Bioinformatics, 2022, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: 19 December 2022
Manuscript Category

Subject Section

Predicting drug response using deep learning

Tamanna Ananna[1,*],

Associate Editor:

Received on 12-19-2022; revised on 12-19-2022; accepted on 12-19-2022

[1]Department of Computer Science, Columbia University
*ta2642@columbia.edu

Abstract

Motivation: The hyperparameter tuning process can have significant impact on the model performance. The model reported by Chalwa et al utilizes gene expression data to predict drug response, achieving highest correlation value reported so far. But this model erroneously overfits on training data due to the hyperparameter tuning process.

Results: I corrected the hyperparameter tuning process and achieved the highest correlation value in drug response using solely gene expression data, higher than the value they achieved by utilizing pathway specific information.

Availability: https://github.com/tamanna-a/CancerDrugResponse
Contact: ta2642@columbia.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Cancer is a highly complex disease. Despite the tremendous increase in number of treatments available and improvement in prognosis, how each patient responds to cancer treatment is often unpredictable. Additionally, drug resistance is also a factor that can lead to reduced patient survival. Thus early inference of drug response can provide critical information.

In recent years, increasingly larger amounts of high-throughput screening data has been made available. Notable among them include Cancer Cell Line Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC). These include data from more than 100 cell lines and hundreds of anticancer drugs. Availability of such data has led to the developments of many personalized predictive models models in oncology research. With recent advances in machine learning and deep learning, the models increasingly adopt deep learning based architectures.

Chalwa et al., use cell line gene expression data to predict treatment response, creating a model they term as "Precily." In addition to using solely gene expression data, they utilize pathway specific information. The reasoning is that gene expressions are not independent of each other but are instead part of pathways and most targeted therapies work through pathways. With the pathway pre-processing, they report achieving the highest correlation value, higher than when using only gene expression data.

They build DNN architecture consisting of 2-6 hidden layers, the exact number of which is determined as a hyperparameter. Hyperparameters are parameters of the model that are not learned during training, such as the number of neurons, the number of layers, the activation functions, and etc. The purpose of hyperparameter tuning is to choose values that lead to a more generalizable model. The authors claim to use 5-fold cross validation on training data to find the best hyper-parameters based on validation loss.

Upon closer inspection of the training loop and cross-validation procedure, I determined that it includes some issues that invalidates the purpose of hyperparameter tuning in the first place. Figure 1 compares the common approach (1a) to the approach taken in the paper (1b).

5-fold cross validation is a common approach where the training set is split into five folds, where 4 folds are the training set and the last fold is the validation set. The RMSE for each fold is averaged to get the performance of that hyperparameter. The hyperparameter combination with the best average RMSE on each fold is chosen.

First, through five-fold cross-validation, instead of creating a more generalizable model, Chalwa et al find hyperparameters and models that are optimized for *that* fold of data. By creating a model optimized for that fold, they end up overfitting to that fold of training data. This leads to models that are even less generalizable than they would have been without the cross-validation.

Second, due to this cross-validation approach, they end up mimicking an ensemble method. After the 5-fold cross-validation, instead of ending up with one model, they output 5 models tuned to that fold of data. Then they fit the whole training data to each of the 5 models. When it comes to predictions on the test set, each of the 5 models do a prediction. The final value comes from averaging the prediction of the 5 models. This mimics the ensemble method where several models are combined to get an optimal prediction. This can reduce the overfitting from the faulty cross-validation practice. If proper cross validation practice is followed, the ensemble step would not be necessary.

In this paper, I build my own model that corrects the cross-validation procedure. Through this, I achieve the highest correlation value in drug response using solely gene expression data, higher than the value they achieved by utilizing pathway specific information.

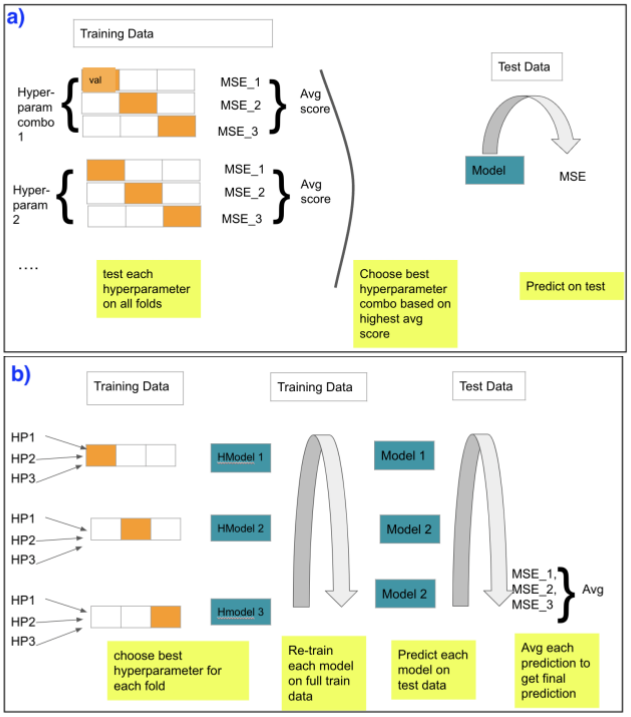

Figure 1: Hyperparameter tuning in a) correct approach b) approach used by paper

2 Methods

2.1 Data

The datasets used in the experiments were from the github repository of the project where the pre-processing was done by Precily group. This includes CCLE and GDSC data. For predicting drug response, they utilized publicly available TPM (transcript per million) normalized RNA-seq gene expression profiles of 1019 Cancer Cell Line Encyclopedia (CCLE) cell lines quantified using the RSEM (RNA-Seq by Expectation-Maximization) software. The corresponding drug response information for the cell lines was sourced from the GDSC2 dataset of the Genomics of Drug Sensitivity in Cancer (GDSC) database9. In the GDSC2 dataset, some drug-cell line pairs have multiple LN IC50 measurements. In such cases, they averaged the LN IC50 values.

They also added pathway activity scores using the Gene Set Variation Analysis (GSVA) R software package to compute GSVA scores based on the log2(TPM + 1) gene expression matrix for selected gene sets from Molecular Signatures Database (MSigDB). The drug response was measured by half-maximal inhibitory concentration (LN IC50). 550 CCLE cell lines overlapped with the GDSC2 dataset. The expression matrix included 54,301 genes and 550 cell lines.

2.2 Model

In replication experiments, I downloaded the models from the github repository using keras model loading functions. In the paper, they build the model in python and test the model predictions in R. To streamline, I adapted the prediction function and the correlation calculation in python.

The genes/pathways and drug features were the independent variables, and LN IC50 was the dependent variable that we tried to predict using the models. In my model, only the genes information was used.

I built my own deep neural network models using open-source software keras. The models included 2 to 4 layers with dropout layers in between. The first two layers were fixed with 600 and 512 units respectively. The number of epochs was set to 30 with a batch size of 128. I used ADAM optimizer and optimized the Mean Squared Error (MSE) as the loss function. The hyperparameters tested included: number of layers, learning rate (0.001 vs 0.0001), activation function (ReLu vs leaky ReLu). The performance was reported on independent test sets.

3 Results

I built the model and report its performance as correlation at the level of drugs (Figure 2) and globally (Figure 3).

3.1 Correlation at the level of drugs

At the level of drug, the correlation was calculated as the mean of the pearson correlation coefficient between predicted and observed for each drug (n = 173).

Utilizing only gene expression information, I achieved mean Pearson correlation of 0.632 (Figure 2, blue bar).. This is the highest correlation value achieved, higher than any of the reported values in the paper, including Precily Pathways that adds pathway information. The model that led to the highest correlation included two layers with 600 and 512 units respectively, then 10% dropout, dense layer with 256 units, 10% dropout, dense layer with 128 units, 10% dropout, and final dense prediction layer. It used ReLu as the activation function and had learning rate of 0.001.

Comparatively, Chawla et al report the results of Precily based predictions of drug response for a)using gene expression (Precily_genes) and b)using pathway score (Precily_pathway). Chawla et al achieved correlation of 0.49 in their "Precily Pathways" model. I attempted to replicate this result by downloading the models provided with the code of the paper. For "Precily Pathways," I found the downloaded models to have a correlation of 0.474, which is within the error range reported in the paper.

In the "Precily genes" model, Chawla et al reported correlation of 0.352. Surprisingly, I was unable to replicate the value reported for "Precily genes" when testing using their model, achieving a much lower value of 0.231 instead of the 0.352 reported. This indicated that there appeared to be some issues with the "Precily_genes" models in the code repository.

By utilizing only gene expression data, I achieved the highest correlation value. The results indicate that the hyperparameter optimization approach utilized by Chawla et al in the Precily models did in fact lead to overfitting and not the best generalizable model.

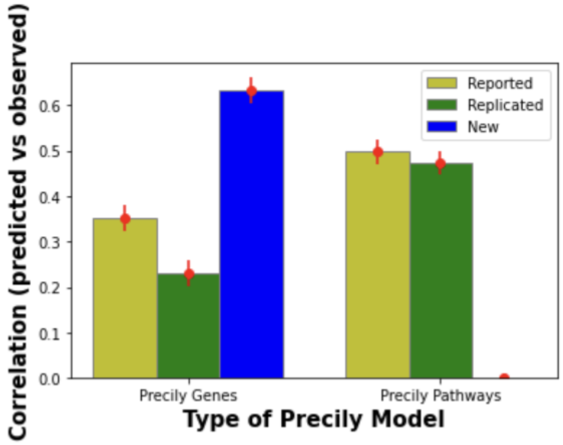

Figure 2: Highest correlation achieved in drug response from the model built in the paper (new in blue)

3.2 Comparable values seen in global response prediction

Using the newly built model that had the highest correlation at the drug level, instead of looking at the correlation in predictions for each drug, we plotted the predicted drug response versus observed drug response globally on the CCLE/GDSC data. We achieved Pearson correlation of 0.88. Chawla et al report Pearson correlation of 0.88, but this was achieved using the pathway model. In my model, R^2 is 0.764 instead of the 0.77 achieved by Chawla et al. Thus we see that similar values were achieved in the model built in this paper as the Precily pathways model. Thus by correcting the hyperparameter optimization process, I achieved comparable global response prediction using only genes expression data. future work, it would be interesting to see if correcting the hyperparameter optimization process of the pathway model leads to an even better model than the model built here.

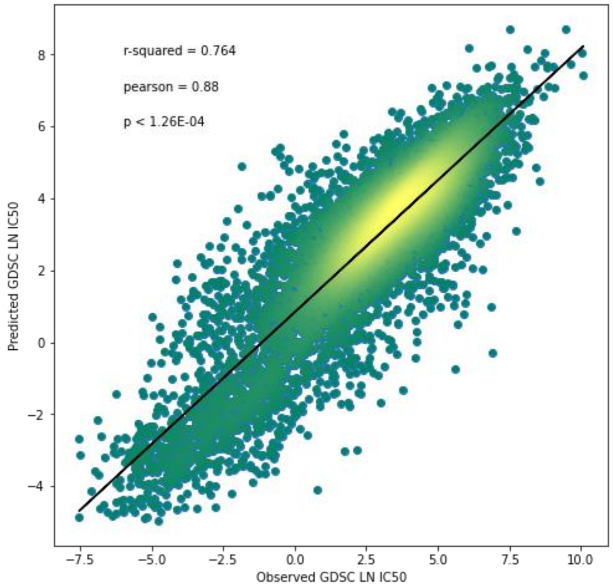

Figure 3: Scatterplot demonstrating model efficiency, comparing observed vs predicted LN IC50 measured by Pearson correlation (p) = 0.88, coefficient of determination (R^2) = 0.76. P-value calculated using two-sided t-test

Code availability

All source codes are available at https://github.com/tamanna-a/CancerDrugResponse

References

1. Chawla, S. et all. (2022) Gene expression based inference of cancer drug sensitivity. *Nature communications* **13**, 5680.
2. Barretina, J. et all. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603–607.
3. Yang, W. et al. (2013) Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucl. Acids Res*. 41, D955–D961 .