Research performance prediction for scientists using LSTM neural networks

Junwan Liu [1] and Rui Wang [1]

[1]*liujunwan@bjut.edu.cn* [1]*wangrui_1996wr@163.com*
[1] Beijing University of Technology, Beijing 100124, China

Abstract: Scientific elites are regarded as engines that, for any nation, drive scientific and technological progress and social development. It is increasingly important to explore the research performance of scientific elites. Over the past studies, the approaches with machine learning or neural networks have been extensively applied in the task of time-series data prediction. We tried to use Long short-term memory (LSTM) neural network for predicting the productivity and influence of scientists who has been selected as academicians of American Academy of Sciences in the field of biology during period of 2008 to 2015. From experimental results, we found that forecasting results with using LSTM neural network can better predict the research performance of academicians. Using neural network approaches to predict research performance is an important attempt in the field of informetric, we will explore and compare more applications of neural networks in research performance prediction of outstanding scientists in the future. Furthermore, we will further study the impact of the election of academicians on the career development of scientists.

Keywords: time-series analysis, Long short-term memory neural network, performance prediction, research performance

Introduction

The scientific elites deserve our attention not merely because they have prestige and influence in science, but also their collective contributions have made difference in the advance of scientific knowledge (Zuckerman 1977). In the field of informetrics or scientometrics, the scientific elites, such as the Nobel laureates, have always been a hotspot among academia (Li et al. 2020). The research performance of these scientific elites is also the topic of many studies that cannot be bypassed, including publications, citations, and their pattern of collaboration (Chan et al. 2015). In the past studies, it had been a focus to investigate research performance and collaboration patterns of scientific elites in the academia (Zhou et al. 2014). We would benefit from understanding the career performance trajectory and development of the scientific elite. For policy makers, predicting future research career trends may provide more perspective for understanding the development potential of young scholars. Thus, how to predict the research performance of scientists is worthy of our exploring.

As we all know, the productivity or influence of scientists is often affected by various factors. The peak productivity in academics' scientific careers is often produced

in specific time periods (Yair and Goldstein 2020). Jones and Weinberg (2011) shows that the nature of life-cycle on research productivity, meanwhile, they also found the strong relationship between scientific creativity and age dynamics. Thus, the research performance is a time-series data that has different manifestations at different stages of a person's life.

Time-series data consists of sampled data points taken from a continuous, real-valued process over time. Obviously, the productivity or influence of scientists has a temporal component. Therefore, the annual number of publications and citations of scientists could be regarded as the time-series data. Analysis of time-series data has been the subject of active research for decades (Dietterich 2002). Traditionally, there are several techniques to effectively forecast the next lag of time-series data such as univariate Autoregressive (AR), univariate Moving Average (MA), Simple Exponential Smoothing (SES), and more notably Autoregressive Integrated Moving Average (ARIMA) with its many variations (Siami-Namini et al. 2018). These models or methods have achieved results among the tasks of time-series predictions and analysis.

Time-series data aims to use historical data to predict future data or trend changes. Deep learning, represented by neural networks, performs well in prediction tasks (LeCun et al. 2015). Meanwhile, machine learning or neural networks is more and more brilliant in interdisciplinary application (Shen et al. 2020). In the neural networks, Long short-term memory (LSTM) networks are a state-of-the-art technique for sequence learning (Fischer and Krauss 2017). Recently, the deep architecture of the recurrent neural network (RNN) and its variant long short-term memory (LSTM) have been proven to be more accurate than traditional statistical methods in modelling time-series data (Sagheer and Kotb 2019). Compared with Recurrent Neural Network (RNN), LSTM can solve complex, artificial long-time-lag tasks (Hochreiter and Schmidhuber 1997). LSTM neural networks often used in time series tasks such as predicting stock prices in financial markets, urban traffic speed (Ma et al. 2015; Fischer and Krauss 2018; Kim and Won 2018). Therefore, it is feasible to predict the research performance of scientific elites in time-series with using LSTM neural networks.

This paper regards the career performance as a time-series prediction problem. We try to use LSTM neural networks to deal with this study. Ultimately, from the experimental results, we found that the LSTM neural network can better predict the research performance of academicians.

Methodology

Methods

This paper adopts LSTM neural networks to predict the career performance of scientists. We measure the annual number of papers and the number of citations per article to represent productivity and influence of scientists. By constructing LSTM neural networks to forecast career performance of scientists, the prediction of the neural network can be improved by adjusting the parameters of the model. Finally, the re-

search performance prediction can be conducted by the LSTM neural network with the historical career publications or citations of scientists.

LSTM neural network

The primary objectives of LSTM neural networks are to model long-term dependencies for time series problems. The LSTM network is a recurrent neural network that is trained using backpropagation through time and overcomes the vanishing gradient problem. Instead of neurons, LSTM networks have memory blocks that are connected through layers. A block has components that make it smarter than a classical neuron and a memory for recent sequences. A block contains gates that manage the block's state and output. A block operates upon an input sequence and each gate within a block uses the sigmoid activation units to control whether they are triggered or not, making the change of state and addition of information flowing through the block conditional. There are three types of gates within a unit: 1) Forget gate: conditionally decides what information to throw away from the block. 2) Input gate: conditionally decides which values from the input to update the memory state. 3) Output gate: conditionally decides what to output based on input and the memory of the block. Each unit is like a mini-state machine where the gates of the units have weights that are learned during the training procedure (Hochreiter and Schmidhuber 1997).

Data

We collected all scientists who have been selected academicians of American Academy of Sciences in the field of biology from online search. Finally, we found a total of 407 scientists who were selected as academicians between 2008-2015. Meanwhile, we searched the publications of these scholars during 1974-2018 from Web of Science (WoS). A total of 79,555 publications were obtained. We also calculated the number of publications and the citations of these scholars each year. Table 1 shows the annual number of academicians from 2008 to 2015. Figure 1 and Figure 2 show the research performance of Levy, R who has been the academician elected in 2008.

Table 1. the number of academicians from 2008 to 2015

Year	Counts	Examples
2008	51	Albright, TD; Carrington, JC
2009	58	Bebbington, AJ; Dougherty, DA
2010	50	Anderson, PW; Feldmann, M
2011	45	Buckingham, M; Gottschling, DE
2012	52	Debenedetti, PG; Simberloff, D
2013	54	Alitalo, Kari; Beverley, Stephen M.
2014	47	Collins, James J.; Davies, Julian
2015	50	Bronner, Marianne E.; Chakravarti, Aravinda,
Total	407	

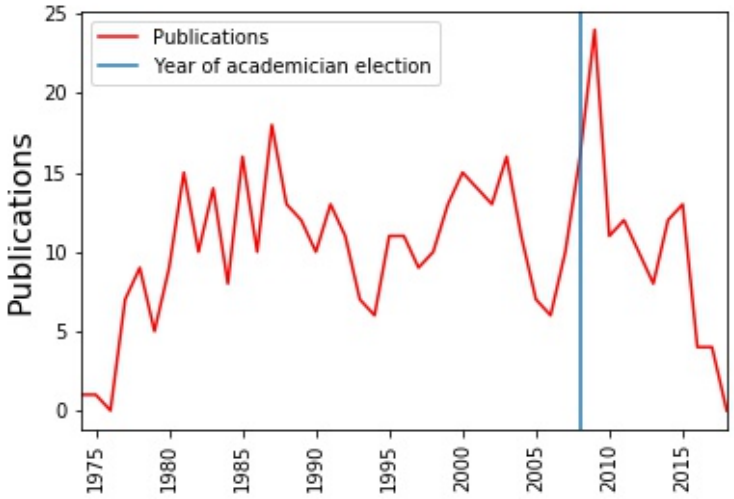

Fig 1 the annual research productivity of Levy, R

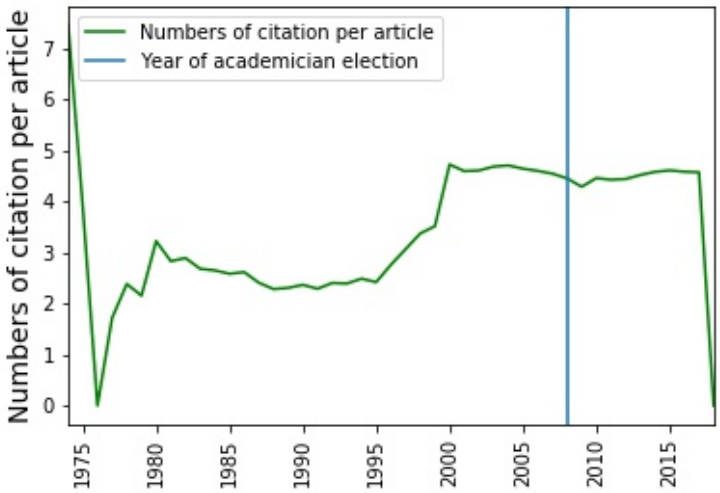

Fig 2 the annual numbers of citation per article of Levy, R

Experimental Results and Discussion

LSTM neural network construction

The LSTM neural network uses the first 10 time-series data as input data to predict the data at the next time. LSTM is sensitive to the scale of the input data, specifically, when the sigmoid or tanh activation functions are used. It can be a good practice to

rescale the data to the range of 0 to 1. We can easily normalize the dataset using the *MinMaxScaler* preprocessing class from the third-party library function. With time series data, the sequence of values is important. We divided the data into training and testing datasets, 90% of which can be used to train our model, and the remaining 10% can be used to test the model. The LSTM network has a layer with 1 input, a hidden layer with 256 LSTM blocks or neurons, and an output layer that makes a single value prediction. The linear activation function is used for the LSTM blocks. The network is trained for 100 epochs and a batch size of 20 is used. We constructed the LSTM neural network to conduct to time-series prediction in python with Keras. For more elaborate and detailed information on LSTM networks in this paper, we uploaded the code to github (https://github.com/wangruiDevin/my_LSTM_paper).

Forecast Results

We used the annual number of publications and citations per article of each scientists selected as academicians from 2008 to 2015 to forecast the career performance on a neural network. Using time-series data trained on LSTM networks, neural network can forecast 'future' productivity and influence based on the past research performance of scientists. Subsequently, by comparing with the real research performance curves of the scientists, we can observe that the prediction curve fit obtained from the neural network shows larger fluctuations before and after the particular event (i.e., elected academicians). Therefore, we suppose that the election of scientists as academicians may affect the development of research performance of scientists. We will focus on it in the future research.

Figure 3 and Figure 4 show the forecast results on LSTM neural networks for the annual number of publications and citations per article. Specially, the vertical line represents the year this scientist was selected as academicians. From Figure 3, we can know that the forecast results of publications on the neural network is greater than the actual number of publications of this scientist before and after he had been elected as academicians. We suppose that it is due to that the occurrence of the special event of being awarded the academician affects the prediction effect of the time-series data, which is not verified now but will be explored in the next work of this paper. We supposed that when the scientist has achieved an academic honor or title, he or she may have a short-lived productivity boost inspired by an academic honor. But the opposite result is obtained from Figure 4, where the actual influences of this scientist after being elected as an academician are greater than the predicted value before and after he had been elected as academicians. Similarly, we suppose that this is also due to a bias in the research performance caused by scientists receiving academic honors. Thus, the impact of academic honors on scientists' career development will be the focus of follow-up research.

In fact, there are three main situations between the forecast results and the true values: 1) The forecast results are greater than the true value; 2) The forecast results are lower than the true value; 3) The prediction result is partly greater than the true value and partly less than the true value.

We analyzed a total of 407 scientists elected as academicians during the period 2008-2015 and counted research performance according to the above three situations.

Ultimately, we found that 40% (n=164) of the scientists had higher number of publications predictions than the true data after being elected as academicians. Meanwhile, we also found that 41% (n=167) of the scientists had higher number of citations per article predictions than the true data after being elected as academicians. Whether the scientists' academic honors cause changes in the predicted curve of research performance of scientists? As to this question, we will also conduct an in-depth study of the above issues in our following work. LSTM neural network forecast results are also not representative of true research performance of scientists, but the academic halo is more of an acknowledgement of past achievements than a propeller for future career development in terms of historical performance of scientists. The existence of special events in time series data forecasting tasks can lead to bias in the prediction results. Therefore, the LSTM neural network needs to be further optimized and able to take into account the impact of special events on the prediction effect as much as possible. It should be emphasized, however, that neural networks to predict research performance are an important attempt to provide ideas for subsequent research.

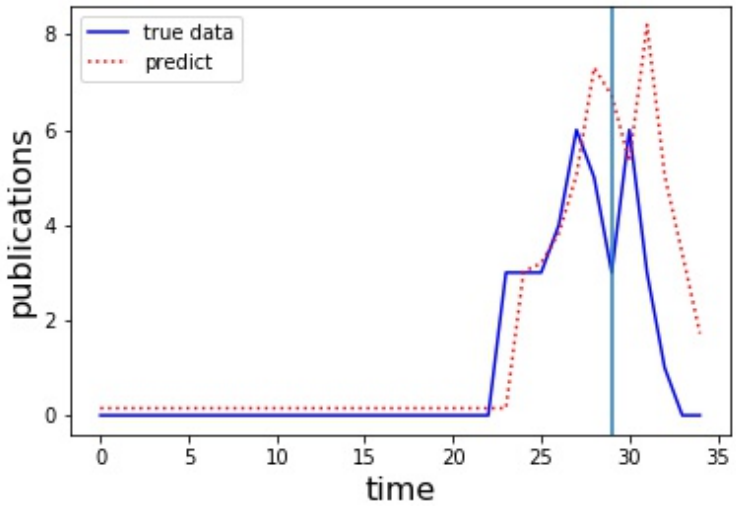

Fig 3 the forecast results of publications on neural networks

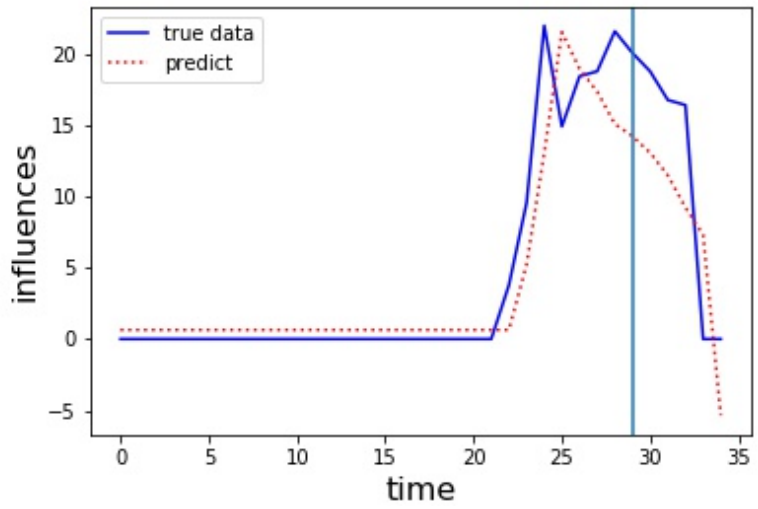

Fig 4 the forecast results of citation per article on neural networks

Conclusions and future work

Based on LSTM neural networks, we predicted the research performance of scientists who has been as academicians from 2008 to 2015. We can find that the neural networks approaches will play a positive role in the research performance prediction of scientists to a certain extent. LSTM neural networks predict the future development of scientists according to their historical research performance. Though the forecast results of LSTM neural networks can reflect the career development of scientists, they are still not objective. In particular, the impact of special events on the forecast results needs further consideration. In the future, we will combine more approaches to further explore the impact of academician election on the prediction of scientists' career performance. We will try to construct other neural networks, such as neural networks with attention mechanism or Gate Recurrent Unit (GRU) to make the forecast results more accurate in the next step. Until now, predicting research performance has been a complex task that is influenced by multiple internal and external factors. We will investigate the application of methods such as neural networks to tasks such as research performance prediction of scientists in the future work.

References

Chan, H. F., Önder, A. S., & Torgler, B. (2015). Do Nobel laureates change their patterns of collaboration following prize reception? *Scientometrics, 105*(3), 2215-2235, doi:10.1007/s11192-015-1738-8.

Dietterich, T. G. Machine Learning for Sequential Data: A Review. In *Structural, Syntactic, and Statistical Pattern Recognition, Joint IAPR International Workshops SSPR 2002 and SPR 2002, Windsor, Ontario, Canada, August 6-9, 2002, Proceedings, 2002*

Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research, 270*(2), 654-669, doi:10.1016/j.ejor.2017.11.054.

Fischer, T., & Krauss, C. J. E. J. o. O. R. (2017). Deep learning with long short-term memory networks for financial market predictions. *270*(2).

Hochreiter, S., & Schmidhuber, J. J. N. C. (1997). Long Short-Term Memory. *9*(8), 1735-1780.

Jones, B. F., & Weinberg, B. A. (2011). Age dynamics in scientific creativity. *Proc Natl Acad Sci U S A, 108*(47), 18910-18914, doi:10.1073/pnas.1102895108.

Kim, H. Y., & Won, C. H. (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. *Expert Systems with Applications, 103*, 25-37, doi:10.1016/j.eswa.2018.03.002.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444, doi:10.1038/nature14539.

Li, J., Yin, Y., Fortunato, S., & Wang, D. (2020). Scientific elite revisited: patterns of productivity, collaboration, authorship and impact. *J R Soc Interface, 17*(165), 20200135, doi:10.1098/rsif.2020.0135.

Ma, X., Tao, Z., Wang, Y., Yu, H., & Wang, Y. (2015). Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. *Transportation Research Part C: Emerging Technologies, 54*, 187-197, doi:10.1016/j.trc.2015.03.014.

Sagheer, A., & Kotb, M. (2019). Unsupervised Pre-training of a Deep LSTM-based Stacked Autoencoder for Multivariate Time Series Forecasting Problems. *Scientific Reports, 9*(1), 19038, doi:10.1038/s41598-019-55320-6.

Shen, Z., Zhang, Y., Lu, J., Xu, J., & Xiao, G. (2020). A novel time series forecasting model with deep learning. *Neurocomputing, 396*, 302-313, doi:10.1016/j.neucom.2018.12.084.

Siami-Namini, S., Tavakoli, N., & Namin, A. S. A Comparison of ARIMA and LSTM in Forecasting Time Series. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018*

Yair, G., & Goldstein, K. (2020). The Annus Mirabilis paper: years of peak productivity in scientific careers. *Scientometrics, 124*(2), 887-902, doi:10.1007/s11192-020-03544-z.

Zhou, Z., Xing, R., Liu, J., & Xing, F. (2014). Landmark papers written by the Nobelists in physics from 1901 to 2012: a bibliometric analysis of their citations and journals. *Scientometrics, 100*(2), 329-338, doi:10.1007/s11192-014-1306-7.

Zuckerman, H. (1977). Scientific Elite. Nobel Laureates in the United States. *196*, 754-755.