Available online at www.sciencedirect.com

ScienceDirect

Cognitive Systems Research 52 (2018) 182–191

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

An effective daily box office prediction model based on deep neural networks

Yunian Ru [a,*], Bo Li [b], Jianbo Liu [a], Jianping Chai [a]

[a] Department of Information Engineering College, Communication University of China, No. 1 East Street, Dingfu Village, Chaoyang District, Beijing 100024, China

[b] Department of College of Science, Communication University of China, No. 1 East Street, Dingfu Village, Chaoyang District, Beijing 100024, China

Received 26 February 2018; received in revised form 13 April 2018; accepted 28 June 2018
Available online 5 July 2018

Abstract

The task of the daily box office prediction model is to build a dynamic prediction model to rolling forecast daily box office. It is a complex task as the movie box office has a short life cycle, and the static data and dynamic data that affect the trend of box office are heterogeneous. This paper proposes an end-to-end deep learning model for daily box office prediction, called Deep-DBP which consists of temporal component and static characteristics component. The temporal component is the main component which uses LSTM to learn the temporal dependencies between data points. The static characteristics component is an auxiliary component and it integrates static characteristics to improve prediction effect. The Deep-DBP can overcome the problems that the ARIMA and traditional ANN model cannot solve. The structure of input and output proposed in the model can well handle short time series prediction problem. It is a successful case in dealing with multi-source and multi-view data, addition of static characteristics component reduces the prediction error by 7%. The prediction error of Deep-DBP is 30.1%, which is better than that of the previous model. The experiment proved that the more training data collected, the better the prediction effect.

Keywords: LSTM; Movie box office; Prediction; Time series; Deep neural network

1. Introduction

With the increasing demand for cultural consumption and the rapid growth of theaters and screens, Chinese film industry continues to show a boom. However, the film industry is a high investment, high risk industry. Therefore, the research of daily box office prediction plays a significant role in avoiding risks. It provides important support for business intelligence decision-making process, such as making, distribution, cinema management, and developing related products. The function of the daily box office prediction model is to build a dynamic prediction model to rolling forecast box office in the days after the premiere. The success of the daily box office is of great significance to the layout and management of the cinema. Compared with the pre-released total gross revenue model, daily box office prediction model's characteristics not only have the basic information of the movie, but also the real-time dynamic data, such as the previous days' box office, the previous days' box office ratio, the previous days' screen count, the micro-blog index and so on.

However, due to the constraints of data acquisition and other factors, there are few studies in prediction daily box office at home and abroad. Jedidi, Krider and Weinberg applied the finite mixture regression method to analyze

* Corresponding author.
E-mail address: ruyn2016@cuc.edu.cn (Y. Ru).

https://doi.org/10.1016/j.cogsys.2018.06.018
1389-0417/

the weekly box office of 102 films, and the movies were divided into four categories, according to the first week's box office and decay rate (Kamel, Krider, & Weinberg, 1998). Bo Li, Fengbin Lu have established the Gamma demand model to analyze the life cycle and the box office trend of the film in Chinese film market (Li, Lu, Zhao, Wang, & Wang, 2010). Andrew Ainslie, etc., have combined the sliding window regression model with the gamma decay model through a hierarchical Bayesian framework to predict the weekly box office, and its Mean Absolute Percentage Error (MAPE) is 40.32% (Ainslie, Dreze, & Zufryden, 2005). Yong Liu established a dynamic model to research the relationship between word of mouth and the box office. The results show that the prediction model based on word of mouth has a good prediction effect on the weekly box office revenue, its prediction error MAPE is 47% (Liu, 2006). Lian Wang investigated the dynamic change of web search and built a weekly box office prediction model based on web search, and the results show that the model has a certain improvement, its prediction error MAPE is 39.9% (Lian & Jian-min, 2014). Taegu Kim, etc., introduced the social network service data and integrated three machine learning algorithms such as SVR, K-NN and GPR to predict the cumulative box office and weekly box office. According to the experimental results, the error MAPE of the single week box office forecast is 44.9% (Kim, Hong, & Kang, 2015). Xiaopeng Luo used the 21 days data of 138 movies to build dynamic panel model and established the box office prediction model by two step system GMM estimation. The model achieved good prediction results, however the model does not introduce factors that reflect the impact of network communication, such as network reviews, and the competition factors in the same period (Luo, Qi, & Tian, 2016). With the rapid development of the movie industry, the life cycle of the movie is shorter and shorter. The daily box office prediction with more granular size is more applicable and valuable than the subsequent week prediction. All the above models have a problem of too many restrictions and low prediction accuracy. Therefore, the daily box office prediction is still a complex and challenging task which is affected by the following issues:

- Various complex factors: There are various factors influencing the daily box office, and these data come from different views, and they are multi view data which can be divided into dynamic and static data. The dynamic data includes the daily box office ratio, daily screen count, and micro-blog index and so on. These data reflect the movie's box-office status, the movie box office competitiveness, network attention and its popularity on the internet. Static data includes word of mouth, distributor, production area and so on. So, it is a valuable research problem as to how to use the data in these different views effectively and reasonably in the model to improve the prediction accuracy rate.
- Short life cycle: Movie box office time series is a special kind of time series, whose life cycle is short, that is, the length of the sequence is short. An individual series usually is too short to be modeled accurately. What is more important is that the data of the previous days are more valuable, so it is unsuitable to use a large amount of previous information to train, and then predict the future data.
- Limitations of the existing algorithm of time series prediction: The prediction of daily box office data is different from the regression model which is a time series prediction problem and it needs to consider the temporal dependencies of data. There are a variety of algorithms for processing data with time information, with their own advantages and disadvantages. How to build a time series prediction model based on the applicable algorithm is also a challenge.

The ARIMA model is the most widely used model for the time series prediction, and the practice also proves that it can get a reasonable prediction result on many problems. But the shortcomings of the ARIMA model is also significant, for example, it's unable to deal with missing values or serious noise pollution data set, unable to deal with the nonlinear relationship or multivariable prediction problem, and it needs a lot of artificial experience in modeling to handle the data to be stationary (Praag, 2003).

Artificial neural network (ANN) is also a popular method of time series prediction for many researchers. The most famous work of using ANN to predict time series is made by Zhang, Patuwo, and Hu (1998). After that, many researchers began to use ANN to predict financial data and got good effects (Ghiassi, Saidane, & Zimbra, 2005; Krauss, Xuan, & Huck, 2016). The advantage of the ANN model is that it is robust to the noise in the input data, and even can be trained and make prediction in the presence of missing values. It is easy to learn linear and nonlinear relations, and it can also support the prediction of multiple time series (Sutskever, Vinyals, & Le, 2014). The disadvantage of traditional feedforward ANN in dealing with time series prediction is that it needs fixed length of context window and extracts limited information.

Unlike standard feedforward neural networks, Recurrent Neural Networks (RNN) retains a state that can selectively hold information from an arbitrarily long context window. RNN is a connectionist model that captures the dynamics of sequences by cycles in the network of nodes (Lipton, Berkowitz, & Elkan, 2015). In RNN, the results of the hidden layer at the present time step are related to the current input and the results of the hidden layer at the last time step. Introducing time information into the model can make the model learn time series correlation, avoid pre-specified time window, and uplift the constraints of the conventional ANN architecture. So, it can satisfy the need for accurate simulation of complex multivariable sequences (Malhotra, Vig, & Shroff, 2015).