Expert Systems with Applications 36 (2009) 6580–6587

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

# Forecasting box office revenue of movies with BP neural network

Li Zhang *, Jianhua Luo, Suying Yang

*School of Electronic and Information Engineering, Dalian University of Technology, 2 Linggong Road, Dalian, Liaoning Province 116023, China*

A R T I C L E I N F O

*Keywords:*
Movies
Box office revenue
Forecasting
BP neural network
MLP
Dynamic threshold
Performance analysis

A B S T R A C T

Forecasting box office revenue of a movie before its theatrical release is a difficult and challenging problem. In this study, a multi-layer BP neural network (MLBP) with multi-input and multi-output is employed to build the prediction model. All the movies are divided into six categories ranged from "blob" to "bomb" according to their box office incomes, and the purpose is to predict a film into the right class. The selections of the input variables are based on market survey and their weight values are determined by using statistical method. As to the design of the neural network structure, theoretical guidance and plentiful experiments are combined to optimize the hidden layers' parameters which include the number of hidden layers and their node numbers. Then a classifier with dynamic thresholds is used to standardize the output for the first time, and it improves the robustness of the model to a high level. Finally, a 6-fold cross-validation experiment methodology is used to measure the performance of the prediction model. The comparison results with the MLP method show that the MLBP prediction model achieves more satisfactory results, and it is more reliable and effective to solve the problem.



## 1. Introduction

Movie is fixed upon by art, economy and science. It is merchandise in art and art in merchandise (De Vany, 2006). There has been over 100 years history after the birth of the movie, and it is undoubtedly the greatest achievement of science and art during the 20th century. The knowledge and research of the movie is becoming deeper along with its development. The movie is not only an important object for people to entertain and relax, but also a significant medium for different countries and areas to do cultural exchanges, and it has become an indispensable part of the world culture. Recent years, an independent and broadness newborn subject named the Movie Study has been formed (Ali, Fernandes, & Paton, 2003).

Movie has becoming a business, and it has huge market profit and potential (Shanklin, 2002). More and more people throw themselves into the film industry, and making movies are their jobs or even their interesting (De Vany & Walls, 2004). Every year, there are thousands of movies made all over the world, and some of them cost tens of million dollars or even hundred million dollars, and their box office is the main income source. There are many cinema line companies and thousands of cinemas in every country and the box office is their only income source. At present, there are about 3000 cinemas join in about 30 cinema line companies in China. The cinema line company has a great effect on the annual box office, so the box office forecast before selecting a movie is an important task of a cinema line company. Only on the base of accurate box office estimation of a film, can we determine the cinema number to show this film, the propaganda cost and the period of showing it in order to get more profit. The policymaker usually meets a series of multitudinous and amorphous problems (Suman & Subimal, 2008; Weng, Chu, & Wu, 2007). Moselhi and Hegazy have defined this kind of problem as "Unstructured, Analogy-intensive Type Problem" (Hegazy & Moselhi, 1994). Internal cinema line companies mainly use the brainstorming method to forecast the box office of a movie, and some researchers used the rend-calibrated movie-demand model to select a movie (To, Wong, & Li, 2000), but these ways lack of correct theory guidance and mature operating model. As a result, the benefit is not good in a movie's showing unit. Therefore, scientific box office forecast system has an important value for the development of the film industry.

Forecasting box office of a particular movie is a typical nonlinear problem. There are too many factors that affect a movie's box office, for example, the quality, showing time, advertising, society environment and weather during showing, and as well as the number of cinema to show a film, socket price, number of movies showing at the same time and so on. It is difficult to model this problem. Sharda and Delen (2006) have summarized the previous researches and the present research situation of this problem. Previous researches mainly focus on forecasting a movie's box office after it had shown a week or more time. They also design a method using multi-layer perceptron (MLP) neural network to solve the problem of predicting a movie's box office before the movie is shown, and obtained better result than other methods. But in their study, the absolute prediction accuracy (36.9%) is

* Corresponding author. Tel.: +86 411 88203433.
*E-mail address:* zhangli@dlut.edu.cn (L. Zhang).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.07.064

not high enough for users to accept it, and the relative accuracy (75.2%) is also not very high.

BP neural networks are the most widely used networks and are considered the workhorse of ANNs (Basheer & Hajmeer, 2000). Because of its simplicity and its power to extract useful information from samples, the application of BP model is very wide recently (Li, Yu, Mu, & Sun, 2006). It allows specification of multiple input criterion, and generation of multiple output recommendations, and no assumption regarding the form of the functions relating input and output variables. BP model eliminates the limitations of the regression method, and establishes the mapping accurately between the input and output variables (Kak, 2002). Due to its strong learning ability and generalization capability, BP networks have been used in a great deal of domains, especially in classification and prediction. Researchers had found the BP ANN model displayed more robust performance than other models in classification problems (Benjamin, Chi, Gaber, & Riordan, 1995; Hudon, Yan, & Kinsner, 1990; Masic & Pfurtscheller, 1993). The BP networks are also successfully used in forecasting some financial problems, for example, predicting stock market returns and price index (Enke & Thawornwong, 2005; Quah & Srinivasan, 1999; Roh, 2007), loan risk warning (Yang, Li, Ji, & Xu, 2001), forecasting bankruptcy firms (Lee, Booth, & Alam, 2005), and areas of decision support systems and management science (Delen, Sharda, & Kumar, 2007).

In our study, we use the multi-layer BP neural network (MLBP) to predict a movie's box office. It is very difficult to forecast the accurate box office value due to the output of a BP network using *S* transfer function is limited between 0 and 1. In order to forecast a movie's box office before its theatrical release, we also convert the prediction problem into a classification problem. The main thought is dividing all movies into several classes according to their box office revenues, and training the neural network with these samples, and then forecasting an unreleased movie by using the trained neural network. After scientific variables selection and successful modeling, we achieve an acceptable result better than the method using MLP neural network. In addition, our research is a longitudinal project, and it is supported by National Natural Science Foundation of China (Grant No. 60573172).

The remainder of this paper is organized as follows. Section 2 mainly describes the data forms of movie information and the selection method of variables. Section 3 gives the details of our methodology by specifically talking about the data, the prediction model, the experiment methodology and the performance measures used in this study. Next, the experimental results along with a comparative study of the MLBP model to the MLP model are shown and explained. Finally, Section 5 of this paper discusses the overall contribution of this study, along with its limitations and further research directions.

## 2. Data forms and variable selection

### 2.1. Movie classes and output form

Our two years box office data between 2005 and 2006 was gotten from Wanda Cinema Line Company in China, and we collected 241 movies after preprocessing with the original data. These sample movies were divided into six classes form blob to bomb, mainly because the number of movies is not big enough to divide more classes. In order to obtain a fair size of each class, we cut the sample set on an average 40 movies for a class. The detail of class rule is shown in Table 1.

The output data adopts discrete form, and each output has six bits. In the process of value assignment, all pseudo representations of an output are given the value of 0, except the one that holds true for the current class, which is given the value of 1.

**Table 1**
Movie classes

| Class no. | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Range (In 10000 RMB) | <4 blob | >=4 <10 | >=10 <30 | >=30 <90 | >=90 <200 | >=200 bomb |
| Number of movies | 33 | 47 | 49 | 46 | 35 | 31 |

### 2.2. Variable selection

It is necessary to determine the network's structure, for which we must determine the input variables first. To simplify the calculation and improve the system's efficiency, it is also necessary to select the input variables in reason. Chua, Kog, and Loh (1997) raised a method to determine important variables by wobble input values. But this method lacks supporting theory, and the description of concrete algorithm is not intuitive. In our study, we select most variables mentioned in the previous researches (Sharda & Delen, 2006), and we also chose some other variables based on combining the reality of Chinese market and our early statistical analysis. We given the weight value of each value depend on statistic data rather than objective surmise. All variables utilize continuous form, because in BP neural networks, using continuous variables can improve its sensitivity. We totally selected 11 variables, and then normalized every one into the [0,1] interval using the following formula except the nation:

$$f(s_i) = \frac{s_i - s_{\min}}{s_{\max} - s_{\min}}, \quad i = 1, 2, \ldots, n \tag{1}$$

in which $n = 241$ is the total number of samples, $s_i$ is some variable's value of the $i$th sample, $s_{\max}$ and $s_{\min}$ are the maximal and minimum value, respectively, of one variable in the whole sample set.

#### 2.2.1. Nation

Our study has shown that a movie's nationality is a factor influencing its box office. In china, movies mainly come from homeland and American, so we just divided this variable into two parameters, which include homemade and import. We have done statistic works to find the influence degree of the nation, and the result indicates that sometimes an imported motion picture has more attraction and box office performance, especially a super movie. We computed three box office summations of different kind movies: all movies, homemade movies and imported movies, and got the ratio of the later two and the total as their weight values. After that, the homemade weight is 0.474 and the other is 0.526.

#### 2.2.2. Star value

In some literatures, an important quality signal that influence the success of box office revenue of a movie (Sharda & Delen, 2006; De Vany & Lee, 2001), and our statistic data also displays the same result. The more super stars in a movie, the higher box office income will be obtained. These super stars can be either directors or players (actors, actresses). We have investigated some people, they expressed that the biggest fascination of a new movie is super star, they like movies directed by great directors or played by super stars and the other information is not very important. So we use both director and player as independent input variables to our model.

In order to get their fames, we searched every star's name in Google search engine, and took the record numbers limited between 50000 and 1200000 as the initial data, finally normalized it using the method mentioned above, the result is the ultimate popularity weight values of stars. The processes for directors and actors/actresses were executed independently. There would be more than one director or famous player in a movie, so we took the weight value summations of directors or main performers related to the movie as the final variable values for a sample.

*2.2.3. Propaganda*

Advertising is a very effective means to popularize a product, and the movie is no exception. The previous propaganda of a motion picture before it is released will has a profound influence to its box office revenue, and some companies even start to publicize a movie before it being made. The most ideal propaganda index is the invest cost of a film firm for a new movie, but this data is a confidential information for a company and cannot be known. To a great extent, the propaganda of movie can be gotten from its fame, so we used the same method of getting the star weight value to obtain its advert weight value.

*2.2.4. Content category*

The content category of movie is also important information of a movie, and our statistical data has shown the income disparity between different type movies is very evident. Internet movie database (IMDB) is the biggest, best, most award-winning movie site on the planet, and we got a movie's director, performers, and content category information all from this web site. We divide our movies into twelve types: Love, Cartoon, Hazard, Thriller, Horror, Warfare, Documentary, Family, Drama, Sci-Fi, Comedy and Action. We summed the total box office of movies related to a type, and then normalized the data into [0,1] interval, and the result is shown in Table 2. The total box office of action movies is too high, so we give it the value of 1.5 to emphasize this kind of movies. A movie maybe belongs to two types or more, and the type weight value is the summation of all the types.

*2.2.5. Showing time*

A movie's released time is also an effect factor to its box office success, and a common film would win a high income in hot season. We summed the total earning of all movies shown during a month, and put them together to do the normalizing option. The process was carried year by year, which means there are two groups of month weight values data in our study, and the final data is the average value of them as shown in Table 3. As to the week weight values, the valuation method is similar to the month, but there are 104 groups of data, and the result is shown in Table 4.

People have custom to celebrate a festival, especially some galas with holidays, and watching movie has becoming a way to enjoy a festive day. There are 18 festivals mentioned in this paper: Spring festival, Dragon boat festival, Children's day, Women's day, Easter day, Thanksgiving day, National day, Teachers' day, Labor day, Star festival (Chinese Valentine's day), Valentine's day, Christmas, Fools' day, New year's day, Lantern festival, Mid-autumn festival, Double ninth festival. Some of them are western festivals, and others are traditional Chinese festivals. The festival weight values were gotten through the same method defining the content category weight values, and the result is shown in Table 5. If there are other festivals or no festival related to a movie, the value of 0 will be given to this variable. In our study, festivals within 7 days before or after the released date of a movie are regarded. If there are two or more festivals related to a movie, the summation is rational.

*2.2.6. Competition*

Sharda and Delen (2006) considered the competition of a movie depends on the month when it is released, but this method is not convincing. A film's real competition comes from movies released at the same period. We counted the number of movies which released within 7 days of a movie's released date as the initial

**Table 2**
Weight value of types

| Type | Love | Cart. | Haz. | Thr. | Hor. | War. | Doc. | Fam. | Dra. | Sci-Fi | Com. | Act. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Weight | 0.14 | 0.323 | 0.293 | 0.455 | 0.028 | 0.038 | 0.009 | 0 | 1 | 0.842 | 0.456 | 1.5 |

Abbreviation captions: Cart. (Cartoon), Haz. (Hazard), Thr. (Thriller), Hor. (Horror), War. (Warfare), Doc. (Documentary), Fam. (Family), Dra. (Drama), Sci-Fi (Science Fiction), Com. (Comedy), Act. (Action).

**Table 3**
Month weight values

| Month | January | February | March | April | May | June | July | August | September | October | November | December |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Weight | 0.242 | 0.215 | 0.017 | 0 | 0.292 | 0.218 | 0.532 | 0.45 | 0.306 | 0.401 | 0.351 | 1 |

**Table 4**
Week weight values

| Week | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday | Sunday |
|---|---|---|---|---|---|---|---|
| Weight | 0.254 | 0.073 | 0.458 | 0.504 | 0.831 | 0.738 | 0.168 |

**Table 5**
Festival weight values

| Festival | Spring festival | Dragon boat day | Children's day | Women's day | Easter day | Thanks giving day |
|---|---|---|---|---|---|---|
| Weight | 0.97 | 0.368 | 0.314 | 0.349 | 0.452 | 1 |
| Festival | National day | Teachers' day | Labor day | Star festival | Valentine's day | Christmas |
| Weight | 0.885 | 0.995 | 0.304 | 0.334 | 0.42 | 0.72 |
| Festival | Fools' day | New year's day | Lantern festival | Mid-autumn day | Double ninth festival | Other |
| Weight | 0.044 | 0.742 | 0 | 0.534 | 0.125 | 0 |

competition data, and then put all the data together to normalize them with the method mentioned above. This data can effectively reflect the strife which a movie will face.

*2.2.7. Cinema information*

Commonly, a cinema line company has more than one cinema, and it usually releases a movie in several cinemas at the same time. In theory, the more cinemas in which a movie released, the higher income will be earned. We got the movie data from Wanda Cinema Line Company in China, so the value is the number of cinemas which adhere to the company and show the same new movie synchronously, and the final weight value is the normalized data.

Besides the number of cinema, the number of screen is also a meritorious contributor for a movie's business success. It provides how many times a movie is shown totally in all cinemas, and also implies its cumulate shown time. And the normalizing option is also adopted. The cinema information can be obtained from the showing plan of a new film made by a cinema line company.

A summary of the above-mentioned and briefly defined decision variables is given in Table 6. In total, 11 decision and six output variables are used in this study.

## 3. Predicting method

In our study, the BP neural network with multi-input and multi-output is employed to build the prediction model. In the following part, the structure and training method of this model will be explained in details.

**Table 6**
Summary of input variables

| No. | Variable name | Value range |
|---|---|---|
| 1 | Nation | Homemade = 0.474<br>Import = 0.526 |
| 2 | Director | [0,1] |
| 3 | Performer | [0,1] |
| 4 | Propaganda | [0,1] |
| 5 | Content category | [0,1] |
| 6 | Month | [0,1] |
| 7 | Week | [0,1] |
| 8 | Festival | [0,1] |
| 9 | Competition | [0,1] |
| 10 | Cinema number | [0,1] |
| 11 | Screen number | [0,1] |

*3.1. The structure of the prediction model*

The design of hidden layer is a very difficult and complex problem, especially to fix the number of hidden layers and their node numbers. Cybenko (1989) has proved that when every processing element utilizes sigmoid transfer function, one hidden layer is enough to solve any discriminant classification problem, and double hidden layers are capable to parse arbitrary output functions of input pattern. Lippmann (1987, 1989) estimated the number of hidden units according to his geometric explanation for multi-layer network function. He pointed out that the node number of the second hidden layer is $M \times 2$, where $M$ is the node number of output layer. As to the first layer, the best proportion of it and the second hidden layer is 3:1 when the input vector is high-dimensional (Gagan & Wei, 1989; Gorman & Sejnowski, 1988; Kung & Hen, 1991; Kung & Hwang, 1991). But the size of hidden layer and its unit number should also be gained from statistical estimation and actual experiment (Fujita, 1998; Hirose, Yamashita, & Hijiya, 1991; Krková, Kainen, & Kreinovich, 1997).

In this model, two hidden layers with sigmoid transfer functions were utilized based on the theories above, and our plentiful experiments also showed that double hidden layers will earn better result than single one. There are 30 nodes given to the first hidden layer, and 10 units for the second one, and the decision was made according to combination of previous researches and actual experiment. This kind of association is more perfect than others. The output layer also uses sigmoid function due to the BP algorithm. The output of the BP neural network is sent to a classifier, which translates it to the classification data of each input. The classifier has dynamic thresholds, and its transformation function is shown as the following formula:

$$f(y_i) = \begin{cases} 1 & y_i = \max[y_1, y_2, \ldots, y_6] \\ 0 & y_i \neq \max[y_1, y_2, \ldots, y_6] \end{cases} \qquad (2)$$

in which $i = 1,2,\ldots,6$, $Y = [y_1, y_2, \ldots, y_6]$ is the neural network output vector of an input. Finally, the prediction model was established successfully, and its structure is shown in Fig. 1.

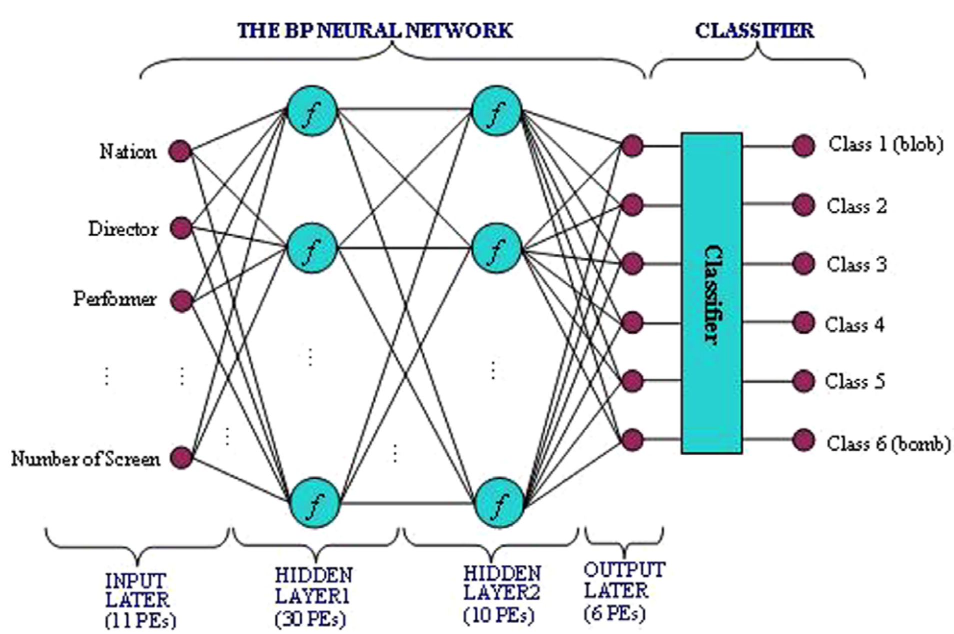


**Fig. 1.** The structure of the prediction model.

### 3.2. Experiment method

The BP algorithm for multi-layer neural network is employed in our model (Yibin, 2000). Let the actual output of the $k$th PE of the output layer is $y_k$, and its input is $net_k$, $y_j$ is the output of the $j$th PE of the second hidden layer, and then:

$$y_k = f(net_k) = f\left(\sum_j w_{kj} y_j\right) \tag{3}$$

here $k = 1,2,\ldots,6$, $j = 1,2,\ldots,10$, , $m = 10$ is the number of PE of the second hidden layer. $w_{kj}$ is the link weight of $k$th PE of the output layer and the $j$th PE of the second hidden layer, and its regulating value is as follows:

$$\begin{aligned} \Delta w_{kj} &= \eta \delta_k y_j \\ \delta_k &= (o_k - y_k) y_k (1 - y_k) \end{aligned} \tag{4}$$

in which $\eta$ is the learning rate, $o_k$ is the expected output of an input. In the same way, the adjustment value of $w_{ji}$ which is the link weight of the $j$th PE of the second hidden layer and the $i$th PE of the first hidden layer is as follows:

$$\begin{aligned} \Delta w_{ji} &= \eta \delta_j y_i \\ \delta_j &= y_j \left(1 - y_j \sum_k \delta_k w_{kj}\right) \end{aligned} \tag{5}$$

$y_i$ is the output of the $i$th PE of the first hidden layer, and the correction method of the link weights between the first hidden layer and the input layer is also in the same way.

Reliable estimates of classification accuracy are important, not only for estimating the true accuracy of a classifier, but also for finding the best classifier from a set of competitive ones (model selection). There is no universal learning algorithm giving the best performance in all possible learning situations (Schaffer, 1994). In this paper, we propose deterministic approaches for $k$-fold cross-validation that construct representative rather than random folds. In $k$-fold cross-validation, also called rotation estimation, the complete dataset ($S$) is randomly split into $k$ mutually exclusive subsets (the folds: $S_1, S_2, \ldots, S_k$) of approximately equal size. The classification model is trained and tested $k$ times. Each time ($t \in \{1,2,\ldots,k\}$), it is trained on all but one folds ($S \setminus S_t$) and tested on the remaining single fold ($S_t$). The cross-validation estimate of the overall accuracy is calculated as simply the average of the $k$ individual accuracy measures. With these methods we attempt to reduce the effects of using fewer instances for training. In our study, all samples are divided into six stratified groups, so a six-fold cross-validation is utilized.

### 3.3. Performance indexes

We used percentage success rate to measure the predictive performance of our neural network approach, and the main performance indexes include absolute accuracy and relative accuracy. Both of them are average percent hit rate (*APHR*) (Sharda & Delen, 2006). The absolute accuracy is the exact (*Bingo*) hit rate (only counts the correct classifications to the exact same class) and the relative accuracy is the within 1 class (1-*Away*) hit rate which reflects the instance that a movie predicted into its adjacent classes. The hit rate measures the average accurate classification rate of the actual outputs and the desired outputs. Algebraically, *APHR* can be formulated as in equations as follows:

$$APHR = \frac{\text{Number of samples correctly classified}}{\text{Total number of samples of a class}} \tag{6}$$

$$APHR_{Bing} = \frac{1}{n} \sum_{i=1}^{C} p_i \tag{7}$$

$$APHR_{1\text{-}Away} = \frac{1}{n} \sum_{i=1}^{C} (p_{i-1} + p_i + p_{i+1}) \tag{8}$$

**Table 7**
Confusion matrix for the BP neural network classification results

| MLBP Predicted Categories | Actual Categories | | | | | | Avg. |
|---|---|---|---|---|---|---|---|
| Class | 1 | 2 | 3 | 4 | 5 | 6 | |
| 1 | **24** | 9 | 3 | 0 | 0 | 0 | |
| 2 | 6 | **25** | 10 | 1 | 0 | 0 | |
| 3 | 3 | 13 | **26** | 5 | 0 | 0 | |
| 4 | 0 | 0 | 13 | **34** | 4 | 0 | |
| 5 | 0 | 0 | 0 | 6 | **24** | 4 | |
| 6 | 0 | 0 | 0 | 0 | 7 | **27** | |
| *Bingo(%)* | 72.7 | 53.2 | 53.1 | 73.9 | 68.6 | 87.1 | 68.1 |
| *1-Away(%)* | 90.9 | 100 | 93.9 | 97.8 | 100 | 100 | 97.1 |

**Table 8**
The results for each year and each group

| Predicted Categories | Class | Actual Categories 1 | 2 | 3 | 4 | 5 | 6 | Avg. |
|---|---|---|---|---|---|---|---|---|
| 2005 | 1 | 10 | 5 | 1 | 0 | 0 | 0 | |
| | 2 | 2 | 13 | 7 | 1 | 0 | 0 | |
| | 3 | 1 | 3 | 16 | 4 | 0 | 0 | |
| | 4 | 0 | 0 | 3 | 15 | 2 | 0 | |
| | 5 | 0 | 0 | 0 | 3 | 11 | 1 | |
| | 6 | 0 | 0 | 0 | 0 | 3 | 8 | |
| | Bingo(%) | 76.9 | 61.9 | 59.3 | 65.2 | 68.8 | 88.9 | 70.2 |
| | 1-Away(%) | 92.3 | 100 | 96.3 | 95.7 | 100 | 100 | 97.4 |
| 2006 | 1 | 13 | 6 | 1 | 0 | 0 | 0 | |
| | 2 | 4 | 14 | 4 | 0 | 0 | 0 | |
| | 3 | 2 | 5 | 10 | 3 | 1 | 0 | |
| | 4 | 1 | 0 | 7 | 15 | 3 | 1 | |
| | 5 | 0 | 0 | 0 | 5 | 15 | 3 | |
| | 6 | 0 | 0 | 0 | 0 | 0 | 19 | |
| | Bingo(%) | 65.0 | 56.0 | 45.5 | 65.2 | 79.0 | 82.6 | 65.5 |
| | 1-Away% | 85.0 | 100 | 95.5 | 100 | 94.7 | 95.7 | 95.1 |
| Group 1 | 1 | 8 | 3 | 1 | 0 | 0 | 0 | |
| | 2 | 2 | 7 | 3 | 0 | 0 | 0 | |
| | 3 | 0 | 4 | 8 | 1 | 0 | 0 | |
| | 4 | 0 | 0 | 4 | 12 | 3 | 0 | |
| | 5 | 0 | 0 | 0 | 1 | 11 | 2 | |
| | 6 | 0 | 0 | 0 | 0 | 1 | 9 | |
| | Bingo(%) | 80.0 | 50.0 | 50.0 | 85.7 | 73.3 | 81.8 | 70.1 |
| | 1-Away(%) | 100 | 100 | 93.7 | 100 | 100 | 100. | 98.9 |
| Group 2 | 1 | 9 | 3 | 0 | 0 | 0 | 0 | |
| | 2 | 2 | 9 | 2 | 0 | 0 | 0 | |
| | 3 | 3 | 4 | 8 | 2 | 0 | 0 | |
| | 4 | 0 | 0 | 2 | 11 | 0 | 0 | |
| | 5 | 0 | 0 | 0 | 3 | 7 | 1 | |
| | 6 | 0 | 0 | 0 | 0 | 4 | 10 | |
| | Bingo(%) | 64.2 | 56.3 | 66.7 | 68.8 | 63.6 | 90.9 | 68.4 |
| | 1-Away(%) | 78.6 | 100 | 100 | 100 | 100 | 100 | 96.4 |
| Group 3 | 1 | 7 | 3 | 2 | 0 | 0 | 0 | |
| | 2 | 2 | 9 | 4 | 1 | 0 | 0 | |
| | 3 | 0 | 5 | 10 | 2 | 0 | 0 | |
| | 4 | 0 | 0 | 5 | 11 | 1 | 0 | |
| | 5 | 0 | 0 | 0 | 2 | 6 | 2 | |
| | 6 | 0 | 0 | 0 | 0 | 2 | 7 | |
| | Bingo(%) | 77.8 | 52.9 | 47.6 | 68.8 | 66.7 | 77.8 | 65.3 |
| | 1-Away(%) | 100 | 100 | 90.5 | 93.8 | 100 | 100 | 97.4 |

where $C$ is the total number of classes (=6), $n$ is the total number of samples which belong to class $i$, and $p_i$ is the total number of samples classified as class $i$, and if $i < 1$ or $i > 6$, $p_i = 0$.

## 4. Results

### 4.1. BP neural network performance

In our study, MATLAB 7.0 is used to realize the neural network model and the algorithm. MATLAB is a powerful simulation platform developed by Mathworks, and it is very adaptive to simulate intelligent algorithms and to solve complex problems. In classification problems, confusion matrix is commonly used to represent the results. This kind of representation method is intuitive and easy to understand, so we also employ this way to show our results. Table 7 shows the aggregated six-fold cross-validation neural network results in a confusion matrix.

We unite the six groups' results into a single matrix, and both performance indexes are given. The columns in the confusion matrix represent the actual classes and the rows represent the predicted classes, and the diagonal elements are the numbers of movies that have been forecasted exactly. In order to explain that our model is a universal method, we also give the predicting results for the samples of 2005 and 2006. Besides, we divided our samples into three groups, and then simulated, respectively. Table 8 exhibits the five groups' result in confusion matrixes, and Fig. 2 gives the intuitive bar chart results. Comparing with the final result, we find that the predicted results do not have very great difference for diverse data types. This suggests that the predictable behavior of the model does not change over time for the variables included in the model and for different data types used in this study.

### 4.2. Comparison to MLP

In previous studies, MLP was the best model to solve the problem of predicting box office, and its result was better than traditional statistical classification methods, such as discriminant analysis and multiple logistic regression, and a popular decision tree method called classification and regression trees (CART) (Sharda & Delen, 2006). The MLP method chose 26 decision (representing seven categories) variables as input vector, and all of them utilized discretization form except the number of screen. We also managed our movie data in this way. In the following section, we present the results of our method compared to the MLP method. We used exactly the same training and testing data set generated by using stratified six-fold cross-validation for the model and the BP neural network model. The MLP aggregated results are shown in Table 9. We also put the results of every class into a bar chart as shown in Fig. 3, which presents *Bingo* and 1-*Away* respectively. As the results indicate, on an average, BP neural network model generates significantly better classification accuracy than the MLP method.

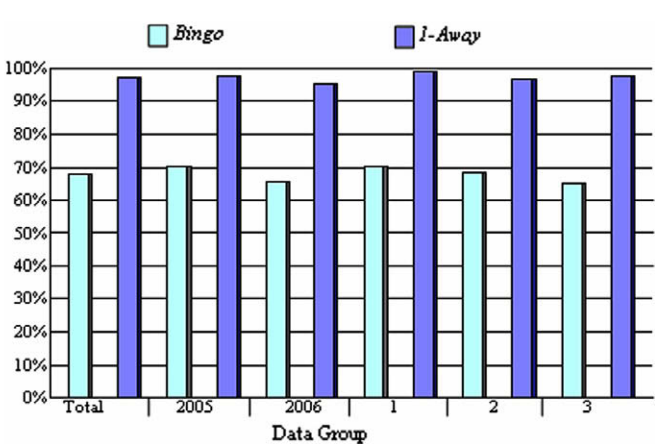


**Fig. 2.** Comparison results of the five groups presented in a bar chart.

### 4.3. Advantages of the prediction model

In order to explain the superiority of our model, we also employed the 26 decision variables used in the MLP method as the input variables of our prediction model. The experiment methods are as same as using our 11 continuous variables. The results are shown in Table 10, and (Fig. 4) gives a result comparison bar chart during MLP, MLBP with 26 input variables and MLBP whit 11 input variables. From the comparison results of above three different methods, we can get the following conclusions: (i) our model is even better than MLP when the same variables are employed, (ii) the variables selected in our model are more reasonable than the MLP method.

**Table 9**
The results of the MLP method

| | | Actual Categories | | | | | | Avg. |
|---|---|---|---|---|---|---|---|---|
| | Class | 1 | 2 | 3 | 4 | 5 | 6 | |
| MLP Predicted Categories | 1 | **10** | 3 | 1 | 1 | 0 | 0 | |
| | 2 | 11 | **15** | 11 | 3 | 3 | 0 | |
| | 3 | 3 | 17 | **16** | 2 | 4 | 0 | |
| | 4 | 6 | 7 | 13 | **20** | 8 | 4 | |
| | 5 | 1 | 5 | 4 | 13 | **15** | 9 | |
| | 6 | 2 | 0 | 4 | 7 | 5 | **18** | |
| *Bingo*(%) | | 30.3 | 31.9 | 32.7 | 43.5 | 42.9 | 58.1 | 39.9 |
| *1-Away*(%) | | 63.6 | 74.5 | 81.6 | 76.1 | 80.0 | 87.1 | 77.2 |

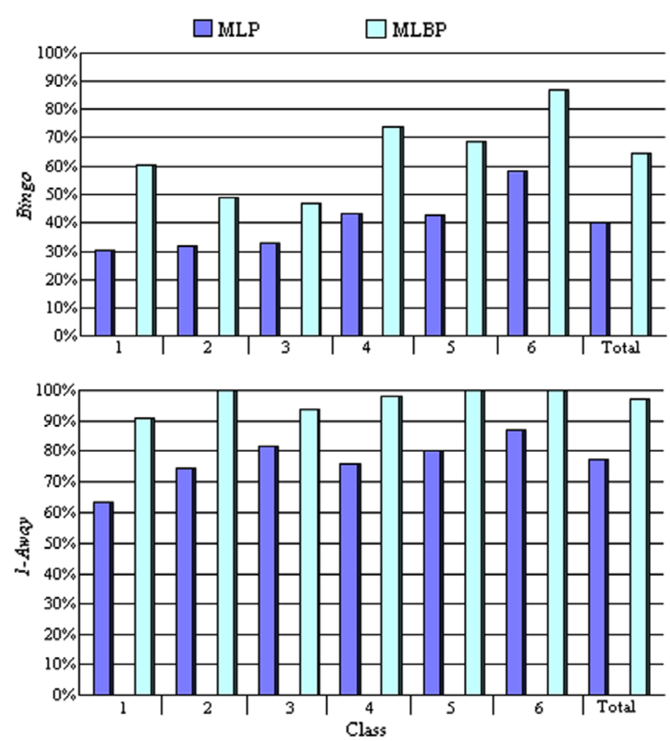


**Fig. 3.** Comparison results of MLP and MLBP presented in a bar chart.

**Table 10**
The results of MLBP with 26 variables

| MLBP (26 Var.) Predicted Categories | Actual Categories | | | | | | Avg. |
|---|---|---|---|---|---|---|---|
| Class | 1 | 2 | 3 | 4 | 5 | 6 | |
| 1 | **12** | 6 | 5 | 2 | 0 | 0 | |
| 2 | 11 | **20** | 6 | 3 | 1 | 0 | |
| 3 | 5 | 11 | **26** | 10 | 4 | 1 | |
| 4 | 2 | 8 | 7 | **16** | 7 | 3 | |
| 5 | 1 | 2 | 4 | 11 | **19** | 9 | |
| 6 | 0 | 0 | 0 | 4 | 6 | **19** | |
| *Bingo*(%) | 38.7 | 42.6 | 54.2 | 34.8 | 51.4 | 65.5 | 47.9 |
| *1-Away*(%) | 74.2 | 78.7 | 81.3 | 80.4 | 86.5 | 96.6 | 82.9 |

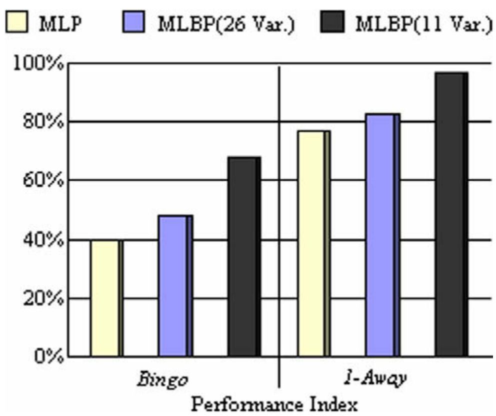


**Fig. 4.** Comparison results of different methods.

In the classifier section of our model, we utilize dynamic thresholds rather than a fixed one to standardize the output. We set the maximum output of the neural network equals to 1, and the others equal to 0. The main advantage of this classification method is that its generalization ability is strong even the training precision is not very high. The training accuracy ($P_{\text{Train}}$) is computed by the equation as follows:

$$P_{\text{Train}} = \frac{\text{Right training sample number}}{\text{Total number of training samples}} \qquad (9)$$

Finally, we did some simulations with different training precisions, and the results are shown in Fig. 5, from which we can see that our model has strong robustness, and it can obtain a high forecasting performance. In our study, the best training accuracy is 95.13%, which corresponds to the final results.

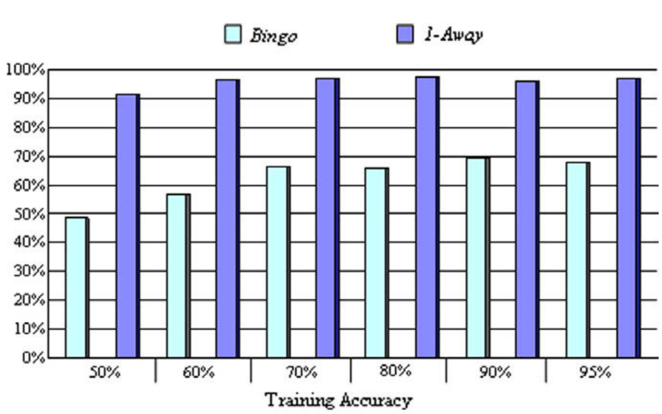


**Fig. 5.** Comparison results of MLBP on different training accuracies.

## 5. Conclusion and discussion

The results show that the multi-layer BP neural networks employed in this study can predict the success category of a motion picture before its theatrical release with pinpoint accuracy (*Bingo*) with 68.1% and within one category (1-*Away*) with 97.1% accuracy. Compared with the MLP method, these prediction accuracies are significantly better. Besides, we also prove the advantages of model on different sides, for example, the stability and superiority of our model, the reasonable of our variable selection method, and the robustness of our model. The experiment results are the most convincing evidence to support our conclusions. The previous studies mainly aimed at predicting the box office of a movie after its initial theatrical release, and they could not supply an effective support for the decision of a cinema line company. The MLP method proposed by Sharda and Delen was the first attempt to foresee the box office income before its theatrical release, and their experiment achievements were also better than any other traditional classification methods.

The MLBP neural network model suggested in this study is also used to forecast the financial success of a movie before putting it into market, and it is a huge improvement of the MLP method by utilizing the architectural parameters such as learning algorithm, number of hidden layers and their own number of PEs, etc. We fixed on the input variables based on market survey and then used statistical method to determine the weight value of every variable. For the design of the neural network structure, it is a very important and pivotal influence factor of the neural network performance, especially the design of hidden layers. So, we did a lot of preparation work on optimizing the hidden layers of the neural network. We referred to some eloquent research achievements on the optimization of neural network structure used in classification problems, especially its hidden layers. In addition, we did not depend on the theoretical guidance only, but also did plentiful simulations to find the best structure parameters of the model. We employed the experiment methodology named *k*-fold cross-validation to train the neural network. Finally, a classifier with dynamic thresholds was firstly used to standardize the output, and it improved the robustness of the model to a high level. After the scientific design, our prediction model achieved satisfactory results.

In any case, the results of this study are very attractive, and prove the continuous value of BP neural networks in solving difficult prediction problems. Beyond the accuracy of our results in predicting box office success, the prediction model could also be adapted to forecast the success rates of other media products. The particular parameters used within the model of a movie or other media products could be altered using the already trained BP neural network model. During this alternative experiment process, the manager of a given entertainment firm could find out, with a fairly high accuracy level, how much a specific actor, a specific release date, or the addition of more technical effects, could mean to the financial success of a film, or a television program.

The number of samples used in this study is limited. It is necessary to enrich the sample movies, and let the classification more logical. Furthermore, using more samples to train the neural network can improve its generalization ability. The accuracy of the neural network model presented in this study can be improved by adding or improving some of the other determinant variables such as exact production budget and advertising budget, which are known to be industry trade secrets and are not publicly released. Some of input variables selected in our research may be unnecessary for the output, or apply less contribution to the final decision. So it is requisite to find a valid theory or technique to reduction the number of variables. Lately, researchers have been developing some knowledge reduction methods to optimize the parameters of neural networks, for example, rough set theory, genetic algorithm, etc. Application of such methods can improve the results that we have obtained in this study.

From an application perspective, once developed to a production system or a commercialization software platform, such a prediction model can be made available (via a web server or as an application service provider) to industrial decision makers, where individual users can plug in their own movie parameters to forecast the potential success of a motion picture before its theatrical release. A neural network model can be designed in a way such that it can calibrate its weights (continuous self learning) by taking into account new samples (movies that are released and determined box office receipts) as they become available. Much additional work, in terms of modeling extensions, further experimentation for testing the performance, and applications to other media product demand forecasting, remains to be done, and this is also the striving direction of our further research.

## References


Ali, M. Akhtar, Fernandes, Alvaro A. A., & Paton, Norman W. (2003). MOVIE: An incremental maintenance system for materialized object views. *Data & Knowledge Engineering, 47*(2), 131–166.

Basheer, I. A., & Hajmeer, M. (2000). Artificial neural networks: Fundamentals, computing, design, and application. *Journal of Microbiological Methods, 43*(1), 3–31.

Benjamin, Colin O., Chi, Sheng-Chai, Gaber, Tarek, & Riordan, Catherine A. (1995). Comparing BP and ARt II neural network classifiers for facility location. *Computers & Industrial Engineering, 28*(1), 43–50.

Chua, D., Kog, Y., & Loh, P. (1997). Model for construction budget performance-neural network approach. *Journal of construction engineering and management, 123*(3), 214–222.

Cybenko, G. (1989). Approximation by superpositions of sigmoidal function. *Mathematics of Control Signals, and System, 2*, 303–314.

De Vany, Arthur (2006). *Chapter 19: The movies handbook on the economics of art and culture* (vol. 1). Amsterdam: Elsevier Press.

De Vany, Arthur, & Lee, Cassey (2001). Quality signals in information cascades and the dynamics of the distribution of motion picture box office revenues. *Journal of Economic Dynamics and Control, 25*(3–4), 593–614.

De Vany, Arthur S., & Walls, W. David (2004). Motion picture profit, the stable Paretian hypothesis, and the curse of the superstar. *Journal of Economic Dynamics and Control, 28*(6), 1035–1057.

Delen, D., Sharda, R., & Kumar, P. (2007). Movie forecast guru: A web-based DSS for Hollywood managers. *Decision Support Systems, 43*(4), 1151–1170.

Enke, David, & Thawornwong, Suraphan (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications, 29*(4), 927–940.

Fujita, Osamu (1998). Statistical estimation of the number of hidden units for feedforward neural networks. *Neural Networks, 11*(5), 851–859.

Gagan, Mirchandani, & Wei, Cao (1989). On hidden nodes for neural nets. *IEEE Transactions on Circuits and Systems, 36*(5), 661–664.

Gorman, R. P., & Sejnowski, T. J. (1988). Analysis of hidden units in a layered network trained to classify sonar targets. *Neural Networks, 1*, 75–89.

Hegazy, T., & Moselhi, O. (1994). Analogy-based Solution to Markup Estimation Problem. *Computing in Civil Engineering, 8*(1), 72–87.

Hirose, Yoshio, Yamashita, Koichi, & Hijiya, Shimpei (1991). Back-propagation algorithm which varies the number of hidden units. *Neural Networks, 4*(1), 61–66.

Hudon, S., Yan, Y., & Kinsner, W. (1990). A comparative study of neural network models. *Mathematical and Computer Modelling, 14*, 300–304.

Kak, Subhash (2002). A class of instantaneously trained neural networks. *Information Sciences, 148*(1–4), 97–102.

Krková, Vra, Kainen, Paul C., & Kreinovich, Vladik (1997). Estimates of the number of hidden units and variation with respect to half-spaces. *Neural Networks, 10*(6), 1061–1068.

Kung, S. Y., & Hen, Hu Yu (1991). A Frobenius approximation reduction (FARM) for determining optimal number of hidden units. *IEEE, II*, 163–168.

Kung, S. Y., & Hwang, J. N. (1991). An algebraic projection analysis for optional hidden units size and learning rate in back-propagation learning. *IEEE, I*, 363–370.

Lee, Kidong, Booth, David, & Alam, Pervaiz (2005). A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms. *Expert Systems with Applications, 29*(1), 1–16.

Li, Qiang, Yu, Jing-Yuan, Mu, Bai-Chun, & Sun, Xu-Dong (2006). BP neural network prediction of the mechanical properties of porous NiTi shape memory alloy prepared by thermal explosion reaction. *Materials Science and Engineering: A, 419*(1–2), 214–217.

Lippmann, R. P. (1987). An introduction to computing with neural nets. *IEEE ASSP Magazine*, 4–32.

Lippmann, R. P. (1989). Pattern classification using neural networks. *IEEE Communications Magazine*, 47–64.

Masic, Nikola, & Pfurtscheller, G. (1993). Neural network based classification of single-trial EEG data. *Artificial Intelligence in Medicine, 5*(6), 503–513.

Quah, Tong-Seng, & Srinivasan, Bobby (1999). Improving returns on stock investment through neural network selection. *Expert Systems with Applications, 17*(4), 295–301.

Roh, Tae Hyup (2007). Forecasting the volatility of stock price index. *Expert Systems with Applications, 33*(4), 916–922.

Schaffer, C. (1994). A conservation law for generalization performance. In *Proceedings of the 11th international conference on machine learning* (pp. 259-267). New Brunswick, NJ: Morgan Kaufmann.

Shanklin, William (2002). What businesses can learn from the movies. *Business Horizons, 45*(1), 23–28.

Sharda, Ramesh, & Delen, Dursun (2006). Predicting box-office success of motion pictures with neural networks. *Expert Systems with Applications, 30*(2), 243–254.

Suman, B., & Subimal, C. (2008). Fast and frequent: Investigating box office revenues of motion picture sequels. *Journal of Business Research, 61*(7), 789–803.

To, T. -P. J., Wong, Koon-Hung, & Li, Chi-Kwong. (2000). Strategic selection and replication of movies by trend-calibrated movie-demand model. In *Proceedings international symposium on multimedia software engineering* (pp. 97–100).

Weng, C., Chu, W., & Wu, J. (2007). Movie analysis based on roles' social network. In *IEEE international joint conference on multimedia and expo* (pp. 1403–1406).

Yang, Baoan, Li, Ling X., Ji, Hai, & Xu, Jing (2001). An early warning system for loan risk assessment using artificial neural networks. *Knowledge-Based Systems, 14*(5–6), 303–306.

Yibin, Song (2000). Quick training method for multi-layer bp neural network and its application. *Control and Decision, 15*(1), 125–127.