Machine Learning Methods on Drug Sensitivity Prediction with One Drug Data

Xiao Weiqi, Zhou Chengyuan, Zhu Chang
Math6380: A Mathematical Introduction to Data Analysis Spring 2017

Introduction

Afatinib works by blocking the function of two closely related proteins, the epidermal growth factor receptor (EGFR) and the human epidermal growth receptor2(HER2). Many cancer cells have unusually large amounts of EGFR and/or HER2 on their surface. These signals lead to excessive cell growth.

Our work reports the relation-ship between binary features as gene mutation status and cells' sensitivity towards Afatinib. Based on these connections, we expect to predict the drug sensitivity data using the machine learning method such as Gradient Boosting Regression and Random Forest Regression.

One Drug Data

To figure out the cancer cell for drug sensitivity to Afatinib, this dataset has contained 642 cancer cell line samples in response to the Afatinib. There are 60 binary features for each cancer cell line.The measurement of sensitivity is IC50.

IC50 values are the dosage amounts of the drug such that after a period (say 24 hours) the experimental cancer cell lines are killed by 50%. Therefore, the lower is the IC50, the more sensitive is the cancer line to the drug.

Above all, the output is drug sensitivity, which is measured by IC50. Inputs are 60 binary features based on specific cancer cell line.

Methodology

➢ Random Forest Regression

Why?

To lower variance of the prediction results from decision trees,
we use bagging method to improve prediction accuracy at the expense of interpretability.

How?

Procedure of random forest:

- Data $(y_i, x_i), i = 1,2, \dots, n$; and a learning method $\hat{f}$.
- Draw a boostrap sample from the data, and compute a $\widehat{f_1^*}$ based on this set of bootstrap sample.
- Draw another boostrap sample from the data, and compute a $\widehat{f_2^*}$ based on this set of bootstrap sample.
- ....
- Repeat M times, obtain $\widehat{f_1^*}, \dots, \widehat{f_M^*}$.
- Produce the learning method with bagging as $\frac{1}{M}\sum_{j=1}^{M}\widehat{f_j^*}$.

➢ Gradient Boosting Regression

Why?

It is different from Bagging which can only improve corresponding to high variance. However, Boosting can make a difference in both controlling bias and variance with higher efficiency.

How?

Procedure of boosting:

- Set $f(x) = 0$ and $r_i = y_i$ for all i in the training set.
- For, $b = 1,2, \dots, B$, repeat:

① Fit a tree with d splits (d + 1 terminal nodes) to the training data (x_i, r_i).

② Update $\hat{f}$ by adding in a shrunken version of the new tree: $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda\widehat{f_b}(x)$.

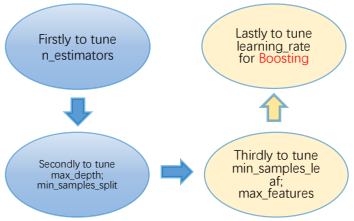
③ Update the residuals, $r_i \leftarrow r_i - \lambda\widehat{f_b}(x) = y_i - \hat{f}(x_i)$.

- Output the boosted model $\hat{f}$. In fact, $\hat{f}(x) = \sum_{b=1}^{B}\lambda\widehat{f_b}(x)$.

➢ Parameters Tuning

In order to improve the prediction accuracy we try to use grid searching method to test various of different values of parameters in the models. And then use mean of cross validation scores to choose the best parameters – the higher the scores the best the parameters.

Workflow For Drug Sensitivity Prediction



Results

➢ Random Forest Regression

n_estimators	Max_depth	Min_samples_split	Min_samples_leaf	Max_features
120	15	3	3	19

Table 1. Best parameters of random forest regression modeling.

A B

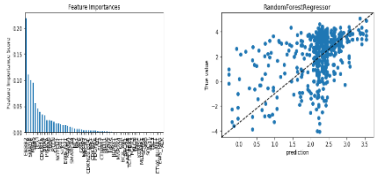


Figure 1. A. Features importance. B. Scatter plot of prediction and true values of response.

Test error of Kaggle in-class competition Drug Sensitivity 2 is **3.19261 measured by MSE (mean squared error),** which is relatively high accuracy compared to the **baseline** whose MSE is **3.56397.**

➢ Gradient Boosting Regression

n_estimators	Max_depth	Min_samples_split	Min_samples_leaf	Max_features	Subsample	Learning_rate
30	4	7	5	19	0.85	0.09

Table 2. Best parameters of Gradient boosting regression modeling.

A B

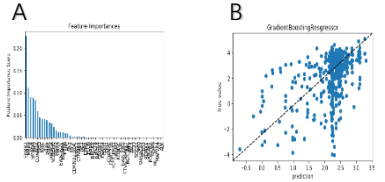


Figure 2. A. Features importance. B. Scatter plot of prediction and true values of response.

Test error of Kaggle in-class competition Drug Sensitivity 2 is **3.24726 measured by MSE (mean squared error),** which is relatively high accuracy compared to the **baseline** whose MSE is **3.56397.**

Conclusion And Drawback

- In this prediction, random forest method is more accurate than gradient boosting.
- Both two methods lead to quite good results which remind us of dealing these kinds of data with machine learning.
- However, the results from trees are not continuous which may lead a less accurate result than other regression models

References

- Scikit-learn.org (tutorials of different functions of machine learning and model selection functions.