Prediction for the Amount of Yields Produced by Bacteria with Syngas

Ruina Chang, SeokHwan Song

Iowa State University

Abstract

Through machine learning models, the amount of five different yields, produced by bacteria with CO_2, CO and H_2 will be able to be predicted. Three different types of models, SVR, Neural networks, and Random forests, are applied. The models are analysed through some experiments and precision and some findings are discussed.

1. Introduction

In the Biochemistry field, one of the important topics is related to syngas. Because syngas contains CO_2 and CO which are one of the reasons for global warming and very common in the air. Therefore, many methods to use the gas have been developed and syngas fermentation is one of efficient ways to conduct the task. The bacteria combined with the harmful gases and H_2 even produce frequently utilized chemicals in diverse ways [1]. The main five different organics are acetate, biomass, butanol, butyrate and ethanol. However, it is difficult to predict the process in the Biochemistry way [2]. Therefore, in this paper, machine learning models give a different point of view to predict the amount of produced yields to help the Biochemistry technology.

2. Methodology

2.1. Data Preprocess

The applied dataset contains 12 different variables including case ID and trial ID. There are four different variables which consist of elements which represent a case, therefore, each trial has different values of elements. Five different variables exist for yields, they are not named which variable is for which chemical but we can figure out the amounts of yields with the variables. Therefore, the variables related to cases and time will be the input vector while the variables about five different yields will be the output vector.

After the dataset is imported and allocated to input and output, StandardScalar is used to normalize the raw data since normalization can make the classifier work better. At the end, splitting data to training and testing sets is conducted.

2.2. Models and Hyperparameters

Three different models were used to find the best model with the best performance. Support Vector Regressors (SVR), neural networks, and random forests are the models and these models are trained and find the best hyperparameters for each model by cross validations. In advance, 10-fold cross validation and leave-one-out cross validation are compared and 10-fold cross validation is chosen since 10-fold cross validations gives better estimation of the hyperparameters and is extremely time-consuming. Moreover, because only one instance is used for leave-one-out cross validation, while R^2 requires at least.

For each model, MultiOutputRegressor is selected for estimators and a hyperparameter variable is made to provide hyperparameter choices for tuning. There are three different options of hyperparameters for each model. Estimator C and $gamma$ have several options for SVR and maximum depth of trees and minimal impurity are applied for random forests. Moreover, the size of the hidden layer, the regularization weight, and maximum number of iterations are used. We insert some possible values for each hyperparameter and through the tuning process they return the best hyperparameters.

Secondly, GridSearchCV from sklearn is provided to build models with the estimators and parameter previously made are used and 10-fold cross validation with R^2 scoring. After then, the models will be fitted with the normalized input and output data.

3. Analysis

3.1. Scores

RMSE, Spearman's correlation coefficient, Person's correlation coefficient, and R^2 are used to evaluate the models' performance with training and testing dataset.

3.2. Tables

Firstly, the values of RMSE of three different models with train set are compared, and random forest shows the lowest value, 1.843. When we check Spearman's correlation coefficient, we can say with the random forest model it has 98.3% correlation between predicted values and ground values in the training set, and it is the highest coefficient. Pearson's correlation coefficients also show that the random forest has the strongest positive correlation of the training set. Moreover, R^2 represents the random forest model can predict 94.8% of the variance of value y of the training set. Therefore, we can say the random forest model performs the best with the training dataset among the three models.

Table 1: Four different scores of train with models

	RMSE mean score (+/-sd)	**Spearman's correlation coefficient** mean score (+/-sd)	**Pearson's correlation coefficient** mean score (+/-sd)	R^2 mean score (+/-sd)
SVR	2.294(+/-0.088)	0.970(+/-0.003)	0.977(+/-0.002)	0.916(+/-0.007)
Random Forest	1.843(+/-0.072)	0.983(+/-0.002)	0.985(+/-0.001)	0.948(+/-0.006)
Neural Network	3.753(+/-0.130)	0.928(+/-0.004)	0.934(+/-0.005)	0.704(+/-0.029)

Secondly, the evaluation of the models with the testing dataset is conducted. The values of RMSE of three different models with train set are compared, and SVR shows the lowest value, 3.60. Therefore, the SVR model works better with the testing dataset than the random forest model. When you check Spearman's correlation coefficient, we can say with the SVR model it has 93.0% correlation between predicted values and ground values in the training set, and the random forest has slightly small values. Pearson's correlation coefficients also show that the SVR has the strongest positive correlation of the training set. Moreover, R^2 represents the SVR model can predict 66.4% of the variance of value y of the training set. While, the values of the random forest are slightly smaller than SVR.

Table 2: Four different scores of test with models

	RMSE mean score (+/-sd)	**Spearman's correlation coefficient** mean score (+/-sd)	**Pearson's correlation coefficient** mean score (+/-sd)	R^2 mean score (+/-sd)
SVR	3.600(+/-0.895)	0.930(+/-0.016)	0.941(+/-0.025)	0.664(+/-0.188)
Random Forest	3.833(+/-1.358)	0.929(+/-0.032)	0.933(+/-0.046)	0.580 (+/-0.322)
Neural Network	4.559(+/-0.673)	0.898(+/-0.020)	0.910(+/-0.016)	0.329 (+/-0.307)

4. Discussion

Through analysis, we can conclude that the random forest model shows very good performance with the training set and the SVR model performs the best with the testing set among the three models. Moreover, with the scores of predictions we can say that the SVR model performs well enough to predict the amounts of yields. Therefore, it could be great to use the models to predict the most efficient combination of time and input variables. It will be a great chance to reduce the syngas and produce yields which are helpful to our life.

Acknowledgments

SeokHwan Song and Ruina Chang cooperated as a group.

References

[1] Ni Wan, Ashik Sathish, Le You, Yinjie J. Tang, and Zhiyou Wen. 2017. Deciphering Clostridium metabolism and its responses to bioreactor mass transfer during syngas fermentation. Retrieved from https://www.nature.com/articles/s41598-017-10312-2?WT.feed_name=subjects_systems-biology

[2] Project idea 1: Organic chemistry. Retrieved from https://github.com/forrestbao/MLClass/blob/master/projects/organic_chemistry.md

Appendix

Location of our Code:

https://colab.research.google.com/drive/1kyaUeFVzRa4yS5gt9AzfVRh2H1iB3Enb?usp=sharing

The source of our data:

https://github.com/forrestbao/MLClass/blob/master/projects/organic_chemistry_data.csv