Forward citation prediction

Shaleen Kalsi

1 Background

1.1 Patent valuation

The downstream task is assessing the value of a patent. This highly nuanced process is currently carried out by industry experts. The goal is to develop a machine learning model capable of predicting a patent's value. Deciding what factors determine the value of a patent is part of the problem.

A proxy for patent valuation currently used in the industry is patent asset index. Based on existing literature [1] and [3], the patent asset index of a company's portfolio is dependent upon the following factors - portfolio size, market coverage, and technological relevance. Portfolio size and market coverage can both be measured by analyzing the company's patents. When it comes to portfolio size we can look at the number of patents and technologies included in a company's portfolio. To quantify market coverage we take into account the GDP of the countries where the patent is protected. However, predicting technological relevance is more nuanced; a patent's forward citation count is one way to measure a patent's technological relevance [2].

2

2.1 Feature selection

For feature engineering, I conducted detailed data analysis and literature review to identify potential candidate features. These features were then evaluated using techniques such as Principal Component Analysis (PCA) to assess their impact on the model's training process. Additionally, backward elimination was applied, systematically removing features one at a time to observe and analyze any changes in the model's performance.

Following features were considered - patent age, abstract and claim text embeddings, the technology field of the patent, the number of independent claims, the number of backward citations, the number of figures, average similarity between the patent and its backward citations (calculated using the abstract and first claim text embeddings), the average number of citations received per patent by the owner of the patent.

2.2 Forward citation count prediction

Initially, the task was modeled as a regression problem; using the above listed features to predict the number of forward citations received by a particular patent 5 years after its publishing. The models that I tried out were - Linear regressions, Support vector regression, Random Forest Regression, and XGBoost Regression.

The best performance I got was using the XGBoostRegressor - RMSE: 15.1299; which is highly unsatisfactory. Given that the downstream task does not require precise predictions of the exact number of forward citations, after discussions I opted to approach the problem as a classification task instead.

2.3 Forward citation count class prediction

2.3.1 Deciding the number of classes

I collected more data and based on the analysis of the distribution of number of forward citations, the following classes were decided -

[0] - class 1

(0, 3] - class 2

(3, 10] - class 3

(10, 100] - class 4

(100+] - class 5

The task is performed in 2 stages:

- The stage 1 model is trained using the previously mentioned features and a ground truth label (0 or 1) indicating whether the patent received any citations.
- The stage 2 model is trained using the same features and a ground truth label (2-5) indicating the class of the patent's forward citation range.

At inference time, the pipeline will receive the extracted features and output the patent forward citation count class.

2.4 Preprocessing

- Since there is class imbalance in the data I performed stratified splitting of the data.
- One-hot encoded the Forward citation class column.
- Experimented with TSNE and PCA. PCA performed better with the models listed below.

2.5 Models

The models I experimented with were -

- Step 1:
  - Logistic Regression
  - ANN with Dense layers for binary classification
- Step 2:
  - Random Forest classifier (n_estimators = 100, max_depth = 2) - Validation accuracy: 0.56

- XGBoost classifier with Randomized Search CV - Validation accuracy: 0.66
- ANN with Dense layers
- ANN with LSTM layers
- ANN with GRU layers
- ANN with biGRU layers

ANN implemetation details: Loss - Categorical Cross-Entropy loss, Optimizer - Adam, Learning rate - 1e-3, EarlyStopping on validation loss with patience - 25.

Training data	**Model**	**Training loss**	**Training Acc**	**Epoch–Val loss**	**Val Acc**
Features	Dense ANN dropout=0.25	0.5590	0.8887	007–0.6501	0.8302

Table 1: Best performing Step-1 Experiments.

Training data	**Model**	**Training loss**	**Training Acc**	**Epoch–Val loss**	**Val Acc**
Features +TSNE	Dense ANN dropout=0.25	0.7824	0.6891	010–0.8901	0.6003
Features +PCA	Dense ANN dropout=0.25	0.7693	0.6990	012–0.8501	0.6403
Features	Dense + GRU ANN dropout=0.25	0.7524	0.6987	011–0.78190	0.6822
Features	Dense + biGRU ANN dropout=0.5	0.7508	0.6974	015–0.78205	0.6881
Features	Dense + LSTM ANN dropout=0.5	0.7454	0.6996	020–0.78241	0.6874

Table 2: Best performing Step-2 Experiments.

3 Further and related work

- Perform further experiments by tweaking the above architectures.

- Deciding the class ratio for step 1, currently is as collected from source data (60% - 0 forward citations, 40% - greater than 0 forward citations).
- Using the number of subgroups in the ipcr as a feature and exploring other features that can be used.
- Patent acceptance prediction: The idea is to discount the patent's value by the probability of its acceptance when performing the valuation of a patent that has not been published yet.

References

[1] URL: https://www.lexisnexis.com/community/insights/professional/b/solutions/posts/what-is-lexisnexis-patentsight.

[2] Leonidas Aristodemou and Frank Tietze. "Citations as a measure of technological impact: A review of forward citation-based measures". In: *World Patent Information* 53 (2018), pp. 39–44. ISSN: 0172-2190. DOI: https://doi.org/10.1016/j.wpi.2018.05.001. URL: https://www.sciencedirect.com/science/article/pii/S0172219017300376.

[3] Holger Ernst and Nils Omland. "The Patent Asset Index – A new approach to benchmark patent portfolios". In: *World Patent Information* 33.1 (2011), pp. 34–41. ISSN: 0172-2190. DOI: https://doi.org/10.1016/j.wpi.2010.08.008. URL: https://www.sciencedirect.com/science/article/pii/S0172219010000864.