RESEARCH ON MOVIE-BOX PREDICTION MODEL AND ALGORITHM BASED ON NEURAL NETWORK

Zexi Di, Jiapeng Xiu, Jinze Lin, Yunzhe Qian

Beijing University of Posts and Telecommunications, Beijing 100876, China
zhaizexi@163.com, xiujiapeng@bupt.edu.cn

Abstract: The paper presents an effective method to predict the movie box-office by Neural network algorithm. The core idea is to use past performance of the similar movie to predict a coming movie. Based on the analysis on Chinese film market, we extracted the elements of a movie like actors, director, screenplay, film genre, technical specs, etc. Some elements were then processed in different quantitative ways and made available for modeling. We extract some new generated indicators as the factors of the neural networks. By using the recent three-year movie data ,we built the prediction model using SPSS software platform and carried out detailed evaluation about the model.

Keywords: Movie box-office prediction; Nonlinear prediction model; Artificial Neural Networks

1 Introduction

In the past three years, the Chinese film market has fuelled an explosive expansion. In particular, the movie ticket sales for 2015 crossed the annual 44 billion yuan for the first time. With year-on-year growth of more than 48 percent, Chinese film market is moving quickly toward surpassing north America[1]. More and more investors are aware of this opportunity and eager to capture it. But, in practice, not all of the them can get their return. High returns also bring high risk. That is, you are likely to become an overnight billionaire or get stuck with the worst stuff. There are a lot of similarities between Chinese film market and the stock market. Only 20 percent of investors can get return, 10 percent of them are breaking even, and the remaining 70 percent of them have a great deficit. So that would be a very practical value for a research about how to reduce the investment risks.

In real life, the huge movie box-office depends on varieties of success factors including familiar actor, attractive screenplay , popular film genre and the special-effects. These complex and nonlinear interactions among these factors bring a great hindrance in modeling this problem. Because of the lack of understanding of its underlying mechanisms, we could not judge which feature is the most essential. In order to solve this problem, we use the neural networks to build our prediction model.

Neural networks are flexible, adaptive learning systems that follow the observed data freely to find patterns in the data and develop nonlinear system models to make reliable predictions[2].

Neural networks have been successfully developed to solve problems in varieties of applied sciences and engineering, such as plant disease identification and prediction of disease spread[6], forecasting inflows into rivers and lakes[7] and electricity load forecasting[8].

Taking the aforementioned applied fields, we start trying to learn the forecasting system behavior through observations and data. Absorbing many researches on Chinese film enables us to have the target to collect data that characterizes the system .Then we try to extract complex nonlinear patterns and relationships embedded in data. As a result ,we selected seven factors of a movie, including actor, director, screenplay, film genre ,technical specs, release date and distribution company.

2 Multilayer perceptron principle

Multilayer perceptron(MLP)is a feedforward artificial neural network model that maps sets of input data onto a set of appropriate outputs. An MLP consists of multiple layers of nodes in a directed graph, with each layer fully connected to the next one as shown in Figure 1. MLP utilizes a supervised learning technique called backpropagation for training the network.[3][4] MLP is a modification of the standard linear perceptron and can distinguish data that are not linearly separable.[5]

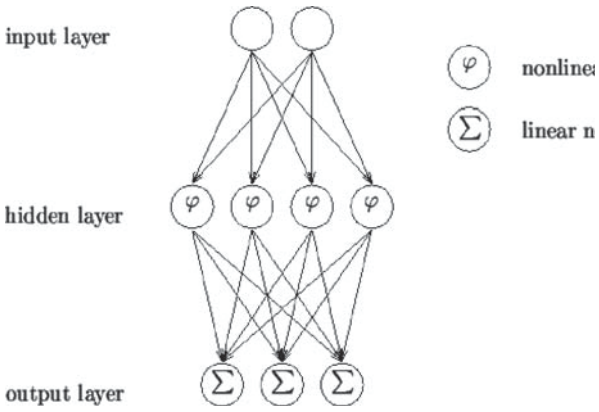

Figure 1 Signal-flow graph of an MLP

The perceptron computes a single output from multiple real-valued inputs by forming a linear combination according to its input weights and then possibly putting the output through some nonlinear activation function. Mathematically this can be written as

978-1-5090-1256-5/16/$31.00 ©2016 IEEE

$$y = \varphi(\sum_{i=1}^{n} w_i * x_i + \mathrm{b}) = \varphi(W^T \mathrm{X} + \mathrm{b}) \qquad (1)$$

Where W denotes the vector of weights, X is the vectors of input, b is the bias and φ is the activation function.

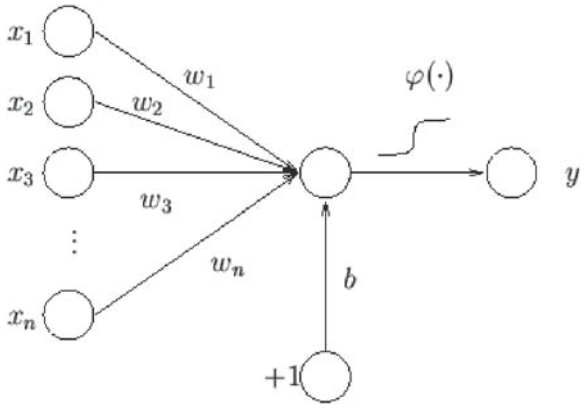

Figure 2 Signal-flow graph of perceptron

Next, the supervised learning problem of the MLP can be solved with the back-propagation algorithm, which consists of two steps. In the forward pass, the predicted outputs corresponding to the given inputs are evaluated as in Equation (1). In the backward pass, partial derivatives of the cost function with respect to the different parameters are propagated back through the network. The chain rule of differentiation gives very similar computational rules for the backward pass as the ones in the forward pass. The network weights can then be adapted using any gradient-based optimisation algorithm. The whole process is iterated until the weights have converged.[9]

3 Modeling

The first step is to collect the detailed data of the movie over the past three years. And then we used different methodologies to process the data. Finally, we used SPSS 23.0 to analyze our data to build our MLP predicting model and made some analysis of the predicted results.

3.1 Sample selection

We gathered the data from some film industry reports and movie official websites such as IMDB(Internet Movie Database). The samples from the most recent three-year movies altogether 842 subjects were taken.

Consider that foreign film may be influenced by varieties of factors which are different from those in Chinese film , the samples of this paper mainly were drawn from Chinese domestic films. It's also worth mentioning that we do not include animated movies. As far as animated movies, their actors are voice actor rather than movie actor. What makes a movie popular with the audience may not consist of actor or director in tradition. It seems that animated movies may influence the model on some level and finally we dropped them.

3.2 Data process

3.2.1 Movie genre

By means of the math prediction models, our research intends to get a possible box-office of a movie in the future. It is important for us to figure out which history movie is to be referenced. We know that there are some similar indexes showing how similar the two movies are. In fact the genre of movies is the most remarkable. So we used the similarity of the genre of movies to represent the similarity between movies.

The genres of movie mainly contain comedy, adventure, fantasy, mystery, thriller, documentary, war, romance, drama, horror, action, sci-fi, music, family, crime. Having made a statistics about the recent three-year released movie of different type, we found that there are a part of genres more common. So we divided the genres of movies into two levels. The genres in first level are more common than that in second level.

Table I The level of different movie genres

No.	Genre	Level
1	Comedy	1
2	Thriller	1
3	Romance	1
4	Horror	1
5	Action	1
6	Family	1
7	Crime	1
8	Adventure	2
9	Fantasy	2
10	Mystery	2
11	Documentary	2
12	Sci-Fi	2
13	War	2
14	Music	2

So, if two movies have many genres in common that belong to first level rather than second level, we'd consider them more similar.

By using Jaccard Coefficient, we can easily calculate the similarity between two movies.

$$J(A,B) = \frac{|A \cap B|}{|\mathrm{A} \cup B|} \qquad (2)$$

$$S(A,B) = J(A_1,B_1) + J(A_2,B_2) \qquad (3)$$

In equation (2), A denotes the first set, B denotes the second set.

In equation (3), S(A,B) denotes the similarity between two movies, A_1 denotes the set of genres of the first movie that belongs to level1, B_1 denotes the set of genres of the other movie that belongs to level1, A_2 denotes the set of genres of the first movie that belongs to level2, B_2 denotes the set of genres of the other movie that belongs to level2.

3.2.2 Movie director and actor

In practice, a good world-of-mouth of the director would bring recognition of a movie. The audience tends to watch a movie directed by their familiar director. In other words , a famous director could bring more concern on the movie than that little known director can bring. And the movie actor also has box office appeal. Then we decide to use the past performance of movies director and actor played to represent the box-office revenue that they could generate.

The director's appeal for box office can get calculated with the following rules.

$$director(x) = \sum_{n=1}^{n=3}(w_n b_n) \quad (4)$$

$$d_n = \text{datep} - date_n \quad (5)$$

$$w_n = \frac{\sum_{n=1}^{n=3} d_n - d_n}{2\sum_{n=1}^{n=3} d_n} \quad (6)$$

Where b_n denotes the office box of the corresponding past movie, d_n denotes interval between the two movie, datep denotes the present movie's release date, and $date_n$ denotes the release date of the corresponding past movie.

The rules for actor is more complex.

$$actor(x) = \sum_{n=1}^{n=3}(g_n i_n b_n) \quad (7)$$

$$g_n = \begin{cases} 1, & if\ this\ actor\ is\ the\ starring\ role \\ 0.5, & else \end{cases} \quad (8)$$

$$i_n = \frac{w_n + s_n}{\sum_{n=1}^{n=3}(w_n + s_n)} \quad (9)$$

In filmmaking, the starring's role is more important than comprimario. So, we consider the whole box office is brought by the starring and comprimario could bring the half box office.

In equation(8), w_n is derived from equation(5), and s_n denotes the similarity between the two movie which is derived from equation(3).

3.2.3 Movie schedule

Like traffic domain, the film market also has a "rush hour" phenomenon. Every year many a large amount of movies was released in time for all kinds of holiday. On holiday, more people prefer to choose to spare their time on watching a movie. And at this moment the competition is very intense. Based on the Chinese holidays, we classified the release date into some main auction slots.

Table II The schedule of movie-dating

No	Schedule name	Starting and Ending time
1	Lunar New Year schedule	2013/1/1-2013/2/24、2013/12/1-2014/2/14、2014/12/1-2015/3/5
2	the Chinese Spring Festival schedule	2013/2/10-2013/2/16、2014/1/31-2014/2/5、2015/2/19-2015/2/25
3	Valentine's Day schedule	2013/2/12-2013/2/16、2014/2/12-2014/2/16、2015/2/12-2015/2/16
4	Labor Day schedule	2013/5/1-2013/5/7、2014/5/1-2014/5/7、2015/5/1-2015/5/7
5	Dragon Boat Festival schedule	2013/6/9-2013/6/14、2014/5/30-2014/6/5、2015/6/18-2015/6/24

We think there is a competition between the similar movies. So we consider to use the similarities between all of the movies in a slot for representing the degree of the competition.

$$competition(x) = \sum_{n=1}^{n}(S_n) \quad (10)$$

Where s_n denotes the similarity between a past movie and this present movie which can be derived from quation(3).

With time going by, the film market is expanding. And we used the number of screens to represent how big the film market is.

3.2.4 Film productions and distribution

A brand of film production company often decides what kind of movie they produce and what quality the movie has. We notice that a good movie is more likely to be product by a big film production company. So we use the amount of the past produced movies to take the place of a coming movie's quality. Very similar, a distribution company's marketing ability would influence the launch of a movie. Like production company, we use the amount of the past produced movies to take the place of a coming movie's spreading scope.

3.2.5 Screenplay and technical specs

Besides the above mentioned aspects, there are also some basic information that embody of a movie.

The first one is the screenplay. In the past years, there

are amounts of IP(Intellectual property) movies having topped the box office. These movies are often adapted from some hot novels. Because of the big-selling of the novel, the readers tend to watch these movies whose characters and story they are familiar with. And some movies adapted from variety shows and cartoons also had good box-office.

The other one is the technical specs of a movie. With more and more foreign films rushing into Chinese film market, audiences are gradually attracted by the vivid and colorful of the movie. Many movies started to make 3D animation. And many people want to get an IMAX ticket to enjoy a better movie-going experience.

3.3 Model building

Firstly, we will introduce the used variable to you. And there are also relevant definition of these variables in table III.And the regulation of computing of these variables can be seen in data processing.

Table III The definition of the variables

No.	Variable	definition
1	box office	Total box-office of this movie and counts in ten thousands
2	Screens	The number of screens at that time
3	Tech_3D	If this movie is 3D, the value will be 1, else the value be 0
4	Tech_IMAX	If this movie is IMAX, the value will be 1, else the value be 0
5	Screenplay	If the screenplay is adapted from IP(Intellectual Property) novel or comic, the value will be 1
6	Actor1	The past performance of the first actor.
7	Actor2	The past performance of the second actor.
8	Actor3	The past performance of the third actor.
9	Actor4	The past performance of the fourth actor.
10	Director	The past performance of the director
11	Production	The number of movies the production company made
12	Distribution	The number of movies the distribution company released
13	Competition	The level of competitive intensity in this schedule

However, besides these variables listed in table III, we also used the genre of movie .Considered the large quantity of movie genres, we don't list them in this table.

After processing of data, we found that a high percentage of movies are completely new for us. Their actor are so new that they just have played movies several times, and some of them even have played movies only once. It increased so many problems that we could not use the past performance to evaluate them. So, we dropped the movie whose director and actor lacked past performance. And it makes us reduce the Data set. The dataset partition are presented in the table IV. The number of samples we use is 104 on total.

Table IV Case Processing Summary

		Number	Percent
Sample	Training	80	76.9%
	Testing	24	23.1%
Valid		104	100.0%
Excluded		0	
Total		104	

The dependent variable of our neural network model is the box office variable. And the other variables are used as factor and covariate of the input layer. The activation function used in hidden layer is hyperbolic tangent, the activation function used in output layer is identity. The error function in output layer is sum of squares.

We build our neural network model on the SPSS 23.0 software platform using MLP . After the learning of the network, the prediction model are obtained whose hidden layer has 11 units. There are so many units in our model that we don't list the graph of network here.

4 Model evaluation

In this section, we would introduce some summary of our Network Model.

Table V Model Summary

Model Summary		
training set	Sum of Squares Error	.013
	Relative Error	.001
	Stopping Rule Used	1 consecutive step(s) with no decrease in error (.001)
	Training Time	0:00:00.08
testing set	Sum of Squares Error	.182
	Relative Error	.108
Dependent Variable: box office		

As can be seen in table V, the sum of squares error in

training set is just 0.013 which means the training of the network is very effective. After the test of our prediction model, we find it acceptable that the relative error is just 0.108.

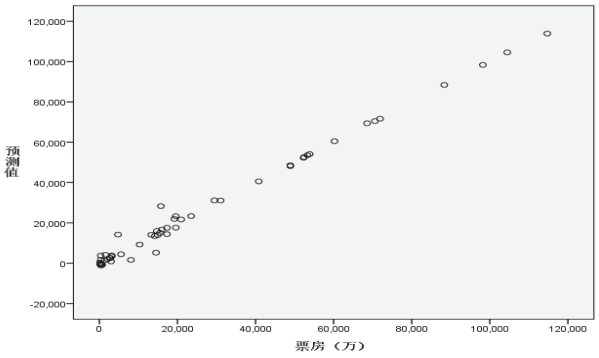

Figure 3 Scatterplot of predicted values

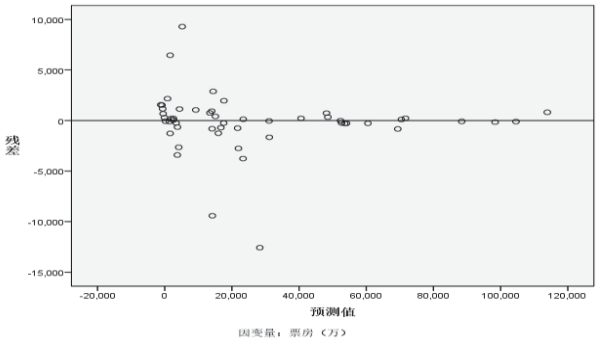

Figure 4 Residuals scatterplot of predicted values

As can be seen from the Figure 3, the scatter plots corresponding to the predicted value and the measured value almost stay in a straight line. And in Figure 4, the scatter plots are distributed around the x axial uniformly. We can draw a conclusion that the training result of network matches the sample well.

5 Conclusions

This paper has absorbed many experts' researches about the influence factors of movie box office. We extracted some indicators from them including actor, director, screenplay, production, distribution film, competition in film market and so on. And then we convert these indicators to the input layer of our neural network. By using the MLP we founded our nonlinear prediction model. And the original intent was realized. This model makes it possible for us to get the box-office output after we input some basic information of a movie after understanding the underlying mechanisms in movie box office.

Actually, this prediction model exists some errors. The factors which have effect to box office are rather more multiple and complex. We could not extract all of the value indicators about a movie only from the basic information. Many things relating to our life are likely to influence the box office. For example, if an earthquake takes place, a disaster movie may well attract a large amount of audience. Movie belongs to culture. If I would like to predict a movie box office, I should further understand the spiritual and physical life of the people on that moment.

Above all, this work let us begin trying to understand and solve problems in real life with the help of neural network.

References

[1] China Film Association.《China Film Industry Report 2015-2016》,2016-5-25.

[2] Kulasiri, D. and Verwoerd, V. Stochastic Dynamics: Modeling Solute Transport in Porous Media, North Holland Series in Applied Mathematics and Mechanics, Vol. 44, Elsevier, Amsterdam, 2002.Rosenblatt, Frank. x. Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms. Spartan Books, Washington DC, 1961.

[3] Rumelhart, David E., Geoffrey E. Hinton, and R. J. Williams. "Learning Internal Representations by Error Propagation". David E. Rumelhart, James L. McClelland, and the PDP research group. (editors), Parallel distributed processing: Explorations in the microstructure of cognition, Volume 1: Foundations. MIT Press, 1986.

[4] Cybenko, G. 1989. Approximation by superpositions of a sigmoidal function Mathematics of Control, Signals, and Systems, 2(4), 303–314.Xin Li,Orchard, M.T.. New edge directed interpolation. Image Processing, 2000. Proceedings. 2000,1:1521-1527.

[5] Simon Haykin.Neural Networks - A Comprehensive Foundation, 2nd ed.Prentice-Hall, Englewood Cliffs, 1998.

[6] De Wolf, E.D. and Francl, L.J. Neural networks that distinguish infection periods of wheat tan spot in an outdoor environment, Phytopathalogy,87, 83, 1997.

[7] Chiang, Y.M., Chang, L.C., and Chang, F.J. Comparison of static feedforward and dynamic feedback neural networks for rainfall-runoff modeling, Journal of Hydrology, 290, 297, 2004.

[8] Chaturvedi, D.K., Mohan, M., Singh, R.K., and Karla, P.K. Improved generalized neuron model for short-term load forecasting, Soft Computing,8, 370, 2004.

[9] Pearl A, Bar-Or R, Bar-Or D. An artificial neural network derived trauma outcome prediction score as an aid to triage for non-clinicians.[J]. Studies in Health Technology & Informatics, 2008, 136(136):253-8.