IJITCE

ISSN 2347- 3657

International Journal of
Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657

Volume 8, Issue 2, June 2020

Prediction of future citation count with machine learning and neural network

Mrs. Mukka Shirisha , Mrs. Gangula Pavani, Mrs. Lakshmi Lavanya Tumu

Abstract—

The ability to forecast a paper's future citation impact is gaining traction in the academic world. To simplify the process of predicting future citation counts, we choose a binary approach in this study. Job that requires categorizing. The research relies on data collected from 2,600 physiology-related publications found on the Web of Science. Only eight bibliometric parameters of papers cited in the first three years following publication were considered. There are three machine learning models and a neural network developed to see how well these features predict future citation counts. The experimental outcome demonstrates the utility of the chosen characteristics in predicting future citation counts. Predicting future citation counts is a challenging task, but machine learning and neural networks can help.

I. INTRODUCTION

Around the globe, nations invest in scientific study, seeing it as a key factor in their own progress. On this backdrop, increased number of scientists spends their time and energy doing research and publishing articles in a wide variety of academic disciplines, each of which adds to the body of literature and has its own unique impact. Assessment criteria are required in the academic community to determine the worth and significance of the works. There are several sorts of \evaluation approach, and citation count is the most popular one \sand widely utilized. There are citation-based studies of scientific impact and innovation [1] [2] [3] [4]. However, citation numbers are inadequate as measures of a paper's quality since they gloss over crucial details. Experts have researched many indicators, such as the h-index [5], g-index [6], r-index [7], and hm-index [8], to help with this challenge. Authorship and cooperation have been the focus of a number of studies [9] [10] [11] [12]. The co-citation network has been the subject of research [13]. Predicting how many times a work will be cited in the future is becoming more important in the academic world. Depending on your goals, this forecast will have more or less significance. Keeping up with the latest developments in an area of study would allow scientists to adequately plan their future investigations [14]. The university or funding agency might then assess the results over time. Recruitment And funding decisions based on an author's research potential [15], [16]. Recent studies have shown that a paper's citations may be predicted by looking at its authors, topics, length, language, and references. Citation counts in the first three years were shown to be predictive of future citation counts by Abram et al. [22].

First, eight bibliometric traits are developed by extracting citations from papers written at an early stage, and their ability to predict citation counts for physiological studies is tested. Second, to check the work with many citations, the job is recast as a simple binary classification. In the third step, four machine learning models (two support vector machines, one logistic regression, and one neural network) are built to see whether they can correctly classify papers based on their citation counts.

1,2,3 Assistant Professor
1,2,3 Department of CSE
1,2,3 Global Institute of Engineering and Technology Moinabad, Ranga Reddy District, Telangana State.

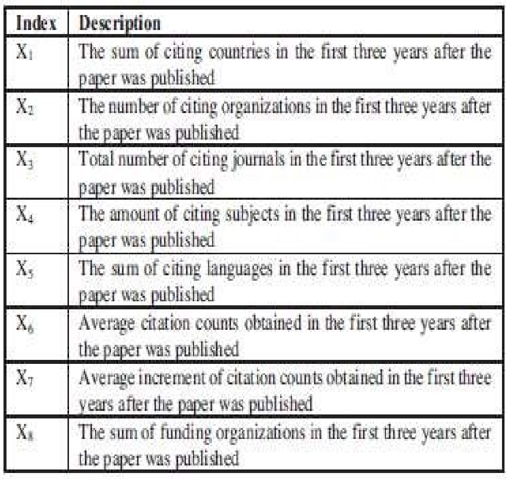
II. DATA SET

Twenty-six Web of Science physiology journals were used for this analysis. There are nine journals chosen from the top tier of the Journal Citation Reports, and they include the Annual Review of Physiology and the Journal of Pineal Research. Journals like "Research," "Comprehensive Physiology," "Reviews of Physiology Biochemistry and Pharmacology," "International Journal of Behavioral Nutrition and Physical Activity," "Act Physiological," "Journal of Physiology—London," and "Exercise and Sport Sciences Reviews" are all included. Included are the journals Hypertension in Pregnancy, Fish Physiology and Biochemistry, Pediatric Exercise Science, Physiological Research, the Korean Journal of Physiology and Pharmacology, Respiratory Physiology and Neurobiology, the Journal of Musculoskeletal and Neuronal Interactions, the Journal of Biological Regulators and Homeostatic Agents, the Journal of Lymphatic Research and Biology, the Archives of Insect Biochemistry and Physiology, and the General and Comparative Physic Web of Science reports that 2,600 downloads have been made from 2011 articles in these journals. For the period between 2011 and 2019, these works are sorted by their rising number of citations. Next, we use the increase in citations to sort the publications into two categories: high-impact papers (TIP) and low-impact papers (LIP). According to the definition of TIP, only the articles in the top 20 percent in terms of citation counts increase. An article is considered LIP if its citation count has increased by at least 20% in the 20% of downloaded papers since its first downloads. The experimental differentiation between TIP and LIP is predicated on this procedure.

III. FEATURE SPACE OF CLASSIFICATION

Eight early-stage bibliometric indices of citation document build the future space for characteristics to forecast whether a publication will be a TIP or LIP. Prominent among the first works of effectiveness predicts citation volume in the future. Early citations show that the article has been accepted by the academic community, increasing the likelihood that it will be referenced in the future [23]. Web of Science's "analyze the result" and "generate citation report" features were used to compile the following metrics. Table I shows the results of X1's analysis of the geographic distribution of citations to scholarly articles as a proxy for the latter's worldwide sway in the academic community. X2 measures the number of organizations of referencing articles which is designed to assess the organization's attention on the paper. X3 displays the number of journals of referencing article which \measures the academic effect on journals in and out the \study area. X4 displays the number of themes which is \scouted from the citing publications. X5 represents the total number \soft languages of cited publications. X6 provides the average citation \scouts in the first three years after the work was published. Index of X7 represents the average growth of citation counts \sin the first three years following the publication of article. X8 \illustrates the quantity of funding organizations in the first three years after the study was published.

TABLE I. FEATURE SELECTION



Index	Description
X_1	The sum of citing countries in the first three years after the paper was published
X_2	The number of citing organizations in the first three years after the paper was published
X_3	Total number of citing journals in the first three years after the paper was published
X_4	The amount of citing subjects in the first three years after the paper was published
X_5	The sum of citing languages in the first three years after the paper was published
X_6	Average citation counts obtained in the first three years after the paper was published
X_7	Average increment of citation counts obtained in the first three years after the paper was published
X_8	The sum of funding organizations in the first three years after the paper was published

IV. MODEL

A. Machine Learning Model

1) SVM

For linear SVM, if $(x^{(i)}, y^{(i)})$ is linear inseparable, a \slack variable $\xi_i \geq 0$ was included to make the sum of\function margin and slack variable equals to 1. Constrain assuming that:

$$y^{(i)}(w \times x^{(i)} + b) \geq 1 - \xi_i \qquad (1)$$

Pay a cost of C for each ξ_i, and the objective function is:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \qquad (2)$$

The optimization objective is:

$$\max_{w,b,\xi_i} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \qquad (3)$$

2) Decision Tree

Let D be a set of samples, the percentage of class k is p_k. the information entropy is:

$$Entropy(D) = -\sum_{i=1}^{m} p_k log_2 p_k \qquad (4)$$

Partition the samples of D with the feature A and the information gain is:

$$igain(D,A) = Entropy(D) - \sum_{P=1}^{mP}\frac{|D_p|}{|D|}Entropy(D)\ (5)$$

3) Random Forest

We will train a decision tree using samples drawn from the training set at random with replacement, with the total size of the training set equaling. Assume the features in the training set are, and Train each decision tree using a random feature drawn from the set of features. The best feature is picked from, and the decision deer is then divided. Each decision tree will only continue to split until all of the training samples for a given node are from the same class. In a decision tree fork, pruning is unnecessary.

4) Neural Network

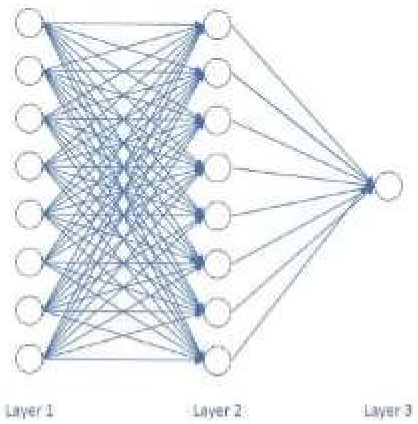


Fig.1. Neural Network

A basic neural network has three layers, as seen in Fig.1. Eight neurons make up the input layer, which is utilized to take in data. The second layer is a completely dense affiliated with the top layer, and containing eight neuronal cells both the first and second layers use a Rectified Linear Unit Activation Function (Rely). In order to optimize this neural network, Adam is used. One Sigmoid-activated neuron sits in the output layer. In this neural network, binary_ cross entropy serves as the loss function. The experiment uses a batch size of 16, and the length of time between epochs is 30.

V. EXPERIMENTAL RESULT

Articles from the TIP group are labeled with a "1," whereas papers from the LIP group are labeled with a "0"; these labels serve as the foundation for the calculations made in the respective models. The information is separated into 80 percent of the time should be spent on training, and 20 percent on testing .A total of 832 trains and 208 tests were used to compile Table. Accuracy is the yardstick by which the performance of the machine learning model and neural network are measured.

TABLE II. TRAIN SET AND TEST SET

TIP	LIP	Train set	Test set
520	520	832	208

The three machine learning classifiers of SVM, Decision \street, and Random Forests and a neural network classifier are\conducted to examine the performance concerning predicting TIP by means of the specified attributes. Table III presents the prediction \results of these attributes based on the preceding four classifiers.

TABLE III. ACCURACY OF CLASSIFIERS

Model	Accuracy
SVM	0.9846
Decision Tree	0.9884
Random Forest	0.9891
Neural Network	0.9942

For this binary classification job with the chosen features, it is clear that all four classifiers perform well. A neural network achieves a 0.9942 precision, which is the highest achievable. A Random Forests machine learning classifier has a 0.9891 accuracy rate. Machine learning's SVM classifier achieves 0.9846 accuracy. In comparison to the SVM classifier, the accuracy of the Decision Tree is 0.9884, which is much better. Over 0.9 accuracy is achieved while running all classifiers. This proves that the chosen characteristics are useful for TIP categorization.

VI. DISCUSSION

In order to categorize the TIP, we choose eight bibliometric parameters to use as our feature space. After that, three machine learning models and a neural network are fed the features. The According to the results of the experiments, it is possible to use these factors characterizing the paper's early performance to foretell its TIP.

Both the machine learning model and the neural network have an accuracy of above 90%, demonstrating that they are able to pick up on the salient characteristics of the job at hand. From the data obtained in experiments, it can be inferred that the chosen characteristics are valid for classifying the TIP. Index X2 is a database that catalogues references to publications published in the first three years. This indicator shows how many universities and colleges are paying attention. Academic institutions may take note of a paper's originality and importance early on if it creates a new area of study or sub-discipline, or proposes a novel approach for addressing an academic topic. In addition, early recognition from a prestigious academic institution might boost a paper's citation count for years to come [24]. It is possible for several groups within the academic community to do research in the same or distinct areas of science. They could learn something new from this recently released research, which would encourage them to go further into the topic and find even more strategies for resolving the situation at hand. On the other hand, some institutions construct alternative hypotheses in order to initiate a study in opposition to the referenced work and open up new topics of research. It's possible that the paper's citation total will rise as a consequence of all these activities. Data from X8 displays the aggregate of the first three years of funding agencies cited in the research. The financing mechanism aids the study that contributes to a major advancement in a subject area, develops cutting-edge technical capabilities, or provides novel approaches to resolving a pressing issue identified in prior research. When a financial agency backs a research project, it shows that they believe it will have a substantial impact in its area and that the project's results will be consistent with those of past studies [25]. This means that the publication might be cited by other researchers interested in contributing to the same field of study.

Number of nations and languages using the work as a reference in the first three years are shown by the X1 and X5 citation indices, respectively. These indicators support earlier work [26] by gauging the extent to which a piece of scientific study has been disseminated internationally. Papers written in other nations citing this one would be published in a variety of languages, with citations included. With ongoing international research, the paper's impact on the academic world is only expected to grow.

Three-year and four-year X3 and X4 indices respectively reflect the total number of publications and topics from which a given work has received citations. The article is representative of a field's research trend when other it may have been cited often in the early days of publication because researchers were eager to reference it from their own work, and the number of citations may rise in the future. Referencing articles from other fields and journals shows that the paper's methods or ideas may be used in other areas of study. Because of this broad applicability, selected papers may be cited and utilized in the future advancement of a variety of research fields. This demonstrates that the journal's and article's topic matter do have an effect on how often it is cited in the future [27].

The attention level in the subject of study is directly reflected in the average citation counts and early age increase of X6 and X7, and this is connected with future citations [23].

VII. CONCLUSION

Using early bibliometric measures, this article treats the prediction of TIP as a simple job of binary classification. Based on the cited study, eight bibliometric indicators are chosen. Within the first 3-year period after the papers initial publication. Binary feature learning is performed using a Support Vector Machine (SVM), Decision Tree, Random Forest, and a neural network. Using these bibliometric indicators of citing papers, we were able to successfully anticipate the TIP and show that a paper's early performance may be utilized to forecast its future citations based on our experiments. Further, the prediction of TIP may make use of machine learning and neural networks. The research is limited in that it relies on physiology publications culled from Web of Science, and not all physiology-related journals are included in Web of Science.

These publications focused on medical research; therefore their findings may not be generalizable to other disciplines. The findings, however, indicate that the bibliometric data of a paper's early age may be used to forecast future TIP, and thus gives a point of departure for further research into citation prediction.

REFERENCES

[1] E. Garfield and C. Emeritus, "The use of journal impact factors and citation analysis for evaluation of science," 41st Annul. Meet. Count. Biol. Ed., 1998.

[2] C. Castillo, D. Donator, and A. Goings, "Estimating number of citations using author reputation," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2007, vol. 4726 LNCS, no. October, pp. 107–117.

[3] D. Wang, C. Song, and A. L. Barabbas, "Quantifying long-term scientific impact," Science (80-.). vol. 342, no. 6154, pp. 127–132, 2013.

[4] H. F. Mode, "New developments in the use of citation analysis in research evaluation," Archive Immunological ET Therapies Experimentalism, vol. 57, no. 1. pp. 13–18, 2009.

[5] J. E. Hirsch, "An index to quantify an individual's scientific research output," vol. 102, no. 46, pp. 16569 16572, 2005.

[6] L. Egghe, "Theory and practice of the g-index," Scientometrics, vol. 69, no. 1, pp. 131–152, 2006.

[7] B. H. Jin, L. M. Liang, R. Rousseau, and L. Egghe, "The R- and A indices: Complementing the h-index," Chinese Sci. Bull., vol. 52, no. 6, pp. 855–863, 2007.

[8] M. Schreiber, "A modification of the h-index: The him-index accounts for multi-authored manuscripts," J. Informer., vol. 2, no. 3, pp. 211– 216, 2008.

[9] X. Liu, J. Bollen, M. L. Nelson, and H. Van De Sompel, "Coauthor ship networks in the digital library research community," Inf. Process. Manage. vol. 41, no. 6, pp. 1462–1480, 2005.

[10] E. Yan and Y. Ding, "Applying centrality measures to impact analysis: a coauthor ship network analysis," J. Am. Soc. Inf. Sci. Technol., vol. 60, no. 10, pp. 2107–2118, 2009.

[11] C. Hsiang and H. Rebecca, "Quantifying the degree of research collaboration: A comparative study of collaborative measures," J. Informer., vol. 6, no. 1, pp. 27–33, 2012.

[12] L. Wildcard, J. W. Schneider, and B. Larsen, "A review of the characteristics of 108 author-level bibliometric indicators," Scientometrics, vol. 101, no. 1, pp. 125–158, 2014.

[13] G. González-Alcaide, A. Calafat, E. Become, B. This, and W. Glaze, "Co-citation analysis of articles published in substance abuse journals: Intellectual structure and research fields (2001–2012)," J. Stud. Alcohol Drugs, vol. 77, no. 5, pp. 710–722, 2016.

[14] T. Amjad, Y. Ding, J. Xu, C. Zhang, and A. Daud, "Standing on the shoulders of giants," J. Informer., vol. 11, no. 1, pp. 307–323, 2017.