Big Data Analysis of TV Dramas Based on Machine Learning

Jiaqi Tan, Feiqiao Mao(✉), Lianghai Yang, and Jiahui Wang

Shenzhen University, Nanhai Ave 3688, Shenzhen 518060,
Guangdong, People's Republic of China
christinatan0704@gmail.com, feiqiao@szu.edu.com, i@silas.hk,
samwanglb@gmail.com

Abstract. Currently, large amount of TV dramas has overwhelmed the demand of TV station which had caused massive waste of resources. This article offers several practical solutions to tackle the above-mentioned problems through building model based on machine learning. Firstly, we build a TV score prediction model with regression and machine learning to rank the most welcomed TV drama. Moreover, we write an Internet worm to collect data from the internet, and build a Star popularity index prediction model by machine learning and regression. And list the much-acclaimed stars based on the popularity index. In conclusion, with the predict score of the TV drama predicted based on machine learning, it can provide a reference for TV station to manage TV programs and with the starring ranking it can help TV drama production team to produce TV dramas in a high quality.

Keywords: Machine learning · Big data · TV dramas · Predict score
Staring ranking

1 Introduction

Currently, with the fast pace development of multimedia, large quantities of TV dramas have overwhelmed the demand of TV station. And simultaneously, the oversupply of TV drama will not only put a potential threat to the development of file and television industry but also leads to massive waste of resources.

In order to lower down the risk of TV investment, improve scrip quality and forecast audience response to ensure maximum benefits. We have predicted the TV dramas' score and rank the starring by their popularity. Based on methods of machine learning (such as Classifier, Clustering, etc.) to extract the Eigen value from the data and use the Eigen value to divide the data into training data and testing data. And subsequently, build an optimized model with these data and use the optimized model to analyze relevant data in order to evaluate and customize file and television.

This study was supported by Guangdong Natural Science Foundation (2016A030313036), Shenzhen Science and Technology Foundation (JCYJ20150324140036842) and Guangdong Graduate Education Project (2015SQXX0).

M. Qiu (Ed.): SmartCom 2017, LNCS 10699, pp. 90–95, 2018.
https://doi.org/10.1007/978-3-319-73830-7_9

2 Concepts and Related Work

Linear regression is popular in tackling some popularity prediction problem, for most of the research that had been done, it will predict data combing various method. According to [1] it made a movie recommendation considering the context sensitive and based on a time-decay model, and in research [2] it has predicted the region-specific crime rate for the future based on statistical auto regressive linear regression modeling.

There also have some related work like creating some raking model, such as a personalized ranking model [3] have make a personalized ranking based on pairwise learning. And in research [4] it has make a Social Media Content ranking based on social computing and user influence.

In this paper, considering that the factors to be use are specific date instead of fuzzy data, according to [5] we should better adapt the tolerance approach for possibilistic linear regression with fuzzy-valued inputs and/or outputs. But for the TV ranking problem in order to predict the score for TV drams, we can find out the most influential element of the model and produce an effective ranking model simply based on linear regression.

3 Score Prediction Model

This model aims for predict the score of each TV drama according to the drama's theme, production team, screen writer and starring. The modeling method is mainly based on linear regression and with this model we can predict the popularity rate of the drama.

3.1 Modeling Process

We build this TV dramas ranking model by these processes:

- **Data division**. Extract data we collected from the internet (which included the TV dramas' theme, ultimate score, director, screenwriter and starring). And then randomly divide these data into training data (at 80%) and testing data (at 20%). Training data is used to make a model to predict the score of all the TV dramas and testing data is used to calculate the standard error of the model and correct the model.
- **Modeling**. Create a model of TV ranking using regression and based on machine learning. Use the score from testing data and the predicted score to calculate the RRS (residual sum of squares) and errors of the model. Use machine learning algorithm to calculate the more accurate value of W (weight) and repeat 'Modeling' and 'Correct the model' step till getting the minimum RRS and error. Finally, we use this evaluation function to predict the score of all TV dramas:

$$\hat{y} = \hat{w}0 + \hat{w}1 \text{ director } + \hat{w}2 \text{ theme} + \hat{w}3 \text{ screenwriter } + \hat{w}4 \text{ starring} \tag{1}$$

$\hat{y}$ is the predicted score and $\hat{w}i(i >= 0,\ i \in N)$ is the weight of each features. (for example, $\hat{w}1$ is the weight of distribution enterprise.)

3.2 Proof Model Effectiveness

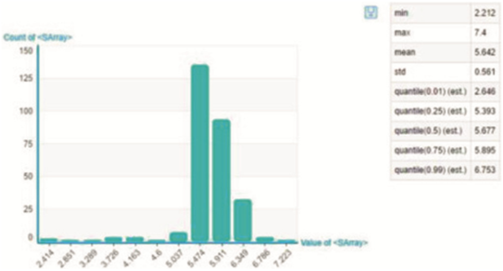
At first, we pick use all the possible influential factors as Eigen to build several models, after comparing the ranking result from our model to the real ranking result we find that the most influential factors are starring and director, thus using director, screenwriter, starring and drama's theme as Eigen value to build a model would be more accurate. As a result, in this model (using director, screenwriter, starring and drama's theme as Eigen value), the max error of this ranking model is 2.358532301405809 and the RRS of the model is 1.6066902 (Tables 1 and 2).

Table 1. The ranking of the latest TV dramas by predicted scores

TV drama	Predicted score	Actual score	Comment number	Theme
Battle of Changsha	9.199973382	9.2	18419	Modern revolution
All Quiet in Beiping	8.799892707	8.8	16084	Modern revolution
Operation Mekon	8.764290332	7.2	468	Modern cases
The Merchants of Qing Dynasty	8.499971847	8.5	1271	
Dating Hunter	8.29978057	8.3	5664	Modern city
Hey Daddy!	8.1997884	8.2	1138	Modern city
A Civic Yuppie in Countryside	8.100002467	8.1	3956	
My Heart is Shining	8.099951785	8.1	234	Other
Dangerous Journey	7.999915658	8	266	Modern cases
Simple World	7.899988534	7.9	5443	Modern countryside

It can be clearly seen from the form that the predicted score of "Operation Mekon" is higher than its original score to 1.5 points, but after searching on the internet, the score of "Operation Mekon" is high in reality (some of the viewer even score it to 9.3). Hence, in conclusion, using director, screenwriter, starring and drama's theme as Eigen value to build a ranking model is accurate and reliable.

Moreover, the spread of the predicted score of TV drama shows like this (see Fig. 1).



min	2.212
max	7.4
mean	5.642
std	0.561
quantile(0.01) (est.)	2.646
quantile(0.25) (est.)	5.393
quantile(0.5) (est.)	5.677
quantile(0.75) (est.)	5.895
quantile(0.99) (est.)	6.753

Fig. 1. The spread of all the TV dramas' scores.

Take the max score 7.4 (more accurate at 7.39978977266) as an example, the TV drama's name is "Soldier" compared to its score from "iQIYI.com" which scores 7.2, so the error of our model is less than 3%.

Lastly, we make a model to compare the predicted score with the original score (see Fig. 2), if the point is close to the middle line, means that it fits the original score well, it can be clearly seen from the graph above that most predict score is equal to its real score.

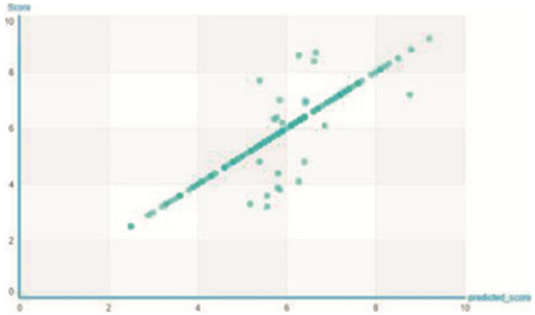


Fig. 2. Proof of consistency of predicted score and actual score

4 Star Popularity Ranking Model

The ranking model can be used to predicted the popularity of stars and rank the star popularity index.

4.1 Modeling Processes

First of all, we have written a web crawler to get data of 201 TV dramas in 2016 from iQIYI.com, which includes 351 stars and scores of relative TV drama. And subsequently, we use the same modeling method in task one (divide the data randomly in to training data (account for 80%) and testing data (account for 20%) using regression and machine learning to build a model, etc.) and the only difference is that we have changed 'star' to target and changed ‘score’ to Eigen value to build a model. And finally use the model to predict the score of every star and get the star popularity index. Finally, we use this evaluation function to evaluate the star popularity index:

$$\hat{y} = \hat{w}0 + \hat{w}1 \text{ score} \tag{2}$$

$\hat{y}$ is the predicted score and $\hat{w}$ 1 is the weight of score.

After predicting the score of every star in our collected data, we can list the top ten popular TV drama star as follow:

Table 2. The ranking of star popularity index

Actor name	Count	Score
Zheng Shuang	4	9.897662804
Gan Tingting	1	9.199681232
GuLi Zhana	6	8.89894841
Zuo Jiani	1	8.700088684
Liu Wenxuan	1	8.700088684
Li Jiaxin	1	8.700088684
Han Rui	1	8.700088684
Wang Zixu	1	8.599943771
Luo Jin	2	8.599815697
Zhao Liying	2	8.540066765

4.2 Proof Model Effectiveness

After using testing data to evaluate the model, we find out the max error of the model is 1.7002398555285776 and the RRS of the model is 0.7214715971395239 which indicates that the model is conformed to the reality. Furthermore, we have compare the star popularity index that we designed to the Chinese Top Searches of Stars in the latest month (the list contains other type of stats like music stars, etc.) from iqiyi.com, Zheng-Shuang who has the highest popularity index in our predicted data, and she ranks the first among other stars we scored according to the list given in iqiyi.com.

5 Conclusions and Future Work

Based on the findings from TV Drama Score Prediction Model we can conclude that screenwriter, starring and drama’s theme affect the score of a drama most. And using these factors to create an evaluation function (1) can predict the score of TV drama. Which can help TV station to set the timetable of TV dramas. Moreover, based on the star popularity ranking model we can find out the most popular star and produce a star popularity ranking list. These predicted data can help us to enhance the quality of TV drama and make a recommend on TV dramas to viewers.

References

1. Selva Priya, S., Gupta, L.: Predicting the future in time series using auto regressive linear regression modeling. In: 2015 Twelfth International (2015)
2. Luminto, D.: Weather analysis to predict rice cultivation time using multiple linear regression to escalate farmer's exchange rate. In: 2017 International Conference, Concepts, Theory, and Applications (ICAICTA) (2017)
3. Guo, W., Wu, S., Wang, L., Tan, T.: Personalized ranking with pairwise factorization machines. Neurocomputing **214**, 191–200 (2016)
4. Ntalianis, K., Salem, A.-B.M., El Emary, I.: Social media content ranking based on social computing and user influence. Procedia Comput. Sci. **65**, 148–157 (2015)
5. Černý, M., Hladík, M.: Possibilistic linear regression with fuzzy data: tolerance approach with prior information. Fuzzy Sets Syst. (2017)