Predicting Drug-Drug Interactions

Gaurav Belani, George Koussa, BP Rimal, Onat Ure

Abstract

We developed several models to predict drug-drug interactions. Our link prediction model (AUCROC: 0.97) is powerful, but requires information on known interactions, which may not alway be available. Our CatBoost (AUCROC: 0.86) model trades performance for convenience, requiring a fewer set of more accessible features. By stacking these models, we achieve link prediction performance and retain CatBoost convenience.

Background

Interactions between drugs can be interpreted as a graph, where the drugs are nodes and the edges are interactions. Link prediction predicts missing interactions from known interactions by using the topology of the interaction graph to identify likely missing edges. Given that it relies only on the interaction graph, it can not find interactions between unconnected components. A classifier trained on separate drug features may be able to predict drug interactions well enough to provide a link predictor enough information to predict interactions between drugs that have few known interactions. For this reason, we develop a link predictor, a classifier, and a stacking model.

Dataset

We extracted all our data from the DrugBank database. For convenience we converted the XML database to JSON. We then identified 25 features worth parsing, which include molecular weight, atc-codes, melting point, etc. The database contains over 15000 drugs, but due to computational constraints we decided to only use 2500. We picked those drugs which had the least missing features.

We then parsed the dataset for drug interactions. The data was initially in the form of a list of names per drug. From this data, we produced the adjacency matrix used by the link predictor.

	name	state	level4	level3	level2	level1	Hydrophobicity	Boiling Point
0	Lepirudin	solid	B01AE	B01A	B01	B	NaN	NaN
1	Cetuximab	liquid	L01FE	L01F	L01	L	-0.413	NaN
2	Dornase alfa	liquid	R05CB	R05C	R05	R	-0.083	NaN
3	Denileukin diftitox	liquid	L01XX	L01X	L01	L	-0.301	NaN
4	Etanercept	liquid	L04AB	L04A	L04	L	-0.529	NaN

Drug Features Dataframe

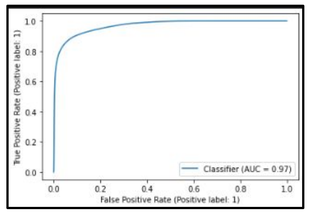
name	Bivalirudin	Leuprolide	Goserelin	Gramicidin D	Desmopressin
Bivalirudin	0	0	0	0	0
Leuprolide	0	0	1	0	1
Goserelin	0	1	0	0	1
Gramicidin D	0	0	0	0	0
Desmopressin	0	1	1	0	0

Adjacency Matrix

Link Prediction Approach

Our link predictor (powered by scikit-network) gets the n nearest neighbors for each node and produces a similarity score based both on cosine similarity and graph distance. This score can be interpreted as the probability of interaction.

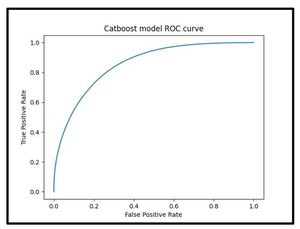
To test our link predictor, we replaced 20% of the edges in our adjacency matrix with 0 and ran the link predictor on the entire graph. Since we replaced edges at random, individual drugs did not lose enough information to hurt the performance of the model. This model achieved an AUCROC score of 0.97.



Classification Approach

Given its ability to elegantly handle both categorical and numerical features, we selected CatBoost for our classifier. The inputs to this CatBoost are two sets of our 25 drug features. The output is a an integer indicating interaction (1) or non-interaction (0).

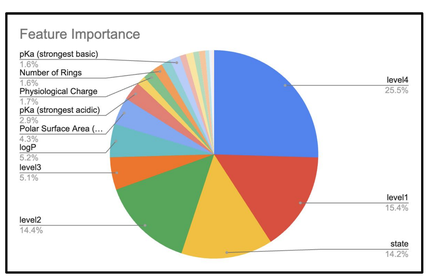
Our training data consisted of all possible pairing of the 2500 drugs we had features for. We withheld 20% of the training data for testing. This model achieved an AUCROC score of 0.86.



Data Visualization

Below is the feature importance we retrieved from CatBoost. Important to note is that the most important features were the categorical ones. In fact, the atc-codes alone (level4, level3,...) contribute more than the rest of features combined.

These categorical features, unlike drug interactions, are easy to obtain. This further suggests the use of classifiers to reinforce link prediction.



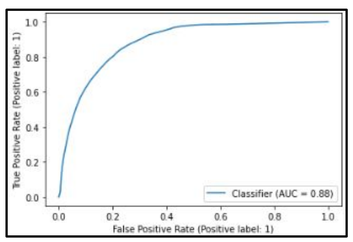
Classifier and Link Predictor

To improve on our CatBoost model, we fed its output into our link predictor. When withholding data from adjacency matrix, instead of simply writing zero, we put the prediction of our CatBoost model. This allows us to use link prediction with drugs that have no known interactions, provided we have the 25 features required by our CatBoost model.

Once again, we replaced 20% of the edges at random with our CatBoost predictions. Since most drugs still had many interactions, this was indistinguishable from simply removing edges.

However, when replacing all edges from a set of drugs, this method achieves an AUCROC score of 0.88 on that set, whereas link prediction alone is unable to predict any interactions.

While this is only a slight improvement over CatBoost alone, we expect this margin to increase as the number of known interactions (while still small) increases.



Conclusion

Given our computational and time constraints, we were not able to take full advantage of the massive dataset provided to us by DrugBank. Certainly, if we used the entire dataset, the scores of each three model should improve. In terms of relative performance, it is clear that our stacking model generally outperforms CatBoost alone and addresses some of the woes of link prediction.

Additionally, the feature importance produced by CatBoost suggests that if we use the entire dataset, we can achieve acceptable performance using only a couple features, namely the atc-codes. This would further improve the usability of our stacking model.