A Machine Learning-Based Method for Predicting unknown Pharmacointeractions

Jayshree Ghorpade Aher, Shreyans Magdum, Nandini Sonkusakle, Parul Jaiswal, Raj Shah

Abstract: ***A lot of research has been done on the efficacy of machine learning algorithms in predicting the pharmacological interference between two drugs. Ordinarily, this interference depends on many factors such as the taxonomical, chemical, pharmacological or genomic similarities between the two drugs. Nevertheless, a lot of adverse events (AEs) are reported every year, due to the simultaneous consumption of two or more drugs. Much research has been conducted on the accuracy of the interference prediction based on these factors, each differing in the algorithms and factors used. In this publication, we propose a machine learning-based approach to predict undiscovered drug-drug interactions based on a few of the impacting factors, for better results and thus, help minimize the potential harm that can be caused to society.***

Keywords: drug-drug interactions, pharmacointeraction, machine learning, DDI

I. INTRODUCTION

Drug interactions are a much-researched field in pharmacology. Drug interactions or pharmacointeractions as they are called, are of many types like drug-food, drug-disease, and drug-drug, all of which could lead to minor or major effects on the consumer. Since most DDIs go undetected even in the premarket phase, an early detection during the post-market phase could prevent potential AEs, either by adding a warning clause on the drug label, or even through drug withdrawal. [6] Drug interactions are mostly identified through in vivo and/or in vitro analysis of interfering drugs. This has resulted in a database of all known interactions between marketed drugs. But, usually till the drug is experimented and found to be interactive, it has already been marketed, this leads to complications as some reactions are fatal enough to force the organization to withdraw the drug, which is extremely difficult after the marketing and delivery of the drug. Hence recently many researches were conducted to predict the unknown interactions to reduce the probability of launching a possibly harmful drug. Many researches were found to be successful that range from drug query systems, and prediction algorithms that use various machine learning algorithms to predict the unknown interactions. Some of the algorithms used only the known interactions to predict the unknown ones, while some used taxonomic and intrinsic properties.

Revised Manuscript Received on December 05, 2019.

Prof. Jayshree Ghorpade Aher, Assistant Professor, Department Computer Engineering, MIT WPU University. Pune, (Maharashtra) India.

Shreyans Magdum, Department Computer Engineering, MIT College of Engineering, Pune, (Maharashtra) India.

Parul Jaiswal, Department Computer Engineering, MIT College of Engineering, Pune(Maharashtra) India.

Nandini Sonkusakle, Department Computer Engineering, MIT College of Engineering, Pune(Maharashtra) India.

Raj Shah, Department Computer Engineering, MIT College of Engineering, Pune(Maharashtra) India.

Machine learning has proven to be a novel and useful way to predict drug interactions and has a lot of potential to be further enhanced to increase its accuracy and reliability.

A. Basic Concepts

1) Jaccard similarity coefficient: Used for measuring the similarity between two finite sets, and is defined as the size of the intersection of the sets, divided by the size of their union.[1]

$0 \leq J(A, B) \leq 1$.

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

2) Hamming distance: It is the number of positions between two strings of equal length, at which the corresponding symbols are different.

3) Multinomial logistic regression: This classification method generalizes logistic regression to multiclass problems, i.e. with more than two possible discrete outcomes. [2][7]

B. Anatomical Therapeutic Chemical Classification System

The Anatomical Therapeutic Chemical (ATC) Classification System is used for the classification of active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties.

It is a hierarchical system, in which drugs are classified into groups at five different levels. [4][8] It is possible for a drug to have more than one ATC codes depending upon the dosage and other factors.

Example:

B01AC06	—	Acetylsalicylic acid
B01AC	—	Platelet aggregation inhibitors excl. heparin
B01A	—	ANTITHROMBOTIC AGENTS
B01	—	ANTITHROMBOTIC AGENTS
B	—	BLOOD AND BLOOD FORMING ORGANS

II. LITERATURE SURVEY

1) Recently, network pharmacology involving a network-based drug development strategy has created a novel paradigm for drug discovery. Hence, leveraging multidimensional drug properties for a machine learning prediction model, could help discover unknown DDIs.

2) A heterogeneous network-assisted inference (HNAI) framework can be used to design a multivariate classifier, for DDI prediction.

Four types of drug-drug similarities are used as features for each drug-drug pair that can be used for predictions: phenotypic, therapeutic, and structural and genomic similarity.

3) A previous study which used five machine learning algorithms as predictive models in the HNAI framework: Naive Bayes (NB), decision tree (DT), k-nearest neighbors (k-NN), logistic regression (LR), and support vector machine (SVM), indicated that HNAI showed observably high performance. [13]

4) A novel signal detection algorithm that identified latent adverse-event signals from spontaneous reporting systems, to create trained models for predicting DDIs, was developed. 8 distinct types of adverse events were the basis for training the models, which made 171, as of then, undiscovered drug interaction predictions. [6]

5) Conducted research also gives insights on the effectiveness of the predictions using taxonomic similarities in the drugs. Taxonomic attributes of the drugs are given by ATC codes. It signifies the part of the human system the drug works on and the 10 therapeutic, pharmacological, chemical subgroups of the drug. [12]

III. RESULTS AND DISCUSSIONS

For predicting the unknown drug-drug interactions, a multi-stage prediction model has been proposed, as shown in the Fig. 1 below.

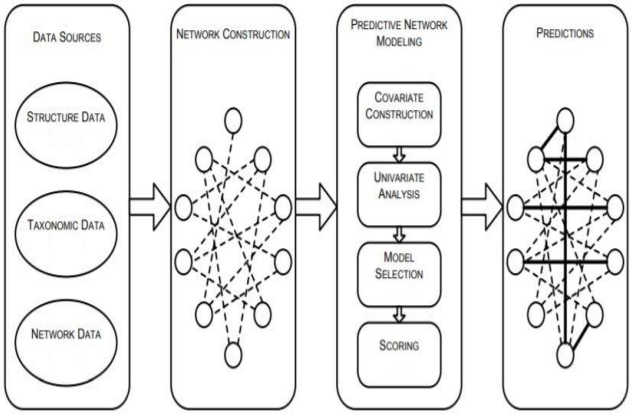

Fig. 1: The Multi-Stage Prediction Model

A. Network Construction

It is intended to construct a database containing data related to the existing network of interacting drugs along with their desired attributes, as mentioned in the DrugBank database version 5.0.11. [3][5][9][10][11]

Each drug will have the following major attributes associated with it:

1) Structure Data: Each drug will have an associated list of similar chemical structures and a count of the number of items on the list.

2) Taxonomic Data: Each drug will have an associated list of ATC codes, used for the identification of the drug.

3) Network Data: Each drug will have an associated list of interacting drugs and a count of the number of items on the list.

Furthermore, these attributes will then be used in the Predictive Network Modeling stage for the generation of co-variates to be used for further analysis.

B. Predictive Network Modeling

In this stage, for each pair of drug-drug interactions to be predicted, three co-variates will be generated via univariate analysis scores of their corresponding attributes, as follows:

1) Scaled Jaccard similarity coefficient of Network Data:

Network Data similarity is calculated as the Scaled Jaccard similarity coefficient of two drugs in terms of their drug-drug interactions similarity. The likelihood of interaction between the two drugs increases with a greater Scaled Jaccard similarity coefficient, as it implies that the two drugs in a pair under consideration, interact with similar sets of drugs.

The Scaled Jaccard similarity coefficient is a modification of the Jaccard similarity coefficient, which also accounts for the relative difference between the sizes of the lists of interacting drugs of each drug in a pair.

$$J_S(X,Y)=J(X,Y)*(MAX(X,Y)/MIN(X,Y)) \quad \ldots(2)$$

Where,
$J_S(X,Y)$ is the Scaled Jaccard similarity coefficient,
$J(X,Y)$ is the Jaccard similarity coefficient,
$MAX(X,Y)$ is the size of the greater of the two sets,
$MIN(X,Y)$ is the size of the smaller of the two sets.

Ordinarily, the set of known drug-drug interactions for a given drug is incomplete, due to factors such as infrequent use of certain drug-drug combinations, lack of availability in a region, etc. In cases like these, the Scaled Jaccard similarity coefficient, thus gives a better understanding of the drug-drug interactions similarity of the two drugs, by eliminating the bias caused due to the difference in the number of drugs with which each of the drugs interact.

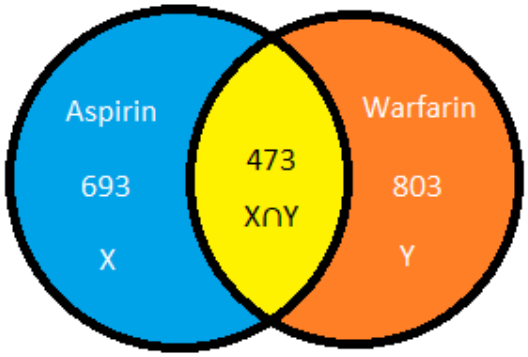

Fig. 2: Set of Interacting Drugs Common to Both Aspirin and Warfarin

Example 1: *Warfarin and Aspirin are known to have a major drug-drug interaction and have 473 drug interactions in common as shown in Fig. 2. Yet, the Jaccard coefficient for the pair (Warfarin, Aspirin) is relatively low at 0.4623.* [3][5][9][10][11]

In this case, the number of common interactions i.e. 473 is a significant portion of the total number of drugs with which Aspirin reacts (693), as compared to Warfarin (803). Here, the Scaled Jaccard similarity coefficient is 0.4623(803/693) = 0.5375, which is a more accurate* measure.

Example 2:

TABLE I.

Drug Pair	Attributes		
	Jaccard Coefficient of DDIs	**Scaled Jaccard Coefficient of DDIs**	**Existence of Interaction**
Warfarin, Acetaminophen	0.1995439	0.643509	Yes
Warfarin, Atorvastatin	0.1969552	0.347594	No

In the above example, it can be seen that the pairs, (Warfarin, Acetaminophen) and (Warfarin, Atorvastatin) have similar values of Jaccard Coefficient, yet one pair is known to interact while the other isn't. In this case, the Scaled Jaccard Coefficient helps distinguish the two pairs.

2) Minimum ATC-Level-Weighted Hamming Distance of Taxonomic Data:

For every drug-drug pair, their corresponding lists of ATC codes will be checked pairwise to obtain the minimum possible level-weighted hamming distance 'd_{atc}', by associating each mismatching character with a weight equal to it's level in the hierarchial ATC system, for a given pair of ATC codes amongst all the possible pairs.

The likelihood of the interaction between two drugs increases with if the minimum level-weighted hamming distance of their ATC codes is small.

Example: *The minimum Hamming distance between the ATC codes of Aspirin (B,01,A,C,06) and Warfarin (B,01,A,A,03) is, 0+0+0+4+5 = 9.*

3) Jaccard similarity coefficient of Structure Data: A greater Jaccard similarity coefficient of the structural data, implies that the two drugs in a pair under consideration, have similar sets of chemically similar structures and are thus, more likely to interact with each other, as chemically similar compounds tend to be metabolized in a similar manner.

Example: *The Jaccard similarity coefficient for structural data of the pair (Warfarin, Dicoumarol), as per the similar structures listed on the DrugBank for each of the drugs at a similarity threshold of 0.7, is relatively high at 0.75. This pair is known to have a major drug-drug interaction due to therapy duplication.* [3][5][9][10][11]

C. Predictions

In this stage, the scores generated in the Predictive Network Modeling stage, resulting in the ordered tuple given by (3),

$$(J_S(X,Y),\ d_{atc},\ J(X,Y)) \quad \ldots (3)$$

Where,

$J_S(X,Y)$ is the Scaled Jaccard similarity coefficient of Network Data,

d_{atc} is the Minimum Hamming Distance of Taxonomic Data,

$J(X,Y)$ is the Jaccard similarity coefficient of Structure Data,

is used as input for a Multinomial Logistic Regression module, which predicts the type of interaction between each drug-drug pair as one of four types, namely: "Major", "Moderate", "Minor" and "None".

This module will need to be provided with a training set of N drugs, where each of the N*(N-1)/2 drug-drug interaction pairs along with their corresponding attributes, are labeled with the appropriate type.

The drug-drug interaction pairs with higher values of $J_S(X,Y)$ and $J(X,Y)$, and lower values of d_{atc}, tend to belong to the "Major" and "Moderate" categories of drug-drug interactions, whereas other combinations tend to belong to the "Minor" and "None" categories.

IV. CONCLUSION

As per the research literature, study and healthcare organizations, it has been found that the drug consumption rate in society is increasing day-by-day. A lot of adverse events are reported every year, due to the simultaneous consumption of two or more drugs. The proposed machine learning-based approach will help the healthcare industry and society in general, to get feasible and effective recommendations against these drug interactions and will thus, serve the society.

REFERENCES

[1] Wikipedia contributors. (2018, June 7). Jaccard index. In Wikipedia, The Free Encyclopedia. Retrieved 06:45, June 8, 2018, from https://en.wikipedia.org/w/index.php?title=Jaccard_index&oldid=844834546

[2] Wikipedia contributors. (2018, May 11). Multinomial logistic regression. In Wikipedia, The Free Encyclopedia. Retrieved 06:46, June 8, 2018, from https://en.wikipedia.org/w/index.php?title=Multinomial_logistic_regression&oldid=840626853

[3] Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res. 2014 Jan 1;42(1):D1091-7

[4] https://www.whocc.no/atc_ddd_methodology/history/

[5] Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A, Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2017 Nov 8. doi: 10.1093/nar/gkx1037

[6] Cami A, Manzi S, Arnold A, Reis BY (2013) Pharmacointeraction Network Models Predict Unknown Drug-Drug Interactions. PLoS ONE 8(4): e61468. doi:10.1371/journal.pone.0061468

[7] Greene, William H. (2012). Econometric Analysis (Seventh ed.). Boston: Pearson Education. pp. 803–806. ISBN 978-0-273-75356-8,

[8] Wikipedia contributors. (2018, June 5). Anatomical Therapeutic Chemical Classification System. In Wikipedia, The Free Encyclopedia. Retrieved 06:49, June 8, 2018, from

https://en.wikipedia.org/w/index.php?title=Anatomical_Therapeutic_Chemical_Classification_System&oldid=844481958
[9] Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A, Banco K, Mak C, Neveu V, Djoumbou Y, Eisner R, Guo AC, Wishart DS. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. Nucleic Acids Res. 2011 Jan;39(Database issue):D1035-41
[10] Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M. DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. 2008 Jan;36(Database issue):D901-6
[11] Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z, Woolsey J. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D668-72
[12] Jonathan R. Nebeker, John F. Hurdle, Jennifer M. Hoffman, Beverly Roth, Charlene R. Weir, Matthew H. Samore, 'Developing a Taxonomy for Research in Adverse Drug Events: Potholes and Signposts', Journal of the American Medical Informatics Association, 2002
[13] Feixiong Cheng, Zhongming Zhao, 'Machine learning-based prediction of drug-drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties', J Am Med Inform Assoc, 2014

AUTHORS PROFILE

Prof. Jayshree Ghorpade-Aher, is working as an Assistant Professor in MIT WPU University. She is pursuing her Ph.D. in Computer Engineering from P.I.C.T., Pune. She has contributed to 60+ publications with 07 Best Paper Awards by IEEE, ACM, Scopus, etc. She was honored with the Prestigious International Individual Award 'Paper Presenter Award at International Conference' by the CSI Awards Committee-2017. Her area of interest are soft computing, data analytics & machine learning.

Shreyans Magdum, is currently a Systems Engineer at Infosys, where he has been working as a part of the Infosys Automation Group for the past year. He graduated from MIT College of Engineering, Pune in 2018. His interests include Computer Vision, Natural Language Processing

Parul Jaiswal, completed his degree from MIT College of Engineering, Pune and has been working in the domain of Business Analytics for the past year. His interests include Data Mining, High-Performance Computing, Big Data Analytics

.