BITS :: Call for Abstracts 2022 - Oral communication

Type	Oral communication
Session	Machine Learning in Bioinformatics
Title	Drug side-effect prediction with Graph Neural Networks
All Authors	Messori E(1), Bianchini M(1), Bongini P(1,2)

Affiliation

(1) Department of Information Engineering and Mathematics, University of Siena, Siena
(2) Department of Information Engineering, University of Florence, Firenze

Motivation

The study of Drug Side-Effects (DSEs) is crucial for the drug discovery process to guarantee that only safe medicines enter the market. Indeed, DSEs are the fourth leading cause of death in the United States, a fact that persists from year to year despite the increasing attention to the problem. Moreover, the significant clinical impact of DSEs requires targeted strategies, aimed at minimizing the risk of the onset of a drug-induced adverse event which, in addition to significantly undermining the patient's health, contributes to considerably increase healthcare costs - in Europe, it is estimated that the percentage of hospitalizations due to drug-taking side-effects is between 2.5% and 10.6%, for a total cost of 706 million euros per year. These hospitalizations accounted for 4% of the hospitals' bed capacity, a percentage that is significant especially in these pandemic years, during which various countries are facing periods of time where hospital capacities are critical due to the Covid-19 infection.

An approach to DSEs that also includes a pharmacogenetic analysis can provide valuable information to prevent or minimize the damage associated with the adverse effects associated with the administration of a specific drug and can shorten the entire drug discovery pipeline, preventing clinical trials of chemical compounds previously classified as toxic.

Methods

In this work, the side-effect prediction problem is taken under consideration in a machine learning framework, modelled as a supervised multi-class classification task that can be tackled through Graph Neural Networks. Indeed, graphs are a simple and natural way to represent complex information, in which some basic entities - drugs, genes, side-effects, in this specific case - are linked together by relationships of different nature. Moreover, also drug molecules are represented by graphs. In fact, considering the drug structure, so that the topology of the molecule is preserved, the loss of information is minimized.

To retrieve associations between drugs and side effects, we have used the public database SIDER, from which some pairs have been filtered out to avoid repetitions in the data and possible sources of noise for the training procedure. Targets necessary for supervised learning are implemented by binary vectors where each entry provides information about the correlation (or lack of) between adverse drug reaction and molecule.

The process leading to the construction of graphs representative of the drug's structure is based on two Python packages, RDKit and NetworkX: through them, we can use SMILES strings (i.e. strings employed to represent both two-dimensional and three-dimensional chemical structures) in order to obtain a GraphObject for each molecule, which is a structure specifically designed for this kind of problem and that includes the target vector associated to the molecule. Constructing the GraphObjects required an analysis of the dataset to retrieve information about atoms and bond types present in the whole dataset. Finally, the experiments were carried out with the original GNN implementation proposed in [1].

[1] F. Scarselli et al. "The graph neural network model", IEEE Transactions on Neural Networks, vol. 20(1), pp. 61–80, 2009.

Results

The performance of the model has been evaluated via a 5-fold cross-validation, using the binary cross entropy as the loss function, minimized by the Adam optimizer. The binary accuracy, the area under the receiver operating characteristic curve (AUC), and the area under the precision recall curve (AUPR) were used as metrics to evaluate the network's ability to correctly predict the true positives in the dataset (considering that false positive and false negative predictions have significantly different implications in this case).

The overall results are close to state of the art, with binary accuracy greater than 95%. Taking into account that the only data used in our experiments are the chemical structures of the drugs, these results are certainly promising. However, it is safe to assume that this method can be improved by exploiting further information in addition to chemical data and known associations of drug side-effects. In fact, since the compound identifiers in SIDER refer to the STITCH database, a reasonable choice is to enrich the current drug information with predicted protein-protein interactions that can be easily retrieved via STITCH. This procedure can produce important improvements due to the impact of protein networks on drug functionality, albeit at the expense of greater computational complexity, since the number of proteins involved could be high.

Info

M. Bianchini is affiliated with CINI-Infolife Lab

filename	-

Figure

-	
Availability	-
Corresponding Author	
Name, Surname	Monica, Bianchini
Email	monica.bianchini@unisi.it
Submitted on	29.04.2022