Minji Kim & Sukyung Seo

Model comparison on identifying Drug-Drug interactions

Introduction:

In the modern age with the rapid development and improvement in the medical area patients are frequently taking multiple different medicines at a time expecting its synergy due to either complexity of their disease or the medicine mechanism. It is efficient for sure if utilizing them appropriately through doctor's prescription, but people often autonomously mix and take them simply based on their symptom with leftovers from the last prescriptions or from drug stores (non-prescription needed drugs). Those cases can cause several problems and it sometimes even goes serious. In the worst case of drug interaction it can occur a death penalty but also overdose or reduce their individual expected effects, even entirely ineffective. Including those reasons, pharmaceutical companies are actively experimenting to discover drug-drug interaction cases to efficiently prescribe medicines to recover symptoms and avoid bad effects.

In this project, we are trying to compare several machine learning models to observe which is the best for distinguishing if a certain pair has bad effects or not. We aim to discover the most effective model that can be used to experiment drug-drug interactions.

Motivation:

Applying Machine learning to drug-drug interactions(DDIs) is a very recent trial that started about 2-3 years ago. Traditional methods of DDIs experiments have been the physical ways, such as injecting choice drug pairs to rats. It is obviously time and cost consuming, therefore many companies in the industry started to use machine learning methods to the organized data that they have which contains various information such as how it works, which part of our body it works. This is the main motivation that we have decided to experiment with the newest machine learning technique. Our idea of the project is to test several machine learning models to observe which model performs the best, and also would like to observe how much we can even more improve the performance through hyper parameter tuning. So that we can have a general insight of how much the hyper parameter tuning effects to improve the model efficiently.

Literature Review:

We have mainly focused on getting ideas regarding the approaches to how the previous works were done, because the bio-information is not very familiar to us. The overall process is using the existing DDIs labels, measuring the drugs similarity based on their properties such as drug substructure, target, enzyme, off-label side effect, etc, and applying machine learning classification. [Rohani et el. 2019] has applied heuristic similarity as a similarity measure and trained Neural Network model, and successfully showed that it is the models' predicted ability. [Mei et el. 2021] also confirmed that both cross validation and independent test showed the drug

target profiles-based machine learning framework outperforms existing data integration-based methods. Also [Feng et el. 2020] employed both graph convolution network (GCN) and deep neural network (DNN), and they reported those methods outperforms four other state-of-the-art methods they experimented.(The two of Vilar's methods, label propagation-based method, Zhang's methods)

From this research we instead of employing the method that was already tested, we intended to try a new approach. We chose Jaccard Similarity and Chemical similarity for our drug data similarity measure, and Random Forest, Neural Network, autoML as our machine learning model. We expect it will still be worth a learning point if the result is not that good enough since this is a new trial that is not publicly reported yet.

Datasets:

We collected diverse information about drugs and unified the drug identifiers into DrugBank ID, Genes to uniprot ID, and phenotypes to MeSH ID respectively.

Drug-drug interaction

To label whether two drugs have an interaction relationship, we collected drug-drug interaction data from Ryu et al [Ryu et al.,2018]. They compiled drug-drug interactions from DrugBank and preprocessed the data with 192,284 interactions and 1,710 drugs [Wishart et al., 2018]. In the original dataset, there are 86 different interaction types between two drugs. However, for simplicity we only denoted whether two drugs have interactions or not.

Indication and Side Effect

Drug indication is one of the important characteristics of drugs which refers to the usage of drugs for treating diseases. We collected drug indications from two different databases: Comparative Toxicogenomics Database [Davis et al., 2021], and SIDER [Kuhn et al., 2016]. The phenotypic identifiers from CTD and SIDER were mapped to UMLS ID. The compiled indication dataset contains 56,919 drug-indication data with 1,933 drugs and 4,830 indications. Also, 104, 996 drug side-effect pairs were collected in SIDER with 1,018 drugs and 5,365 side effects.

Chemical Structure

Since drugs with similar chemical structures show similar biological responses, considering chemical characteristics is crucial. Each drug can be denoted by a chemical structure that contains a graph with nodes (atoms) and edges (bonds). In this study, we collected drug chemical structures as a simplified molecular-input line-entry system (SMILES) with 2,509 drugs from DrugBank. In the structure feature, we computed the tanimoto similarity between two drugs using OpenBabel [O'Boyl et al., 2011].

Enzyme

Enzymes are crucial for the metabolism of drugs, which can affect the drug response. In this study, we collected enzyme data from DrugBank and compiled enzymes related to 1,643 drugs.

ATC

Anatomical Therapeutic Chemical (ATC) code is a unique code that represents how a drug works in an organ or system. We collected 16,341 drug-atc codes with 3,224 drugs and 1,093 ATC codes.

Target

When a drug is absorbed in a body, it encounters certain target genes that affect the therapeutic or non-therapeutic effects. In this study, we collected 20,704 drug-target pairs with 7,132 drugs and 4,709 targets from DrugBank.

Side effect Anatomical hierarchy

Wadhaw et al. manually classified the side effect's hierarchy according to where they affect the organ, subsystem, and system in the body [Wadhwa et al., 2018]. In this study, we mapped drug's side effects and the hierarchy information to represent.

Approaches:

We first gathered as much data as we could from several public drug data websites. Each of the websites are using different drugID to identify the drugs, so that we unified it as an ID from Drugbank. Also several preprocesses were needed such as synonyms of data name and also to construct a specific structure that we can work on. We unified them all on a table. The seven properties we will use are Chemical, Target, ATC, SideEffect, Indication, Anatomical and Enzyme.

Drug1	Drug2	Class	Chemical	Target	ATC	SideEffect	Indication	Anatomica	enzyme
DB00350	DB01080	0	0.15873	0	0	0.057522	0	0.292683	0
DB00350	DB01201	0	0.476744	0	0	0.057692	0	0.295455	0
DB00350	DB01114	0	0.333333	0	0	0.058824	0	0.392157	0
DB00350	DB00698	0	0.405063	0	0	0.053571	0.044444	0.433962	0
DB00350	DB01100	0	0.567164	0	0	0.087591	0	0.338235	0
DB00350	DB06589	1	0.4125	0	0	0.069264	0	0.325	0
DB00350	DB00647	0	0.211268	0	0	0.102273	0	0.339623	0
DB00350	DB00454	0	0.358209	0	0	0.056075	0.01087	0.368421	0
DB00350	DB08893	1	0.410959	0	0	0.086022	0	0.4	0
DB00350	DB01119	1	0.325	0	0.176471	0.112245	0.082353	0.490566	0
DB00350	DB01149	1	0.56	0	0	0.05	0	0.295455	0
DB00350	DB01018	1	0.25	0	0.2	0.057377	0.020408	0.359375	0
DB00350	DB01172	0	0.347222	0	0	0.051724	0	0.432432	0
DB00350	DB00590	1	0.581081	0	0.2	0.050209	0.024691	0.305882	0
DB00350	DB01001	1	0.263889	0	0	0.055556	0.1	0.311688	0
DB00350	DB01198	0	0.5	0	0	0.02381	0	0.317073	0

Figure 1. Part of our data table

Also with our assumption of similarities between two drugs affecting their existence of drug interaction, we applied two different similarities on each data point. Then training and gaining results with the models of Random Forest, Logistic Regression, Neural Network, and autoML.

Experiment Design:

Calculating Similarities

Jaccard Similarity was employed to measure similarities between two drugs using. By using Jaccard we expected the grouping effect of similar drugs. Jaccard coefficient takes the ratio of intersection over union as follows equation (1).

$$J(Da,\ Db) \ = \frac{|\ Pa \cap Pb\ |}{|\ Pa \cup Pb\ |} \quad (1)$$

Where D_a, D_b means drug *a* and drug *b*, respectively. P_a means set of a property of drug *a*, P_b means a set of a property of drug *b*. For example, if P_a is {S1, S2, S3, S4}, and P_b is {S2,S3,S5} then $P_a \cap P_b$ is {S2, S3} and $P_a \cup P_b$ is {S1, S2, S3, S4, S5}.

For a chemical property, we used tanimoto similarity which is a standard way of calculating similarity of chemical.

Building Tables

In this study, we used drugs that are only existed in seven features and drug-drug interaction data. Using these drugs, we made all possible drug-drug pairs and if a pair is in drug-drug interaction data, we assigned class 1 ‘already known interaction’ and class 0 if it is not in the data. Then, we assigned similarity values for each feature.

Applying Machine Learning and Deep Learning Algorithm

In this section, we sampled 70,000 positive cases, and 70,000 negative cases to build balanced table for machine learning. Then we split the train and test as 8:2 ratio and applied four different algorithms: random forest, logistic regression, neural network, and autoML. We reported the model’s accuracy, precision, recall, Area Under the Curve (AUC), and F1 score to evaluate the model.

Prediction of candidate drug-drug interactions

To gain candidate drug-drug interactions, we used 4,275 samples that were not used in the train and test dataset. We chose the autoML model to predict because it shows the best AUC among four models.

Evaluation:

We first measured AUC/Accuracy/Precision/Recall/F1-score to identify the best classifier.

	AUC	Accuracy	Precision	Recall	F1
Logistic Regression	0.7104	0.6595	0.6816	0.5896	0.6323
Random Forest	0.7269	0.6704	0.6681	0.6676	0.6679
Neural Network	0.7299	0.6772	0.6666	0.6997	0.6828
AutoML	0.7436	0.6852	0.6783	0.6960	0.6870

Figure 2. Results from each models

As in the Figure 2, AutoML indicated the best result and we assume it is because the AutoML choosed mainly tree-based models. For AutoML, XGBClassifier, GradientBoosting, and ExtraTreesClassifers were selected and they performed better than others.

After identifying the best model, we process predicting DDIs candidates only with the model for the best prediction.

DrugBank ID	DrugBank ID	Predicted Class	Negative probability	Positive probability
DB00705	DB00949	1	0.05821198	0.94178802
DB00528	DB00938	1	0.075738987	0.924261013
DB00656	DB06700	1	0.083838245	0.916161755
DB00763	DB01173	1	0.089162967	0.910837033
DB00502	DB00850	1	0.0917965	0.9082035
DB06616	DB08875	1	0.092017367	0.907982633
DB00203	DB00334	1	0.095774058	0.904225942
DB00357	DB06209	1	0.097508917	0.902491083

Figure 3. Sample of prediction results

For the Class=1 it implies that DDIs exists and for Class=0 non-DDIs. Based on the result we chose 10 of drug pairs as candidates to confirm that it is correctly predicted. The candidates were selected according to the highest probability.

Drug a	Drug b	Probability
Delavirdine	Felbamate	0.94178802
Lercanidipine	Salmeterol	0.924261013
Trazodone	Desvenlafaxine	0.916161755
Methimazole	Orphenadrine	0.910837033
Haloperidol	Perphenazine	0.9082035
Bosutinib	Cabozantinib	0.907982633
Sildenafil	Olanzapine	0.904225942
Aminoglutethimide	Prasugrel	0.902491083
Hydrocortisone	Rifapentine	0.900175784
Rifabutin	Sitaxentan	0.899995581

Figure 4. List of 10 candidates of Drug-Drug interactions

For those cases we tried to confirm that the four cases have interactions through a web drug interaction checker, to access the latest DDIs reported. As a result, we found there are 4 DDIs cases out of the 10 drug pairs. If a patient takes Trazodone and Desvenlafaxine together, this can increase the risk of a serious condition called serotonin syndrome. Haloperidol and Perphenazine or Bosutinib and Cabozantinib together can affect the risk increase of irregular heart rhythm that can be related to life-threatening situations. Rifapentine and Hydrocortisone are taken together might affect the efficacy of Hydrocortisone. For the rest of drug pairs the web checker either denotes “0 interaction found” or even was not able to search for some drugs which means the site doesn’t have any information at all reported about the drug. Since the web checker does not clearly show that if it means no interaction between them or no reported experiment result, we can not certainly check for them, at least yet.

Conclusion:

Data is always the key for every machine learning project. This factor has been the main challenge through this entire project, because of the feature of this area is data starvation. In reality numerous drugs exist, but access to their data is very limited. The companies that developed the drugs rarely open its information because of the security, since it is directly connected to their earning. Even some open resources have missing certain information regarding drug properties such as side effects, enzymes, etc., if they are not investigated. These limited open data sources cause difficulty to represent diverse characteristics of drugs during progress.

Despite those difficulties, our project is still meaningful for the aspect that it can be used as a precaution for the further experiment which does not have DDIs result yet, through experiment. Thus, this study can be a guideline to clinicians to detect unexpected DDIs.

For further works, trials to find different drug-related data by investigating other biological literatures can help to overcome these difficulties. Thus, further experiments with accumulated data will definitely be helpful to improve the performance with this algorithm to many drug-drug interactions. And this also implie our project can be very meaningful if a certain company in the industry is available to use a variety of data.

References:

[Rohani et el., 2019] Rohani, Narjes, and Changiz Eslahchi. "Drug-drug interaction predicting by neural network using integrated similarity." *Scientific reports* 9.1 (2019): 1-11.

[Mei et el., 2021] Mei, Suyu, and Kun Zhang. "A Machine Learning Framework for Predicting Drug-drug Interactions." (2021).

[Feng et el., 2020] Feng, Yue-Hua, Shao-Wu Zhang, and Jian-Yu Shi. "DPDDI: a deep predictor for drug-drug interactions." *BMC bioinformatics* 21.1 (2020): 1-15.

[Ryu et al.,2018] Ryu, Jae Yong, Hyun Uk Kim, and Sang Yup Lee. "Deep learning improves prediction of drug–drug and drug–food interactions." *Proceedings of the National Academy of Sciences* 115.18 (2018): E4304-E4311.

[Wishart et al., 2018] Wishart, David S., et al. "DrugBank 5.0: a major update to the DrugBank database for 2018." *Nucleic acids research* 46.D1 (2018): D1074-D1082.

[Davis et al., 2021] Davis, Allan Peter, et al. "Comparative toxicogenomics database (CTD): update 2021." *Nucleic acids research* 49.D1 (2021): D1138-D1143.

[Kuhn et al., 2016] Kuhn, Michael, et al. "The SIDER database of drugs and side effects." *Nucleic acids research* 44.D1 (2016): D1075-D1079.

[O'Boyl et al., 2011] O'Boyle, Noel M., et al. "Open Babel: An open chemical toolbox." *Journal of cheminformatics* 3.1 (2011): 1-14.

[Wadhwa et al., 2018] Wadhwa, Somin, et al. "A hierarchical anatomical classification schema for prediction of phenotypic side effects." *Plos one* 13.3 (2018): e0193959.