Predicting the future influential researchers in big scholarly network

Siyi Wang
F1503001
515030910005

***Abstract*—**

In this paper, we study both theoretically and practically to predict the future influential researchers based on the Citation Network data set. We first create the diffusion threshold model to prove the attenuation of information diffusion. And then we come up with regression models including linear and nonlinear regression algorithm, linear SVM and decision tree(regression) to predict the total citation number of an author. Results show that some factors do have great influence on an author's citation, and by comparing the errors and judge the performance of models, we also obtain some optimal models to do the prediction.

I. Introduction

Not only scientists are curious about their future influence but also the world does, because whether to be accepted or identified by top organizations often depends on the influence of papers(or works). One of the essential factors that indicate the influence of a scientific work is its frequency of citation.

Predicting the future influence in scholarly network has been very popular lately. In [1], a model of machine learning is used to predict an author's h-index(which has already become a highly-acclaimed factor to measure and predict an author's future impact). In [2], a mechanistic model for the citation dynamics is derived to predict the long-term citation. Unavoidable challenges to do such a prediction include: a) Relationship between factors contributing to a paper's influence is complicated; b) Most features don't have numeric values, thus can't be directly used for fitting and prediction.

In this project, we firstly introduce a threshold model to get some ideas of the information diffusion in real condition. And then we introduce some machine learning models to fit the features and citation using regression algorithm to implement the prediction.

A. Threshold Model

During the process of information diffusion (e.g.scientific impact of a paper), the diffusion of newly activated neighbors affects the nodes most, while the diffusion information of the historical nodes will be attenuated.

Let's suppose node u spreads to node v with the influential weight $w_{v,u}$. If successful, node v can be transformed into an active node; if not, the effect on node v is not $w_{v,u}$ anymore, but

$\gamma w_{u,v}$, which means the diffusion of information has been attenuated. Each node v has an information acceptance threshold θ_v and is affected by all of its active neighbor nodes A(v). The influence value is represented by weight w, which satisfies

$$\sum_{u \in A(v)} w_{v,u} \leq 1 \tag{1}$$

The relationship of nodes can be shown as Figure 1.

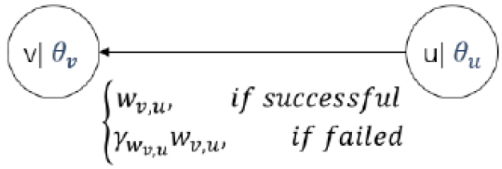

Fig. 1. Diffusion Threshold Diagram

Node v will be activated when

$$\begin{aligned} &\textstyle\sum_{u \in At}(v) w_{v,u} + \gamma_1 \sum_{u \in At-1}(v) w_{v,u} \\ &+ \ldots + \gamma_{t-1} \textstyle\sum_{u \in A1}(v) w_{v,u} \geq \theta_v \end{aligned} \tag{2}$$

B. Machine Learning

Apart from theoretical analysis, in this project we also tried to use machine learning, which means to train some regression models to predict the results. And by comparing the performance, to finally obtain the optimal model.

The following parts of the paper are organised as (2) Data Preprocessing (3) Model Description (4) Experiment and Result (5) Conclusion

II. Data Preprocessing

A. Citation Network Dataset

Citation Network Dataset from Aminer is designed for research purpose only. The citation data is extracted from DBLP, ACM, MAG (Microsoft Academic Graph), and other sources. Each paper is associated with features including abstract, authors, title, publication year and venue, which are important factors for the future impact of a paper.

The data set can be used for clustering with network, studying influence in the citation network, finding the most influential papers, predicting scientific future impact, etc. And in our experiment, we choose a data set containing 566,591 scientific papers.[3]

B. Preprocessing Method

In this part, we aims at preprocessing the data set, separating the features from the original text and representing them using numeric values. After obtaining the preprocessed data set(”output.csv”), it will be more convenient for us to train the machine and implement the algorithms to predict the author citation.

Generally, each paper has features of five fields: title, authors, year, publication venue, and abstract, whereas the citation number will be the influence metric. Firstly, we separate and extract them from the text file into several arrays for future usage. To implement the following predicting work, we convert the ’author’ and ’publication venue’ features into numeric values by accumulating the corresponding papers’ total citation. Thus, a specific value represents an author or a conference.

While for the content included in the title and abstract, it is quite difficult to convert them into numeric values. Thus, we import gensim in python and

use the Word2Vec function to obtain the distributed characterization of word vectors. The main idea of it can be shown in Figure 2. The model is trained to predict the distributed weights of each word, each dimension of a vector represents a feature vector that acts on all words rather than a simple mapping between elements and values. In this way we can abstractly express the ”meaning” of the title and abstract.

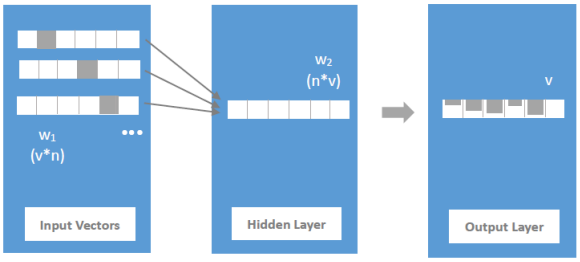

Fig. 2. Word2Vec Mechanism

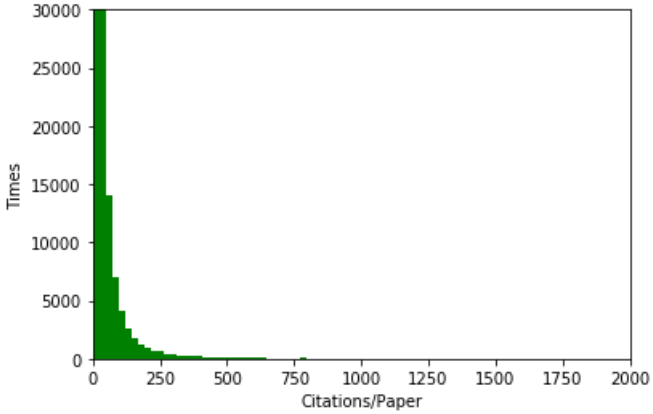

Fig. 3. Frequency of Citation

We then sketch some plots to show the processed features more clearly. In Figure 3, most of the papers have citations less than 500 times; in Figure 4, we get the idea that the citation is distributed nearly evenly along with years; citation of different publication conferences is shown in Figure 5.

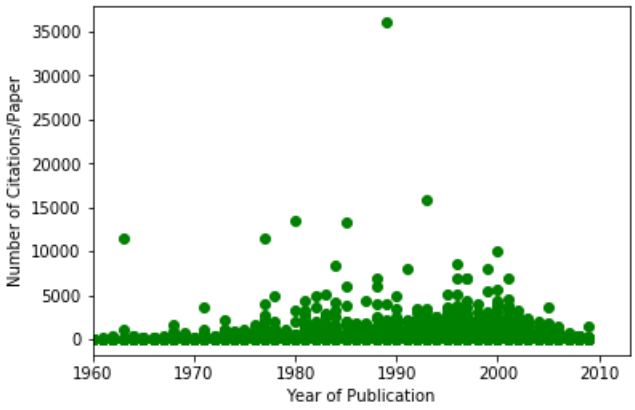

Fig. 4. Distribution of Year and Citation

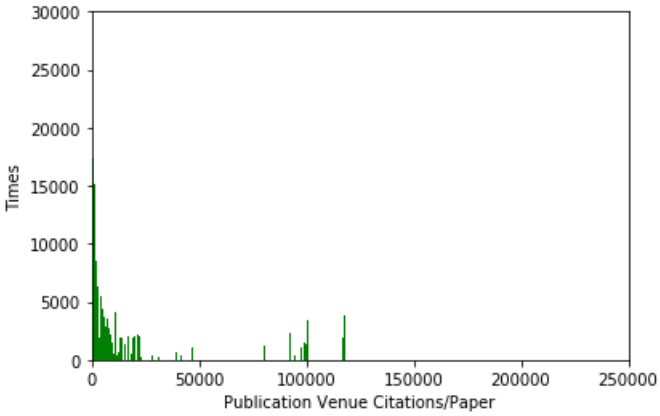

Fig. 5. Citation Frequency of Public Venue

III. Model Description

We choose different models to train the machine and predict the future impact: (a)Linear Regression (b)Non-linear Regression(depends on one or more independent variables) (c)Linear SVM (support vector machine) (d)Decision Regression Tree(binary tree where each branching node is split based on the values of a column of feature matrix.).

There are several common steps: (1)Firstly we import the data(”output.csv”), extracting the citation column as the result for training. (2)Then we split the data set into training set and testing set(train—the first 350,000 items; test—the rest part). (3)We

choose different models to fit the feature and citation and obtain the weight vector. (4)After obtaining the predicted citation value of single paper, we accumulate the citation of the same author and use the total value to represent an author's future impact. (5)Finally we calculated the errors between the real citation value and the predicted result to compare the performance.

The algorithms are shown below:

Algorithm 1: Predicting Procedure

Input: Preprocessed feature matrix"output.csv"
Output: "predict.csv"(Single paper's predicted citation)

```
input=importdata('output1.csv');
Extract features except for citation;
cit = input(:,3);
for i=1:10 do
    trCit = cit[1:350,000];
    trFea = feat([1:350,000],:);
    tCit = cit[the rest index];
    tFea = feat([the rest index],:);
    b = model_type(e.g.regress)(trCit,trFea);
    pred_cit[:,i] = tFea*b;
csvwrite('predict.csv',mean(pred_cit[:,i]));
```

Algorithm 2: Prediction of Author Citation

Input: Prediction of single paper's citation $p_1, p_2, ...p_i$, and the original dataset "acm_output1"
Output: 'output1.csv'

```
f=open('acm_output2.txt');
reader = csv.reader(open("predict.csv"));
lines=read_all_lines;
for l in lines do
    if l[2:] == "#@" then
        authors=l[2:];
    for j,rows in reader do
        if j==i then
            pre=rows;
    names = authors.split(',');
    record = 'authors':
    authors,'p_citations' : p_author_citation;
    names = authors.split(',');
    for name in names do
        if name in p_author_citation then
            p_author_citation[name] += float(pre[0]);
return record;
```

TABLE I
PERFORMANCE OF DIFFERENT ALGORITHMS

Model	MMAE	MMSE
Linear	18.96	136.80
Non-linear	18.48	136.32
SVM	17.74	138.42
Decision Tree	20.27	147.54

IV. EXPERIMENT AND RESULTS

The predicting error of citation is represented as mean absolute error(MAE) and mean squared error(MSE) between the predicted author citation value and the appreciable value. MSE is used to measure deviations between observations and true values while MAE can better reflect the actual situation of the prediction error. The performance of different models is shown as Table 1 and Figure 6.

As is shown, we found that the LR algorithm gets relatively higher MAE and MSE value, it is quite reasonable because the interior relationship between the features and citation numbers are complex. Besides, NLR has the minimum MSE and the SVM gets the minimum MAE, which are more optimal than other two models.

V. CONCLUSION

In this project, our purpose is to predict the future influential researchers in big scholarly network. We quantify the impact as the citation times of a researcher and the problem is converted to obtain the future citation numbers of an author. We firstly study theoretically based on a Diffusion Threshold Model and derive the formula of information dif-

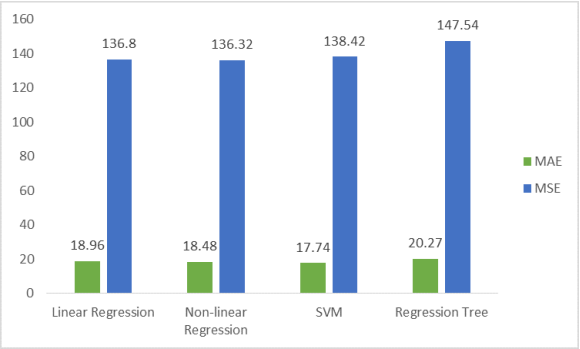

Fig. 6. Comparison of Different Models

fusion with attenuation. Then we proposed several machine-learning regression models to implement the prediction based on the preprocessed data set. By comparing the performance of the different models, we find that the Non-linear regression and SVM models obtain the relatively better predicted results among the four and we are expecting to implement more complicated and accurate algorithms in the future to deeply study the future impact prediction of a researcher.

REFERENCES

[1] E. Acuna, Stefano Allesina, and P. Kording, ”Predicting scientific success” Nature.489(7415): 201C202, September 2012.

[2] D.Wang, C. Song, and A.-L. Barabasi, Quantifying long-term scientific impact, Science, vol. 342, no. 6154, pp. 127C132, 2013.

[3] Introductions of data set, https://www.aminer.cn/citation, 2018.