PROJECT MACHINE LEARNING I
PREDICT MISSING CITATIONS IN THE CITATION NETWORK

NGUYEN Tu Anh, PHUNG Lam Binh, D'ALMEIDA Benoit
Team name: PHUNG

January 10, 2019

1 Introduction

In this project, we will study a classification model on the citation network. Given the citation data on a set of papers, our objective is to predict the missing links in the network based on the different features of a pair of papers. We firstly investigate the given dataset and then study our model (features selection, choice of classifier,...) in order to improve it.

2 First investigation on the data

Our dataset contains 27,770 articles from the period 1992-2003, each contains 6 information of the paper including:

(1) paper unique ID
(2) publication year (from 1992 to 2003)
(3) title
(4) authors
(5) name of journal
(6) abstract

The training data consists of 615,512 rows, each contains a node pair with an associated value 1 if there is a citation between the 2 nodes, 0 if not. From the information of the papers, we can count the number of papers published in each year. We can see in Figure 1 that the number of published papers is indeed increasing every year, as expected in [3].

Using the data on the citation network, we also observe another interesting trend as in [3], that the number of papers cited n times (in one single year) decrease exponentially as n increase (Figure 2).

We can also see that the most cited papers has a large gap with other papers, this phenomena is called *the rich get richer*.

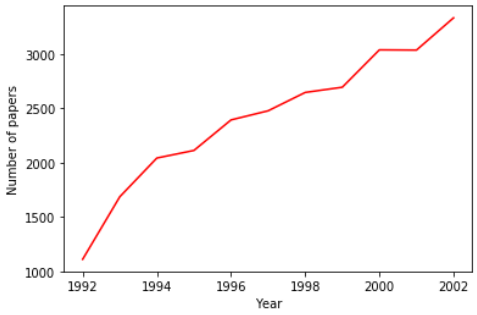

Figure 1: Number of published papers from 1992 to 2002

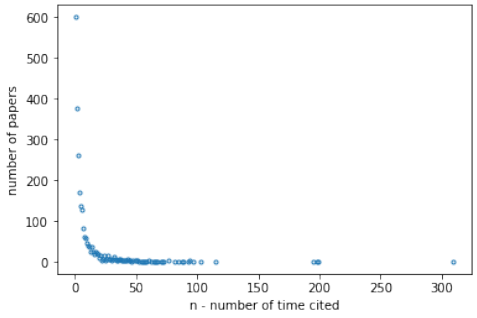

Figure 2: Number of papers cited n times in 2002

3 Investigation on the model

3.1 Features selection

Our ideas of choosing features basically come from [1] and [2], we also try to add some additional features in order to improve our model.

Our features can be divided mainly into 2 categories, the first one is textual features which only concerns the similarity between the textual information of the two papers. The graphical features come from the graphs generated by the given training data.

The selected features are:

Textual features
(1) Cosine similarity of abstracts
(2) Number of overlap words in abstract
(3) Cosine similarity of titles
(4) Number of overlap words in title
(5) Number of overlap words between target's title and source's abstract
Graphical features
Citation graph
(6) Number of common neighbourhoods
(7) Link-based Jaccard coefficient
(8) Adamic-Adar index
(9) Preference attachment
(10) Difference in betweenness centrality
(11) Difference in the number of in-links
(12) Number of times target cited
(13) Pagerank of source
(14) Pagerank of target
(15) Minimal distance*
(16) Is the same cluster?
Author collaboration graph
(17) Number of common neighbourhoods
(18) Link-based Jaccard coefficient
(19) Preference attachment
(20) Adamic-Adar index
Other features
(21) Difference in publication year
(22) Journal popularity of target
(23) Is the same journal?
(24) The number of common authors
(25) Is self-cited?

The choice of textual features is natural because the higher the 2 papers are correlated, the higher chance they will cite each other. We add the feature (5) as there's a chance that the source paper will cite the target in their abstract.

The graphical features come from 2 networks. The first one is the citation network and the second one is the author collaboration network. We believe that the collaboration network can have an effect on the citation network. An author is more likely to cite authors who have a *relatively close* collaboration with him, or at least who work in the similar domain. The feature *Minimal distance** is the length of the shortest path between the two nodes excluding the edge connecting them (if exists). This is to prevent the overfitting in the training process.

We also add some other features such as the temporal distance of 2 papers, the number of common authors (as authors tend to cite them-self) and the popularity of journal (as papers in a more popular journal are more likely to be cited).

The Figure 3 shows the feature space of the first 1000 training data.

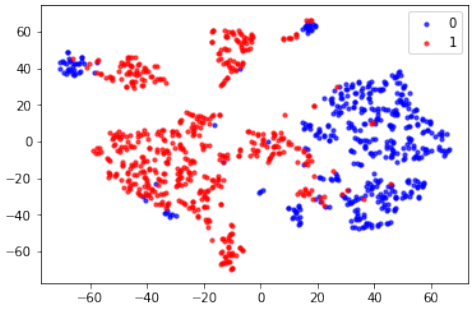

Figure 3: t-SNE visualization of feature space (first 1000 instances from training set)

3.1.1 Correlation of features

We use the library *pandas* to calculate the correlation of the features. The results are shown in Figure 4.

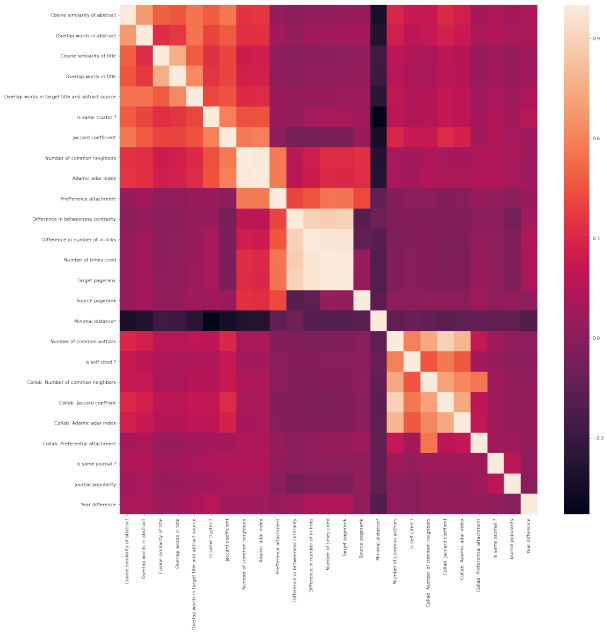

Figure 4: Correlation of all the features, lighter means more correlated

We see that there is a relative correlation between the features in the same context. The textual features seem to correlate with the graphical features which related with the clustering, which is reasonable since papers in the same domain will have similar textual information. There is an other group of graphical features that are very correlated: *Difference in betweeness centrality, difference in*

number of in-links, number of times cited and target pagerank, they are indeed all related to the popularity of the target paper. The features from the collaboration network are quite correlated, and they indeed correlate with the features *Self-cited* and *Number of common authors*. We see that the feature *Minimal distance** is strangely uncorrelated with all other features, it can be suggested that this feature may have a positive effect to the model.

3.1.2 Importance of features

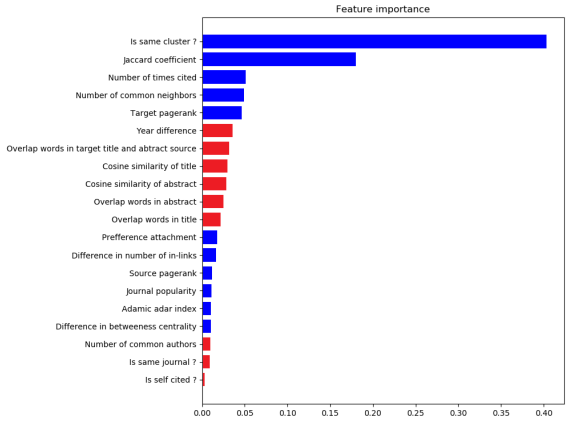

Figure 5: The importance of all the features, the graphical features are in blue

We can see from the figure that the graphical features dominates other features. The feature *Is same cluster* shows the largest importance, which can be seen that 2 papers are much more likely to cite each other if they have the same working field. The feature *Self-cited* is the least importance in this case.

3.2 The choice of classifier

In this project, we try the following classifiers for comparison : ExtraTree, Adaboost, LogisticRegression, LinearSVM, RandomForest, and NeuralNetwork. We use sklearn library with default parameters in each classifier for implementations, except the case of Neural network, we use Keras library with sequential model of 3 hidden layers (30 nodes each), with batch size of 128 and 20 epochs. We use 95 percent of training data for training, and calculate the performance of each classifiers on the rest 5 percent. The 25 features can be classified to 3 types : the ones without graph features, citation graph features and author collaboration graph features. The charts below show the performance comparison between different models, and different combination of features:

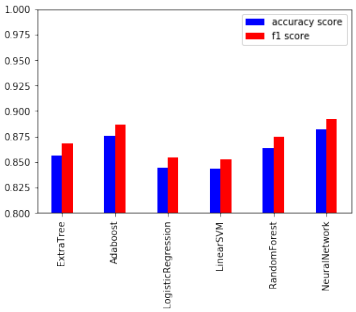

Figure 6: Performance of models with non-graph features

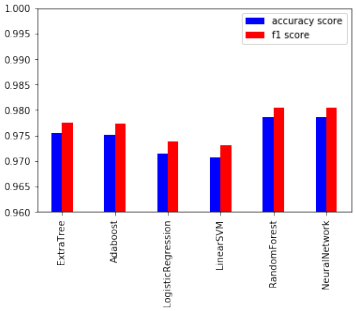

Figure 7: Performance of models without author collaboration graph features

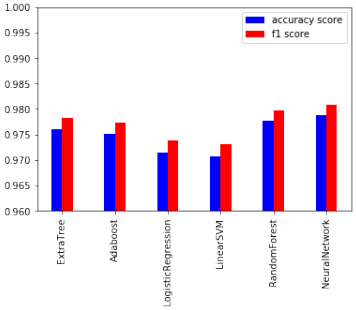

Figure 8: Performance of models with all features

We see that adding citation graph features makes a significant improvement for the performance of all models (accuracy raised from about 0.88 to 0.976). The author collaboration ones contribute slightly to some methods like Neural network, but not every model. The three best models are Neural network, random forest and boosting. We choose Neural network model to make the submissions in Kaggle, which performs the best accuracy and f1 score. The models Extra tree or Logistic Regression, with smallest running times, can be used in practice.

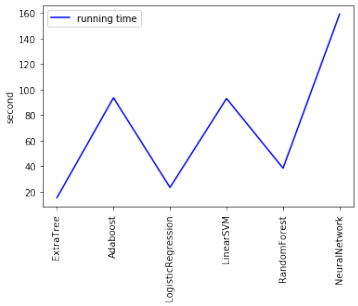

Figure 9: running time of each model

3.3 Number of training data

In this part, we'll see how the data affect the final result. We compare the results when we use 80% and 95% of the data. We use the Neuron network model in this case.

	80%	95%
Precision	0.96988	0.97544
F1 score	0.97212	0.97735

Table 1: Result when increase training data from 80% to 95%

We see that by increasing the training data, we receive a much better result.

4 Conclusion

In this project, we have studied how to use a machine learning classification model to predict missing links in the citation network. We have chosen features in 3 main categories: Textual features, citation network features and author collaboration network features and have experimented on several classifiers. Our results are quite satisfying. However, there are still several things that we haven't had chance to experiment yet, like the associated author citation network (which is different from the collaboration network) or some other network features (Katz, simrank, weights,...) as mentioned in [2].

References

[1] Shibata Naoki, Kajikawa Yuya and Sakata Ichiro. *Link Prediction in Citation Networks*, JASIST, DOI: 10.1002/asi.21664 (2012).

[2] D. Liben-Nowell, J. Kleinberg. *The Link Prediction Problem for Social Networks*, http://doi.acm.org/10.1145/956863.956972 (2004).

[3] Derek J. de Solla Price. *Networks of Scientific Papers*, Science: vol 149, p510-515 (1965).