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

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 1 2 3 4 5</td>
<td>0 9 2 5 1 1 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 3 1 2 2 3 0</td>
<td>2 5 7 2 2 7 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 4 7 3 1 1 2 1</td>
<td>4 5 2 3 1 1 2 3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference</th>
<th>Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 7 5 2 3 7 1 9</td>
<td>5 7 5 2 3 7 1 9</td>
</tr>
</tbody>
</table>

Accuracy: 0.4815
Random forest

Accuracy: 0.3915
XGBoost

Accuracy: 0.25
SVM

Accuracy: 0.16
Random guessing