Classification of microorganisms

By: Jing Lee

Abstract

In this project, I am classifying microorganisms from a dataset containing features of microscope images. The target is 1 if it is 'Ulothrix' and 0 for 'Other'. Machine learning classification models were used in this project and the best model was selected through validation. The dataset was split into a training and testing set for proper evaluations. Models were scored using cross-validation. Testing set that was unseen was used to evaluate the best model at the end of the model selection. F1 score was used to select the best model and to optimize the model's parameters since both true positives and true negatives are of importance to the goal of the project. Random forests was the best model with an accuracy of 90% on the train set. The model suffered from some overfitting, and the test accuracy was 80% with better prediction on the 0 class (Other).

I. Design

My client will be microscope companies. The automated classification feature can be used alongside the microscope software to speed up overall image analyses while reducing bias and error.

II. Data

The dataset was obtained from Kaggle. It contains 30,000+ datapoints of microscope images analysis (numerical data) with 24 features. The target was Ulothrix (1) or Other (0)

III. Algorithms

Feature engineering

- Target labels were converted to numerical values of 1 and 0
- Duplicated entries across all columns were removed

Models

K-nearest neighbors, Logistic regression, Random forests, and Gradient boosted machine.

Model evaluation and selection

The entire training set split into 80% train and 20% test. The validation method used was cross validation with StratifiedKFold. To account for class imbalance, minority class was oversampled.

Best model: Random Forests 5 fold cross-validation scores

- Accuracy 0.897
- Precision: 0.879
- Recall 0.923
- F1: 0.900

IV. Tools

- Modeling: Sklearn
- Visualizations: pandas, matplotlib, seaborn

V. Communication

- In addition to the slides and visuals presented, all information can be found on the following GitHub repo: https://github.com/lee-jin81/metis_project_4_classification
- Below is the roc curve for comparing the models: