Title: Predicting cell type specific functional consequences of non-coding variation using deep learning

CS230-WINTER 2019

Santosh Kumar
Sangjukta Kashyap
Contents

- Team
- Predicting
- Data
- Features
- Models
- Results
- Discussion
- Future
- References

Stanford University
Team

Santosh Kumar
Department of Computer Science
Stanford University
ksantosh@stanford.edu
https://www.linkedin.com/in/kumsantosh/

Sangjukta Kashyap
Department of Computer Science
Stanford University
sanju321@stanford.edu
https://www.linkedin.com/in/sangjukta-kashyap-99901b5/

Mentor

Hoormazd Rezaei
Stanford University
https://www.linkedin.com/in/hoormazd-rezaei/

Stanford University
Predicting

- Predicting the functional consequences of genetic variants in non-coding regions is a challenging problem.
- We used here a deep learning approach, to jointly utilize experimentally confirmed regulatory variants (labeled variants), unlabeled variants genome-wide, and more than a thousand cell/tissue type specific epigenetic annotations to predict functional consequences of non-coding variants.
- Through the application to several experimental datasets, we demonstrate that the proposed method gets good prediction accuracy,
Data

The dataset we got from Dr. He from Stanford University is ENCODE. The Encyclopedia of DNA Elements (ENCODE) Consortium is an international collaboration of research groups funded by the National Human Genome Research Institute (NHGRI).
Features

127 * 8 = 1016 Features:

Features example:
DNase-E001 where,
DNase is feature and E001 is Cell type/Tissue type

Labels
0 or 1

No. of training data with Label 0 and 1
• 0 - 22384
• 1 - 693

No. of test data with Label 0 and 1
• 0 - 1451
• 1 - 525
Models

Classifier:
1. LogisticRegression
2. DecisionTreeClassifier
3. KNeighborsClassifier
4. BernoulliNB
5. LinearDiscriminantAnalysis
6. GaussianNB
7. RidgeClassifier
8. SGDClassifier
9. Support Vector Machines

Neural Network
1. Non-regularized NN model
2. NN model with L2 Regularization
3. NN Model with Dropout
## Results

<table>
<thead>
<tr>
<th>Classifiers</th>
<th>Mean ROC (AUC Value)</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>LogisticRegression</td>
<td>0.72</td>
<td>95%</td>
</tr>
<tr>
<td>DecisionTreeClassifier</td>
<td>0.57</td>
<td>94%</td>
</tr>
<tr>
<td>KNeighborsClassifier</td>
<td>0.57</td>
<td>95%</td>
</tr>
<tr>
<td>BernoulliNB</td>
<td>0.5</td>
<td>94%</td>
</tr>
<tr>
<td>LinearDiscriminantAnalysis</td>
<td>0.69</td>
<td>96%</td>
</tr>
<tr>
<td>GaussianNB</td>
<td>0.68</td>
<td>95%</td>
</tr>
<tr>
<td>RidgeClassifier</td>
<td>0.55</td>
<td>97%</td>
</tr>
<tr>
<td>SGDClassifier</td>
<td>0.72</td>
<td>95%</td>
</tr>
<tr>
<td>Support Vector Machines</td>
<td>0.65</td>
<td>95%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NN Model</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-regularized</td>
<td>96.8%</td>
</tr>
<tr>
<td>L2 Regularization</td>
<td>96.3%</td>
</tr>
<tr>
<td>Dropout</td>
<td>97.1%</td>
</tr>
</tbody>
</table>
Discussion

- Dataset is with very less Label 1 and most of them are label 0.

- We tried with different classifiers, we got very good accuracy but average ROC(AUC) Curve
  - LogisticRegression Classifier gives best result

- We also tried with Neural Network
  - Dropout gives best result

- We would like thanks out mentor for this project Mr. Hoormazd Rezaei who guided us during each step

- We would also like to thanks Dr. He who has given us the dataset and project requirements.
Future

We plan to continue exploring the project with below strategy

1. Work with larger dataset

2. Try convolution network

3. Also try other neural network along with convolution network
References


Link to youtube video

https://youtu.be/OCi2CX_GVCk