Enzyme Classification with Embedding Methods

Ilya Tataurov

Problem

Enzymes can be very complex

Enzymes need to be classified accurately

Goal

Make predictive enzyme graph classification model

Methodology

With graph data extract features. Train different types of machine learning models on data. Use models to predict on out of sample data. Evaluate the performance of each model.

Data/Features

- 17 features

-600 data points

-Features mined in Python with NetworkX

Models

-Random Forest

-XGBoost

-SVM

Conclusions

-Tree based model > linear models

-Random forest is best

-Difficult to get high accuracy

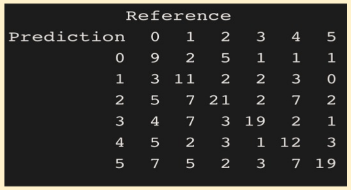
Future work

-Feature engineering

-Lasso and Neural networks

-Hyper parameter optimization

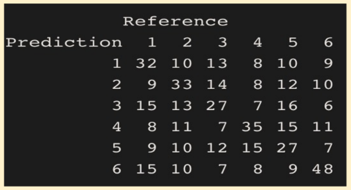
Results



```
                  Reference
Prediction   0   1   2   3   4   5
         0   9   2   5   1   1   1
         1   3  11   2   2   3   0
         2   5   7  21   2   7   2
         3   4   7   3  19   2   1
         4   5   2   3   1  12   3
         5   7   5   2   3   7  19
```

Accuracy: 0.4815

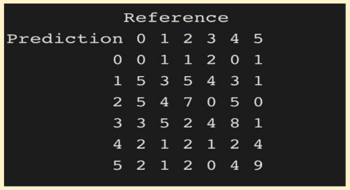
Random forest



```
                  Reference
Prediction   1   2   3   4   5   6
         1  32  10  13   8  10   9
         2   9  33  14   8  12  10
         3  15  13  27   7  16   6
         4   8  11   7  35  15  11
         5   9  10  12  15  27   7
         6  15  10   7   8   9  48
```

Accuracy: 0.3915

XGBoost



```
          Reference
Prediction 0 1 2 3 4 5
         0 0 1 1 2 0 1
         1 5 3 5 4 3 1
         2 5 4 7 0 5 0
         3 3 5 2 4 8 1
         4 2 1 2 1 2 4
         5 2 1 2 0 4 9
```

Accuracy: 0.25

SVM

Accuracy: 0.16

Random guessing