Predicting Paper Quality in the Biological Sciences

Matthew Denton (mdenton), Jose Hernandez (josehdz), Debnil Sur (debnil)

Abstract

Predicting the "success" of a paper is an important problem in the growing field of bibliometrics and one with relevance to researchers, journals, and academic institutions. An accurate predictor of a paper's citation count would enable better academic indexing and expedite the research process. In this paper, we investigate the following question: given data about a paper in the biological sciences, can we predict its citation count? We use a paper's journal impact factor, author history, topic, age, and number of references. To build a classifer, we use supervised and unsupervised learning techniques, including latent Dirichlet allocation, softmax logistic regression, support vector machines, cross-validation, and grid-search. The method accurately classified 73% of testing data into a "bucket" of citation counts modeling its impact. Feature analysis demonstrated that age, author history, and references most strongly influenced a paper's eventual success. Topic clustering had lower accuracy than analysis of the entire data set, suggesting the strength of interdisciplinary research in biology.

1 Introduction

1.1 Motivation

An article's citations are considered a measure of the scientific recognition the study has received and thus indicate its value and impact on the scientific field [1]. Researchers commonly aim to publish articles that will attract citations and thus be regarded to have a high scientific impact, as this may be associated with career advancement [2]. Similarly, citations are the main factor determining a journal's scientific impact, denoted by the journal impact factor [3]. Accurately predicting citation counts, then, helps institutions better understand what determines a paper's ultimate success and provides guidance for funding allocation; enables researchers to more effectively publish their work; and generally creates more efficient research processes by finding papers likely to succeed.

1.2 Past Work

Previous studies have considered the correlation of factors intrinsic to a paper, such as its age, journal impact factor, and author history, with its eventual success. Though holistically performed in other subjects, such as chemistry or the environmental sciences, bibliometric studies of biomedical and life sciences literature have not used the repository of open access journals to perform a similar analysis [5-6]. Instead, they focus on specific topics or regions, which removes the potential for comparative analysis and results in smaller data sets that sacrifice robustness [4, 7-8].

1.3 Our Work

Our research has three major design advantages over previous bibliometric studies in biology.

Scale: as mentioned, prior studies utilized a small data set, on the order of 25-400 papers. This prevents the generalization of findings for meaningful predictions. The advent of open-access publication has meant that data no longer bottlenecks statistical analysis as in older papers [3]. Utilizing the PubMed Central database provides over a million journals for feature analysis. This helps understand general trends in biology, as opposed to specific fields; in particular, scholars can see if a particular feature especially influences citation counts in general and in topics.

Classification: previous studies used a regression based approach to predict specific citation counts of papers [6, 8]. Instead of predicting citation counts, we classify a paper's impact into one of a group of "buckets" based on citation count. A paper with 0, 1, or 2 citations has not had a wide impact, and one with 50 or 200 citations is being cited a significant amount. Accordingly, we still quantify the impact of a paper while eliminating sources of skew. A larger sample lets us better classify papers into individual "buckets" and increases the robustness of our findings.

Topic Analysis: rather than handpicking topics from certain journals or subjects, we utilize a version of latent Dirichlet allocation called SciReader, written by the Pritchard Lab in the Stanford School of Medicine (unpublished). Given a paper's full text, this algorithm returns the probability that a paper is in one of 150 topics in the biological sciences. Using this, we can see how topical clustering affects the prediction of citation counts for biomedical and life sciences literature.

2 Dataset

We utilized the Open Access subset of PubMed Central, a database of biomedical and life sciences journal literature [9]. The subset as a whole contains 1.2 million journals, and we used a sample of 300000 articles. For a given paper, the database provides publication date; information about authors, journal, and references; and the abstract and full-text. We found the Journal Impact Factor of each paper through CiteFactor [10]. We excluded journals lacking an impact factor, which decreased to 100000 samples. The PMC Web Service helped us find a paper's citation count and the career citation count of the paper's principal investigator [9]. The features for each paper were as follows: paper age, journal impact factor, number of authors, number of PI citations, and number of references. After analyzing the entire data set, the SciReader algorithm assigned a topic number to each paper. Techniques were then run within each topic.

3 Methodology and Results

3.1 Initial Classification

3.1.1 Softmax

The softmax approach generalizes logistic regression for classification problems where $y \in \{1, ..., k\}$, each representing a different category that the hypothesis function can select. Specifically, the hypothesis will estimate $p(y = i|x; \theta)$, for every value of $i = 1, ..., k$. For our preliminary models, papers were assigned a citation count "bucket" number $y \in \{1, ..., 4\}$, where $y = 1$ represented 0-2; 2, 3-7; 3, 7-15; 4, 15+. We chose the buckets with two criteria in mind: papers in the same bucket have similar "impact," and each bucket has enough samples. Inspection of the data set indicated that many papers had very few citations, while few papers had very many. Thus, the choice of citation counts generate buckets of a significant number of papers with similar impacts. Each bucket had adequate samples for analysis, ranging from $y = 1$ with 62,137 samples, to $y = 4$ with 11,006 samples.

The softmax regression, run with 50000 and 100000 papers, had training and testing accuracies clustered around 62±1%. Since the training accuracy is too low, softmax regression is likely optimizing the wrong function, indicating that our data may not be linearly separable.

3.1.2 Support Vector Classifier

Since our data may not be linearly separable, we continued with a one-vs-one scheme multi-class SVM classifier, which used a Gaussian kernel. We varied our sample size and separated our training and testing data with a 70/30 random split to analyze the behavior of the training and testing accuracies. Our SVM aimed to optimize the function on the right, with $C = 1$, $\gamma = 0.2$, and $m =$ sample size.

$$\arg\min \left[\sum_{i=1}^{k} K(w_i, w_i) + \frac{C}{m} \sum_{i=1}^{m} \xi_i \right]$$

$$\text{s.t. } \forall y \in [1, \ldots, k] : K(x_1, w_{y_1}) \geq K(x_1, w_y) + \gamma * [1 - 1\{y = y_1\}] - \xi_1$$
$$\vdots$$
$$\forall y \in [1, \ldots, k] : K(x_m, w_{y_m}) \geq K(x_1, w_y) + \gamma * [1 - 1\{y = y_m\}] - \xi_m$$

3.1.3 Results

Sample size	Test accuracy	Train accuracy
1,500	54.7%	98.6%
10,000	60.2%	94.9%
25,000	63.5%	93.2%
50,000	67.0%	91.5%
75,000	70.3%	91.0%
100,000	73.6%	90.5%

Table 1: As the sample size increased, the training accuracy decreases at a slower rate than the testing accuracy increases.

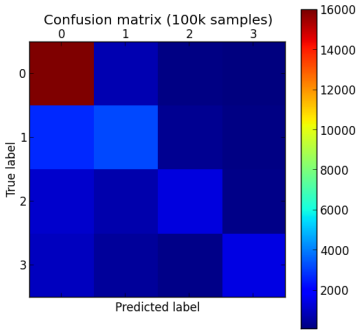

Figure 1: Shows the confusion matrix for the testing data for the 100k model

The SVM performed consistently better than softmax. Training accuracies are asymptotically approaching 90% (Table 1). However, the quick increase of test accuracy suggests that more data may see continued increase. The disparity between the training and testing accuracy, indicating high variance, suggests that we may need either more data or a different set of features.

The confusion matrix, for the testing data, shows that most samples for each class were classified correctly. Misclassification tends to predict a lower class, likely due to the data's left skew.

To explore this high variance, we ran the SVM classifier on four of the five features, because a smaller set of features might improve our accuracy. Additionally, we used Principal Component Analysis (PCA) to project our data to four principal component vectors.

Trial	− PI citation count	− # authors	− journal IF	− age	− # references	PCA
50k (train)	75.2%	85.7%	87.7%	84.2%	77.1%	84.2%
50k (test)	66.7%	67.2%	65.2%	62.2%	66.1%	62.3%
100k (train)	74.1%	84.8%	86.8%	83.4%	76.4%	83.0%
100k (test)	68.2	70.8%	71.1%	66.9%	68.2%	66.6%

Table 2: training and testing accuracies for modeling using one step of backwards search, indicated by − *feature*, and PCA to project data to four components.

3.2 Additional Techniques

3.2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA), a generative model for natural language processing, allows sets of observations to be explained by unobserved groups; these groups, in turn, explain similarities between parts of the data [11]. If it observes words collected into documents, it posits each document as a mixture of a small number of topics and attributes each word's creation to one such topic (shown on right). We utilized a version of the algorithm that assigns a given paper the probability it appears in 150 different biology topics. We then assigned each paper to the topic that it most likely belonged to and ran the above classification methods in each topic.

Input: words $\mathbf{w} \in$ documents $\mathbf{d}$
Output: topic assignments $\mathbf{z}$ and counts $n_{d,k}, n_{k,w}$, and n_k
begin
  randomly initialize $\mathbf{z}$ and increment counters
  foreach *iteration* **do**
    for $i = 0 \to N-1$ **do**
      $word \leftarrow w[i]$
      $topic \leftarrow z[i]$
      $n_{d,topic}$-=1; $n_{word,topic}$-=1; n_{topic}-=1
      for $k = 0 \to K-1$ **do**
        $p(z = k|\cdot) = (n_{d,k} + \alpha_k)\frac{n_{k,w}+\beta_w}{n_k+\beta \times W}$
      end
      $topic \leftarrow$ sample from $p(z|\cdot)$
      $z[i] \leftarrow topic$
      $n_{d,topic}$+=1; $n_{word,topic}$+=1; n_{topic}+=1
    end
  end
  return $\mathbf{z}$, $n_{d,k}, n_{k,w}, n_k$
end

3.2.2 Validation and Model Selection

To cross-validate, we placed 70% of the data into 8 different folds, using every combination of 7 folds to train 8 different models. We then chose the model that performed best on the odd fold out and tested it on the remaining 30% of our data. Because of computational and time limitations, we only utilized this method on the smaller sized clusters to add robustness.

We also used grid search to improve model selection. Some parameters of the model, such as choice of the constants C, γ, and the kernel function in SVM, are not selected by training but rather are design decisions. Grid search tries every possible combination of these parameters and trains each model separately, using 5-fold cross-validation accuracy as a scoring mechanism. We then chose the model with the highest accuracy.

Softmax grid searched over the combinations of the following parameter values: penalty function, L1 and L2; and C=0.05, 0.1, 80, 1000. SVM chose the following parameters for the Gaussian kernel: γ used 6 points on a logarithmic scale from 10^{-6} to 10^{-1}; C=1, 10, 100, 1000.

3.2.3 Results

Method	Test accuracy	Train accuracy
Softmax	57.9%	63.1%
Softmax (CV)	63.9%	66.4%
Softmax (GS)	63.9%	66.0%
SVMC	57.5%	96.2%
SVMC (CV)	60.1%	94.6%
SVMC (GS)	62.4%	71.4%

Table 3: Average topic accuracy, weighted by sample size per topic. (CV) indicates cross-validation; (GS) indicates grid-search.

The numbers displayed on the left are the average accuracy between samples weighted by the number of papers per topic. Note that disparities exist between topics in accuracy. Highest accuracies were, for SVMC, 99% train/80% test and, for softmax, 92% train/85% test. Lowest accuracies were for SVMC 91% train/11% test and, for softmax, 51% train/14% test. Cross-validation and grid-search slightly helped training accuracy.

4 Discussion

Predicting article citation counts helps explain dynamics of academia. Previous bibliometric studies use small data sets, utilize regression to predict specific counts, and handpick certain topics. This research provided a novel method to estimate the citation count of biological sciences literature. The method accurately classified 73% of papers into a "bucket" of citation counts, modeling impact. It utilized a data set of unprecedented size in biological bibliometrics; approached it using classification, not regression; and understood the impact of topical clustering on classification accuracy.

Feature analysis helps demonstrate which features more significantly determine accuracy. Principal component analysis lowered training and testing scores, suggesting that no pairs of our features have a strong linear relationship. Additionally, a step of backwards search showed that the paper's age, PI career citation count, and number of references strongly indicate classification. Removing features showed that the other two features (journal impact factor and number of authors) are not strong class indicators, because testing accuracy remained near the original model's. Since PCA indicated no strong linear relationship between these two features and the others, they likely were independent features and did not significantly affect the model. This suggests that at least within biology, neither the reputation of the journal (as measured by impact factor) nor the number of authors significantly impacts a paper's citation count – two significant findings in bibliometrics. These tests also suggest that we may benefit from changing our features. Because removing features didn't help the testing accuracy converge, more data is required to optimize the SVM classifier.

Interestingly, topic clustering did not improve classification test or train accuracy. In reality, there was a 15% decrease in testing accuracy from the entire data set to average of the topic based models. Even in the largest topic, with around 1500 training examples, testing accuracy was only 47%, 7% less accurate than running our original model on 1500 samples. This suggests that topic clustering did not help build a better model, because it should group papers within similar circles of academic research and activity and therefore provide better accuracy. One potential implication is that there may be a large number of interdisciplinary work between biological topics: even if one topic has significantly more papers, citing work in other topics would somewhat evenly distribute citations across topics. More research is needed in examining links between topics to understand this distribution. However, more papers in each topic may improve model accuracy. Since there are 150 topics and only about 110000 papers in our final data set, the vast majority of topics only had a few hundred papers. Consequently, though test error was high, larger data sets increased the model's accuracy over the entire data set, and significantly increasing the number of papers in each topic may have a similar effect.

Finally, note that even though we used previous years for feature data and age as an input variable, the model can still be used for prediction. Because the paper's age is given to the model, to predict the number of citations of a newly published paper in x years, we can input x as the paper age. Additionally, other variable feature data, such as the principal investigator's career citation count, will only increase into the future – so at worst, our model generates a lower bound for paper performance.

5 Future Work

Generally, more data will increase the robustness of our findings and potentially create a more accurate classifier. Notice that as the size of training data increased when using the SVM classifier, testing accuracy increased faster than training accuracy decreased and never converged. More papers would thus help build a better classifier over the entire data set and within each topic.

Changing features may also benefit the model's accuracy. Characteristics of the paper's form, like its syntactical structure (active versus passive voice), paper length, and number of images, could all influence citation [2]. Additionally, other common bibliometric methods could be used. An example is the H-index, a quantification based on the set of the scientist's most cited papers and the number of citations received in other publications [1]. Furthermore, some quantification of the author's institution's reputation may provide interesting results. The output variable could also be changed. Rather than counting immediate or "primary" citing papers, we could also count "secondary" citations that cite "primary" papers, and determine some heuristic to combine this with "primary" citation count.

Finally, in response to the seeming ineffectiveness of topics, social network analysis could help understand the relationships between papers of different subjects. This would demonstrate which topics are strongly related and thus find the most popular subjects for interdisciplinary research.

6 References

1. Cheek J, Garnham B, Quan J (2006) What's in a number? Issues in providing evidence of impact and quality of research(ers). Qual Health Res 16: 423–435. doi: 10.1177/1049732305285701

2. Falagas ME, Zarkali A, Karageorgopoulos DE, Bardakas V, Mavros MN (2013) The Impact of Article Length on the Number of Future Citations: A Bibliometric Analysis of General Medicine Journals. PLoS ONE 8(2): e49476. doi:10.1371/journal.pone.0049476

3. Garfield E (1996) How can impact factors be improved? BMJ 313: 411–413. doi: 10.1136/bmj.313.7054.411

4. Willis DL, Bahler CD, Neuberger MM, Dahm P (2011) Predictors of citations in the urological literature. BJU Int 107: 1876–1880. doi: 10.1111/j.1464-410x.2010.10028.x

5. Van Raan A (2006) Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. Scientometrics 67:3 491-502. doi: 10.1007/s11192-006-0066-4

6. Kulkarni AV, Busse JW, Shams I (2007) Characteristics associated with citation rate of the medical literature. PLoS One 2: e403. doi: 10.1371/journal.pone.0000403

7. Filion KB, Pless IB (2008) Factors related to the frequency of citation of epidemiologic publications. Epidemiol Perspect Innov 5: 3. doi: 10.1186/1742-5573-5-3

8. Bhandari M, Busse J, Devereaux PJ, Montori VM, Swiontkowski M, et al. (2007) Factors associated with citation rates in the orthopedic literature. Can J Surg 50: 119–123.

9. Europe PubMed Central. EBI Europe PMC Web Service 3.0.1. Online.

10. CiteFactor. 2014 Impact Factor List. Online.

11. Blei D, Ng A, Jordan M (2003) Latent Dirichlet allocation. ML Research 3:4-5. p. 993-1022. doi:10.1162/jmlr.2003.3.4-5.993