Predicting Patent Outcomes with Text and Attributes

CS230

Ryan Lee & Marcos Torres

Problem Statement

Patents are expensive and can take years to be issued. Despite the large investments companies make producing patents, there are no predictors of patent outcomes that account for both the text of the patent and the attributes of the patent. We implement two models, one utilizing the text of the patent and one utilizing the metadata, or general attributes of the patent. We focus on optimizing our models to predict whether a patent will be issued or not.

Metadata Examples

We derive our metadata, or the attributes surrounding the patent, from PatEX, which comprises of records of over 9 million patents with qualitative information such as invention characteristics, applicants, attorneys, and the status of the patent. Examples of metadata features include:

1. Number of patents assigned to your examiner's art unit
2. Number of patents in the example's USPC class
3. "Allowance rate" of the USPTO examiner
4. "Success rate" of a given patent author or firm

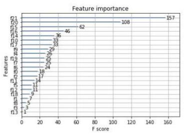
Text Based Examples

We derive the patent text from the USPTO bulk downloads, keeping the first 500 words of claims text.To obtain the correct labels for our text-based dataset, we used the patent application number in the text file to find and attach the corresponding labels from PatEx.

Discussion

Our benchmark algorithm, the XGBoost tree algorithm yielded an F1 score of 0.838 for the classification of issued or not, comparable to the performance of our metadata neural network and text CNN on the same classification problem (0.83 and 0.776).

We see that the most predictive features are the firm success rate and examiner allowance rate.



CNN for Text-based Prediction

Our CNN learns embeddings of 256 features for each encountered word. These embeddings are run through parallel 1D convolution filters of size 2, 3, 4, and 5. The outputs of these convolutions are fed through maxpool, a dense layer, ReLU, and a final dense layer before softmax activation and cross entropy loss.

Deep Neural Network for Metadata Prediction

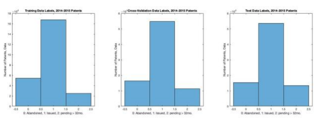
A fully connected neural network trained and evaluated consists of an input layer of 22 features, a 6 hidden ReLU layers, and an softmax output layer of size 3.

Cost Function & Gradient

The cost function and gradient for the neural network is defined below:

$$J(\Theta) = -\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{K}\left[y_k^{(i)}\log((h_\Theta(x^{(i)}))_k) + (1-y_k^{(i)})\log(1-(h_\Theta(x^{(i)}))_k)\right] + \frac{\lambda}{2m}\sum_{l=1}^{L-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}}(\Theta_{j,i}^{(l)})^2$$

Label Distribution



XGBoost Benchmark

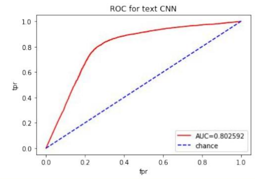
As a benchmark for performance of the neural network model, we trained and evaluated an XGBoost model. XGBoost stands for extreme gradient boosting, and is a variant of a decision tree algorithm optimized for model performance and speed. With the XGBoost decision tree algorithm, we were also able to extract a measure of feature importance to the classification task.

	precision	recall	f1-score
0	0.703	0.569	0.629
1	0.789	0.870	0.828

CNN Test Performance

F1 = 0.776

AUC = 0.80



Deep Neural Net Test Performance

F1 = 0.832

AUC = 0.82

Test Performance	precision	recall	f1-score
Abandoned or Not	0.493	0.156	0.237
Issued or Not	0.748	0.937	0.832
Pending or Not	0.563	0.435	0.491
micro avg	0.709	0.709	0.709
macro avg	0.601	0.510	0.520
weighted avg	0.670	0.709	0.665

Figure 4: F1 Score of 0.832 for 2014-2015 Metadata Neural Network

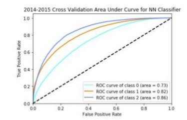
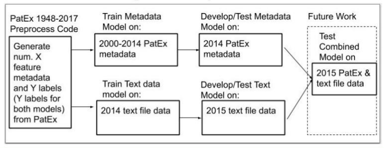


Figure 5: ROC Area Under Curve of 0.82 for 2014-2015 Metadata Neural Network

Future Works

Future work will involve combining the models for the prediction task and experimenting with different NLP models.



References

I. Atigui and E. Atlani , "Patent analysis using machine learning." https://funginstitute.berkeley.edu/capstone-project/patent-analysis-using-ml/

D. Winer, "Predicting bad patents: Employing machine learning to predict post-grant review outcomes for us patents," EECS Department, University of California, Berkeley, May 2017