IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 20, 2018, accepted February 23, 2018, date of publication March 6, 2018, date of current version March 16, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2812804

Predicting the Number of Publications for Scholarly Networks

XIAOMEI BAI
Computing Center, Anshan Normal University, Anshan 114007, China
(xiaomeibai@outlook.com)

ABSTRACT Scholarly networks have attracted great attentions, such as scholarly impact evaluation, scholarly impact prediction, scholarly recommendation, co-author relationships analysis, and team identification. Ranking research institutions as an important aspect of scholarly impact research is of great significance for decision makers, such as funding allocation, promotion, and transfer. There has been much debate about the scientific correctness behind those rankings. Predicting the number of accepted conference papers of research institutions next year is proposed by KDD Cup 2016, which aims to measure the impact of research institutions. To accurately predict the impact of different institutions in the eight top conferences: FSE, ICML, KDD, MM, MobiCom, SIGCOMM, SIGIR, and SIGMOD, a novel model was proposed in which the number of accepted papers of each institution, country, and time factors driving the impact of institution change are used as training features. Correspondingly, a hybrid model of support vector machine and neural network is constructed for resolving the predictive task. The experimental results show that the proposed method is better than the Markov model and the neural network model in terms of normalized discounted cumulative gain.

INDEX TERMS Ranking research institution, predictive model, support vector machine.

I. INTRODUCTION

Internet of Things can increase the ubiquity of the internet and has important application [1]–[3]. Scholarly networks have attracted great attentions, such as scholarly impact evaluation, scholarly impact prediction, scholarly recommendation, co-author relationships analysis and team identification. For example, with the rapid growth of academic big data at an exponential rate, it is of great guiding significance to predict the influence of institutions. It mainly includes the following three aspects: directing government agencies to make decisions, helping institutions recruit new members and guiding the awards of higher education institutions.

The impact of institutions evaluating methods mainly focus on the existing impact of the institutions, and predicting the impact of institutions is to discover its future impact. The ultimate problem of predicting the impact of institutions is determining which institution will become the most influential in the coming year of years and how to predict the impact of institutions. In general, the future impact of each institution is inseparable form the changing trend of previous impact of each institution. Previous researchers evaluated the impact of institutions based on the assumption that each author makes an qual contribution to a paper, and if an author has multiple institutions, each affiliation also contributed equally [4], [5]. Kuan *et al.* [6] proposed using H-index and c-descriptor to characterize the research performance of institutions. The fractional counting methods mainly consider the best journal and the best paper rate indicators and highly-cited papers [7]. Myers *et al.* [8] ranked institutions by article citations distinguishing different topical area. Abramo and D'Angelo [9] ranked research institutions by the number of highly-cited articles and applied this indicator to measure performance of Italian universities in each research field.

However, predicting the number of accepted papers of research institutions is challenging. First, the issue of predicting the future impact of institutions is an open problem, and the ground truth is not known beforehand. Second, abundant available heterogeneous information also increases the difficulty of resolving the problem. In this paper, given the published historical records of each institution, the paper's aim is to introduce several types of models for predicting the next year's impact of each institution. Here this work attempts to predict the impact of institutions on the basis of several important features found in the experiments: the ground truth of each institution, time and space information. The ground truth of each institution is determined by the full research papers accepted in a given top conference. Time and space information mainly are used to weight constructed predictive models. In the predictive models, if the ground truth of an institution is more adjacent to the ground truth of

the institution in a predictive year, the close data will be given higher time weight. In terms of space information, this paper primarily considers the country feature, namely the publications status of each country. For each institution, the more its home country published the papers, the higher the weight given to its institution. Based on these important features, this paper leverages Markov [10], neural network (NN) [11], and support vector machine (SVM) [12] models to construct three categories of predictive models: Markov predictive model, NN predictive model, and a hybrid predictive model based on SVM and NN (SVM_NN). Time and space information are used to improve the predictive performance in the NN and SVM_NN models. Specially, SVM_NN model first leverages SVM to classify the historical data of each institution, and then gives specific predictive value by the NN model. To accurately predict the impact of each institution, the main novelty of the proposed models lies in mining the important features related to the future impact of institutions. This paper integrates these features with several machine learning models. The experimental results demonstrate that the prediction performance of the SVM_NN model is better than the Markov and NN models in terms of NDCG. Besides, The models have good flexibility, and can be used to predict the impact of scholars, university, and country. The models can be used in other fields, such as economics, industry, and environment.

The rest of the sections of this paper is organized as follows. Section 2 introduces the related works. Section 3 introduces the proposed models for predicting the impact of each institution. The experimental results are shown in the Section 4, and Section 5 presents a discussion.

II. RELATED WORK

Scholarly impact research can be categorized into two aspects: evaluation and predication. Scholarly impact evaluation mainly contains a scholar' impact, the impact of paper, journal, institution, and country. Scholarly impact prediction mainly includes a scholar' impact, paper impact and institution impact. Scholarly impact evaluation focuses on evaluating the previous impact of scholarly entities. While scholarly impact prediction stresses on predicting future impact of scholarly entities, which has more guiding significance compared to evaluation. Since the founding of the Institute of Scientific Information (ISI) by Eugene Garfield in 1960, the impact of scientific papers has been captured in numbers, namely, the number of publications and the citations of academic work. During the past decade, the study of scientific impact has progressed dramatically, from H-index [13] to its many variants [14], [15] and from Journal Impact Factor(JIF) [16] to Eigenfactors [17]. Throughout the evolution of measuring scholarly impact, Rank-based metrics [18] with multiple dimension characters have been rapidly rising instead of evaluation metrics with a single-dimension character.

Predicting the impact of scholarly impact has attracted wide attention. Most previous researchers focus on predicting the impact of a paper and an author. In general, predicting the impact of a scientific paper can be divided into two categories: citations-related and citations-unrelated. The citations-related studies are listed as follows. A simple data analytic method was developed to predict future citations of a paper across different disciplines using short-term historical citation data and its published journal [19]. Journal impact factor and early citations were used to predict the long-term citation impact of a publication [20]. Otherwise, previous researchers explored the prediction of the impact of a paper by leveraging citations-unrelated features. The same universal temporal pattern has been discovered for characterizing the citation dynamics of papers by identifying three fundamental mechanisms: preferential attachment, aging and fitness [21]. The centrality of coauthoring networks was used to predict scientific success [22].

Due to the fact that H-index is the most widely known metric for scholar's impact, predicting a scholar's impact mainly focuses on scholar's future H-index. A cumulative measure model was proposed using the following four features: the number of years since a scholar's first publication, current H-index, author's number of articles, and the number of articles published in high impact journals [23]. Under certain specific topics, an author's authority and venue of a paper were two crucial factors for increasing the primary author's H-index [24]. Similarly, topical authority and publication venue were considered to distinguish whether a new publication can enhance its primary author's future H-index [25]. Another novel method was to predict a scholar's H-index using characteristics of the co-author network [26]. The H-index of a junior professor and a senior professor was predicted by using the cost-sensitive naive Bayes method [27].

In contrast to above studies, more specific interest of this work is to predict the impact of research institutions at conferences. Previous researchers mainly focus on evaluating the impact of institutions. Currently, the methods for assessing the impact of institutions can be divided into two categories: full counting method and fractional counting method. The former distributes the impact of a scholarly paper to different signed authors' institutions equally; the latter allocates it according to the best journal, best paper rate or highly-cited papers. However, these methods face the limitations. Since most current measures consider all citations as having the same importance, the result is an unfair evaluation of the impact of institutions. Also, the future impact of an institution is more significant than the past impact of an institution, but these methods cannot foresee their impact in the future.

The paper addresses the limitation that current measures cannot foresee impact of institutions in future by considering the close correlation between past measures and future measures. In the predictive models, the most important feature is the ground truth of each institution in different years. Two other features, country and time information, need to be paid more attention to enhance the accuracy of prediction.

III. METHODS

A. DATASET DESCRIPTION

The experiments are conducted on the Microsoft Academic Graph (MAG), which is a freely available dataset (https://kddcup2016.azurewebsites.net/Data) and contains 56, 677 publications. Each paper includes paper ID, original paper title, normalized paper title, publish year, publish date, paper Document Object Identifier (DOI), original venue name, normalized venue name, Journal ID mapped to venue name, conference ID mapped to venue name, and paper rank. A separate dataset provides the citation relationship list, including paper ID and paper reference ID. About eight percent of the publications with incomplete data were removed from the experimental dataset.

1) TRAINING DATASET

The following the eight top conferences-SIGIR, SIGMOD, SIFCOMM, KDD, ICML, FSE, MoBiCom, and MM-from the MAG dataset are used as training dataset, spanning a period of 16 years from 2000-15. The feature factors-ground truth of each institution, weighted time and weighted country-are considered as the input of training dataset.

2) TESTING DATASET

The *2016 KDD Cup Selected Papers* are used to as the testing dataset which are from aforementioned eight top conferences between 2011-15 year.

B. FIRST-ORDER MARKOV MODEL

The work's goal does not lie in providing an estimated value of future impact of each affiliation. Instead, this paper focuses on predicting a probability distribution for the impact of each affiliation.

Formally, the first-order Markov model is given by

$$S^{(n)} = S^{(n-1)} \times P = S^{(0)} \times P^n \tag{1}$$

where $S^{(n)}$ denotes the nth year state probability, $S^{(n-1)}$ represents the $nth - 1$ year state probability, P is transition probability, and $S^{(0)}$ is the initial year state probability vector. Figure 1 shows a Markov model with three states. In the experiment, the initial state probability vector is constructed by the ground truth data of each affiliation in 2000. In the first-order Markov, this paper adopts three states: rise, stay the same, and decline. According to the initial state probability vector $S^{(0)}$ and the transition probability P of each affiliation in eight conferences at $nth - 1$ year, state of each affiliation at nth year can be predicted.

Next this paper uses t distribution with $n - 1$ degree of free to estimate the prediction interval of affiliation impact at nth year. The prediction interval formula of future certain observed value I_{n+1} is as follows:

$$\bar{I} \pm t_{a/2} s \sqrt{1 + \frac{1}{n}} \tag{2}$$

Further, the prediction interval of each institution impact is amended according to the preceding prediction trend of each institution impact. For example, if a future observed value I_{n+1} is on the rise, the prediction interval ranges from the average value $\bar{I}$ to $\bar{I} + t_{a/2} s \sqrt{1 + \frac{1}{n}}$. Otherwise, if it presents the downward trend, its prediction interval ranges from $\bar{I} - t_{a/2} s \sqrt{1 + \frac{1}{n}}$ to the average value $\bar{I}$.

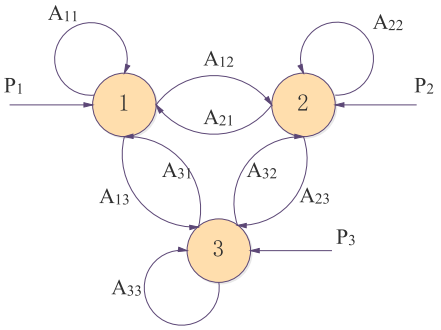

FIGURE 1. Markov model with three states. This is a stochastic automaton, where Pi is the probability that the system starts at state Si, and Aij is the probability that the system shifts from state Si to Sj.

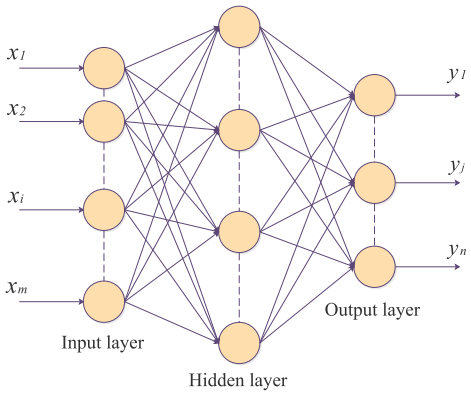

FIGURE 2. Neural network model for predicting impact of institution.

C. NEURAL NETWORK MODEL

Given the historical data of scientific publications, the various feature components under different models are trained to obtain the best predictive ability. The piecewise training of each individual component is adopted. In this paper, 18, 822 papers extracted from MAG dataset are leveraged to construct a feed forward neural network (see Figure 2), abbreviated as NN. In particular, this paper trains feed forward neural network learners using the following features: the ground truth of each institution, weighted time and weighted country. The input of training data is classified as follows:

1) Considers a feature: the ground truth of each affiliation.
2) Considers two features: the ground truth of each affiliation and weighted time.
3) Considers two features: the ground truth of each affiliation and weighted country.
4) Considers three features: the ground truth of each affiliation, weighted time and weighted country.

The inputs of the training dataset are applied to predict the impact of institutions. The inputs of the testing dataset are

also divided by four categories accordingly. The ground truth of each institution between 2011 and 2014 is used to predict the impact of affiliation for 2015, as the output of testing dataset.

Take, for example, Case 1. The impact prediction for affiliations is described as follows: Given n training data $(x_1, \cdots, x_4, y_1), (x_2, \cdots, x_5, y_2) \cdots, (x_n, \cdots, x_{n+3}, y_n)$, where n is the year numbers for the impact of certain affiliation, x_i indicates ground truth data, corresponding with each institution, and y_i indicates the ground truth of the impact of certain affiliation in next year keeping pace with its previous four years. In order to predict the impact of each affiliation at the *nth* year, the predictive formula is defined as follows:

$$y_n = F(x_{n-4}, x_{n-3}, x_{n-2}, x_{n-1}) \tag{3}$$

To Case 2, in addition to the ground truth of each affiliation as inputs, the cumulative sum of time-weighted ground truth of each affiliation is also considered as an input. If a year of training data is closer to predictive year, a higher weight will be given to the ground truth of the affiliation. To Case 3, the ground truth of each country first is calculated, and then accumulate the ground truth of all countries to obtain the ratio of each country. At last, because some affiliations may cover different countries, an accumulative sum of the ground truth of each affiliation in certain year times the weight of its countries in the same year serves as an input to predict future impact. In Case 4, Case 2 and Case 3 are integrated.

D. A HYBRID MODEL OF SUPPORT VECTOR MACHINE AND NEURAL NETWORK

To improve the performance of predicting the impact of each affiliation, the use of SVM for pattern recognition is designed, and then rank the affiliations using feed forward neural network. Figure 3 shows the hybrid model of predicting the impact of an institution. To recognize a pattern, a function $f : R^N \rightarrow \{\pm 1\}$ is estimated. This paper first constructs a reasonable training dataset, which is N-dimensional Patterns x_i and class labels y_i. The formula is as follows:

$$(x_1, y_1), \cdots, (x_k, y_k) \in R^N \times \{\pm 1\} \tag{4}$$

The SVM model can implement the following idea: the input feature vectors are mapped into a high dimensional feature space and construct an optimal separating hyperplane, aiming at maximizing the distance between the hyperplane and the nearest data points for each class in the space. Different mappings can construct different SVM models. The mapping can be finished by a kernel function. The function is defined as follows:

$$f(\overrightarrow{x}) = sgn(\sum_{i=1}^{N} y_i \alpha_i \cdot K(\overrightarrow{x}, \overrightarrow{i}) + b) \tag{5}$$

where the coefficients α_i can be obtained by the solving the convex Quadratic Programming (QP) problem:

$$Maximize \sum_{i}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \cdot y_i y_j \cdot K(\overrightarrow{x_i}, \overrightarrow{x_i})$$

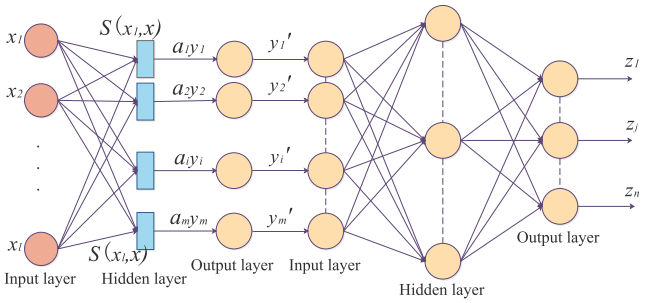

FIGURE 3. A hybrid model for predicting the impact of institutions.

$$subject\ to\ 0 \leqslant \alpha_i \leqslant C \sum_{i=1}^{N} \alpha^i y_i = 0 \quad i = 1, 2, \cdots, N. \tag{6}$$

where C is a regularization parameter controlling the trade-off between the margin and misclassification error. If the corresponding $\overrightarrow{x_i} > \overrightarrow{0}$, these $\overrightarrow{x_j}$ are called Support Vectors. In this paper, the radial basic function kernel is used and is defined as follows.

$$K(\overrightarrow{x_i}, \overrightarrow{x_j}) = exp(-\gamma \left\| \overrightarrow{x_i} - \overrightarrow{x_j} \right\|^2) \tag{7}$$

Compared with NN, the advantage of SVM lies in the solution of the *QP* problem, which is globally optimized while NN only finds a local minima. In addition, SVM can effectively avoid overfitting. In this paper, the class labels include 11 types, and each of five institutions for top 50 affiliations are allocated a class label. Namely, affiliations of ranking top $1 - 5$, which class label is 1, and so on. The class label 11 is assigned to the affiliations for more than the top 50 affiliations such that function f will correctly classify testing data. For example, the same potential probability distributions $P(z, y)$ from the training data are assigned to different data (z, y) of the testing dataset. A predictive value for each affiliation is then estimated by relying on the feed forward neural network. In this model, in order to investigate the best predictive capacity, the following the three features: ground truth of each affiliation, weighted time, and weighted country are considered and present the 16 cases as TABLE 1.

IV. RESULTS

To accurately predict the impact of each affiliation in the top eight conferences in the next year, this paper first extracts article data from the top eight conferences in the MAG dataset, then calculate the ground truth of each affiliation, which is defined to determine each affiliation ranking by the following simple rules:

1) Each accepted paper is equally important.
2) Each author has equal contribution to a paper.
3) If an author has multiple affiliations, each affiliation also contributes equally.

In this paper, we adopted the above-mentioned ground truth which is KDD2016 official rules (https://kddcup2016.azureweb sites.net/Rules).

In order to effectively evaluate the performance of different ranking measures of affiliations impact, the Normalized Dis-

TABLE 1. A hybrid method of SVM and NN.

	SVM_NN1	SVM_NN2	SVM_NN3	SVM_NN4	SVM_NN5	SVM_NN6	SVM_NN7	SVM_NN8
SVM	a	a	a	a	a,c	a,c	a,c	a,c
NN	a	a,b	a,c	a,b,c	a	a,b	a,c	a,b,c
	SVM_NN9	SVM_NN10	SVM_NN11	SVM_NN12	SVM_NN13	SVM_NN14	SVM_NN15	SVM_NN16
SVM	a,b,c	a,b,c	a,b,c	a,b,c	a,b	a,b	a,b	a,b
NN	a	a,b	a,c	a,b,c	a	a,b	a,c	a,b,c

Notes: a, b, c represent different features, respectively. a represents the ground truth of affiliation. b represents weighted time. c indicates weighted country.

FIGURE 4. NDCG of different prediction methods in SIGIR conference.

counted Cumulative Gain (NDCG) is utilized as the metric to measure the relevance. NDCG is a normalized measure method of Discounted Cumulative Gain (DCG). DCG is defined as follows:

$$DCG_n = \sum_{i=1}^{n} \frac{rel_i}{log_2^{i+1}} \tag{8}$$

where DCG_n represents a weighted sum of relevant degree of ranked entities, and its weight is a decreasing function varying according to ranked position. Variable i indicates the rank of an affiliation, and rel_i represents the relevance score of the affiliation of ranked ith position.

To normalize DCG values, NDCG@N is defined by

$$NDCG_n = \frac{DCG_n}{IDCG_n} \tag{9}$$

where $IDCG$ is a an ideal DCG, which is identified as the simple DCG measure with best ranking results. Therefore, the probability score of NDCG measure always ranges from 0 to 1. In this paper, the NDCG represents the importance of an affiliation in the given relevant top conference. If an affiliation has not appeared in the results, its NDCG value will be assumed as 0.

Figure 4 shows the predictive performance of SIGIR in 2015 year. The predictive performances of SVM_NN models are slightly higher than NN or Markov models in terms of the NDCG in general. For example, in a series of SVM_NN models, the best predictive result with 0.7705 is given by SVM_NN12, which is higher 0.0107 than NN4 for NDCG@20. For NDCG@5, the best predictive result with 0.7740 is given by SVM_NN4, SVM_NN12, SVM_NN15, which is higher 0.089 than NN1, NN3, and NN4. In terms of NDCG@20, NDCG@15, NDCG@10, and NDCG@5, SVM_NN12, SVM_NN2, SVM_NN2 and SVM_NN12 obtain the best predictive results respectively, which are 0.7705, 0.7736, 0.7812 and 0.7740 respectively, compared with other's models.

Figure 5 shows the predictive performance of SIGCOMM in 2015. The predictive performance of SVM_NN models is also slightly higher than NN and Markov models in terms of NDCG in general. For example, in a series of SVM_NN models, the best predictive result with 0.7988 is given by SVM_NN4, which is higher 0.0199 than NN4 for NDCG@20. For NDCG@15, the best predictive result of SVM_NN models is 0.8126 provided by SVM_NN13, and in NN models, NN1 obtains the best predictive result, which is 0.7984. For NDCG@5, the difference of predictive performance between SVM_NN, NN,and Markov is relatively large. In SVM_NN models, the highest is 0.7101, and the lowest is 0.6550.

Figure 6 shows the predictive performance of KDD in 2015 year. The predictive performance of SVM_NN models is slightly higher than NN and Markov models in terms of NDCG in general. For example, in a series of SVM_NN models, the best predictive result with 0.8109 is given by SVM_NN15, and the second is 0.8106 generated by SVM_NN2 for NDCG@20. The SVM_NN4 model obtains the best predictive result reaching 0.8602 for NDCG@15, while in NN models, the best is NN4 with 0.7762. To KDD, if the value of NDCG is relatively large, the predictive performance is better. For example, the predictive results under the evaluation metric of NDCG@20 and NDCG@15 are better than the predictive results by relying on NDCG@10 and NDCG@5. In terms of NDCG@5, the best predictive result only is 0.51789.

Figure 7 shows the predictive performance of ICML in 2015. Under the four evaluation metrics, all the predictive results are higher than 0.640. Compared with NN and Markov models, the SVM_NN models can gain better predictive capacity. To ICML, the best predictive results of Markov is slightly higher than NN models for NDCG@20,

FIGURE 5. NDCG of different prediction methods in SIGCOMM conference.

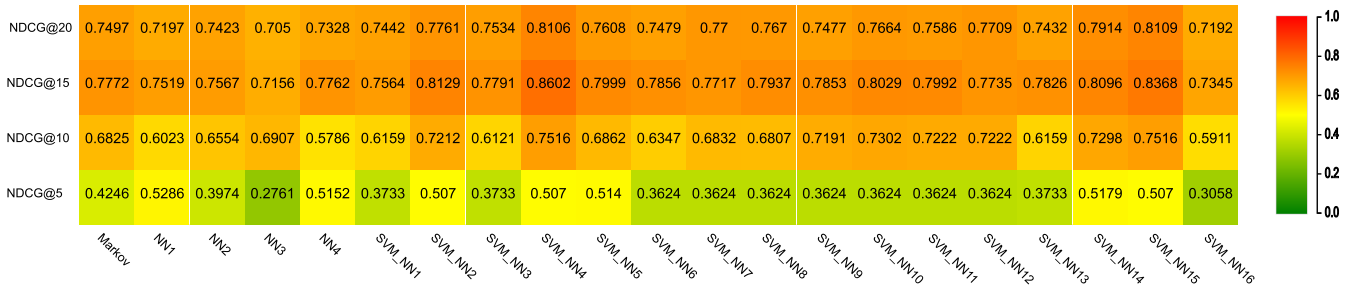

FIGURE 6. NDCG of different prediction methods in KDD conference.

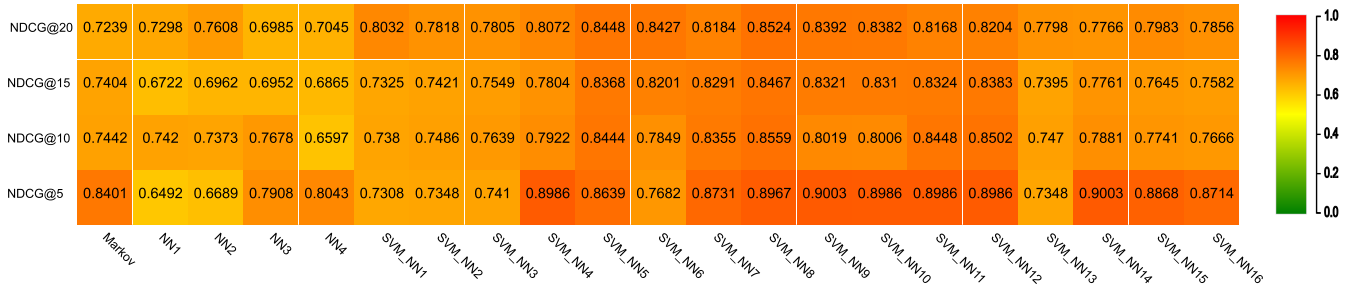

FIGURE 7. NDCG of different prediction methods in ICML conference.

NDCG@15 and NDCG@5. In terms of NDCG@20, the best predictive result is 0.8524, obtained by SVM_NN6 model, which is higher than 0.0916 compared with the best predictive result obtained by NN models. Under the evaluation method of NDCG@5, the best predictive performance is beyond 0.9000, such as SVM_NN9 and SVM_NN14 models reach 0.9003.

An interesting phenomenon can be observed that the value of NDCG@5 is relatively large between KDD and ICML. The reason perhaps derives from the historical data distribution. By comparing the different NDCG, this paper finds that the historical ground truth of each institution is a crucial feature for predicting the impact of institution. Meanwhile, the weighted time and weighted country also can enhance the predictive capacity of each institution impact in some cases.

V. DISCUSSION

How to better predict the impact of an institution? To address this question, this paper conducted the exploration research by relying on machine learning techniques, such as Markov, NN and SVM models. Based on these models, a hybrid model of SVM and NN was developed, which allowed us to compare individual Markov and NN models for efficiently predicting the impact of each institution. In this paper, to fully characterize the historical ground truth of each institution, time and space information on the impact of institution in future, a series of NN models and SVM_NN models are constructed. By comparing the NN models and Markov model, a slight increase in predictive performance was found to some extent. In order to improve the predictive ability of NN models, SVM_NN models are constructed, which are to first classify the historical data, then predict the impact for each institution. By comparing SVM_NN and NN models, there is a significant increase in predicting ability. Instead of using the individual NN models, SVM_NN models aim to improve the accuracy of categories which establishes the foundation for predicting the impact of institutions accurately. The SVM_NN models are not restricted to the institutions of conferences. The proposed approaches can be naturally applied to all kinds of fields, such as medical treatments, weather, industry and agricultural production.

The experimental results provide objective evidence for predictive performance of different methods. The predictive performance of the SVM_NN is better than the other two categories machine learning methods, illustrated by the metric of NDCG on different conferences. In this experiment, two interesting phenomenon are found. (1) Given the same predictive method, the predictive abilities are also different for different conferences. For example, in terms of SVM_NN16 model, NDCG@5 of ICML is much higher than one of KDD. This indicates that different historical data is intimately correlated with the predicting ability of models. (2) The features of weighted time and weighted country can increase the predictive capacity of models to some extent, but the increasing degree is also related with different conference. For example, the NDCG values of SVM_NN with weighted time and weighted country for ICML are higher than ones for KDD. This indicates that the predictive abilities are different mainly because the distributions of the ground truth of each institution in each conference are different.

There are two major limitations in this study. First, the data only are derived from several top conferences in the computer science field. Second, other important features were not considered in current models, such as topic trends of previous years' conference papers, previous years' conference top authors' impact factor based on the citation graph, co-author factor and related information from other conferences and journals. In future, the predictive ability of the proposed approach for different disciplines will be explored. Is there a big difference between the disciplines in predicting institution impact? Furthermore, the effectiveness of the predicted methods in other fields will be investigated, such as economic and weather fields. At last, data mining techniques can be better leveraged to identify more effective features to improve the proposed prediction models.

REFERENCES

[1] F. Xia, L. Yang, L. Wang, and A. Vinel, "Internet of Things," *Int. J. Commun. Syst.*, vol. 25, no. 9, pp. 1101–1102, 2012.
[2] X. Liu *et al.*, "RFID estimation with blocker tags," *IEEE/ACM Trans. Netw.*, vol. 25, no. 1, pp. 224–237, Feb. 2017.
[3] T. Qiu, A. Zhao, F. Xia, W. Si, and D. O. Wu, "ROSE: Robustness strategy for scale-free wireless sensor networks," *IEEE/ACM Trans. Netw.*, vol. 25, no. 5, pp. 2944–2959, Oct. 2017.
[4] S. J. Bensman, "The evaluation of research by scientometric indicators," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 62, no. 1, pp. 208–210, 2011.
[5] P. Vinkler, *The Evaluation of Research by Scientometric Indicators*. Amsterdam, The Netherlands: Elsevier, 2010.
[6] C.-H. Kuan, M.-H. Huang, and D.-Z. Chen, "A two-dimensional approach to performance evaluation for a large number of research institutions," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 63, no. 4, pp. 817–828, 2012.
[7] L. Bornmann, M. Stefaner, and F. de Moya Anegón, and R. Mutz, "Ranking and mapping of universities and research-focused institutions worldwide based on highly-cited papers: A visualisation of results from multi-level models," *Online Inf. Rev.*, vol. 38, no. 1, pp. 43–58, 2014.
[8] N. Myers, N. Snow, S. L. Summers, and D. A. Wood, "Accounting institution citation-based research rankings by topical area and methodology," *J. Inf. Syst.*, pp. 33–62, 2015.
[9] G. Abramo and C. A. D'Angelo, "Ranking research institutions by the number of highly-cited articles per scientist," *J. Informetrics*, vol. 9, no. 4, pp. 915–923, 2015.
[10] R. A. Jarrow, D. Lando, and S. M. Turnbull, "A Markov model for the term structure of credit risk spreads," *Rev. Financial Stud.*, vol. 10, no. 2, pp. 481–523, 1997.
[11] K. Kasiviswanathan, K. Sudheer, and J. He, "Quantification of prediction uncertainty in artificial neural network models," in *Artificial Neural Network Modelling*. Cham, Switzerland: Springer, 2016, pp. 145–159.
[12] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. Appl.*, vol. 13, no. 4, pp. 18–28, Jul./Aug. 2008.
[13] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 46, pp. 16569–16572, 2005.
[14] L. Egghe, "Theory and practise of the *g*-index," *Scientometrics*, vol. 69, no. 1, pp. 131–152, 2006.
[15] M. Schreiber, "A variant of the *h*-index to measure recent performance," *J. Assoc. Inf. Sci. Technol.*, vol. 66, no. 11, pp. 2373–2380, 2015.
[16] E. Garfield and I. H. Sher, "New factors in the evaluation of scientific literature through citation indexing," *Amer. Documentation*, vol. 14, no. 3, pp. 195–201, 1963.
[17] C. T. Bergstrom, J. D. West, and M. A. Wiseman, "The eigenfactor metrics," *J. Neurosci.*, vol. 28, no. 45, pp. 11433–11434, 2008.
[18] Y. Wang, Y. Tong, and M. Zeng, "Ranking scientific articles by exploiting citations, authors, journals, and time information," in *Proc. AAAI*, 2013, pp. 933–939.
[19] X. Cao, Y. Chen, and K. R. Liu, "A data analytic approach to quantifying scientific impact," *J. Inform.*, vol. 10, no. 2, pp. 471–484, 2016.
[20] C. Stegehuis, N. Litvak, and L. Waltman, "Predicting the long-term citation impact of recent publications," *J. Inform.*, vol. 9, no. 3, pp. 642–657, 2015.
[21] D. Wang, C. Song, and A.-L. Barabási, "Quantifying long-term scientific impact," *Science*, vol. 342, no. 6154, pp. 127–132, 2013.
[22] E. Sarigöl, R. Pfitzner, I. Scholtes, A. Garas, and F. Schweitzer, "Predicting scientific success based on coauthorship networks," *EPJ Data Sci.*, vol. 3, no. 1, p. 1, 2014.
[23] O. Penner, R. Pan, A. Petersen, K. Kaski, and S. Fortunato, "On the predictability of future impact in science," *Sci. Rep.*, vol. 3, p. 3052, Oct. 2012.
[24] Y. Dong, R. A. Johnson, and N. V. Chawla, "Will this paper increase your h-index? Scientific impact prediction," in *Proc. 8th ACM Int. Conf. Web Search Data Mining*, 2015, pp. 149–158.
[25] Y. Dong, R. A. Johnson, and N. V. Chawla, "Can scientific impact be predicted?" *IEEE Trans. Big Data*, vol. 2, no. 1, pp. 18–30, Jan. 2016.
[26] C. McCarty, J. W. Jawitz, A. Hopkins, and A. Goldman, "Predicting author h-index using characteristics of the co-author network," *Scientometrics*, vol. 96, no. 2, pp. 467–483, 2013.
[27] A. Ibáñez, P. Larrañaga, and C. Bielza, "Predicting the h-index with cost-sensitive naive Bayes," in *Proc. 11th Int. Conf. Intell. Syst. Design Appl. (ISDA)*, Nov. 2011, pp. 599–604.

XIAOMEI BAI received the B.Sc. degree from the University of Science and Technology Liaoning, Anshan, China, in 2000, the M.Sc. degree from Jilin University, Changchun, China, in 2006, and the PhD degree from the Dalian University of Technology, Dalian, China, in 2017. Since 2000, she has been with Anshan Normal University, China. Her research interests include computational social science, science of success, and big data.