Small and multi-peak nonlinear time series forecasting using a hybrid back propagation neural network

Xuefan Dong [a,b], Ying Lian [a,b], Yijun Liu [a,*]

[a] Institutes of Science and Development, CAS, No. 15 ZhongGuanCunBeiYiTiao Alley, Haidian District, Beijing 100190, PR China
[b] University of Chinese Academy of Sciences, No.19A Yuquanlu, Beijing 100049, PR China

ARTICLE INFO

ABSTRACT

Gushes of online public opinions may trigger unexpected incidents that significantly affect social security and stability. Number of posts published per time interval, which is a time series dataset featured with multiple small-scale peaks and nonlinearities, is a simple and direct indicator of how severe the situation is and how much attention has been attracted. Thus, it is of great interest and significance to be able to accurately forecast this type of time series datasets. In this paper, a hybrid Back Propagation Neural network (BPNN) model is proposed to predict the features of this kind of time series datasets. Specifically, a modified Particle Swarm Optimization (PSO) algorithm combined with an Information Entropy (IE) function is used to optimize the weights and thresholds of the network, and the Bayesian Regularization is applied during the training process. Two real online public opinion cases are investigated to verify the effectiveness of the proposed model. Results showed that the proposed model has better performance in accuracy and stability, compared with Levenberg–Marquardt (LM) based BPNN, PSO based BPNN, Bayesian Regularization (BR) based BPNN, Stochastic Gradient Descent (SGD) based BPNN and Least Squares Support Vector Machines (LS-SVM) models.

© 2017 Elsevier Inc. All rights reserved.

1. Introduction

In recent years, the internet has become one of the most important channels through which people accessing information and expressing their own attitudes and opinions about certain social events [40]. The online public opinion itself now has become a significant factor which will interfere with the development of social events. For instance, online public opinions have been proved to be an effective tool for the political candidates to win out in the 2016 United States Presidential Election [19]. Real-world social security can also be greatly affected by online public opinions [34]. The "Occupy Wall Street" movement in 2011 was mainly facilitated by Internet social media, among which Twitter played a crucial role in facilitating interactions and opinion sharing among the participants [4]. Rapid spreading speed, open access and high degree of connectivity among netizens make it easy to call on a huge number of supporters to drive a certain social movement [6]. Therefore, a deep understanding of the online public opinions has been a research focus for sociologists and data scientists in recent years. Previous studies on this issue mainly focused on six aspects: event analysis [18], text sentiment classification [41], online social network (OSN) analysis (Kim et al., 2015) [48], opinion leader identification [25,46], rumor spreading [35] and monitoring system [3,21]. All these studies shed light on understanding the online public opinions. However, in most cases,

---

* Corresponding author.
  E-mail addresses: dongxuefan15@mails.ucas.ac.cn (X. Dong), lianying15@mails.ucas.ac.cn (Y. Lian), yijunliu@casipm.ac.cn (Y. Liu).
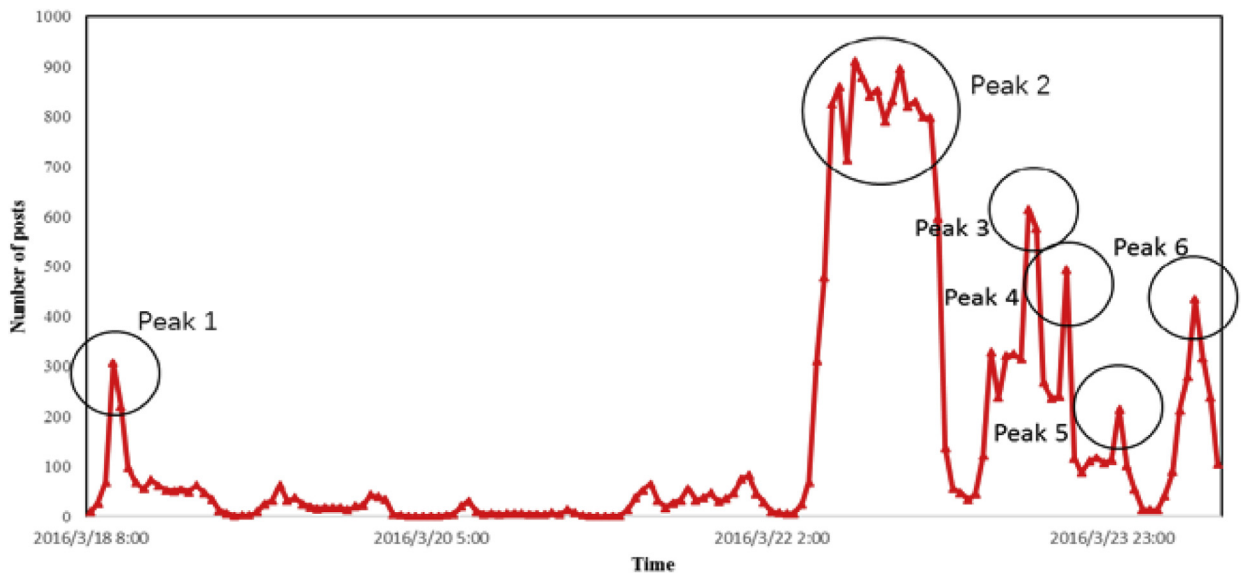
Fig. 1. Changes in number of posts published versus time in a real online public opinion case.
Notes: The online public opinion case shown above was caused by the "Illegal vaccine scandal in Shandong, China", happened on March 18, 2016 (For details, see http://www.bbc.com/news/world-asia-china-35859927). Data are selected from the Weibo website (China), with time interval one hour. The timeline is from 08:00, March 18, 2016 to 13:00, March 24, 2016. All microblogs hashtagged "illegal vaccine; Shandong Province" are collected as the dataset. It can be noticed that there are six significant peaks in this graph. In addition, compared to other kinds of time series, the duration of this online public opinion case is relatively short.

no timely guiding suggestions can be offered to the decision makers during the development of the event, especially in social and political areas. In order to formulate interventions in advance, it is curial for decision makers engaged in social and political management to know what is likely to happen as early and precise as possible.

As mentioned above, in most previous study programs related to Internet public opinions, focus had been mainly laid on the contents of those posts so as to know what have happened. In order to access the state more accurately and quickly, some techniques have been proposed. However, before analyzing the contents of those posts, it is necessary to keep an eye on the number of posts firstly as it can provide the most direct and simple information indicating the extent to which the situation is and how much attention has been attracted. In general, in most specific Internet public opinion cases, the social authorities or corporation managers will not pay attention on the contents until related posts have amounted to a quite large number. In addition, like in the case of the Occupy Wall Street movement, social group events tend to be caused only when a considerable number of participants are involved. Moreover, to a large extent, for both social and economic benefits, the earlier an accurate and comprehensive prediction on the number of posts is conducted, the more likely effective decisions on conducting intervention can be made. Indeed, a tremendous amount of attention has been attracted in previous studies [18]. In addition, most real data sets or real events employed in the past literature always involved a great number of participants (in other words, posts are very large in number). Therefore, it is very significant to propose an approach that can be adopted to accurately predict the number of posts related to certain Internet public opinion cases. However, as far as our knowledge is concerned, there is no paper focusing on such a point. A general form for the development trend of the Internet public opinions is shown as Fig. 1, which is featured with three main characteristics: nonlinearity, a small scale, and multiple peaks. It can be easily noticed from Fig. 1 that the data is relatively fewer compared to some other kinds of time series, such as financial time series. The reason lies in that focus is only laid on the number of posts related to the Internet public opinion case while simply ignoring the specific contents of respective posts. Hence, the number of posts at each time point can be set as one data. Moreover, as most specific Internet public opinion cases do not last long, only small data sets can be obtained. In addition, the nonlinear characteristic can be seen from the Fig. 1 as well, so we will not explain it in detail. With respect to the last characteristic, it can be noticed from Fig. 1 that there are six significant peaks. Such a phenomenon is mainly due to the uncertainty of posting behaviors and the random number of participants. Such features may lead the traditional neural network models to easily fall into the local optimum, thus tending to have a negative effect on the accuracy of the forecasted results, to a large extent.

As for methodologies used to forecast nonlinear time series, those models related to the neural network (NN) are commonly applied thanks to the NN's advantages of flexibly nonlinear transfer function and great self-learning ability [36]; thus, their promising applications in various areas, such as economics [27], finance [1], chemistry [29] have been illustrated. It has been revealed in several comparison research programs that compared to other models, neural network models have a generally better accuracy performance in nonlinear time series prediction (see the works of Lee et al. [20] and Yesilnacar and Topal [42]). Moreover, reviews in that area, mainly respecting the studies of Gooijer and Kumar [9] and Zhang et al. [45], have boosted our knowledge of that subject. However, certain inherent weaknesses of NN models have also been pointed out in previous studies. For example, the operation of complex criteria, by choosing the mean square empirical risk as the