Prediction of Film Score: Based on Character Relations

Yijin Zeng *

School of mathematics and statistics, Beijing Institute of Technology, Beijing 100081, China

yijinzeng1998@163.com

*Corresponding author

Keywords: linear model, random forest, relationship network of the characters, 5-fold cross validation

Abstract: Film industry plays an important role in people's leisure life, and it has been booming even more important in recent years. With increasingly amount of film and television practitioners joining the industry, the competition in the film industry is becoming progressively fierce. Therefore, the significance of writing good movie scripts has appeared in the field. In this article, we constructed the character relationship network with software python and extracted some information from the network and constructed a linear model with least square method based mainly on 45 randomly selected films to help the relevant practitioners build a better character relationship network in the movie script. The appropriate linear model M2 is selected by using the 5-fold cross-validation method. Then we built the random forest model MR and also used 5-fold cross-validation method to test the model. The results show that the model MR has a much better predictive effect on movie scoring compared with model M2. The model reveals that some parameters of the character relationship network in the script are able to determine the final score of the film to some degree, which will provide a reference value for the creation of the script in the film production.

1. Introduction

Network model is a very powerful and effective model, which can concertize abstract problems and express them as physical entities or abstract entities [1, 2]. The advantages of the network are obvious, for example, it has good visual performance, and the use of the network can be very convenient for quantitative analysis of abstract problems, so as to establish a suitable model. In fact, the application of grid is very extensive, especially in the field of Social Sciences [3, 6].

It is particularly important but difficult to choose an appropriate model to analyze the network after the establishment of the relationship network. Multivariate linear regression is a basic and practical statistical model [7]. It can predict dependent variables by the optimal combination of multiple independent variables, so as to explore whether there is a linear relationship between independent variables and dependent variables and the strength of the linear relationship. Linear regression has simple structure and strong explanatory ability, and it has a very wide range of applications in practice [8, 10]. Random forest is a classifier that can be used to train and predict samples by using multiple trees. Breiman first proposed it in 2001 [11]. Random forests are welcomed by scientific researchers because of their high accuracy, good data processing effect and the ability to evaluate the importance of variables. Thus, it also has a very wide range of applications in many fields, especially for prediction [12, 14]. In fact, random forests can also show good performance when the number of observations is relative small [15]. In this paper, we consider using the above two models to analyze the relationship network, but it is still difficult to choose a better one between them. This is due to the fact that their indicators may have their own advantages and disadvantages, so we introduce the cross-validation method to select the model. The cross validation method uses most of the data to model and a small part of the data to forecast, and the cross validation method can make full use of each data, which shows great advantages when the sample size is small. Because of its superiority, it

DOI: 10.25236/ICHCAD.2019.006

is used by researchers to apply linear models and various nonlinear models to evaluate models. [16, 17]

As a matter of fact, there are many methods to predict the score of movies [18, 21]. Most of them predict the popularity of movies from outside of the movie itself, such as by extracting the discussion of a movie on social network media to predict the score. However, at present, there are little literature predicting the movie score from the script itself. From this point of view, this article establishes a network of characters through the script, so as to make a reasonable prediction of the final score of the film.

2. Data Processing and Network Establishment

2.1 Data and code availability

The real scores of 40 movies are selected from https: //www. imdb. com/. The scores of movies may fluctuate slightly with the change of date. The scripts are mainly from the website: https: //transcripts. f andom. com/wiki/Transcripts_Wiki. All the code and data in the paper can be found at https: //github. com/yij inzeng/ code.git.

2.2 Dealing with scripts

We have noticed that a large number of scripts are in the form of the following formats: "Name" and and "Dialogue”. So we used always almost the similar formats to reduce the workload in the process of searching for scripts. In the process of using Python to process the script, we first dealt with all the letters of the script in lowercase, which is to avoid the problem of the same character's name capitalization in the script. Subsequently, we read all the characters with “:”, and according to our judgments of the script format, the colon signifies the appearance of role dialogue. Here, it is worth mentioning that it is not always the person's name with “:”but also the time, such as 6:00. However, according to the processing methods of our subsequent scripts and the data extracted, the occurrence of this situation will not affect our final results, so next we will not consider other situations besides names. At the same time, in the scripts we selected, there are also some movie scripts in the formal of “person name”+ “blank line”+ “dialogue”. At this time, we merely need to extract a line of names. After reading the colon or a separate column of name to get the names sorted in the sequence in the script, we can store these names in an alphabetic string A. Through the sequence and number of names in A, we can judge the dialogue between the characters. Next, we deal with A.

2.3 Establishment of network

Firstly, we count the number of different elements in A. A new alphabetic string B with the size of N by assuming the number of different elements in A is n is set up, in which the non-repetitive elements in A are stored in order in B, and the matrix O of blank n*n is established. Next, we use the following methods to establish the Person Relational network R: Let any one of the individuals in A be Q and the other be P. Consider the edge RQP of Q and P in the network. When the distribution of QPQP (PQPQ) appears in A, we consider that there is a conversation between Person Q and Person P. That is to say, in network R, edge RQP (RPQ) exists. In B, the locations of Q and P are retrieved separately. The locations of Q and P in B are I and J respectively. In this case, the locations corresponding to matrix B in matrix O are I,J, so that OU = OU + 1 and OJI = QJI + 1. At this point, we get a complete network of people, in which the elements in B represent each person, and the position of matrix O is bigger than 0 representing the person corresponding to the row and column having a dialogue.The visual effect of the grid is shown in Fig.1.

2.4 Data extraction

Data extraction: Through the established network, we extract the following information: NC: The number of “main”role clusters, in which “main” refers to having more than two consecutive dialogues with oneself or other characters in the play. The value is equal to the number of clusters in the graph. DG: The diameter of the graph. It refers to the maximum number of vertexes that must be passed from any vertex to another vertex in a graph. MD: The maximum degree of a node in one graph. The degree of a node is the number of other points directly related to that node. In this paper, the actual meaning of degree is the number of roles that communicate directly with a role. MI: The maximum intermediary centrality. The intermediary centrality of a node refers to the number of times a node acts as the shortest path strength bridge between the other two nodes. RL: The number of roles in the largest cluster in a network diagram. CC: The number of clusters of all the characters in the play. MR: The number of "main" roles. DN: The number of edges in a network graph. CP: The number of characters in the play. CZ: Number of characters with zero degree in the play. VL: The vertexes in the longest path in order. VS: The vertexes in the shortest path in order. MP: The maximum proximity to centrality. Maximum proximity centrality reflects the proximity between one node and other nodes in the network. The reciprocal sum of the shortest path distances from one node to all other nodes represents proximity centrality. MC: The minimum proximity to centrality.

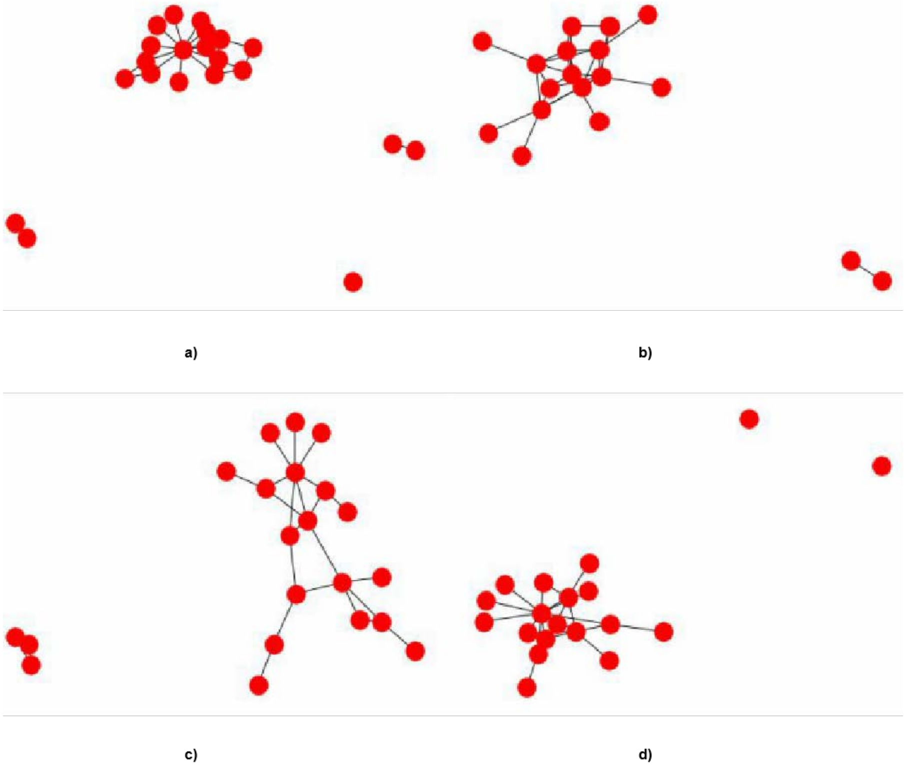

Figure 1. These four pictures are all derived from the grid created by Python software based on the above method: a): The Character Network of the Movie "Thor" b): The Character Network of the Movie "The Avengers" c): The Character Network of the Movie "Finding Nemo" d): The Character Network of the Movie "Zootopia"

3. Model selection

3.1 Linear prediction model

After obtaining the above data we came to a question on how to build a model to make the established model having a better prediction effect on movie score? To this end, we consider the use of 5-fold cross validation method: First, we randomly and equally divided 40 data into five groups, of which five groups of data are selected to generate random numbers based on 1-40. They are G1 (28, 27, 34, 12, 19, 6, 40, 9), G2 (22, 38, 33, 1, 39, 29, 5, 26), G3 (36, 10, 14, 17, 4, 8, 3, 35), G4 (21, 32, 15, 11, 37, 23, 7, 16), G5 (20, 18, 24, 30, 13, 25, 2, 31). And each of these groups contains 8 elements. When scoring the prediction effect of the model, we make the standard deviation between the real value and the predicted value of the test group, and take the standard deviation as the final score. Besides, all of the following linear models are trained by least squares method.

First, the full model Ml is considered. The full model Ml, which is modeled by linear least squares method, uses all variables except CC, which has a very serious collinearity with other variables and is rejected by the soft SPSS automatically. Let us take G1 as the test set, G2, G3, G4, G5 as the evaluation set. Let the real score of G1 be SG1i, then SG11=7.60, SG12=7.30, SG13=7.50, SG14=7.80, SG15=7.70, SG16=7.00, SG17=8.40, SG18=7.00, using G2, G3, G4, G5 after training with linear least squares method, assuming the predictive score of the G1 group was sG1i, then sG11=8.42, sG12=5.56, sG13=7.09, sG14=7.74, sG15=7.61, sG16=7.88, sG17=9.07, sG18=9.08. The relationship between predicted and real values is shown in Fig.2 a). The standard deviation of the score is 0.3970. Let us suppose that the score of this group is SG1. Then we have SG1=0.3970. Similarly, we get the score of SG2=0.8324 with G2 as the test set, G1, G3, G4, G5 as the training set, SG3=0.9436, with G3 as the test set, G1, G2, G4, G5 as the training set, SG4 =0.5436, with G4 as the test set, G1, G2, G3, G5 as the training set, SG5=0.6539 with G5 as the test set, G1, G2, G3, G4 as the training set. By weighting all the scores mentioned above, the final score of the whole model is

$$S_{M1} = 1 - \frac{\sum_{i=1}^{5} SG_i}{5} = 0.3259$$

Fig.2 b) is a scatter plot of all the predicted values (Predict) and all the real values 5 (Real) obtained by model Ml using a 5-fold cross validation method. Table 1 illustrates their linear correlation.

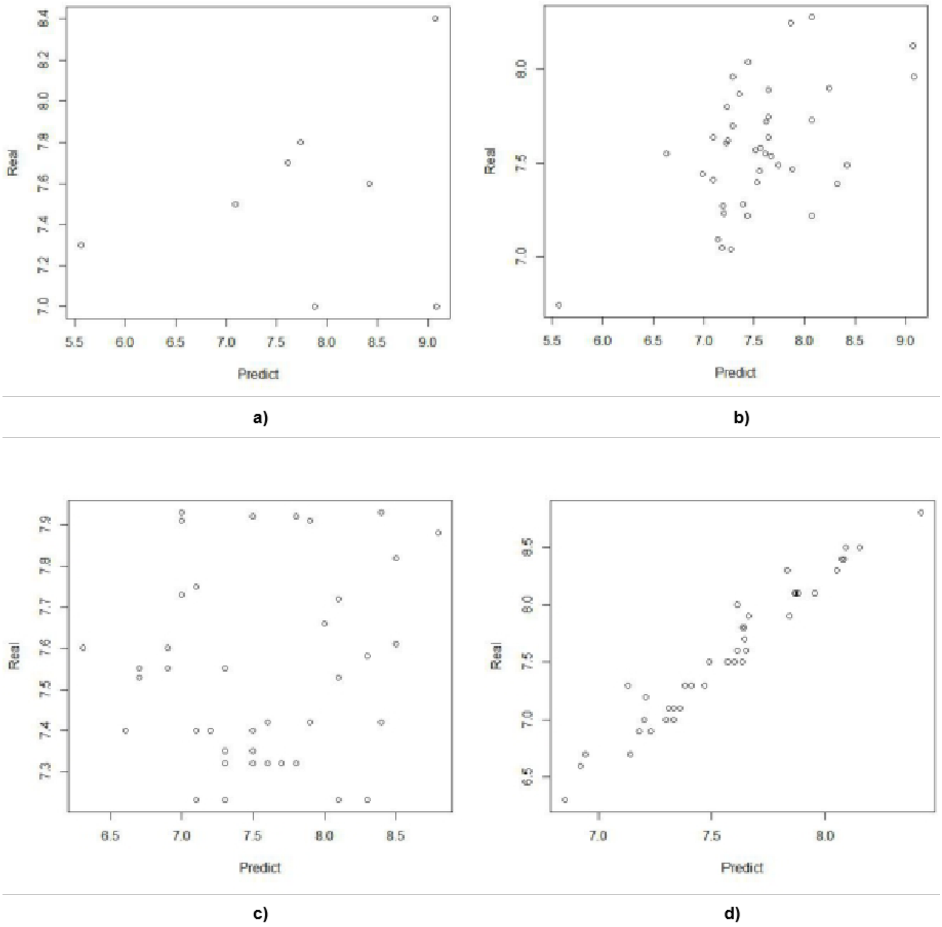

Figure 2. These four pictures all reflect the true score of the movie and the predicted score obtained by cross-validation of the five-st ep method, a): Real Score and Predictive Score in Group G1 of Ml b): 5-fold Cross Validation Method of Ml c): 5-fold Cross Validation Method of M2 d): 5-fold Cross Validation Method of MR

Next, we consider the selected models. To consider the selected models, we should consider deleting some unnecessary variables or those with strong correlation to prevent the occurrence of over-fitting. In fact, this is not a simple problem, and there is currently no general method for optimizing variables. In this paper, in order to solve the problem of variable selection, we first consider the traditional forward, backward and stepwise method. They all determine whether a variable enters or leaves the fitting function by evaluating the F value of the variable in the fitting process. First of all, let us consider the model M2 obtained by stepwise regression method (probability of F-to-enter≦0.10, probability of F-to-remove≥0.20): through the stepwise regression method, only one variable MC enters the function, and the adjusted R-side is 0.149. This shows that in this model, the explanatory power of the number of clusters to the score of movies is 14.9%. However, this outcome with only one independent variable is undoubtedly out of line with reality. Researchers often invoke stepwise ordinary least squares regression to explain, predict or classify practical problems or theoretical constructs in social research.

Unfortunately, most of the current applications of stepwise regression should be rejected [22]. As the same as the scoring method as model Ml, we get S_{M2}=0.4297. The relationship between predicted and real values shown in Fig.2 c). From the score, we can see that the prediction error of M2 model is smaller than that of Ml model. As we discussed before, the M2 is not a realistic model, but it still perform better than Ml. However, we can find from the graph that the predicted value of model M2 has almost no correlation with the true value, and this non-linearity is even more obvious than that of model Ml. The results obtained by the forward method (probability of F-to-enter ≦0. 10) after treatment are the same as those obtained by the stepwise regression method, which will not be repeated here. Next, we consider the model M3 obtained by the backward method (probability of F-to-remove≧0.20): through the backward method, the final model contains variables MC MD RL CP CZ In order to better judge the predictive effect of the model, we still use the 5-fold cross-validation method to evaluate the predictive ability of M3, getting S_{M3}=0.3460. From the above results, we can see that the bigger the score, the smaller the deviation between the predicted value and the real value is. Therefore, in the linear model, M2 is the best model for predicting. How-ever, as we discussed above, M2 is a model that is not very common sense, which indicates that the relationship between dependent and independent variables may not be linear. To solve this problem, we need to introduce more complex models to explore their relationships.

Table 1. Linear correlation between the predicted values of the two models and the true values of the movie

Models	Correlation Coefficient	Sig.
M1	0.536	0.000
M2	-.390	0.060
M3	-.147	0.186
MR	0.967	0.000

3.2 Random forest

Random forest is an excellent algorithm in regression and prediction. It is simple, easy to implement and has low computational cost. However, its regression and pre-diction results are amazing. At the same time, random forest can effectively prevent the occurrence of over-fitting phenomenon, and multiple decision trees constructed in random forest can prevent extreme points from pairing the whole model. In this paper, as a comparison with the linear model, we introduce the random forest model MR. All the data in MR model are the same as those mentioned above, and the independent variables are the same as those introduced in Ml. The basic parameters are nTree = 1000 and mtry = 1. As mentioned above, we also used the 5-fold cross validation to process the model. The results are shown in Fig.2 d). At the same time, we also scored the prediction ability of the model by the standard deviation between the predicted value and the real value, and the final score of the model MR was SMR=0.7416. The scores of all models are shown in Table 2. Obviously, model MR scored much higher than all linear models. Therefore, we can draw a preliminary conclusion: based on the current data set, the prediction accuracy of random forest is higher than that of linear regression model, and it is the fittest model of all mentioned.

Table 2. Score of All Models

Models	Scores based on five-step method
M1	0.3259
M2	0.4297
M3	0.3460
MR	0.7416

4. Analysis of MR Model

We have selected the best MR model through cross-validation. Next, we will discuss the significance of MR model, that is, whether MR can predict movie scores to some extent and whether M3 can reveal some link between movie scores and the network of characters in movies.

4.1 The predictive ability of MR model

In order to test the predictive ability of the model M3, we randomly selected five movies as the test set. The real scores and expected scores of five films are shown in the Table 3 below.

From the results, it can be found that the prediction ability of the model is strong. From the figure, we can see that the deviation of three movies is lower than 0.3 points, one movie is 0.7 points and one movie is 1.1 points. It is worth mentioning that in the process of predicting films with a real score of 8.6, there are two main reasons for the relatively larger gap between the real and the expected: Firstly, from the perspective of the model, the score of 8.6 belongs to the extreme situation in the model, the highest score in the training concentration is 8.8, and the rest is lower than 8.6. So it's clear that the model will make more unreliable predictions for this more extreme situation. Secondly, from the point of view of the film itself, the factors to be considered in getting the score of 8.6 are not only scripts, but also more aspects, such as casts, dialogues, plots and so on. However, we can also see in the table that even though the absolute deviation of the prediction is large, the predicted score based on the relationship between the characters is still relatively high of the five films, which is up to 7.5 points. As a contrast, we also used the M2 model, which are proved to be the best linear model by the 5-fold cross validation method, to predict the score of five-step movies. We can see that in most movies, the prediction effect of the M2 model is worse than the MR model. To be specific, In addition to the movie se7en, the MR model has made more accurate predictions of movie ratings.

Table 3. Real Score and Predicted Score of Test Set

Models	Real score	Expected score (M2)	Expected score (MR)
Cars	7.1	8.0	7.8
se7en	8.6	7.9	7.5
Monsters University	7.3	7.7	7.4
Star Trek Into Darkness	7.7	7.4	7.4
Titanic	281.0	7.4	7.6

4.2 The relation between individual variables and dependent variables and the importance of individual variables

Another important feature of the model is that it reveals the relationship between movie ratings and some independent variables based on the script. First, we use software R to rank the importance of each variable in MR of random forest model.

From Fig.3, we find that the important ranking of independent variables based on the two criteria is not entirely consistent. But if we consider Group one (MC, MD, MI, MR, DN, VL) and Group two (DG,RL,CP,CZ,VS,MP,MC), we can clearly see that in the two graphs, the values of all variables in both Groups are close to each other. And the proximity of variables in group one is significantly higher than that in group two. Based on this, we grouped the importance of all independent variables, that is, group one is an important variable, and while group two is a sub-important variable.

Table 4. Partial Correlation of Group one

Independent Variable	Correlation Relationship	Sig.
NC	-.182	0.182
MD	-.574	0.001
MR	0.289	0.072
MI	0.334	0.044
DN	0.389	0.022
VL	0.157	0.218

We care more about the group that is more important for movie rating. Therefore, we consider the partial correlation between all the independent variables and dependent variables in Group one, in order to explore whether there is a significant linear relationship between independent variables and dependent variables. The results are shown in Table 4, which we can see that the independent variables pass the unilateral test at the significance level of 0.05, that is, they may have a linear relationship with the dependent variable movie score.

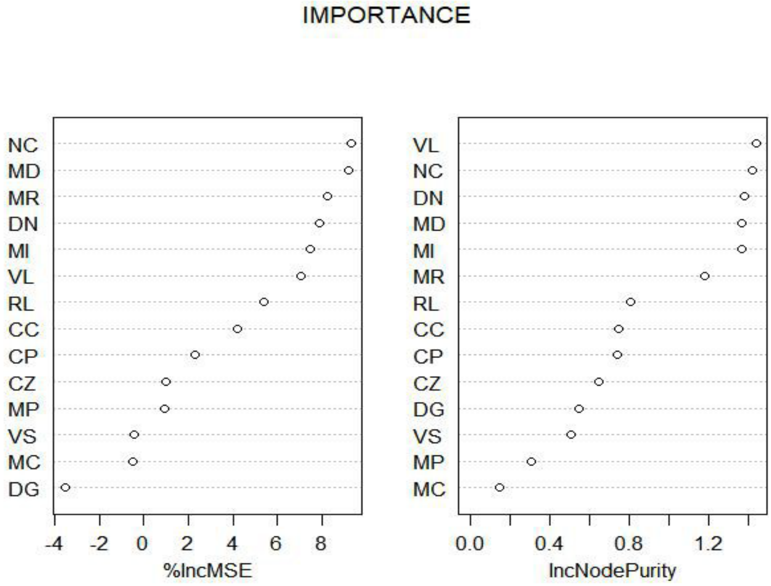

Figure 3. Importance of all variables based on random forest

5. Discussion

The main purpose of this paper is to design a model for the prediction of the score of a movie and exploration of the relationship between the movie score and the relationship network of characters in the script. We firstly abstracted the characters relationship network in this article, then extracted the key features of the network structure, and constructed several models to solve the above problems.

In this paper, the MR model demonstrated good predictive ability of movie script score. Firstly, we can see that the predicted value is very close to the real value from the 5-fold cross validation method. Then we randomly selected five movies and used the established the MR model to predict the scores of the five movies. We can see that the predicted scores are close to the real movie scores. Therefore, the model provides the possibility for the script creators to make a preliminary test on the quality of their scripts. We suggest that after script creation, script creators can use the MR model to make a preliminary prediction of the final possible score of the film, so that they can further modify the script. However, the creators should also pay attention to the fact that we have neither thoroughly analyzed the importance of the character relationship network in the whole script composition nor compared the importance of the characters relationship network with the plot, dialogue or other factors in the script. Thus, in the process of script creation, we should not deliberately modify the plot in order to build a good relationship network while ignoring the other factors. After all, the result can only be suggestive, not decisive.

Meanwhile, according to our final M3 model, we are allowed to analyze which variables may be more important in the process of script creation. From the chapter 4.2, We can see that NC, MD, MI, MR, DN, VL play more important roles in terms of the score of a movie. On the basis of random forest model, we did partial correlation analysis for each important variable and get the following results: variable NC, MD is negatively correlated with movie scores in data, while other important variables are positively correlated with the movie scores. It may indicate that a character should not have direct conversations with too many people. At the same time, too many characters appear in the script is also a disadvantage for the movie rating. From the film point of view, the relationship constructed in this way may seem unorganized and without deep consideration. Meanwhile, we noticed that the values of variables MI, MR, DN, and VL are positively correlated with movie scores. These positive correlations may indicate that a movie with more characters and more dialogues will be more popular. Besides, variable MI and variable VL are positively correlated with the movie score, indicating that the relationships between the characters in the play should be more "deep”, that is, the important characters in the play assume the bond of other characters7 dialogue, rather than directly with other characters. This is understandable since films that are presented such way tend to be more carefully thought out and designed in terms of plots.

However, there are still some shortcomings in this article. Firstly, from the artistic aspect, the evaluation criteria of a film is far beyond the relationship between the characters in the play, but also the consideration of the actors7 acting skills, the richness of the plots, special effects and so on. Therefore, no matter how deep the relationship network of characters in the script is excavated, the predictive and explanatory abilities of a movie score are limited in the end. Moreover, it is not a simple job to mine the persona relationship network. It is undeniable that there may be some improper handling of the script, such as the change of the name of the same character in the script. The name of the character Iron Man in the movie “The Avengers: Endgame” has the name “stark”, “tony”, and so on. of course, we can deal with a particular movie to avoid this problem, but when the sample size is enlarged, it is undoubtedly unrealistic to deal with each sample in detail. More-over, in the process of dealing with the relationship network, we can neither judge the relationship between the characters nor make a good statistical analysis of a one-time dialogue between the characters. All these lead to the limitations of our model.

References

[1] Newman, M. E. The structure and function of complex networks^ SIAM review 45, 167 - 256 (2003).

[2] Newman, M., Barabasi, A.-L. & Watts, D. J. The structure and dynamics of networks (Princeton University Press, 2011).

[3] David Liben-Nowell, Jon Kleinberg. The link-prediction problem for social networks. Journal of the American society for information science and technology 58 (7), 1019 - 1031, 2007.

[4] Arent Greve, Janet W Salaff. Social networks and entrepreneurship. Entrepreneurship theory and practice 28 (1), 1 - 22, 2003

[5] Waumans, M. C., Nicodeme, T. & Bersiui, H. Topology analysis of social networks extracted from literature.PZoS tme 10, e0126470 (2015).

[6] Liu, Wenlin & Sidhu, Anupreet & Beacom, Amanda & Valente, Thomas. (2017). Social Network Theory. 10.1002/9781118783764. wbieme0092.

[7] Alexopoulos, Evangelos. (2010). Introduction to Multivariate Regression Analysis. Hippokratia. 14. 23 - 8.

[8] Raphael A Mrode. Linear models for the prediction of animal breeding values. Cabi, 2014.

[9] Kun war P Singh, Shikha Gupta, Atulesh Kumar, Sheo Prasad Shukla. Linear and nonlinear modeling approaches for urban air quality prediction. Science of the Total Environment 426, 244 - 255, 2012.

[10] John Makhoul. Linear prediction: A tutorial review. Proceedings of the IEEE 63 (4), 561 - 580, 1975.

[11] Breiman, L. Machine Learning (2001) 45: 5. https://doi.Org/10.1023/A: 1010933404324

[12] Zou, Zhi & Peng, Hong & Luo, Lin. (2015). The Application of Ran-dom Forest in Finance. Applied Mechanics and Materials. 740. 947-951. 10.4028/www.scientific.net/AMM.740.947.

[13] Kim, Seongjin & Ahn, Hyunchul. (2016). Application of Random Forests to Corporate Credit Rating Prediction. The Journal of Business and Economics. 32. 187 - 221.

[14] Song, J. & Gao, Q. & Li, Z.. (2016). Application of random forests for re-gression to seismic reservoir prediction. 51. 1202 - 1211. 10.13810/j.cnki.issn. 1000 - 7210.2016.06.021.

[15] Biau, Grard & Scornet, Erwan. (2015). A Random Forest Guided Tour. TEST. 25. 10.1007/sll749-016-0481 - 7.

[16] Stone, M.. (1974). Cross- Vo,lido,tory Choice and Assessment of Statistical Pre-dictions. J R Stat Soc Series B Stat Methodol. 36. 111 - 113. 10.1111/j.2517- 6161. 1974. tb00994.x.

[17] Li, T. & Chen, C. & Cheng, W. & Li, X.. (2007). Application of cross validation in power quality de-noising. 31. 75 - 78.

[18] Jennifer Golbeck. Generating predictive movie recommendations from trust in social networks. International Conference on Trust Mo,no,gern,ent^ 93-104, 2006.

[19] Chung-Yi Weng, Wei-Ta Chu, Ja-Ling Wu. Rolenet: Movie analysis from the perspective of social networks. IEEE Transactions on Multimedia 11 (2), 256 - 271, 2009.

[20] Xiaojun Wang, Leroy White, Xu Chen. Using Twitter data to predict the performance of Bollywood movies. Industrial Manageme'rd & Data Systems, 2015.

[21] Mehreen Ahmed, Maham Jahangir, Hammad Afzal. Using Crowd-source based fea-tures from social media and Conventional features to predict the movies populari-ty. 2015 IEEE International Conference on Smart City/Social Cow, /Sustain Com (S-martCity), 273 – 278.

[22] Douglas A Henderson, Daniel R Denison. Stepwise regression in social and psychological research. Psychological Reports 64 (1), 251 - 257, 1989.