Neural Network Based Movie Rating Prediction

SU YuMin
School of Science and
Technology Communication
University of China
Beijing China
cuc_sym@cuc.edu.cn

ZHANG Yuan
Computer Network Center
Communication University of
China
Beijing China
yuanzhang@cuc.edu.cn

YAN JinYao
Key Laboratory of Media Audio &
Video
Communication University of
China, Beijing China
jyan@cuc.edu.cn

ABSTRACT

Movie rating is a numerical comprehensive evaluation of movie viewers. It represents an objective basis for general audience and advertisers to choose high quality movie, and is a crucial factor affecting row piece and estimate final box office receipts. As a result, predicting the ratings of a new movie is a great convenience for the audience, cinema, advertisers and investors. This paper identifies the movie rating prediction factors, designs a series of movie evaluation metrics, and proposes a movie rating prediction model based on the neural network algorithm. After preliminary data preprocessing, our algorithm can reach a movie rating prediction accuracy up to 70%. Through the improvement of the starring metric, the prediction accuracy can be further improved to 88.8%. Compared with other related work of movie rating prediction and movie recommendation systems, our method has high accuracy and does not require user information. We compare our solution with several existing methods. Evaluation results show that the prediction accuracy would improve by 89.8%, 48.0% ,50.5% and 37.2% compared with the knn algorithm, the decision tree algorithm, the SVM algorithm, the NBC algorithm respectively.

CCS Concepts

• CCS → Computing methodologies → Machine learning → Machine learning approaches → Instance-based learning

Keywords

Prediction, neural network, movie rating, data processing

1. INTRODUCTION

With the development of movie technology and the needs of modern social life, in 2007, the annual output of movies in the world has reached about 4000[1]. Statistics show that there are about 10 to 300 thousand movies in the world [1]. In the multitude of movies, how to choose the movie we need? Movie rating prediction, as a core of movie evaluation, has important reference value and application value to cinema, investors and ordinary audience. For movie theaters, movies of higher rating tend to have higher attendance and better returns [2]. Therefore, the movie rating prediction has important application value for the arrangement of movies. For investors, predicting the box office trend according to the prediction of movie rating has a high reference value for the investment of the latter decisions of the investors [3]. For the general audience, a movie of high rating usually means it somehow has a value to watch [2]. The movie rating prediction also acts as a reference for audience choice.

This paper will provide a movie rating prediction algorithm to accommodate the needs from various aspects. Through the analysis of the influencing factors in the early stage of the movie release, this paper designs a movie rating prediction model with high prediction accuracy. Compared with other related work, our method has high accuracy and does not require user information.

The following sections of this paper are arranged as follows: the second section introduces the related work and the application of the corresponding scene; the third section introduces a movie rating metric evaluation and prediction system based on neural network algorithm; the fourth section explains the source of the data, and gives a detailed description of the three steps data preprocessing method; the fifth section analysis the experimental results on different data and make comparison with several existing algorithms. At last we summarize our work and discuss several open issues.

2. RELATED WORK

There are several existing studies on movie rating prediction and movie recommendation systems. Yew Jin Lim et. al have studied the application of Bayesian method of singular value decomposition to the prediction of movie rating [4]. They proposed a comprehensive use of variational inference on a priori parameters. Experiments were conducted with the user's scoring data, and the results showed that their proposed Bayes method achieved a 2% improvement and handled the unobserved entries properly.

Jennifer Golbeck et. al have proposed a social network based movie prediction recommender system [5]. It is based on a movie trust system which used the user's average influence on the comments to improve collaborative filtering in social network. Experimental results show the accuracy rate can reach up to 50%.

A prominent study on movie rating prediction is movie grade prediction model on user attributes proposed by Wang Jiang et. al [6]. By analyzing the movie rating data, [6] finds that users with similar attributes tend to share similar preferences, and users with special attributes have their own preferences. Based on these two observations, the potential model is used to prediction the rank, and the average absolute error is reduced to 71%.

Hu Miaoyuan uses the latent variable model as the uncertain

ICBDC '18, April 28–30, 2018, Shenzhen, China
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6426-3/18/04…$15.00
http://doi.org/10.1145/3220199.3220204

knowledge representation, and he builds a user behavior preference model which uses the domain knowledge. Using the movielens data set which contains the attribute of movie and user, he can get the movie rating through UPM model, and the prediction accuracy rate can reach 72% [7].

Tian Maogen proposed a recommendation algorithm based on the super network model, which used the super network model to model movie rating of users [8]. Compared with the traditional recommendation algorithm, the recommendation algorithm based on the super network model fully excavates the user's own interest factors, and the average absolute error can be as low as 16%.

All the studies above have used the attributes of user data, but in our study, we only use the attributes of movie data to predict the movie rating. So, we have to design a new method to suit our data.

3. MOVIE RATING PREDICTION MODEL

3.1 Movie Evaluation Metrics

Many factors would influence the movie rating. In this paper, we evaluate the significance of a movie from three aspects, namely the type of the movie’s production, the level of the movie’s publicity, and the amount of attention it draws.

From a perspective of a movie's production type, this paper selects the release date as a measure of the timeliness of the movie.

From a perspective of a movie’s publicity, the influence factors are selected as three metrics: attention degree, publicity degree and soft lift degree. Number of raters is chosen as a measure of attention. the trailer number is chosen as a measure of publicity degree. and we select star metrics the measure of the rise of soft movie.

From a perspective of the amount of attention a movie draws, the influence factors are selected as two indicators: attraction degree and hot degree. The number of people who want to see the movie is the attraction degree, and comprehensive number of the brief comment and the reviews is on behalf of the movie's hot degree.

Based on the above analysis, the impact factors of the movie rating prediction value can be divided into three layers, and the bottom floor has 7 factors. This paper selects 7 basic data corresponding to the 7 factors mentioned above, and designs the movie evaluation metrics according to the above idea. The overall scheme chart is shown in figure 1.

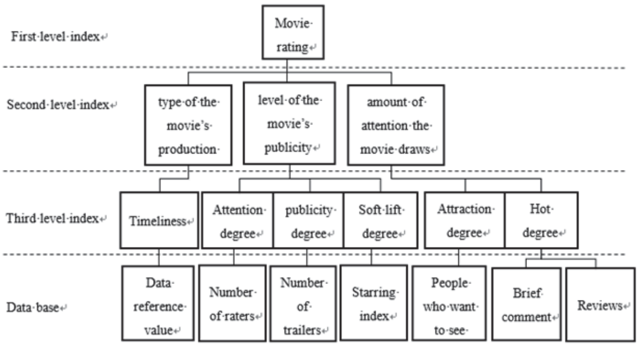

Figure 1 movie evaluation metrics

The indicators of the evaluation metrics are as follows:

On the data base level, there are the date reference value, the number of raters, the number of trailers, people who want to see the movie, the brief comment and the reviews are the original captured numerical data.

On the third level, the timeliness is taken as the preprocessed data of the released date. Attention degree, publicity degree, attraction degree and hot degree are the post value or weight value after data pretreatment.

The starring metrics the comprehensive rating of the movie star, which has been improved by data preprocessing.

The second level metrics are the classification and synthesis of the three indicators, and the rating of the movie in the three major categories is obtained. The formulations are as follows.

$$MTP= tl \ (1)$$

$$MLP= atte + pub +sl \ (2)$$

$$MAD = attra+ hot \ (3)$$

Here *MTP* represents the type of the movie’s production, *tl* represents timeliness, *MLP* represents the level of the movie’s publicity, *atte* represents attention degree, *pub* represents publicity degree, *sl* represents soft lift degree, *MAD* represents the amount of attention a movie draws, *attra* represents attraction degree, *hot* represents hot degree.

The third level metrics the synthesis of the second level metrics, and the result is the final movie rating value. Formula is as follows.

$$Movie\ rating = MTP+MLP + MAD \ (4)$$

3.2 Movie Rating Prediction Based on Neural Network Algorithm

3.2.1 Movie Rating Prediction Model

On the base of the movie rating evaluation metrics and the neural network algorithm, this paper designs a movie rating prediction model. The model scheme is shown in figure 2.

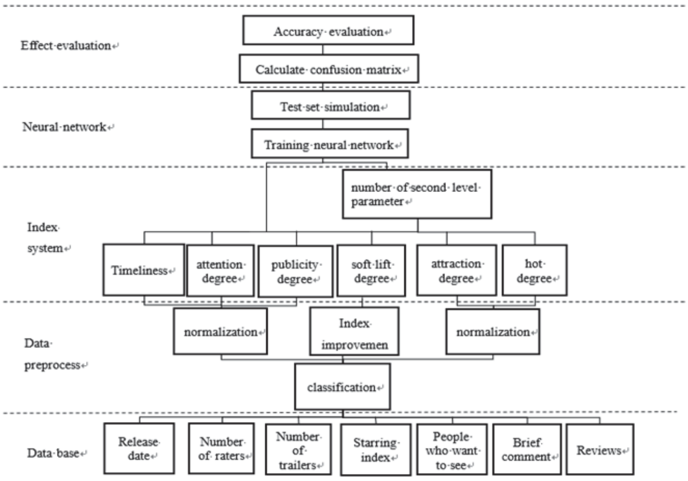

Figure 2 movie rating prediction model

The model is illustrated as follows:

The model is divided into five layers. The first layer of the model is the data base, including 7 specific data, which is grabbed directly from the network and preliminary cleaned.

The second layer is the data pretreatment layer, through the classification, data normalization, metric improvement and other steps, the data from data base gets processed, specific processing methods and standards are detailed in the fourth chapter.

The third layer is the metrics, and the preprocessed data are put into the metrics to get the third level metric value of each movie and the number of the second level metric parameters.

The fourth layer is the neural network layer. After the data have been randomly divided into training set and test set, the training set is used to train the neural network, and the adjustment of parameters are continuously improved. The test set is simulated, and the prediction results are obtained.

The fifth layer is the effect evaluation layer, the confusion matrix is calculated by the prediction results of the neural network layer, the prediction accuracy is obtained by the confusion matrix, and the effect of the model is evaluated.

3.2.2 Movie Rating Prediction Based on Neural Network Algorithm

In the light of the characteristics of the movie rating prediction, 7 variables are analyzed. After repeated experiments on the same set of data, the optimal algorithm flow and default parameters are determined, and the pseudo code of the algorithm is expressed as follows:

```
Input:
Traindata[n][10]: N movie for training, the traindata[i][]=
{date reference value, number of raters, number of trailer, starring metric, number of who want to see, number of brief comment, number of reviews, movie rating};
Testdata[m][9]: M movie for test, the testdata[i][]=
{date reference value, number of raters, number of trailer, starring metric, number of who want to see, number of brief comment, number of reviews};
Reference[m]: the true rating of the M movie for testing
E: single maximum error allowed
Count: the number of consecutive predictions
Output:
Prerating[m]: the predictive rating of the M movie for testing
C.M: the total confusion matrix of the M movie in the test set
BEGIN:
//Start algorithm
E0 = init (e);
/ / initialize error
Traininput = pretreata (traindata);
//Get training set of third level metric
Testinput = pretreata (testdata);
//Get test set of third level metric
N.neurons = lenth (traininput[i][10] - 1);
N.hide = pretreatb (traininput);
//Get the number of parameters in second level metric
FOR i = 1 to M
    FOR j = 1 to count
        IF e0 > e
/ / initialize neural network
Net = init (n.neurons, n.hide, learning.rate);
Res = train (net, traininput);
/ / train neural network
            Pre[i][j] = sim (RES, testinput[i][]);
//test set Simulation
            C.M[i][j] = confusionmatrix (Reference, Pre)
//Get single confusion matrix
            E0 = Get (C.M)
//get the prediction error from the confusion matrix
            ELSE
                Prerating[i] = Pre[i][j]
/ / predict movie rating
        END
    END
    C.M = C.M[m][j-1]
//Get the final confusion matrix
END
/ / the end of the algorithm
```

According to the above algorithm, the final prediction rating of movie i can be expressed by formula (5).

Prerating[i]=bmrpmodel (testdata[i][ ] (5)

Where testdata[i][ ] represents the data value of the 7 previous factors of the movie i, bmrpmodel is the movie rating prediction algorithm proposed in this paper, and prerating[i] represents the final rating predictive value of the movie i.

4. DATA PROCESSING

4.1 Data Sources

Based on the 7 basic data that affect the movie rating prediction, the required information is captured from the douban movie details page and the star details page [9]. We Grab all the movies shown in China mainland in 2016 and delete the movie entries with incomplete information. In the end, we obtain 447 movies with complete information. We grab 7 star indicators of the selected movie from douban star page, which is the starring age, gender, awards, the latest works rating, praise works rating, micro-blog number and the fan number.

4.2 Data Preprocessing

4.2.1 Data Normalization

The difference of movie quality is too large that the data still has high volatility value after initial cleaning. We normalize the data to further study the impact of each metric on the movie rating prediction.

Release date: We use 2016/12/31 minus the release date, then divide the result by the number of days in 2016. The final value is between [0,1].

Number of raters: we define the number of raters between $[0{,}5000)$ as for 1, between $[5000{,}10000)$ as for 2, between $[10000{,}20000)$ as for 3, no less than 20000 as for 4.

Similarly, watch intention, trailer number, number of brief comment and number of reviews are empowered to 1 to 4.

4.2.2 Starring Metric

To determine the best feature of the starring metric, the movie data of 2016 are used to integrate each of the starring indicators. The starring indicators include the average rating of the last five productions, the average rating of the five highest productions, awards, gender, age, number of Micro-blogs, number of Micro-blog fans. Details are as follows.

The average rating of the last five productions and the five highest productions are both in the range from 1 to 5. We use 2 for the actors/actresses/directors with awards, and 1 for no award. The weight of men is 2, the weight of women is 1, the weight of unknown is 0. We define age between $[1{,}16)$ as for 1, between $[16{,}35)$ as for 2, between $[35{,}48)$ as for 4, no less than 48 as for 3, unknown as for 0. And number of Micro-blogs, number of Micro-blog fans are empowered to 1 to 4 in the same way.

The average number of actors is counted and expanded by 100 times, which is used as the starring metric of the corresponding movie. Formula is as follows.

$$\textit{starring index}=100\ X \quad \frac{\textit{(latest work score+acclaimed work score+Awards+Gender} \atop \textit{+age+Microblog number+Microblog fans)}}{7} \qquad (6)$$

5. EXPERIMENTAL RESULTS AND ANALYSIS

The experimental results are measured by accuracy. The accuracy is defined as follows:

$$accuracy = \frac{pre}{ref} \qquad (7)$$

Here Pre represents the number of movies in which the results of the test set are consistent with the reference results and ref represents the total number of movie entries in the test set.

5.1 Preliminary Prediction Results

We randomly disrupt the data and select 90% of the entries as training set and the remaining as test set to feed our model. Table1 is the confusion matrix of prediction result. From Table 1, we can see the accuracy is around 77%.

When the normalized data is applied to the movie rating prediction model, the accuracy can reach 77%, which is close to the level of practical application.

Table 1 confusion matrix of prediction result

confusion matrix and statistics		
	reference	
prediction	**0**	**1**
0	154	26
1	26	19
accuracy	76.89%	

5.2 Improved Prediction Results of with Starring Metric

To improve the prediction accuracy, we use the starring metric data processing method in 4.2.2. We still randomly disrupt the data and select 90% as training set and the remaining 10% as test set. We repeatedly select the training and testing sets for 20 times, and the accuracy rate is always between 80%~90%. Table 2 shows part of the average prediction results of 20 tests.

Table 2 partial prediction results of improved starring metric

reference	Pre(BMRP)
2	3
3	3
2	2
4	3
2	2

The first column is the actual rating of the movie, and the second is the prediction result of the BMRP algorithm. The confusion matrix of the average of 20 times predictions is shown in table 3:

Table 3 the confusion matrix of prediction results after starring metric improved

confusion matrix and statistics		
	reference	
prediction	**0**	**1**
0	158	14
1	10	33
accuracy	88.83%	

As can be seen from Table 3, the prediction accuracy and be improved using additional starring information. And the average accuracy is above 80%, this prediction result has a high reference value for the real rating.

5.3 Compare with Other Algorithms

We compare our neural network based prediction model with several classical prediction methods. Specifically, we compare with knn, decision tree, SVM and Naive Bayesian Classifier (NBC) methods based on the douban data set.

We still use the 2016 movie data to evaluate the movie rating algorithms. We run the algorithms on three randomly selected training and testing data sets, and list the results in table 4.

Table 4 accuracy comparison table of three algorithms

	Test1	Test2	Test3	Improve
BMRP (%)	83.1	85.4	90.2	
knn (%)	38.6	47.6	50.1	89.8
TREE (%)	54.8	58.1	61.9	48.0
SVM (%)	57.1	55.3	59.5	50.5
NBC (%)	61.9	66.3	60.4	37.2

From the tables above, we can see that in terms of accuracy, compared to the other four kinds of algorithm, the movie rating prediction accuracy of the model are greatly improved, the average increase is more than 37%, thus, the prediction accuracy of the model in the rating was significantly higher than that of knn algorithm, decision tree algorithm, SVM model and NBC model.

In the aspect of complexity, the prediction model proposed in this paper only needs to do the preprocessing of the data, obtain the layer parameters of the neural network, and then calculate the prediction value through the Metrics and neural network. The decision tree prediction model needs many experiments to determine the optimal decision tree, and finally gets the rating prediction value of the movie. Although the knn prediction model is simple, the accuracy rate is too low. The mathematical calculation of SVM model is complex, so it is necessary to find the appropriate kernel function. The NBC model is suitable for the classification prediction of simple attributes. When the number of attributes is large or the correlation between attributes is large, the classification efficiency is low, and it is not sensitive to missing data. In contrast, in the scene of movie rating prediction, the model constructed in this paper is simple, and the complexity of manual operation is small, the accuracy is high.

In the aspect of applicability, the prediction model proposed in this paper is based on the recent data captured in the actual test on the network, the captured data are relatively regular, and can be captured even in a small amount of direct or single site, it is closely combined with the practical application, and the data capture difficulty is small. To improve the accuracy, decision tree prediction model need data from multiple web crawling with large dimensions, and the initial treatment is difficult. KNN prediction model is more suitable for overlapping sample sets, and the prediction accuracy of other types of sample sets is greatly reduced. It has some limitations. Like neural networks, SVM model is a learning mechanism, but unlike neural networks, SVM uses mathematical methods and optimization techniques, and prefers to theoretical calculations of small samples rather than practical applications. NBC model requires that the attribute is independent of each other, and it is difficult to find all data attributes independent of each other in the actual scene. Therefore, the applicability of the four prediction algorithm models in the actual movie rating prediction scene is less than the model proposed in this paper.

6. CONCLUSION

This paper studies the prediction model and method of the movie rating prediction. The work first identifies 7 antecedent factors that affect the movie rating prediction results. On this basis, a movie evaluation metrics system is designed. We design a neural network based movie rating prediction model with a preliminary prediction accuracy of 70%. To improve and the prediction accuracy, we make fusion on weight starring metric. The movie rating prediction accuracy with starring metric can then be improved to 88%. Compared with other related work of movie rating prediction and movie recommendation systems, our method has high accuracy and does not require user information. Compared with the knn algorithm, decision tree algorithm, SVM algorithm and NBC algorithm, the accuracy rate in this paper has increased dramatically. In the further work, we seek to study the impact of director and multiple actors on the movie rating prediction and the adjust weight across years. We will also work on the influence of movie reviews on movie rating and box office.

7. REFERENCE

[1] https://www.douban.com/doulist/1275845/comments/

[2] http://www.sohu.com/a/111029913_476313

[3] Wang Xiang. Research on the relationship between movie website rating and movie box office [D]. Nanchang University, 2016.

[4] Lim Y J, Teh Y W. Variational Bayesian Approach to Movie Rating Prediction[J]. Proceedings of Kdd Cup & Workshop, 2007.

[5] Golbeck J. Generating Predictive Movie Recommendations from Trust in Social Networks[J]. Lecture Notes in Computer Science, 2006, 3986:93-104.

[6] Wang J, Zhu Y, Li D, et al. Joint User Attributes and Item Category in Factor Models for Rating Prediction[C]// International Conference on Database Systems for Advanced Applications. Springer International Publishing, 2016:277-296.

[7] Hu Miaoyuan. Movie rating data analysis and user behavior preference modeling [D]. Yunnan University, 2016.

[8] Tian Maogen. Research and application of super network parallel in the movie rating prediction problems[D]. Chongqing University of Posts and Telecommunications, 2016.

[9] https://movie.douban.com/subject/25662329/