A Temporal Metric-Based Efficient Approach to Predict Citation Counts of Scientists

Saumya Kumar Dewangan(✉), Shrutilipi Bhattacharjee, and Ramya D. Shetty

Department of Information Technology, National Institute of Technology Karnataka, Surathkal, Mangaluru, India
{saumyakumardewangan.212it025,shrutilipi, ramyadshetty.207it004}@nitk.edu.in

Abstract. Citation count is one of the essential factors in understanding and measuring the impact of a scientist or a publication. Estimating the future impact of scientists or publications is crucial as it assists in making decisions about potential awardees of research grants, appointing researchers for several scientific positions, etc. Many studies have been proposed to estimate publication's future citation count; however, limited research has been conducted on forecasting the citation-based influence of the scientists. The authors of the scientific manuscripts are connected through common publications, which can be captured in dynamic network structures with multiple features in the nodes and the links. The topological structure is an essential factor to consider as it reveals important information about such dynamic networks, such as the rise and fall in the network properties like in-degree, etc., over time for nodes. In this work, we have developed an approach for predicting the citation count of scientists using topological information from dynamic citation networks and relevant contents of individual publications. This framework of the citation count prediction is formulated as the node classification task, which is accomplished by using seven machine learning-based classification models for various class categories. The highest average accuracy of 85.19% is achieved with the XGBoost classifier on the High Energy Physics - Theory citation network dataset.

Keywords: Citation networks · Citation count · Node classification · Directed and weighted networks · Temporal networks

1 Introduction

Citation analysis is a method of measuring the importance or influence of an author or published articles by counting the number of times other works cite this author or publication. It is analyzed for various purposes, such as to evaluate the impact of a particular work or a scientist, how much the related research

Published by Springer Nature Switzerland AG 2023
I. Maglogiannis et al. (Eds.): AIAI 2023, IFIP AICT 675, pp. 343–355, 2023.
https://doi.org/10.1007/978-3-031-34111-3_29

area is impactful in the future. An essential objective of citation analysis is to make decisions about giving grants, accepting appointments, etc. Citation count is a well-known measure of such scientific impact. The h-index and i10-index are crucial metrics for the impact analysis of researchers or research outcomes, which are based on the citation count [1]. The citation is a consequence of referring to some article and can be thought of as directed links between the referred and the referencing objects and eventually constructing networks. Citation networks can be broadly categorized into two networks; in a paper-based citation network, the graph is always acyclic because an article can refer to another article that is already published. However, in an author-based citation network, the graph can be cyclic also, as two authors can cite each other's work reciprocally in the same time frame. A self loop may also exist if authors cite their previously published work. These networks are dynamic in nature, such that nodes may get added/removed, and the structure and the weights on the links change from one time frame to the other. The existing works on citation count prediction [2–4] mainly use the content information in the publication, such as *abstract, title, keywords*, etc., for the analysis. However, since such networks are evolving over time, the topological properties [5,6] of such networks are also changing, which can reveal an essential pattern for the citation count prediction task. It also tells about the increasing and decreasing trends in the citation count of the authors over time. In this work, a new approach is proposed for predicting the citation count of the scientists by utilizing the temporal metrics from the topological properties of the author-based citation network. Then, the prediction of citation count is formulated as a node classification problem, accomplished using various machine learning (ML)-based classification models, such as logistic regression, decision tree, random forest, nearest neighbor, support vector machine, multilayer perceptron, and XGBoost.

1.1 Background Study

It is found from the existing literature that the dynamic and complex citation networks have drawn a lot of interest in fields like mining and evaluating scientific activities, promoting authors and papers to researchers, estimating the number of citations an author or paper will receive, etc.

Some studies [2–4] have considered the content present in the published articles for the task of citation count prediction. They use the information in the papers, like *title, abstract, index terms*, etc., for the prediction. Bhat *et al.* [4] proposed methods based on classification and created a predictive technique for predicting the citation count of scientists based on classification models. Using different characteristics, they analyzed how *author influence, author interdisciplinarity, and title terms* affected citation counts. They achieved a training accuracy of 88.7% with the classification tree model.

Some studies [7,8] have used the information present in the graph structure of the citation networks. They have used features like closeness centrality, betweenness centrality, etc., to predict citation counts. Zhu *et al.* [8] suggested a citation count forecasting model based on academic network characteristics.

They have considered multiple features, such as the paper feature, author feature, network feature, etc., and examined the importance of each feature for the task of citation count prediction. Then, they compared the performance with different prediction algorithms and found that the SVM was the best model for their dataset and achieved an 88.87% coefficient of determination.

Some studies [9,10] have addressed the problem of citation count prediction in dynamic citation networks. They have used the link prediction technique for the prediction of citation count. Kaya *et al.* [9] proposed an approach for predicting the citation count of the scientist in a directed, weighted, and dynamic citation network. They introduced a dynamic proximity metric for the classifiers to predict citation count and some basic topological properties. The dynamic metric is based on rising and falling trends in citation networks throughout transitional time frames. They achieved an area under curve (AUC) score of 0.836 with a random forest classifier for the Aminer-Citation network. Bütün *et al.* [10] presented an approach based on the link prediction problem for the scientist's citation count prediction using a supervised learning method. They have developed a temporal link prediction measure using topological properties in complex networks at the local and global levels. Additionally, they compared how well the suggested link prediction measure performed in anticipating new links in complex networks with five other widely used link prediction measures. The highest area under the receiver operating characteristic curve (AUROC) value of 0.872 was achieved with a random forest model for the Aminer-Citation network.

It has been found that many studies are reported for the citation analysis of the papers, and limited studies have addressed the citation analysis of the authors or scientists. Many works [2–4] have considered the content information in the publications to predict its future impact and did not utilize the network's topological properties. Some studies [7,8] have also examined the dynamic network's structural characteristics for predicting the scientists' citation count. Few works [9,10] have used the link prediction method for citation count prediction. In this study, we have made the following contributions:

1. Our study has considered the dynamic structure of the citation networks and relevant contents of the publications to predict the citation count of the scientists.
2. We have created a temporal metric based on various temporal events happening in the dynamic citation network.
3. The problem of predicting the citation count of scientists is formulated into a node classification problem. The temporal metric is used for the node classification task. The node classification task is accomplished by using and comparing different ML-based models.
4. The citation count prediction is also considered a regression task and analyzed for different feature sets using a linear regression model.

The rest of the paper is organized as follows. Section 2 discusses the proposed methodology, and Sect. 3 consists of the empirical setup and results. Finally, Sect. 4 concludes the work.

2 Proposed Methodology

This study has a sequence of stages for predicting the citation counts of scientists. Figure 1 represents the outline of the proposed approach for solving the problem step by step. In the first stage, we have a paper-based citation network. Then, it is converted to an author-based citation network in the second stage. The temporal metric is calculated for all the authors in the third stage, as discussed in Sect. 2.3. The past citation count of the scientists is utilized over time for creating the temporal metric. The temporal metric and the content based information from the publications, such as *title and abstract*, are utilized as the features of an author. In the fourth stage, the problem is framed as a node classification problem, which is achieved using ML-based classification models. A thorough description of these steps are discussed in the following subsections.

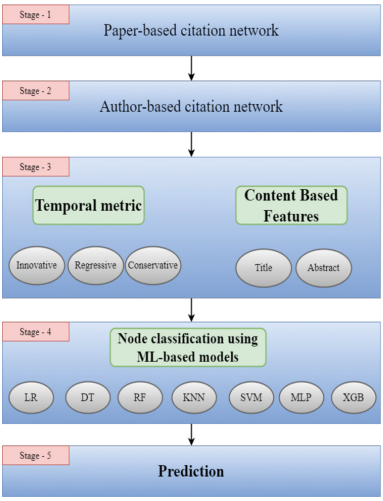


Fig. 1. Proposed methodology for the prediction of citation counts of the scientists

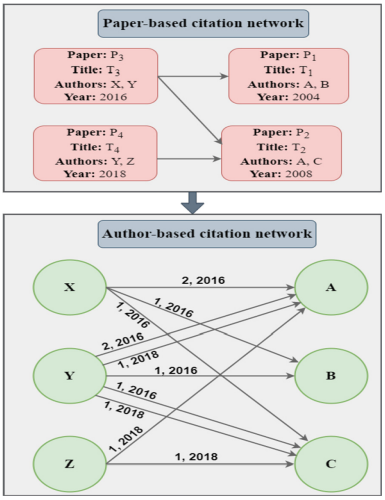


Fig. 2. Conversion of paper-based citation network to author-based citation network

2.1 Paper-Based Citation Network

In a paper-based citation network, the papers published in various conferences, journals, etc., constitute the network nodes, and the directed edges show the citation among the papers. The details present in an article are the *paper title, authors, year of publication, etc.*, which act as the node's features in the paper-based citation network. An example is shown in Fig. 2, consisting of four papers, P_1, P_2, P_3, and P_4, and there is a directed link from paper P_3 to paper P_1, which indicates the paper P_3 has cited paper P_1. The papers as the nodes and their mutual citations as the links construct the entire paper-based citation network

from the given datasets. Since our objective is to predict the citation counts of the scientists rather than predicting the citation count of the articles, the paper-based network is converted into an author-based citation network.

2.2 Author-Based Citation Network

To predict the citation count of the scientists, the author-based citation network is generated from the paper-based citation network. In an author-based citation network, the scientists are represented as vertices, and the directed edges show the citations between authors. The directed links include multiple properties, such as weight and time, as shown in Fig. 2. The weight attribute indicates the number of times a scientist has cited the papers of another scientist. Consider an example paper-based citation network shown in Fig. 2. There is an edge from author X to author A with weight and time attributes as 2 and 2016, respectively; it indicates that author X has cited author A two times in the year 2016. The author-based citation network is defined as $G_t(N, E')$ in the time frame t, which is a directed and weighted network, where N is the set of scientists and E' is the set of edges. Each link (i, j) present in E' represents a quadruple of the form (i, j, w, t), where $i, j \in N$, w is the weight attribute, and t is the time instance. Here, every i^{th} node in N is represented by N_i. The citation count of each node i in time frame t is represented by $CC(N_i)_t$. $G_{t,t'}(N, E)$ is the author-based citation network from time frame t to t'; hence it is a directed, weighted, and temporal network. There will be a total of T time frames (where, T depends on the time frame considered in a given dataset) in $G_{t,t'}(N, E)$, each with a window size of s.

The approach to making an author-based citation network from the paper-based network (as shown in Fig. 2) is explained in the following example. If there is an edge from the paper P_3 to a paper P_1, then there will be an edge from all the authors of paper P_3 to all the authors of paper P_1, and the time attribute for the edges will be the time of publication of the paper P_3. Considering the network with four papers, P_1, P_2, P_3, and P_4. For each unique author present in the paper-based citation network, there is a node in the author-based citation network, represented as authors A, B, C, X, Y, and Z. Let us consider the relationship between the authors Y and A. Author Y is present in two papers, P_3 and P_4, published in the year 2016 and 2018, respectively. The weight attribute corresponding to the edge in the year 2016 is 2 because author A is present in the 2 papers cited by Y in 2016, and the time attribute is 2016. Similarly, the weight attribute for the edge in 2018 is 1 because author A is present in 1 paper cited by Y in 2018.

2.3 Temporal Metric (M)

The concept of temporal events is studied in [11], mainly for the link prediction task, where temporal events are based on the increase or decrease in the weights of the links over time. In this work, the idea of temporal events is extended for the nodes, and variation in temporal events is based on the rise and fall in

citation count of a node over time. The temporal metric calculates the score for a node which is based on the temporal events occurring to the node in the network. The various temporal events, with respect to a node, are described as follows:

- **Innovative Event (I)**:
  An innovative event states that the node (the author) has gained citations in the current time frame t while it did not have any citations in the previous time frame $t-1$. Hence, this event is positively scored by multiplying with a positive constant i, and the innovative score, $I(N_i, t)$, is formulated as follows:

$$I(N_i, t) = i * CC(N_i)_t \qquad if \quad CC(N_i)_{t-1} = 0 \wedge CC(N_i)_t > 0 \quad (1)$$

- **Regressive Event (R)**:
  A regressive event shows that a node has gone through a complete loss in its citation count (zero in the current time frame t) compared to the last state (citation count greater than zero in the time frame $t-1$). This event is negatively scored by multiplying with a negative constant r because the node has lost all its value in the ongoing time frame, and the regressive score, $\mathrm{R}(N_i, t)$, is calculated as follows:

$$R(N_i, t) = r * CC(N_i)_{t-1} \qquad if \quad CC(N_i)_{t-1} > 0 \wedge CC(N_i)_t = 0 \quad (2)$$

- **Conservative Event (C)**:
  Conservative event shows the continuation of acquiring the citations by a node from other nodes in the current state from the previous state. If a node has a citation count greater than zero in the current time frame t and a citation count greater than zero in the last time frame $t-1$, this kind of event is considered a conservative event. Since no complete loss in the node's citation occurs in this event, it is positively scored by multiplying with a positive constant c. In this event, three cases may arise, citation count of a node may increase, decrease, or remain the same in the transition from the time frame $t-1$ to time frame t. If the citation count increases, then the event is rewarded by the proportion of the increase in the citation count from time frame $t-1$ to t, and the conservative score is calculated from Eq. (3). If the citation count decreases, a penalty with the proportion of decrease in the citation count is applied due to this event, and the conservative score is calculated from Eq. (4). If the citation count remains the same, then the event is scored positively, and conservative score is calculated from Eq. (5).

$$C_i(N_i, t) = c * (CC(N_i)_{t-1} + \frac{CC(N_i)_t}{CC(N_i)_{t-1}}) \qquad if \quad CC(N_i)_t > CC(N_i)_{t-1} \quad (3)$$

$$C_d(N_i,t) = c * [CC(N_i)_{t-1} + \frac{CC(N_i)_{t-1} - CC(N_i)_t}{CC(N_i)_{t-1}}] \quad if \quad CC(N_i)_t < CC(N_i)_{t-1} \tag{4}$$

$$C_u(N_i,t) = c * CC(N_i)_t \quad if \quad CC(N_i)_{t-1} = CC(N_i)_t \tag{5}$$

The total score for a node, N_i at time frame t is calculated as follows:

$$T(N_i,t) = I(N_i,t) \; or \; R(N_i,t) \; or \; C(N_i,t) \tag{6}$$

The temporal metric for each node, N_i is calculated as follows [10]:

$$M(N_i) = \sum_{t=2}^{n} log(t+1) * T(N_i,t) \tag{7}$$

2.4 Content-Based Features

This study also considers the content-based information in the publications for the citation count prediction. The content present in the papers, i.e., *title and abstract*, are also taken as features of the authors, along with the temporal metric. NLP-based text preprocessing tasks have been carried out for these feature extraction, such as tokenization, removal of stop words, and stemming [12].

2.5 Node Classification

This work is formulated as a classification problem of the scientists based on their citation count. The ML-based classification models are used for the node classification task. The class categories are established at various citation count intervals, as shown in Table 1. The temporal metric is calculated for every single node that acts as a feature for the classification models, and class categories act as labels for the classifiers. The task is to predict the class of a node (author) for a time frame according to the temporal metric for that node, which is calculated based on the events with respect to that node over time frames.

Table 1. Class Categories

Citation count interval	Class category
0	C_0
[1, 50]	C_1
>50	C_2

The idea for deciding the citation count intervals for the class categories is to distinguish between the nodes based on their citation counts. If a node falls in the class category C_0, it means it has not received any citation(i.e., it has a citation count of zero) and is significantly less influential for the time being. If a node belongs to the class category C_1, it is an effective node with a citation count between 1 and 50, both inclusive. If a node exists in category C_2, it has citations above 50 and is highly influential. As required, the number of class categories can vary according to the citation count intervals.

3 Empirical Set-up and Results

This section briefly discusses the dataset, and different machine learning algorithms used in our study and analyzes the results of various machine learning models.

3.1 Dataset Details

1. HEP-TH (High Energy Physics - Theory):
   This network dataset [13] consists of 352,807 links and 27,770 nodes. The paper information is available from January 1993 to April 2003. It also provides meta-information descriptions of the articles like *paper title, abstract, author details, publication date, etc.* The statistics of the dataset are shown in Fig. 3. It shows that citation among scientists has increased over time except for 2003 because of the partial data availability. The dataset has 27,770 papers, but we have obtained meta information of around 26,600 papers from [14].

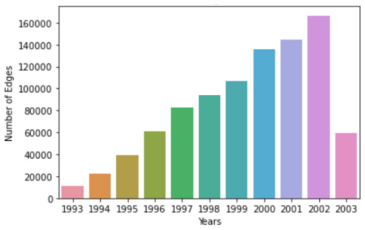


Fig. 3. Statistics of the HEP-TH dataset

3.2 ML-Based Models

The task is to predict the class to which the nodes belong based on the temporal metric and content based features. In this study, we have used seven models, such as logistic regression, decision tree, random forest, nearest neighbor, support vector machine, multilayer perceptron, and XGBoost [15], for the node classification task.

3.3 Experimental Results

Four resampled datasets, DS_1, DS_2, DS_3, and DS_4, are created from the HEP-TH datasets by taking six consecutive years for training and the next one year for prediction. Each resampled dataset has the same time window of size of one year and has different time frames in creating the temporal metric and predicting the citation count. The details of the sampled datasets are given in Table 2.

Table 2. Sampled Datasets Generated from HEP-Th Network

Dataset	Training years for predicting Temporal Metric	Prediction year	Time window (year)
DS_1	1993–1998	1999	1
DS_2	1994–1999	2000	1
DS_3	1995–2000	2001	1
DS_4	1996–2001	2002	1

The experiments are conducted for a different combination of features. The feature set FS_1 consists of the *title* and *abstract* of the publications by the authors, FS_2 consists of the *title, abstract*, and *citation count* of all the previous time frames, and FS_3 contains *title, abstract*, and *temporal metric.* The metrics utilized to evaluate the performance of classifiers are accuracy, precision, recall, and F1 score [16]. There are ten experiments conducted for each of the four resampled datasets, as mentioned in Table 2 with three feature sets. All experiments' train test split is taken as a 75:25 ratio. The average values of the evaluation metrics are taken from the four resampled datasets for each of the three feature sets, and the results are shown in Table 3 and Table 4.

After experimenting with different sets of values of the constants, i, r, and c, we have reported the results with the following values of constants i, r, and c (please refer to Sect. 2.3):

- For the innovative events, the positive constant $i = 1$ is used to score the event positively.
- For the regressive events, the negative constant $r = -0.5$ is used to score the event negatively.
- For the conservative events, the positive constant $c = 0.5$ is used to score the event positively.

The results of the experiments are recorded in Table 3 and Table 4. The average of the results obtained with feature set FS_1 for all the four resampled datasets is presented in Table 3. The XGBoost classifier has given the highest accuracy and precision of 76.39% and 68.32%, respectively, and multilayer perceptron has given the highest recall and F1 score of 61.08% and 61.34%, respectively.

Table 3. Result obtained with FS_1 and FS_2

Models	Results with FS_1				Results with FS_2			
	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Logistic Regression	74.51	67.38	52.99	56.20	82.45	82.97	70.45	74.96
Decision Tree	73.90	60.39	57.28	58.53	82.05	75.50	73.06	74.14
Random Forest	68.58	65.91	37.32	34.21	74.05	73.77	44.28	44.75
Nearest Neighbor	67.41	66.70	37.51	34.68	81.69	79.98	71.78	74.92
Support Vector Machine	74.18	63.55	54.50	57.21	80.17	70.49	56.70	57.97
Multilayer Perceptron	75.25	61.87	61.08	61.34	83.18	77.73	74.75	76.02
XGBoost	76.39	68.32	57.67	61.03	85.19	83.49	75.76	79.00

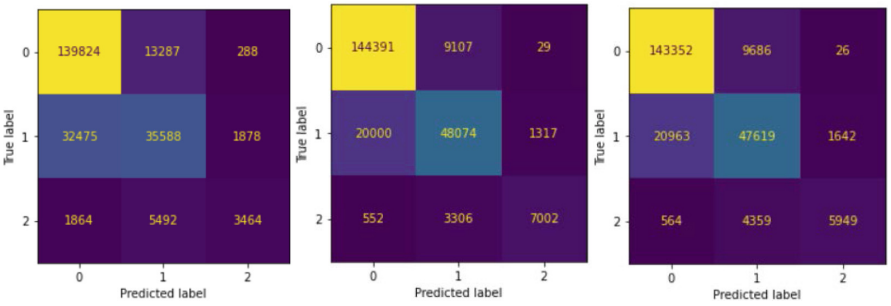
The average of the results obtained with feature set FS_2 for the four resampled datasets are given in Table 3. It has been observed that the XGBoost classifier has performed better than the other models in terms of accuracy, precision, recall, and F1 score and has given the highest accuracy of 85.19%. The results of citation count prediction have improved with the feature set FS_2, as compared with the results of the feature set FS_1.

The average of the results obtained with feature set FS_3 are recorded in Table 4. The XGBoost model has outperformed the rest of the classifiers in terms of all the evaluation metrics used. The XGBoost model has achieved the highest accuracy of 84.09% with feature set FS_3, which consists of the *title, abstract*, and *temporal metric*.

Table 4. Result obtained with FS_3

Models	Accuracy (%)	Precision (%)	Recall (%)	F1 score (%)
Logistic Regression	81.95	80.00	67.29	71.80
Decision Tree	80.69	71.50	69.12	70.15
Random Forest	70.11	68.47	39.09	36.99
Nearest Neighbor	80.75	76.26	66.21	69.86
Support Vector Machine	79.80	64.66	53.31	53.16
Multilayer Perceptron	81.97	74.19	71.50	72.65
XGBoost	84.09	80.64	71.92	75.41

Confusion matrix is displayed to analyze the performance of the classifiers for each of the three classes considered in this study. Figure 4 represents the combined confusion matrix obtained from the XGBoost model over all the four resampled datasets. The confusion matrix corresponding to feature set FS_1 is shown in Fig. 4a. With FS_1, the accuracy achieved for class C_0 is 91.15%, but the accuracy for class C_1 and C_2 is comparatively less. It can be inferred from Fig. 4b with FS_2 that the results are better than FS_1. With FS_2, accuracy for all the classes, i.e., C_0, C_1, and C_2, have better accuracies of 94.04%, 69.27%, and 64.47%. However, results with the feature set FS_3 are comparable with FS_2 for all the classes, and better than the results of FS_1, as shown in Fig. 4c.



(a) FS_1 (b) FS_2 (c) FS_3

Fig. 4. Confusion matrix for the feature sets FS_1, FS_2 and FS_3 respectively

Table 5. Results of a linear regression model with previous six years' *citations* as the features

Dataset	MAE	RMSE
DS_1	4.88	16.76
DS_2	7.30	24.48
DS_3	5.99	16.65
DS_4	7.33	23.43
Average	6.38	20.33

Table 6. Results of a linear regression model with the *temporal metric* as feature

Dataset	MAE	RMSE
DS_1	6.31	26.24
DS_2	9.05	31.93
DS_3	8.67	27.05
DS_4	9.40	29.55
Average	8.36	28.69

From the experiments, it has been found that the results with feature set FS_3 are better than the results with the feature set FS_1, with an increase in accuracy of 7.7%. However, results with the feature set FS_2 and FS_3 have marginal difference in accuracy (1.1%), with FS_2 having better results. The feature set FS_2 has more features as it has *citation count* of all the time frames, and FS_3 has a single feature *temporal metric*, other than *title* and *abstract*. Therefore, FS_3 is better as compared to FS_2 in terms of reduced number features and time complexity for classification models, incurring similar accuracy.

The prediction of citation count of the scientist is also considered as a regression problem, and a linear regression model is used for this task. The metrics used for the performance evaluation of the regression model are mean absolute error (MAE) and root mean squared error (RMSE). These errors, MAE and RMSE, represent the comparison between the predicted and the actual citation counts. The results are reported in Table 5 and Table 6, respectively.

The average MAE of 6.38, and RMSE of 20.33 is achieved with the citation counts of previous six years as the features, and the average MAE of 8.36, and RMSE of 28.69 is achieved with temporal metric as the feature.

4 Conclusions

In this study, an approach for predicting the citation count of the scientists in a dynamic citation network has been developed, which utilizes the topological structure of the dynamic citation network using temporal metric, and content information of the publications. The temporal metric is created to capture the temporal events occurring in the network. The problem is conceptualized as a node classification task to understand the future citation classes of the scientists. The classification is accomplished using multiple ML-based models for different class categories. The XGBoost classifier has achieved the highest scores for all the evaluation metrics with the feature set FS_2, consisting of *title, abstract, and citation count* of previous years, and FS_3, consisting of *title, abstract, and temporal metric*. The XGBoost model has achieved the highest average accuracy of 85.19% with FS_2. It is observed that the best score for different evaluation metrics is obtained from the feature set FS_2. However, results with feature set FS_3 are better than those with FS_1, and also comparable with FS_2. Since FS_3 has *title, abstract, and temporal metric* as its features and FS_2 has *citation count* of every time frame as its feature along with *title and abstract*, FS_3 is advantageous over FS_2 in terms of less number of features and time complexity for classification. The task of citation count prediction is also considered as a regression problem, and linear regression model is used for the prediction. The average MAE of 6.38 and 8.36 is achieved with the previous year's citation count and temporal metric as the features, respectively. These results are also comparable, having a difference of 2.02 of MAE. Hence, the temporal metric can be an essential feature for the prediction of the citation count of scientists. The author-based features like the number of publications, area of research, etc., can also be used along with the temporal metric, and deep learning models can be applied in the future to enhance the accuracy of the citation count prediction process.

References

1. Ibáñez, A., Larrañaga, P., Bielza, C.: Predicting the h-index with cost-sensitive naive Baye. In: 11th International Conference on Intelligent Systems Design and Applications, pp. 599–604 (2011). https://doi.org/10.1109/ISDA.2011.6121721
2. Fu, L.D., Aliferis, C.F.: Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. Scientometrics **85**, 257–270 (2010). https://doi.org/10.1007/s11192-010-0160-5
3. Chakraborty, T., Kumar, S., Goyal, P., Ganguly, N., Mukherjee, A.: Towards a stratified learning approach to predict future citation counts. In: IEEE/ACM Joint Conference on Digital Libraries, pp. 351–360 (2014). https://doi.org/10.1109/JCDL.2014.6970190
4. Bhat, H.S., Huang, L.H., Rodriguez, S., Dale, R., Heit, E.: Citation prediction using diverse features. IEEE Int. Conf. Data Mining Worksh. (ICDMW) **2015**, 589–596 (2015). https://doi.org/10.1109/ICDMW.2015.131

5. Shetty, R.D., Bhattacharjee, S., Dutta, A., Namtirtha, A.: GSI: An influential node detection approach in heterogeneous network using covid-19 as use case. IEEE Trans. Comput. Soc. Syst. (2022). https://doi.org/10.1109/TCSS.2022.3180177
6. Shetty, R. D., Bhattacharjee, S.: A weighted hybrid centrality for identifying influential individuals in contact networks. In: 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, pp. 1–6 (2022). https://doi.org/10.1109/CONECCT55679.2022.9865749
7. Chen, J., Zhang, C.: Predicting citation counts of papers. In: 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), pp. 434–440 (2015). https://doi.org/10.1109/ICCI-CC.2015.7259421
8. Zhu, X.P., Ban, Z.: Citation count prediction based on academic network features. In: 2018 IEEE 32nd International Conference on Advanced Information Networking and Applications (AINA), pp. 534–541 (2018). https://doi.org/10.1109/AINA.2018.00084
9. Bütün, E., Kaya, M., Alhajj, R.: A supervised learning method for prediction citation count of scientists in citation networks. IEEE/ACM Int. Conf. Adv. Soc. Netw. Anal. Mining (ASONAM) **2017**, 952–958 (2017)
10. Bütün, E., Kaya, M.: Predicting citation count of scientists as a link prediction problem. IEEE Trans. Cybernet. **50**(10), 4518–4529 (2020). https://doi.org/10.1109/TCYB.2019.2900495
11. Soares, P.R.S., Prudêncio, R.B.C.: Proximity measures for link prediction based on temporal events. Exp. Syst. Appl. **40**(16), 6652–6660 (2013). https://doi.org/10.1016/j.eswa.2013.06.016. ISSN 0957-4174
12. P. Ganeshkumar, A. K. BR, S. Padmanabhan and V. A, "Social Media Personal Event Notifier Using NLP and Deep Learning", 2022 International Conference on Power, Energy, Control and Transmission Systems (ICPECTS), Chennai, India, 2022, pp. 1–5, doi: 10.1109/ICPECTS56089.2022.10047710
13. HEP-Th Dataset. https://snap.stanford.edu/data/cit-HepTh.html. Accessed Sept 2022
14. HEP-Th Dataset Metadata. https://www.kaggle.com/datasets/tayorm/arxiv-papers-metadata. Accessed Sept 2022
15. Sarker, I.H.: Machine learning: Algorithms, real-world applications and research directions. SN Comput. Sci. **2**, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x
16. Nikam, U.V., Deshmuh, V.M.: Performance evaluation of machine learning classifiers in malware detection. In: IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) 2022, pp. 1–5 (2022). https://doi.org/10.1109/ICDCECE53908.2022.9793102