Citation Prediction Bioinformatics

Luke Davis, Mouna Kalidindi, Darpan Jhawar, Charles Stamey, Logan Hornbuckle
CSC 505-01 Fall 2018 UNCG

After the 3rd presentation, we started working towards applying Machine Learning on our dataset.

We first had to perform NLP to convert all textual columns to numeric. We had 3 such columns. We first removed all stop words, performed stemming and then used TFIDF. We used 50 topics to train the model. For each row we used only the topic which had the highest probability. We also added a column where we mentioned the number of topics we got for each row.

The next task was to decide which Machine Learning algorithms to use. We decided to use both classification and regression models. We figured out that for our dataset Random Forest and XGB were the best algorithms.

We first tried regression and the results were not very good. So we went ahead and tried classification models. To divide up the classes we tried different techniques. They were:

1. Use pandas qcut
2. Do a Box Cox transformation and then do a Z score
3. Do a log transformation and then a Z score
4. Used Quartiles(Divided citations based on quartile ranges)

We tried all combinations of dividing the classes and ML algorithms and the highest accuracy we received was 74%. This was when we used Log Transformation with XGB.