하였다. Negative세트는 표본 편향을 피하기 위하여 무작위로 5세트를 선택하고 클래스 F(False)를 부여하였다. 하나의 세트는 Positive세트와 Negative 1세트를 합쳐서 만들어졌다.

3.5 테이블 생성 및 성능 평가

이 단계에서는 위에서 생성된 세트와 유사도를 이용하여 테이블을 생성하였다. 하나의 샘플인 약물

-부작용 쌍에 대하여 샘플의 부작용과 알려진 관계가 있는 약물을 Training 세트의 Positive 세트에서 찾아낸다. 찾아낸 약물들과 샘플 약물의 유사도를 3.4절의 방법을 통하여 계산한 후, 최대 유사도를 해당 특성에 대한 샘플 약물-부작용 쌍의 값으로 사용하였다. 따라서, 하나의 약물-부작용 쌍은 7가지의 특성 값을 가지게 된다.

생성된 테이블에 5가지 기계학습 방법(random forest, naive bayes, XGBoost, logistic regression, neural network)을 적용함으로써, 약물-부작용 관계를 예측하는 분류기를 구축하였다.

본 연구에서는 제안된 방법의 성능을 평가하기 위하여 10 fold 교차 검증을 수행하였다. 따라서 하나의 테이블은 10등분 되며, 9개 등분은 Training 세트로 사용하고 나머지 1개 등분을 Test 세트로 사용하는 방법을 10번 반복하였다. 하나의 테이블 당 하나의 AUC_{merge}값을 구하고, 최종적으로 5개 테이블을 통하여 얻어진 AUC_{merge}값의 평균을 구하였다.

Ⅳ. 결 과

본 연구에서 약물의 화학적, 생물학적, 표현형 특성을 이용하여 약물-약물 사이의 유사도를 계산하였다. 약물 간 유사도를 기반으로 기계학습 기법을 적용하여 약물이 특정한 부작용을 가지고 있는 가를 식별하는 분류기를 구축하였다. 먼저, 기존에 사용되던 화학적, 생물학적 특성과 약물 적응증 특성이 개별적으로 사용되었을 때 약물-부작용 관계 식별을 위한 분류기의 성능을 보인 후, 다양한 특성들을 함께 사용하였을 때 성능이 향상됨을 보였다.

본 연구는 각 특성에서 약물 유사도 값이 없는 약물은 제외하였기 때문에 258개의 약물과 3065개의 부작용을 사용하였다.

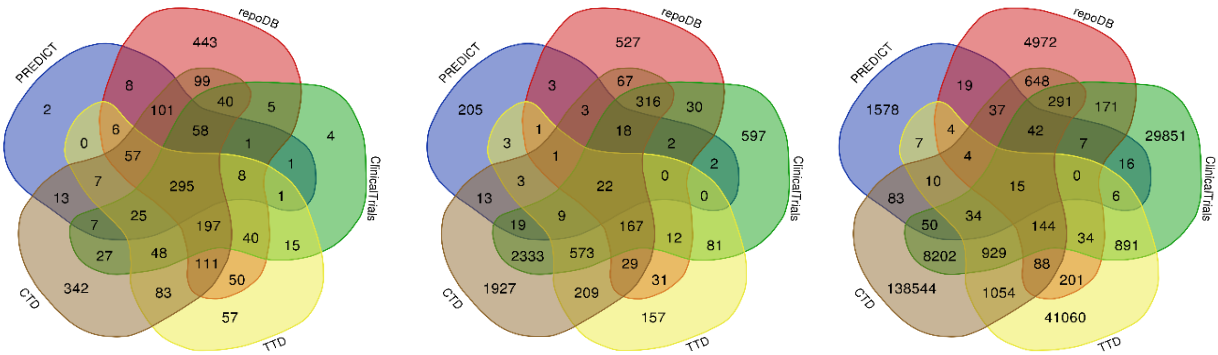
4.1 약물 적응증 데이터베이스

본 연구에서는 약물 적응증 정보를 5가지의 데이터베이스에서 수집하였다. 그림 2는 약물 적응증 데이터베이스가 포함하고 있는 약물과 적응증에 대한 벤다이어그램이다.

각각의 데이터베이스는 약물과 적응증 정보를 수집하는 방법에 있어 차이를 갖고, 이로 인하여 데이터베이스 별로 포함하고 있는 약물과 질병의 종류가 다름을 그림 2를 통하여 확인할 수 있다. 각 데이터베이스에서 포함하고 있는 약물 종류의 개수를 그림 2(a)에, 약물 적응증 종류의 개수를 그림 2(b)에, 약물-적응증 쌍 개수를 그림 2(c)에 나타내었다. 본 연구에서는 데이터베이스 별 차이를 고려하여 데이터베이스 각각을 특성으로 사용한 경우와 하나로 통합한 경우를 비교하였다.

표 2. 약물 적응증을 특성으로 이용한 실험 결과
Table 2. Experimental results using drug indications as features

	Random forest	Logistic regression	Neural network	XGBoost	Naïve bayes
Combined data	0.8560	0.8148	0.8203	0.8530	0.8195
Separated data	0.8699	0.8516	0.8583	0.8763	0.8517



(a) 약물의 개수 (b) 약물 적응증 (c) 약물-적응증 쌍 개수

그림 2. 약물과 적응증 정보에 관한 벤다이어그램

Fig. 2. Venn diagram of drug and indication information, (a) Number of drugs, (b) Number of indications, (c) Number of drug-indication pairs

표 2를 통하여 약물 적응증은 데이터베이스를 각각의 특성으로 활용하는 것이 모든 기계학습 방법에서 좋은 결과를 얻음을 확인하였으며, 따라서 본 연구에서는 약물 적응증 정보를 데이터베이스 별로 구분하여 활용하였다.

4.2 성능 평가

본 연구에서는 기계학습 방법으로 약물-부작용 관계를 식별하는 분류기를 구축함에 있어 화학적, 생물학적, 약물 적응증 정보가 독립적으로 보이는 성능을 테스트하였다.

그림 3은 각각의 특성 정보를 개별적으로 사용하여 분류기를 구축하고 AUC_{merge}를 계산하는 과정을 5회씩 반복한 후 평균한 값이다. 모든 기계학습 방법에서 약물 적응증, 생물학적 특성, 화학적 특성 순으로 좋은 성능을 보였다.

표 3은 위에서 사용한 개별적인 특성을 함께 사용하여 학습했을 경우의 성능을 보여준다. 전통적으로 사용되던 약물의 화학적, 생물학적 특성을 함께 사용했을 경우에 해당 특성을 개별적으로 사용하는 것보다 좋은 성능을 보였다. 또한 약물의 적응증을 추가할 경우 결과가 향상됨을 확인하였다. 가장 좋은 분류기 성능을 보인 기계학습 방법은 XGBoost로 0.8856의 평균 AUC를 보였다.

또한, 본 연구가 제시한 특성정보가 기존 연구에서 제시된 특성에 비하여 약물-부작용 관계를 더 잘 예측한다는 점을 보이기 위하여 Zhao의 연구와 비교하였다. Zhao는 약물-부작용 쌍의 관계를 ATC code, 약물 표적 유전자, 화학적 구조, 약물 fingerprint, 문헌 정보에 다양한 기계학습 방법을 적용하여 후보 부작용을 제시한 연구이다[19]. 해당 연구는 random forest, nearest neighbor, dagging, support vector machine와 같은 4가지 알고리즘을 적용하였고, random forest를 사용하여 구축된 분류기에서 가장 좋은 분류 성능을 얻었다. 표 4는 Zhao et al.에서 최종적으로 선택된 분류기의 성능과 사용된 약물과 부작용의 수를 나타낸다.

표 3. 적응증 특성이 추가되었을 때의 실험 결과
Table 3. Experimental results when indication features are added

	Random forest	Logistic regression	Neural network	XGBoost	Naïve bayes
Chemical +Target	0.8237	0.8215	0.8302	0.8412	0.8307
Chemical+ Target+ Indication	0.8744	0.8613	0.8739	0.8856	0.8595

표 4. Zhao 연구와 비교 실험
Table 4. Comparsion with Zhao's study

	Zhao's (random forest)	Our study (XGBoost)
AUC	0.8492	0.8856
Drug	841	258
Side effects	824	3,065

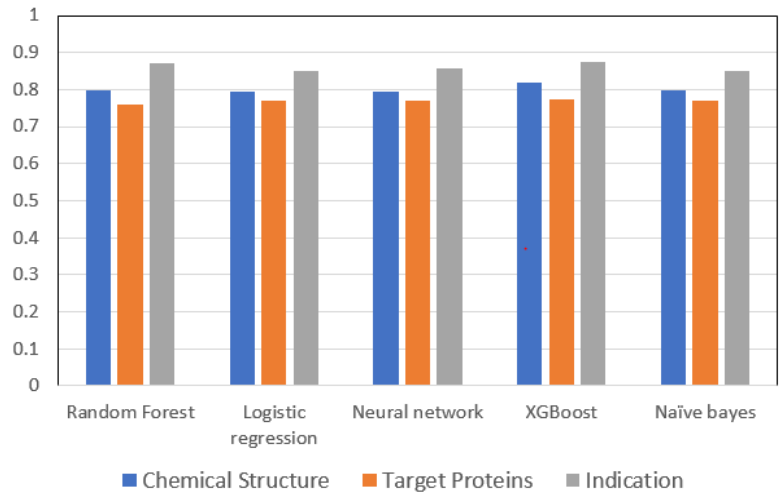


그림 3. 3가지 데이터 카테고리 별 실험 결과
Fig. 3. Experimental results for three data categories

표 4는 본 연구에서 제안한 특성을 사용하여 구축된 분류기가 Zhao의 연구에서 제안된 특성을 사용한 분류기보다 우수한 성능을 보인다는 것을 보여준다. 또한 본 연구는 기존에 수행된 연구에 비하여 보다 많은 부작용에 대하여 예측이 가능하다는 장점을 갖는다.

Ⅴ. 결론 및 향후 과제

본 연구는 약물에 대한 정보를 이용하여 약물의 부작용을 예측하였다. 기존 연구에서 주로 고려한 약물의 화학적, 생물학적 특성 이외에도 다양한 적응증 데이터베이스로부터 수집한 약물의 표현형 정보를 추가하였다. 따라서, 본 연구는 다양한 적응증 정보를 개별 특성으로 추가하였을 때 화학적, 생물학적 정보만을 사용한 것보다 모델의 성능이 향상됨을 보였다.

향후 연구에서는 다른 방법으로 수집된 약물 적응증 데이터베이스를 추가함으로서 정확도를 더 개선시킬 것으로 생각한다. 뿐만 아니라 표현형 데이터인 적응증 정보 이외에도 다른 생물학적, 화학적 정보를 추가할 수 있을 것으로 기대한다.

References

[1] World Health Organization, "International drug monitoring: the role of national centres, report of a WHO meeting [held in Geneva from 20 to 25 September 1971]", World Health Organization, 1972.

[2] T. Kennedy, "Managing the drug discovery/ development interface", Drug discovery today, Vol. 2, No. 10, pp. 436-444, Oct. 1997.

[3] S. Whitebread, J. Hamon, D. Bojanic, and L. Urban, "Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development", Drug discovery today, Vol. 10, No. 21, pp. 1421-1433, Nov. 2005.

[4] N. P. Tatonetti, T. Liu, and R. B. Altman, "Predicting drug side-effects by chemical systems biology", Genome biology, Vol. 10, No. 9, p. 238. Sep. 2009.

[5] Y. Yamanishi, E. Pauwels, and M. Kotera, "Drug side-effect prediction based on the integration of chemical and biological spaces", Journal of chemical information and modeling, Vol. 52, No. 12, pp. 3284-3292, Dec. 2012

[6] M. Liu, Y. Wu, Y. Chen, J. Sun, Z. Zhao, X. W. Chen, M. E. Matheny and H. Xu, "Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs", Journal of the American Medical Informatics Association, Vol. 19, No. e1, pp. e28-e35, Jun. 2012.

[7] A.S. Brown and C.J. Patel, "A standard database for drug repositioning", Scientific data, Vol. 4, p. 170029, Mar. 2017.

[8] O. Ursu, J. Holmes, J. Knockel, C. G. Bologa, J. J. Yang, S. L. Mathias, S. J. Nelson, and T. I Oprea, "DrugCentral: online drug compendium", Nucleic acids research, p.gkw993, Oct. 2016.

[9] T. Chen, T. He, M. Benesty, V. Khotilovich, and Y. Tang, "Xgboost: extreme gradient boosting", R package version 0.4-2, pp. 1-4, Aug. 2015.

[10] G. Forman and M. Scholz, "Apples-to-apples in cross-validation studies: pitfalls in classifier performance measurement", ACM SIGKDD Explorations Newsletter, Vol. 12, No. 1, pp. 49-57, Nov. 2010.

[11] Y.H. Li, C.Y. Yu, X.X. Li, P. Zhang, J. Tang, Q. Yang, T. Fu, X. Zhang, X. Cui, G. Tu, and Y. Zhang, "Therapeutic target database update 2018: enriched resource for facilitating bench-to-clinic research of targeted therapeutics", Nucleic acids research, Vol. 46, No. D1, pp. D1121-D1127, Nov. 2017.

[12] W. A. Kibbe, C. Arze, V. Felix, E. Mitraka, E. Bolton, G. Fu, C. J. Mungall, J. X. Binder, J. Malone, D. Vasant, and H. Parkinson, "Disease Ontology 2015 update: an expanded and updated

database of human diseases for linking biomedical knowledge through disease data", Nucleic acids research, Vol. 43, No. D1, pp. D1071-D1078, Oct. 2014.

[13] A. P. Davis, C. J. Grondin, R. J. Johnson, D. Sciaky, R. McMorran, J. Wiegers, T. C. Wiegers and C. J. Mattingly, "The comparative toxicogenomics database: update 2019", Nucleic acids research, Vol. 47, No. D1, pp. D948-D954, Sep. 2018.

[14] A. Gottlieb, G.Y. Stein, E. Ruppin, and R. Sharan, "PREDICT: a method for inferring novel drug indications with application to personalized medicine", Molecular systems biology, Vol. 7, No. 1, Jan. 2011. https://doi.org/10.1038/msb.2011.26.

[15] D. A. Zarin, T. Tse, R. J. Williams, R. M. Califf, and N. C. Ide, "The ClinicalTrials. gov results database—update and key issues", New England Journal of Medicine, Vol. 364, No. 9, pp. 852- 860, Mar. 2011.

[16] Y. Wang, T. Suzek, J. Zhang, J. Wang, S. He, T. Cheng, B. A. Shoemaker, A. Gindulyte, and S.H.Bryant, "PubChem bioassay: 2014 update", Nucleic acids research, Vol. 42, No. D1, pp. D1075-D1082, Nov. 2013.

[17] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, and N. Assempour, "DrugBank 5.0: a major update to the DrugBank database for 2018", Nucleic acids research, Vol. 46, No. D1, pp. D1074-D1082, Nov. 2017.

[18] M. Kuhn, M. Campillos, I. Letunic, L. J. Jensen, and P. Bork, "A side effect resource to capture phenotypic effects of drugs", Molecular systems biology, Vol. 6, No. 1, Jan. 2010. https://doi.org/10.1038/msb.2009.98

[19] X. Zhao, L. Chen, and J. Lu, "A similarity-based method for prediction of drug side effects with heterogeneous information", Mathematical biosciences, Vol. 306, pp. 136-144. Dec. 2018.

저자소개

서 수 경 (Sukyung Seo)

2018년 2월 : 가천대학교 컴퓨터공학과(공학사)
2018년 3월 ~ 현재 : 가천대학교 DataBio Lab 연구원
관심분야 : 데이터마이닝, 네트워크 바이올로지, 바이오인포매틱스

이 태 건 (Taekeon Lee)

2018년 2월 : 가천대학교 컴퓨터공학과(공학사)
2018년 3월 ~ 현재 : 가천대학교 DataBio Lab 연구원
관심분야 : 데이터마이닝, 네트워크 바이올로지, 바이오인포매틱스

윤 영 미 (Youngmi Yoon)

1981년 : 서울대학교 자연과학대학 (학사)
1983년 : 오하이오 주립대학 수학과(학사 수료)
1987년 : 스탠포드대학교 컴퓨터 과학과 졸업(이학석사
2008년 : 연세대학교 컴퓨터과학과 졸업(공학박사)
1987년 5월 ~ 1993년 5월 : IntelliGenetics Inc., California, USA, Software Engineer
1995년 2월 ~ 현재 : 가천대학교 컴퓨터공학과 교수
관심분야 : 데이터베이스 시스템, 데이터 마이닝, 바이오인포매틱스, 소셜미디어 데이터 마이닝