Cancer Drug Efficacy Prediction (Erdös Institute) – Executive Summary

Team: Zhenyu Wu, Mingming Abundo, Palak Arora, Kwok Wai Ma

Github: https://github.com/PalArora94/Drug-Effectivity-Prediction

1. Overview

Drug resistance accounts for nearly 90% of failures in cancer treatment. We build and compare between various ML models to improve the prediction of drug efficacy on different tumors.

- Stakeholders: Oncologists, biologists, pharmaceutical manufacturers
- KPI: Root mean squared error (RMSE) between the true and predicted drug efficacies

2. Data Collection and Explanation

Different cells have different RNA sequences (each sequence can be viewed as a collection of weights among different genes). These weights are the defining features of a tumor. The IC50 of a drug is the concentration required to inhibit a specific biological or biochemical function by 50%. Hence, the drug efficacy on different tumors could be inferred based on the IC50 data.

- RNA-sequence data of cell lines from CCLE database: online data source
- Drug IC50 data of cell lines from GDSC: online data source

3. Approaches

Since there are more than 17000 features (genes) in the data set but only 805 instances (tumors), we reduced the number of features before the training process based on the following two approaches:

- Method 1: Keeping the 50 genes with the largest variances among all 805 tumors.
- Method 2: Keeping the 50 genes with the highest SHAP (Shapley Additive Explanations) importance from XG Boost. The approach requires a training on the data set first and evaluates the importance of each feature towards the prediction.

An $80-20$ training split is used in training our ML models. We used Scikit-Learn to build linear regression, supporting vector regression (SVR) with the radial basis function (RBF) kernel, and XG Boost models. A simple three-layer neural network (Three-layer NN) is built using TensorFlow. Hyperparameter tuning is performed using the GridSearchCV package.

4. Results

We only focus on the drug "Docetaxel" in the IC50 database, which has the largest variance in its efficacy among all 805 instances in the data set. We obtain the following results:

- For the simple linear regression model, we observe heteroscedasticity from the residual plot. Hence, linear regression model is not a suitable model to fit the data set.
- By comparing the results between the two different feature selection approaches, we conclude that the SHAP importance method leads to a better prediction in most of the ML models. This is somehow expected as the variance based selection approach ignores the actual correlation between the features and the targets that we need to predict.
- The XG Boost model with the SHAP-importance feature selection gives the lowest RMSE in the result

ML Model \ Approach	Variance selection	SHAP importance
Benchmark	0.192	0.192
Linear Reg.	0.175	0.170
SVR (rbf kernel)	0.168	0.162
XG Boost	0.167	0.158
Three-layer NN	0.171	0.176

Table 1: The root mean squared errors (RMSE) in predicting the efficacy for Docetaxel by different ML models. Here, the RMSE for the test set is reported. The benchmark is the corresponding standard deviation of the drug efficacy among all 805 tumor cells.

5. Future Directions

There are several future directions that worth to explore:

- Collecting more data from clinical researches to develop better ML models
- Seeking professional advice on feature selection to improve ML performance
- Classifying tumors in different groups (such as lung cancer) before building ML models
- Exploring the possible generalization of the ML models to other possible drugs
- Develop user friendly interface for practical and clinical applications