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/© 2018 Elsevier B.V. All rights reserved.

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).

Long Short-Term Memory networks (LSTM) is a special type of RNN that can learn long-term dependence information. LSTM is proposed by Hochreiter (1996) and has been improved and popularized by Alex Graves in the near future (Graves, 2012). LSTM has achieved considerable success on many problems that could not be solved before. The LSTM based system can learn to translate language, predict disease, predict click rate, and predict stock forecast (Salehinejad et al., 2017). However, in the field of time series prediction, the application of LSTM model is very limited, especially for box office time series prediction (Gamboa, 2017).

To address the above challenges, this paper proposes an end-to-end *deep* learning model for *daily box* office *prediction*, named Deep-DBP. The main contributions of this paper are as follows:

- Deep-DBP consists of temporal component and static characteristics component, and it successfully processed the influence factors from different views, such as dynamic and static. It is a successful case in dealing with multi-source and multi-view data. In temporal component this paper uses LSTM to learn the temporal dependencies between data points. The static characteristics component integrates static factors, including word of mouth, distributor, and so on. The learned latent patterns in static characteristics component are fed into temporal component to improve prediction effect. It is found through experiments that the addition of static characteristics component reduces the prediction error of 7%.
- Another advantage of Deep-DBP is that it can overcome the problems that the ARIMA and traditional ANN model cannot solve and need less manual experience. It can deal with nonlinear relations and multivariable problems, and it does not need to manually determine time window, and also does not need to wait until the time window length data are accumulated before it starts to predict. It has been proven by experiments that Deep-DBP performs better than ANN and SVR when using the same characteristics.
- In temporal component, in order to addresses the challenge of short life cycle of daily box office sequence, an effective input and output structure of LSTM network is designed in model training. Compared with the one input and one output structure, the structure designed in this paper is more effective in dealing with short time series prediction problems.
- We conduct extensive experiments on Chinese film market, the prediction error MAPE is 30.1% which is obtained by using 80 films as training dataset and 34 films as test dataset. The effect of the model is better than that of the previous models, and the error is reduced by at least 10%. More notable is that this model has good scalability and applicability. The experiment proved that the more training data collected, the better the prediction effect.

2. Model architecture

In this section, we first describe the prediction problem of daily box office formally. Then, we introduce the architecture of the proposed Deep-DBP model.

We define B^m as dynamic data of the movie m, $B^m = \left(b_1^m, b_2^m, \cdots, b_T^m\right), m = 1, 2, \cdots, N$. where N equals the total number of the film. Each point b_i^m represents the data in day i of film m (where $i \in \{1, 2, \cdots T\}$,T is the length of the film life cycle). The b_i^m contains the following dimensions: the box-office(box), Micro-blog index(W-index), box office ratio (box_r), screen count(sc), the day of week (week), that is $b_i^m = \left(\mathrm{box}_i^m, \mathrm{W_{index}}_i^m, \mathrm{box_r}_i^m, \mathrm{sc}_i^m, \mathrm{week}_i^m\right)$. Furthermore, each movie m has its static characteristics and they can be expressed as: Word of mouth(WOM^m), distributor($\mathrm{distributor}^m$), origin country(OC^m), gene(gene^m), where $m \in \{1, \cdots, N\}$.

Given the information of movie opening day b_1^m and static characteristics: WOM^m, $\mathrm{distributor}^m$, OC^m, gene^m, the object of our prediction task is to forecast the box office of film m in the day after the opening day, that is, box_i^m, $i = 2, \ldots, T$.

The Deep-DBP consists of temporal component and static characteristics component, as shown in Fig. 1. Temporal component is the most important component in Deep-DBP and it uses LSTM network to learn the temporal dependencies between data points. The static characteristics component is an auxiliary component. It integrates static characteristics to learn the latent patterns and feeds them into temporal component to improve prediction effect.

2.1. Static characteristics component

Daily box offices are affected by many complex factors, such as genre, word of mouth, distributor, origin country and so on. This paper builds a simple yet effective component named static characteristics component to incorporate these characteristics into the Deep-DBP. As we known, the movies with different genres have different trends. Domestic films and Hollywood films also have different trends.

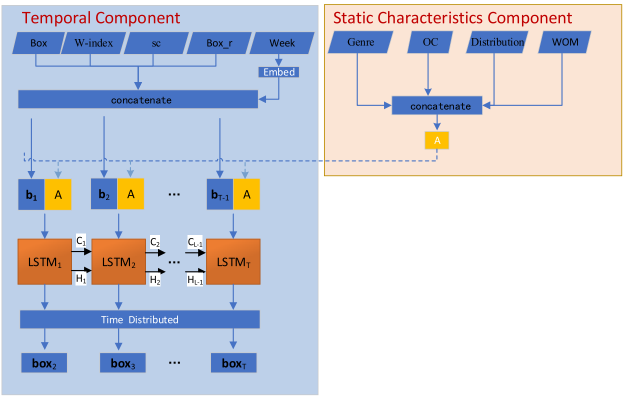

Fig. 1. The architecture of Deep-DBP.

The marketing ability of distributor also has a certain impact on box office. However, these useful characteristics are categorical values, so they cannot be fed to the neural network directly. In order to successfully use these data, this model transforms them into scalars. The specific transformation method will be introduced in detail in the third chapter. The word of mouth directly reflects the audience's affection for the film, and many audiences will choose whether to watch the movie according to the quality of word of mouth.

The static characteristics component integrates these static factors to learn the latent patterns and feeds them into temporal component to improve prediction effect. In order to connect these static factors to temporal component, this paper copies these static factors to the same length with life cycle of the movie as is shown in Fig. 1.

2.2. Temporal component

Temporal component is an important part of Deep-DBP model. As is shown in Fig. 1, it processes and integrates dynamic data to a vector B through concatenate components, and then split b_i with vector A to a new vector x_t. A is the output of static characteristics component. After that, temporal component uses RNN with LSTM units to learn the temporal correlations of data in the new sequence X. Finally, we apply a TimeDistributed dense layer which is the same as a dense layer but with multiple inputs and outputs to process the data output from LSTM network. Therefore, LSTM plays an important role in temporal component and even the whole Deep-DBP model.

The structure of LSTM network and its input and output structure play a crucial role in the whole model. Meanwhile, Understanding the details of forward and backward propagation of LSTM network and its implementation in the TensorFlow framework will play an important role in designing the LSTM input and output structure. So, in this part, we first introduce the LSTM algorithm, and then introduce the LSTM input and output structure designed in this paper.

2.2.1. Forward and backward propagation of LSTM

LSTM is a special type of RNN that can learn long-term dependence information. LSTM is proposed by Hochreiter and Schmidhuber and has been improved and popularized by Alex Graves in the near future. The LSTM model is the replacement of the RNN cells in the hidden layer into LSTM cells, which has the ability to memory for a long time, and it also can solve the vanishing and exploding gradient problem of RNN.

As is shown in Fig. 2 (refer to Christopher Olah's blog), LSTM has three gates to protect and control the cell state.

The first step is to determine what information we will discard from the cell state. The decision is done by a layer called the forgetting gate. The gate reads h_{t-1} and x_t, and outputs a number between 0 and 1 to each of the cells in the cell state C_{t-1}. 1 means "complete reservation" and 0 means "completely abandonment".

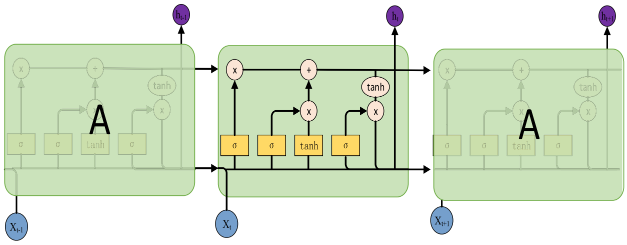

Fig. 2. LSTM internal structure.

$$f_t = \sigma(W_f[h_{t-1},x_t] + b_f) \tag{1}$$

The second step is to determine what kind of new information is stored in the cell state. It contains two parts. First, the sigmoid layer calls the 'input gate' to determine what value we are going to update. Then, a tanh layer creates a new candidate value vector that will be added to the state. The calculation process is shown as follows.

$$i_t = \sigma(W_i[h_{t-1},x_t] + b_i) \tag{2}$$

$$\widetilde{C}_t = \tanh(W_c[h_{t-1},x_t] + b_C) \tag{3}$$

The third step is to update the state of the old cells. C_{t-1} updates to C_t, as shown by the formula (4).

$$C_t = f_t * C_{t-1} + i_t * \widetilde{C}_t \tag{4}$$

The fourth step determines what value is output.

First, we run a sigmoid layer to determine which part of the cell will output. Next, we process the cell state through tanh (get a value between −1 and 1) and multiply it with the output of sigmoid gate. Finally, we will only output the part that we determine to output.

$$O_t = \sigma(W_o[h_{t-1},x_t] + b_o) \tag{5}$$

$$h_t = O_t * \tanh(C_t) \tag{6}$$

In addition, there are many variations in the LSTM model, and one of the most successful is the Gated Recurrent Unit (GRU). GRU model is a simplified version of the LSTM model but retained the long-term memory of the LSTM model.

For LSTM, the output at current time step depends on all previous inputs. So, it is trained with a variation of the backpropagation algorithm named Backpropagation Through Time(BPTT). For BPTT algorithm, if the sequence is very long, it will make back propagation computation difficult, its efficiency will be very low. In order to train the network more easily, a common used algorithm is truncated BPTT algorithm. This algorithm has a fixed steps approximation of backpropagation in time, which reduces the difficulty of training.

An example of the training process based on the truncated BPTT algorithm is shown in Fig. 3. The forward propagation step K1 is 6, and The back propagation step K2 is 3 (R2RT Blog, 2016).

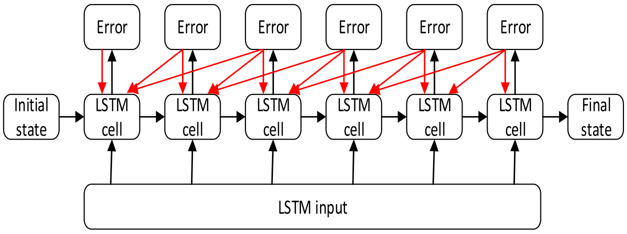

Fig. 3. Truncated BPTT algorithm (K1 = 6, K2 = 3).

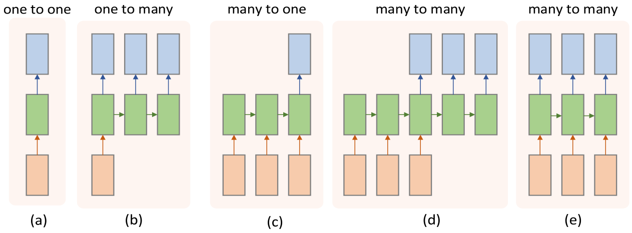

Fig. 5. Input and output structure of LSTM.

In this paper, we use the TensorFlow framework to implement the prediction model. In TensorFlow, the implementation of BPTT algorithm is different from the truncated BPTT algorithm introduced above. In this framework, K1 equals K2 and expressed with the parameter timesteps. A specific example is shown in Fig. 4 (R2RT Blog, 2016).

Therefore, the choice of timesteps will affect the internal state accumulated on the forward propagation and the weights updating on the backward propagation. And it also decided the structure of the input and output data.

2.2.2. *Input and output structure of LSTM in temporal component*

This part introduces the input and output structure of LSTM network designed in this paper and discusses the details of Deep-DBP model training, and how to rolling forecast the sequence.

Another significant advantage of LSTM is the flexibility of its input and output structure, compared to other deep learning networks, such as the convolution neural network which they use a fixed size vector as input, and then output a fixed size vector. As is shown in Fig. 5 (refers to Andrej Karpathy blog), the LSTM allows us to operate on the vector sequence: sequences in the input, the output, or in the both (R2RT Blog, 2016). When using Keras framework to implement Deep-DBP model, input and output structure are controlled by parameter timesteps.

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 this paper is unable to use a large amount of previous information to train, and then predict the future data. Therefore, the commonly used input and output structure cannot be well applied to the problem in this paper. Based on the need to solve the problem, we adopt the structure of sequence input and sequence output (Fig. 5(e)) in the training phase.

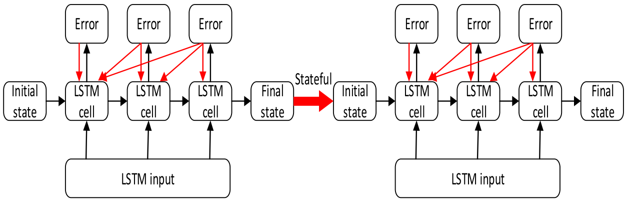

Fig. 4. Truncated BPTT algorithm (K1 = K2 = 3) in TensorFlow.

As is shown in Fig. 1, In training phase, After the concatenate module, the daily box office and other data of 20 days of each film (sequence *A* and *B*) are concatenated into a sequence of $X^m = (x_1^m, x_2^m, \cdots, x_T^m)$, and each point x_t^m, in the sequence is a vector with 9 dimensions. The 2nd to 21th days of the movie box office is taken as the target sequence Y, that is, $\mathrm{Y} = (\mathrm{box}_2^{\mathrm{m}}, \mathrm{box}_3^m, \cdots, \mathrm{box}_{21}^m)$. Thus, the loss is updated through 2–21 days forecast value and true value. In order to implement this type of input and output structure, this paper uses the Keras framework and sets timesteps to 20 and input_dim to 9. Meanwhile, this paper applies a TimeDistributed dense layer which is the same as a dense layer but with multiple inputs and outputs.

In the prediction phase, in order to rolling forecast the daily box office sequence, due to the flexibility of LSTM input and output, this paper copies the weights from the trained Deep-DBP model and to creates a new Deep-DBP model with the pre-trained weights. In the new model, we set parameters of LSTM network timesteps to 1 and input_dim to 9. It is important to note that the parameter stateful needs to be set to True and batch size needs to be set to 1. After anti standardization and exponential calculation processing, the box office data are predicted.

3. Data

3.1. *Dynamic data*

With the advent of the big data era, online score and Internet attention has become an important factor affecting the box office, and these data are constantly changing in the movie life cycle. The data that reflect the competitiveness of the film market, such as box office, screen count, and seat rate, are also changing in the life cycle of the film. These dynamic movie market information plays different roles in the life cycle of movie, so the box office prediction should consider the influence and value of dynamic information.

In order to illustrate the relationship between the microblog index and the box office change, this paper uses the data of the film *The BFG* to make an example, which is shown in Fig. 6. In order to show the relationship between

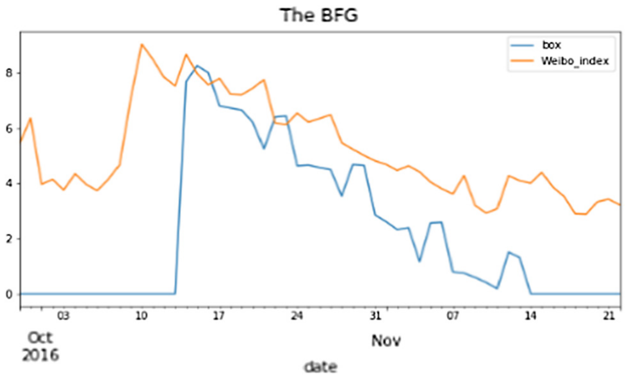

Fig. 6. The trend of micro-blog index and box office by day for film *The BFG.*

the two variables more clearly, this paper makes a log transformation to the box office and the micro-blog index data. As it can be seen from the figure, the micro-blog index has a great value before the movie shows, which shows that the marketing effect of the movie is pretty good. It has a high degree of attention and topic on micro-blog, and micro-blog index has reached the maximum value fifth days before the movie were screened. It can be seen that there is a strong correlation between the micro-blog index and the movie box office.

The following figure shows the correlation between the dynamic data used in this paper and the movie box office data. In order to show the relationship between the two variables more clearly, this paper makes a log transformation to these data.

As is shown in Fig. 7, It can be seen that there is a strong correlation between the previous day's box office, screen count, micro-blog index and the next day's box office.

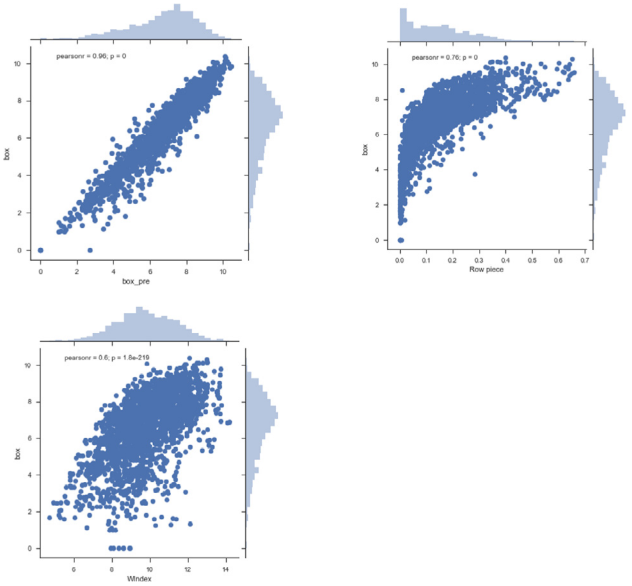

Fig. 7. The correlation between characteristics and box office.

3.2. Static data

3.2.1. Static data and index description

This paper adopts a reasonable and effective way to complete the collection of related factors of movie box office influence factors, clean and integrate the collected data, standardize the data obtained from different platforms, and quantify the non-numeric data. Based on the actual situation of the Chinese mainland film market, this paper has collected the following static data.

3.2.2. Static data preprocessing

A. Genre

Convert the categorical variable of genre into a numerical variable: the value of the genre is the box office ratio of the genre of movie.

$$\mathrm{Genre}_i = \frac{G_i}{\sum_i^l G_i} \tag{7}$$

If the first genre of the movie is genre i, then the value of the genre variable in the prediction model is Genre_i. Where, G_i is the total box office for all films whose first genre is genre i, and l is the total count of genre.

B. Origin country

Convert the categorical variable of origin country (production area) into a numerical variable.

$$\mathrm{OC}_i = \frac{R_i}{\sum_i^l R_i} \tag{8}$$

If the production area of the film is i, then the value of the origin country variables in the prediction model is Region_i. R_i is the total box office for all films whose production area is i, and l is the total count of production area.

C. Distributor

Convert the categorical variable of distributor into a numerical variable.

$$\mathrm{Distribution}_i = \frac{D_i}{\sum_i^l D_i} \tag{9}$$

If the distributor of the film is i, then the value of the distributor variables in the prediction model is $\mathrm{Distribution}_i$. D_i is the total box office for all films whose distribution is i, and l is the total count of distributions.

D. Word of mouth

This paper introduces the variable of internet word of mouth. The word of mouth score comes from the Douban.-com, the micro-blog and the Mtime. These three websites are the most popular movie reviews in China at present. They have authority and credibility and can cover all the films. Therefore, this paper obtains the film score data from three platforms and takes its mean as the network word of mouth of the film.

$$\mathrm{WOM} = \frac{\sum_{i=1}^{p}\mathrm{Critics_socre_i}}{\mathrm{p}} \tag{10}$$

3.3. Experimental data preparation

The purpose of this paper is to build an effective daily box office prediction model. The Chinese film market shows 200–300 films a year, but only a few films can get good box office income, most films are very low in the box office and the life cycles are very short, and some films have a life cycle less than a week. At the same time, these films not only have no research value, but also are easy to be disturbed by noise which increase the difficulty of modeling. Therefore, from the point of view of research value and accuracy of the model, we only study films that exceed 80 million of the box office. When the movie is released for more than three weeks, most of them no longer attract audiences. So, this paper only chooses the daily box office for the first 21 days as the object of research. This paper processed the box office data in 2015 and 2016, and deleted the missing data, and got 114 films altogether.

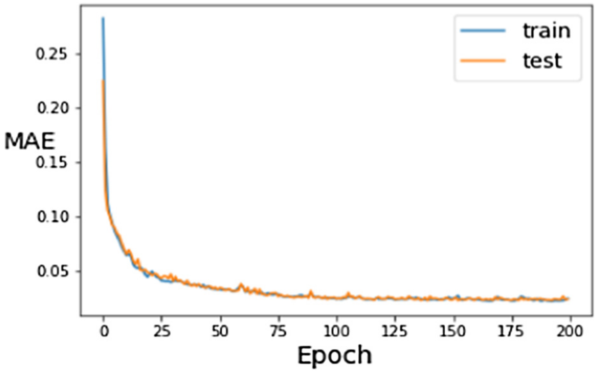

Fig. 8. Training and testing errors vary with epochs.

4. Experiment

4.1. Measurement

In this paper, MAPE is used to measure the prediction effect of the model. MAPE calculates the percentage of the error relative to the real value, and its calculation method is as follow:

$$\mathrm{MAPE} = \frac{1}{N}\sum_{t}^{N}\left(\frac{|Actual_t - Predict_t|}{|Actual_t|}\right) \tag{11}$$

where $Actual_t$ and $Predict_t$ are the observed values and the model output values of the t moment, respectively, N is the number of data points. MAPE is often used as an indicator of prediction accuracy, which is easy to calculate and understand, especially when the predicted data do not have unit. However, MAPE is more sensitive to scope, so it is not appropriate to use it in a case of very small real values. This is because the real value will become a denominator in the calculation process, so when the real value approaches zero, MAPE will approach infinity.

4.2. Experimental result performance comparison

In this paper, the prediction accuracy of the Deep-DBP model is evaluated by calculating the MAPE value on the test set.

It can be found from Fig. 8 that the Deep-DBP model has been converged after 200 training. And the result of the test set is consistent with the training set, and there is no phenomenon of over fitting.

Fig. 9 is a part of the prediction effect. It can be seen from the figure that the Deep-DBP model can predict the daily box office well and can also recognize the change pattern.

On Chinese film market dataset, Deep-DBP and common regression algorithms are compared. As the life cycle of daily box office data is too short, ARIMA is not suitable for this problem, so this paper does not take it as a comparison object. In order to get objective and fair comparison results, all algorithms used the same characteristics and

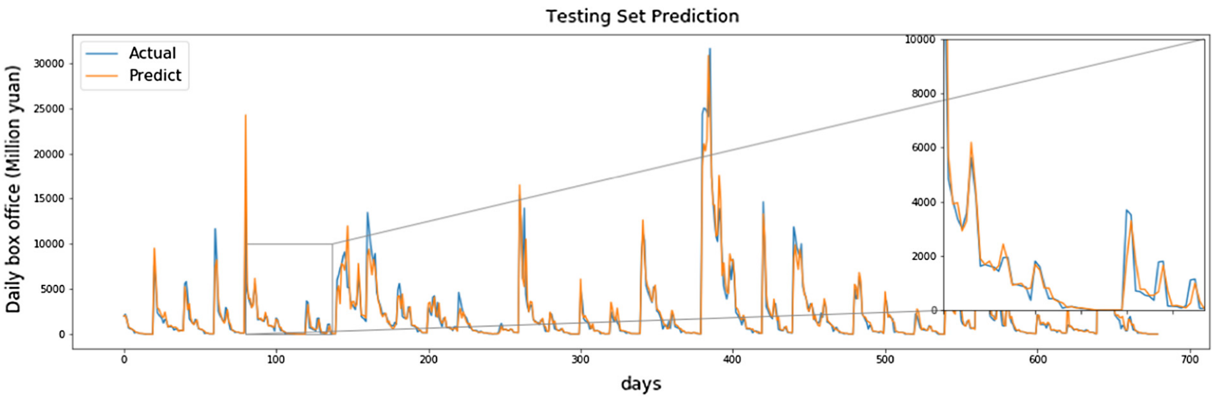

Fig. 9. Rolling prediction on test set.

set the time window length to 3. ANN algorithm is implemented by using MLPRegressor model(hidden_layer_sizes =256, activation = relu) in scikit-learn which is a free software machine learning library. SVR is implemented using the SVR model in scikit-learn.

As is shown in Table 1, it can be found by experiments that Deep-DBP performed best when using the same characteristics and it can fully use dynamic market information and static movie data to predict the 21 days movie box office. Deep-DBP can also predict box office from second days, and other algorithms begin to predict box office from fourth days, as they need to prepare 3 days' data to predict the next day.

Deep-DBP was compared with the previous box office prediction models as is shown in Table 2. The comparison showed that the model achieves the best prediction effect. However, because of the complexity of the box office prediction, the accuracy of the forecast is still only 30.1%, so there is still a lot of room for promotion.

4.3. Effect of input and output structure of LSTM in temporal components

LSTM is the main component of the temporal components. The box office data life cycle is short, and at the same time, we hope to predict the box office from the second day of the movie, and we also want to get good predictions. This paper compared the following two input and output structures.

Structure 1: one input and one output.
Structure 2: many input and many output in training phase, and recreate a new network with the pretrained weights, then rolling forecast the daily box office in the forecast phase.

Experiments show that the structure designed in this paper has great advantages in prediction effect, as is shown in Fig. 10.

Table 1
Compared with the common regression algorithms.

Model	MAPE
ANN (MLP)	38.6%
SVR	41.9%
Deep-DBP	30.1%

4.4. Effect of components

How to design a successful model to process multi-source and multi-view data is the key content of this paper. The daily box office prediction model needs to deal with dynamic data and static data. The Deep-DBP model designed in this paper also includes components for processing dynamic data and static data respectively. This paper also makes an experimental analysis on the effect of these two components on the prediction accuracy of the model. In order to ensure the reliable experimental results, each experiment was repeated 20 times. The parameters of the following experiments are as follows: Epoch is 200, Batch size is 8, learning rate is 0.001.

In order to get the reasonable results, this experiment uses the following 2 schemes:

scheme a: Temporal components
scheme b: Temporal components with static characteristics component

Table 2
Compared with the previous box office prediction models.

Model	Paper	MAPE
Ainslie et al. (2005)	*Modeling movie life cycles and market share*	40.32%
Liu (2006)	*Word of mouth for movies: its dynamics and impact on box office revenue*	47%
Lian and Jian-min (2014)	*Forecasting box office performance based on online search: Evidence from Chinese movie industry*	39.9%
Kim et al. (2015)	*Box office forecasting using machine learning algorithms based on SNS data*	44.9%
Deep-DBP	This paper	30.1%

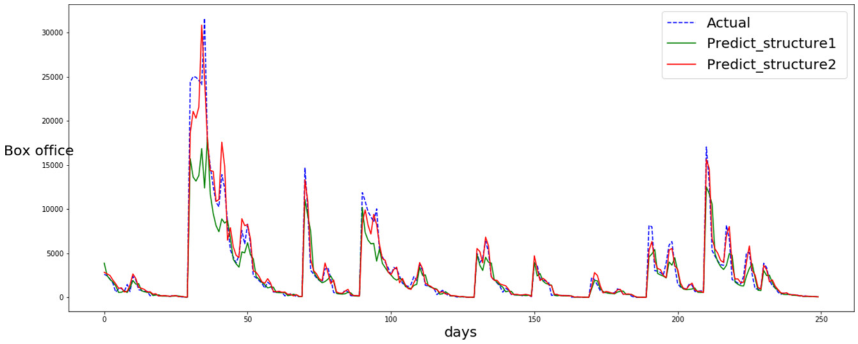

Fig. 10. Effect of Input and output structure of LSTM.

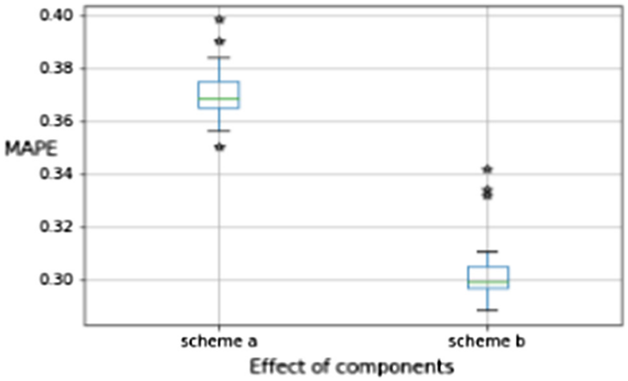

Fig. 11. Effect of components.

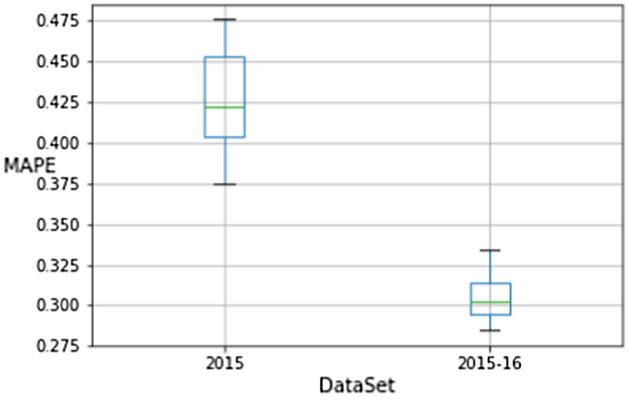

Fig. 12. Box line diagram of prediction error of model with different dataset.

As is shown in Fig. 11, through the experiment it can be found that only using temporal components the prediction accuracy of the model can reach 37%. After adding the static characteristics component, the effect of the model increased by 7%. This experiment fully shows that Deep-DBP model can successfully process multi-source and multi-view data. Static characteristics component plays a positive role in forecasting results.

4.5. Extensibility of the Deep-DBP model

The following experiment verified the extensibility of the model. In this paper, we used 61 movie data in 2015 and 114 film data in 2015 and 2016 years to train the model.

As is shown in Fig. 12, through the experiment, it can be found that as the amount of data increases, the prediction effect of the model is getting better and better, which is up to at least 10%. This fully illustrates the applicability and extensibility of the model to this problem. As more and more data are collected, the accuracy of the prediction will be enhanced.

5. Conclusion

In this paper, a new daily movie box office prediction model Deep-DBP was proposed. It is a successful model in dealing with multi-source and multi-view data and fully uses dynamic market information and static movie data to predict the 21 days movie box office. Through the experiment it was established that only using temporal components, the prediction accuracy of the model can reach 37%. After adding the static characteristics component, the effect of the model increased by 7%.

Deep-DBP model can overcome the problems that traditional time series prediction method (ARIMA) cannot solve, such as dealing with nonlinear relations and multi-variable problems, and it require little artificial experience. This model can also overcome the disadvantages of traditional ANN model that needs to specify the time dependent length. Its advantage is also very obvious. It not only leaves out need to manually determine the time window, but also can start from second days, and does not need to wait until the time window length data is accumulated before it starts to predict. It has been proven by experiments that Deep-DBP performs better than ANN and SVR when using the same characteristics.

In temporal component, in order to address the challenge of short life cycle of daily box office sequence, an effective input and output form of LSTM is designed in model training. Compared with the one input and one output structure, the structure designed in this paper is more effective, and the accuracy of the model is higher.

In this paper, the prediction error of 30.1% (MAPE) was obtained by using 80 film training models and finally tested in 34 films. The effect of the model is obviously better than that of the previous model, and MAPE has been reduced by at least 10%. Through the experiment, it was discovered that as the amount of data increases, the prediction effect of the model is getting better and better, which is up to at least 10%. This fully illustrates the applicability and extensibility of the model to this problem. As more and more data are collected, the accuracy of the prediction will be enhanced.

This paper attempts to fully use dynamic market information and static movie data to solve the complex problem of daily box office prediction. The successful prediction of the daily box office by this model will be of great guiding significance to the management of the cinema.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.cogsys.2018.06.018.

References

Ainslie, Andrew, Drèze, X., & Zufryden, F. (2005). Modeling movie life cycles and market share. *Marketing Science, 24*(3), 508–517.

Gamboa John Cristian wwBorges (2017). Deep learning for time-series analysis. arXiv:1701.01887 [cs.LG]. URL: paperuri:(37591ee21a0484a28e4cef0229b7c9ea).

Ghiassi, M., Saidane, H., & Zimbra, D. K. (2005). A dynamic artificial neural network model for forecasting series events. *International Journal of Forecasting, 21*, 341–362.

Graves, A. (2012). *Supervised sequence labelling with recurrent neural networks*. Berlin Heidelberg: Springer.

Hochreiter, S. (1996). LSTM can solve hard long time lag problems. In *International conference on neural information processing systems* (pp. 473–479). MIT Press.

Kamel, Jedidi, Krider, R., & Weinberg, C. (1998). Clustering at the Movies. *Marketing Letters, 9*(4), 393–405.

Kim, Taegu, Hong, J., & Kang, P. (2015). Box office forecasting using machine learning algorithms based on SNS data. *International Journal of Forecasting, 31*(2), 364–390.

Krauss, Christopher, Xuan, A. D., & Huck, N. (2016). Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500. *European Journal of Operational Research, 259.2*.

Li, Bo, Fengbin, Lu, Zhao, Xiujuan, Wang, Qian, & Wang, Shouyang (2010). Chinese movies' life cycle model and empirical analysis. *Systems Engineering-Theory & Practice, 30*(10), 1790–1797.

Lian, Wang, & Jian-min, Jia (2014). Forecasting box office performance based on online search: Evidence from Chinese movie industry. *Systems Engineering-Theory & Practice, 34*(12), 3079–3089.

Lipton, Zachary C., Berkowitz, J., & Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning. *Computer Science*.

Liu, Yong (2006). Word of mouth for movies: its dynamics and impact on box office revenue. *Journal of Marketing, 70*(3), 74–89.

Luo, Xioapeng, Qi, Jiayin, & Tian, Chunhua (2016). Box office forecasting after premiere. *Statistics &Information Forum, 31*(11), 94–102.

Malhotra P., Vig L., Shroff G., et al. (2015). Long short term memory networks for anomaly detection in time series. ESANN 2015 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 22–24 April 2015, i6doc.com publ., ISBN 978-287587014-8. Available from http://www.i6doc.com/en/.

Praag, C. M. Van. (2003). 25 years of IIF time series forecasting: a selective review. Tinbergen Institute Discussion Paper, No. 05-068/4.

R2RT Blog (2016). Written memories: Understanding, deriving and extending the LSTM - R2RT. https://r2rt.com/written-memories-understanding-deriving-and-extending-the-lstm.html.

R2RT Blog (2016). Written memories: Styles of truncated backpropagation-R2RT. https://r2rt.com/styles-of-truncated-backpropagation.html.

Salehinejad Hojjat, et al. (2017). Recent advances in recurrent neural networks. arxiv.org/abs/1801.01078v3. url:paperuri:(e09235a0b4c14df97b976bb8f00bf9bc).

Sutskever Ilya, Vinyals O., Le Q. V. (2014). Sequence to sequence learning with neural networks 4, pp. 3104–3112. arCiv:1409.3215v3[cs.CL]. url:paperuri:(6ea32fc3658b700f44c578a528346e80).

Zhang, Guoqiang, Patuwo, B. E., & Hu, M. Y. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting, 14*(1), 35–62.