Learning Approaches to Predicting Future Citation Counts

Sreeram Sreekar
BTech 4th year
CSE, IIT Hyderabad
cs13b1008@iith.ac.in

Akshita Mittel
BTech 4th year
CSE, IIT Hyderabad
cs13b1040@iith.ac.in

Surya Teja Chavali
BTech 4th year
CSE, IIT Hyderabad
cs13b1028@iith.ac.in

ABSTRACT

In this paper, we attempt to propose a solution to the problem of predicting the citation count of a research article in a certain number of years. We make minimalistic assumptions on the dataset and train a sequence-to-sequence LSTM encoder-decoder model. We adopt the Aminer dataset and the High-Energy-Theoretical-Physics dataset(link) to perform our experiments. The experimental results show a significant improvement(upto 15%) over the baseline models. Further, the proposed solution extends to any dataset, as no assumptions over the data-set are made.

Keywords

Citation count prediction; regression models; data engineering

1. INTRODUCTION

The explosive increase in the volume of scientific publications in recent years is expected to continue for the foreseeable future. Citation count of a paper is widely used to indicate the popularity/impact of a scientific papers. Accurate and automatic prediction of citation counts would provide a powerful method for evaluating articles and faster identification of promising articles could accelerate research and dissemination of new knowledge. Accurate models for citation count prediction would also improve our understanding of the factors that influence citation.

Citations are a noisy, indirect quality measure, and accumulation rates vary unpredictably between articles. Breakthrough papers can stop receiving citations after review articles replace them or the subject matter becomes common knowledge. Predictions based on current data assume that citation behavior will not change in the future, and this assumption may be violated in fast-paced research fields such as Computer Science.

The very limited number of solutions suggested have mostly modeled the problem as a learning task âĂŞ given a set of features and a particular time interval, a regression model is trained on the entire set of the training population, and accordingly, the future citation count of a query paper is estimated.

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

2. PROBLEM STATEMENT

In this section we present a mathematical representation of the citation count prediction problem.

Citations: Given a corpus L, the citation count of a p paper in year y, C is given by:

$$cites(a,b,y) = \begin{cases} 1, & \text{if a cites b and b was published in year y} \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

$$C(p,y) = \Sigma(cites(x,p,y)),\ \forall x \in L \quad (2)$$

Problem:

Given a representation of a paper , our goal is to learn a predictive model to predict the citation count of a paper Δt years after its publication. Thus,

$$\text{Learn: } \Psi(d, F, \triangle t) \rightarrow C(d, \triangle t)$$

The contributions of our paper are as follows:

- We develop an end-to-end encoder-decoder learning model that takes as input a feature vector representing the paper at the time of its publication and output a sequence of length delta t, predicting the the citation count received by the paper in each year after its publication.
- We show that features related to paper-topic diversity and author-topic diversity are crucial for the purpose of predicting the citation counts. This is significant, because the diversity features were computed unsupervised, by employing a Latent Dirichlet Allocation model.
- Additionally, we show that including features of a paper two years after its publication, drastically improves the prediction accuracy in subsequent years.

3. COMPARITIVE STUDY

We include the features and learning models for each of the papers studied along with our novel method.

3.1 Features

1. Towards a Stratified Learning Approach to Predict Future Citation Counts:

(a) Author centric features
   i. Author productivity
   ii. Author H-index
   iii. Author diversity
   iv. Sociality of author

(b) Venue centric features
   i. Long term venue prestige
   ii. Short term venue prestige
   iii. Venue diversity

(c) Paper centric features
   i. Team size
   ii. Reference count
   iii. Reference diversity
   iv. Keyword diversity
   v. Topic diversity

2. Citation Count Prediction: Learning to Estimate Future Citations for Literature

   (a) Topic Rank
   (b) Diversity
   (c) Recency
   (d) H-Index
   (e) Author Rank
   (f) Productivity
   (g) Sociality
   (h) Authority
   (i) Venue Rank
   (j) Venue Centrality

3. Sequence to Sequence Learning for Citation Count Prediction:

   The feature set used in our methodology is the same as the Stratified Learning Approach.

3.2 Learning Models

1. Towards a Stratified Learning Approach to Predict Future Citation Counts:

   Two-stage prediction model

   In the first stage, the model maps a query paper into one of the six categories, and then in the second stage a regression module is run only on the subpopulation corresponding to that category to predict the future citation count of the query paper.

2. Citation Count Prediction: Learning to Estimate Future Citations for Literature

   Linear regression(LR), k-NN, SVR, CART models were all used and compared against each other. The SVR model gave the best performance.

3. Sequence to Sequence Learning for Citation Count Prediction:

   The learning model is described in section 4.2.

4. PROPOSED SOLUTION

All previous methods have not been able to leverage training data successfully. During training, features for a particular paper were extracted only at the time of publication and the learning model is provided no information about the evolution of the features through successive years. Our solutions addresses that problem.

4.1 Proposed solution

The problem is re-phrased as a sequence-to-sequence learning model that during training time, learns to map a sequence of feature vectors with sequence of citation counts. More concretely, for each year i after the the paper has been published, feature vector F_i is extracted. Sequence of feature vectors F_1, F_2, F_3, F_4, F_5 is extracted and mapped to C_1, C_2, C_3, C_4, C_5 where Ci is the citation count received by the paper in the ith year after publication.

During testing phase, though, the network is only provided the features of a paper at the time of its publication. The network outputs a 5-length sequence predicting the citation count for each year succeeding the year of publication.

The sequence-to-sequence learning model that we use is the same one as that used by [6]Sutskerver et al. The network uses a multi-layered LSTM(Long Short-term Memory) network to map the input sequence to a vector of a fixed dimensionality, and then another deep LSTM to decode the target sequence from the vector.

A useful property of the LSTM network is that learns to map an input sequence of variable length to a fixed dimensional vector representation. This is particularly useful in our case, where the input in the training set is a sequence of length of 5, while the input for testing is a sequence of length only 1.

4.2 Learning model

A simple strategy for general sequence learning is to map the input sequence to a fixed-sized vector using one RNN, and then to map the vector to the target sequence with another RNN (this approach has also been taken by Cho et al. [7]). While it could work in principle since the RNN is provided with all the relevant information, it would be difficult to train the RNNs due to the resulting long term dependencies. However, the Long Short-Term Memory (LSTM) is known to learn problems with long range temporal dependencies, so an LSTM may succeed in this setting.

Our actual models differ from the usual LSTM networks in three important ways. First, we used two different LSTMs: one for the input sequence and another for the output sequence, because doing so increases the number model parameters at negligible computational cost and makes it natural to train the LSTM. Second, we found that deep LSTMs significantly outperformed shallow LSTMs, so we chose an LSTM with four layers.

5. EVALUATION

5.1 Dataset

We use the well-studied Arnet Miner Dataset. The citation data is extracted from DBLP, ACM, and other sources.

The first version contains 629,814 papers and 632,752 citations. Each paper is associated with abstract, authors, year, venue, and title. It has been extensively used for the purposes of clustering with network and side information, studying influence in the citation network, finding the most influential papers, topic modeling analysis, etc. We used V1 and V2 as training sets, V3 for validation, and V4 for testing.

We further evaluated our proposed model on the HEP dataset that was originally release as part of the KDD Cup 2003.

5.2 Evaluation metric

We seek to compare the approaches to this problem by way of using the Coefficient of Determination(R^2) statistic. It provides a measure of how well observed outcomes are replicated by the model, based on the proportion of total variation of outcomes explained by the model. This is defined by the equation:

$$R^2 = \frac{SSR}{SST} = \frac{\sigma(\hat{y}_i - \bar{y})}{\sigma(y_i) - \bar{y}}$$

Further, we use MSE which is the mean square sum of the differences of all predictions with the ground truth.

6. RESULTS

6.1 Yan et al

Comparison with Yan et Al				
	Method	R^2 Value reported in paper	R^2 Value reported by our implementation	MSE
1-year CCP	LR	0.664	0.620	0.8197
	KNN	0.607	0.587	0.8976
	SVR	0.625	0.606	0.8597
	CART	0.683	0.534	1.8082
5-year CCP	LR	0.706	0.618	6.9491
	KNN	0.640	0.580	7.6715
	SVR	0.719	0.657	7.3733
	CART	0.752	0.718	11.8492
10-year CPP	LR	0.767	0.403	34.6031
	KNN	0.767	0.403	34.6031
	SVR	0.755	0.416	36.2493
	CART	0.786	0.418	45.2084

As is evident our implementation is commensurate with the papers although the data set used was different.

6.2 KGP

Comparison with KGP				
Time	Paper: R^2	Implem: R^2	Paper: MSE	Implem: MSE
$\Delta t = 1$	0.62	0.57	2.84	-
$\Delta t = 2$	0.51	0.55	3.11	-
$\Delta t = 3$	0.44	0.52	3.85	-
$\Delta t = 4$	0.38	0.50	4.32	-
$\Delta t = 5$	0.31	0.45	5.87	-

6.3 Overview of the Comparison

SET	Papers	MSE(our approach)	Yan et al MSE	KGP MSE
Train data	225,746	0.018	5.84	3.76
Validation data	111,189	0.014	6.21	4.44
Test data	165,952	0.021	6.94(best)	5.87

The above is on the V1 version of the Aminer dataset.

7. REFERENCES

[1] Towards a Stratified Learning Approach to Predict Future Citation Counts, Tanmoy Chakroboty et al, JCDL '14, Pages 351-360

[2] Citation Count Prediction: Learning to Estimate Future Citations for Literature, Yan et al, CIKM '11, Pages 1247-1252

[3] Models for Predicting and Explaining Citation Count of Biomedical Articles,Lawrence D. Fu, Aliferis, AMIA Annual Symposium Proceedings. 2008; 2008: 222-226.

[4] Predicting Citation Counts Using Text and Graph Mining, Avishay Livne et al, Modern Advances in Applied Intelligence, Volume 8482, Pages 109-119

[5] ArnetMiner: Extraction and Mining of Academic Social Networks, Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. In Proceedings of the Fourteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD'2008). pp.990-998

[6] Sequence to sequence learning in Neural networks, Ilya Sutskerver et al, Advances in Neural Information Processing Systems 27 (NIPS 2014)