Predicting Movies User Ratings with Imdb Attributes

Ping-Yu Hsu, Yuan-Hong Shen, and Xiang-An Xie

Department of Business Administration, National Central University, Jhongli City, Taoyuan County, Taiwan (R.O.C.)
pyhsu@mgt.ncu.edu.tw
{edwardshen1976,ho2009}@gmail.com

Abstract. In the era of Web 2.0, consumers share their ratings or comments easily with other people after watching a movie. User rating simplified the procedure which consumers express their opinions about a product, and is a great indicator to predict the box office [1-4]. This study develops user rating prediction models which used classification technique (linear combination, multiple linear regression, neural networks) to develop. Total research dataset included 32968 movies, 31506 movies were training data, and others were testing data. Three of research findings are worth summarizing: first, the prediction absolute error of three models is below 0.82, it represents the user ratings are well-predicted by the models; second, the forecast of neural networks prediction model is more accurate than others; third, some predictors profoundly affect user rating, such as writers, actors and directors. Therefore, investors and movie production companies could invest an optimal portfolio to increase ROI.

Keywords: User rating, prediction model, classification, linear combination, convex combination, neural networks, multiple linear regression, stepwise regression, IMDb.

1 Introduction

Since the 20th century, movies have been an essential and important recreation to human beings. According to a survey by Motion Picture Association of America (MPAA), global box office for all films released in each country around the world reached $35.9 billion in 2013, up 4% over 2012's total [5]. More than 4000 movies were produced within one year in the whole world, and only the top 5 movies box office exceeded US$100 million and these movies gained 14% of the gross box office[6]. It is a winner-take-most industry. In early 21st century, the production cost of one movie already reaches US$65 million while the advertisement and marketing budget also reaches US$35 million [6]. An investigation with 281 movies produced in the period from 2001 to 2004 pointed out, the return on investment (ROI) of movies ranges from −96.7% up to over 677%, an average ROI at −27.2% [7]. The rigorous circumstances exposed movie investors and production companies to higher financial risks. However, to the industry practitioners, forecasting the box office of a specific movie is a difficult mission because of some uncertain characteristics. Therefore, the industry practitioners relied heavily on traditional wisdom and simple empirical rules

D. Miao et al. (Eds.): RSKT 2014, LNAI 8818, pp. 444–453, 2014.
DOI: 10.1007/978-3-319-11740-9_41

to make their decision in the past [8]. Most movie investments seem like a gambling. Therefore, the engagement of a forecast in box office is an imperative and challenging study issue to scholars and the industry practitioners.

User rating is a kind of Word of Mouth (WOM), it simplified the procedure which consumers express their opinions about a product. User rating is highly important to a certain product or service, because it reflects the wisdom of crowds. Undoubtedly user rating is a great indicator to predict the future sales performance of a product. Movie industry specialists agree that it is a key success factor of movie and help movie production company and investor gain a financial success [1-4].

The purpose of this study was to develop an accurate user rating prediction models which based on the early information. We used classification technique of data mining to develop prediction model. First, developed one learning algorithm to identify a linear combination (convex combination) model that best fits the relationship between the attribute set and class label of the input data. Second, testing data was used to estimate the accuracy of the model. In the meantime, we employed other techniques (multiple linear regression and neural networks) to develop comparison model. The results of comparison model were used to illustrate how effective these attributes are.

2 Related Works

2.1 Internet Movie Database (IMDb) Voting (User Rating)

IMDb (www.imdb.com) is the largest movie database in the world. The service was launched in 1990. The website had 2.8 million titles (includes episodes) and 5.9 million personalities in its database on May 2014.

IMDb registered users can rate every movie in the website (rating scale from 1 to 10). User can rate one movie as many times as they want but each rating overwrite the previous rating for the same movie. The rating shown in IMDb is not an average rating of the original data by every voting user but a kind of weighted average of an undisclosed calculation method. IMDb applies various filters to screen the original data, the objective is to present a more representative rating which is immune from abuse by subsets of individuals who have combined together with the aim of influencing (either up or down) the ratings of specific movies; IMDb keeps the mystery of rating calculation method, without disclosing whether/when/how to perform a weight for certain ratings, to provide a more objective rating.

2.2 Movie Box Office and User Rating Prediction

During the past 20 years, marketing scholars have developed some prediction models and decision support tools to increase the accuracy of forecast. One mainstream in which is to use multiple linear regression, by making the box office of movie as the dependent variable while the independent variable as the predictors with an impact on box office forecast, to establish a forecast model [1, 9-15]. [16] points out some

production and marketing characteristic factors influence the financial performance of a movie. [17] used neural networks in predicting the financial performance of a movie. They compared their prediction model with models that used other statistical techniques; it is found the model built by neural networks do a better job of predicting box office.

2.3 Linear Combination (Convex Combination) and Prediction

Linear combination model is a decision rule for deriving a linear combination that predicts some criterion of interest. This method is intuitive and easy to understand to decision makers[18]. A linear combination is constructed from a set of terms by multiplying each term by a constant and adding the results. The constants were considered as weights when the linear combination model was used for decision-making or predictive purposes. Given a finite number of predictor variables $x_1, x_2, \ldots, x_n$, a linear combination of these predictor variables (independent variables) is a criterion variables (dependent variables) of the form.

$$w_1x_1 + w_2x_2 + \cdots + w_n; \text{ where the constant } w_i \geq 0 \text{ and } \sum w_i = 1;\ i = 1,2,\ldots,n$$

A proper linear combination model is a linear equation which predictor variables are given optimal weights to optimize the relationship between the prediction and the criterion [19]. However, some authors pointed that it is a misunderstanding to interpret the weights as measures of the importance [18, 20]. The value of weight is dependent on the range of predictor variable values; in other words, a weight of a predictor variable can be different by increasing or decreasing the range of observed value of predictor variable. In this study, all the observed values (score) of predictor variable were average user rating which comes from IMDb user voting. Furthermore, the average user rating is interval scale and ranges from 1 to 10.

3 Data Preprocessing

3.1 Data Collection

Data for this study were collected from IMDb. We collected the user rating and attributes of all movies released from 2002 to 2012. We obtained a data set of 32968 movies. The data set consists of attributes: actors, as known, country, directors, episodes, film locations, genres, IMDb id, IMDb URL, language, plot, plot simple, poster, rated, rating, rating count, release date, runtime, title, type, writers, year, opening weekend, gross, filming dates, budget, weekend gross, copyright holder. In this study, the structure of data set is listed below. Attributes are factors that related to movies (e.g., user rating, genre, actor). Element is a subgroup of attribute (e.g., action is one kind of genre). In other words, at the high level are the attributes which can be defined in terms of more elements.

3.2 Data Cleaning, Transformation and Reduction

For the purpose of our analysis, we need to remove or reduce the noise and missing values from test data. This step reduce confusion to derive more useful classification rules[21]. We remove the irrelevant, weakly relevant or redundant attributes according to previous research conclusion and IT scholars' opinions. Besides, runtime is continuous type data. For research purpose, we convert runtime to categorical nominal type data and divide runtime data into four groups. Table 1 presents the attributes (independent variables) which we used in this study.

Table 1. Description of selected attributes

	Attributes	Attribute Types	Number of element	Literature
1	genres	Categorical nominal	24	[11, 12, 16, 17]
2	directors	Categorical nominal	8,880	[1, 11, 15]
3	actors	Categorical nominal	98,116	[1, 11, 13, 15-17]
4	writers	Categorical nominal	13,447	[7]
5	country	Categorical nominal	117	[22]
6	film_locations	Categorical nominal	1,220	-
7	runtime	Categorical nominal	4	[16]

3.3 Calculate the Score of Elements and Attributes

In this study, element is a quantifiable indicator of the extent to user rating. We collected movies that related to a certain element, and then we averaged user rating of the movies. The average user rating is the score of element. For example, Ang Lee is a director of Brokeback Mountain (2005), Hulk (2003), Talking Woodstock (2009). The user ratings of these movies are 7.6, 5.7, and 6.6. The score of Ang Lee is (7.6 + 5.7 + 6.6) / 3 = 6.63.

As noted in the previous section, element is a subgroup of attribute. We calculated the score of attribute after we had calculated the element score. We averaged the score of elements that belong to a certain attribute, and then the result was the attribute score. For example, there are five elements (animation, action, adventure, family, and mystery) which belong to the genre of The Adventures of Tintin (2011). The elements scores are 5.89, 5.97, 5.88, 6.87, 6.29, the genre score of The Adventures of Tintin is (5.89 + 5.97 + 5.88 + 6.87 + 6.29) / 5 = 6.18. The other attributes (actors, writers, country, film locations, runtime) use the same method to calculate the score.

4 Develop Prediction Models

We collected the user rating and attributes of all movies released from 2002 to 2012. Total dataset is including 32968 movies, 31,506 movies were used to be training data, and others were testing data. In section 3.3, we calculated all attribute scores and element scores, and then used training data to generate prediction rules. The rules can be used to predict future data. Methods used to develop the prediction models are represented below:

4.1 Linear Combination (Convex Combination) Model with Enumerating Value

The predicted user rating is derived by a linear combination of the scores of the attributes. The attributes may have different weights in deriving the predicted user rating. The computation method is as follows:

$$\text{Predicted user rating} = w_1 \times \text{Score}_{\text{genres}} + w_2 \times \text{Score}_{\text{directors}} + w_3 \times \text{Score}_{\text{actors}} + w_4 \times \text{Score}_{\text{writers}} + w_5 \times \text{Score}_{\text{country}} + w_6 \times \text{Score}_{\text{f\_location}} + w_7 \times \text{Score}_{\text{runtime}} \quad (1)$$

$$\text{Weights: } w_1, w_2, \ldots, w_7 \in [0,1];\ w_1 + w_2 + w_3 + w_4 + w_5 + w_6 + w_7 = 1$$

To find the optimal line combination, we tested all combinations of $w_1, w_2, \ldots, w_7$ by enumerating the values systematically in increments of 0.01 range from 0 to 1. When the accumulated difference between predicted user rating and actual user rating is the smallest, it can be considered as an optimal line combination. We use algorithm 1 and algorithm 2 to find the optimal weight combination. The algorithm and prediction model is as follows:

```
ALGORITHM 1: List all linear combinations
OUTPUT:
weight
PROGRAM:
For w1 = 0 To 1 Step 0.01
For w2 = 0 To 1 Step 0.01
    For (…)
      If w1 + w2 + w3 + w4 + w5 + w6 + w7 = 1 Then
        weight(weightcount, 1) = w1
        weight(weightcount, 2) = w2
        (…)
        weightcount ++
      End If
    End For
  End For
End For
```

```
ALGORITHM 2: Calculate the forecast error to yield the optimal
linear combination
INPUT:
movie
GenresScore, DirectorsScore, ActorsScore, WritersScore,
CountryScore,
LocationsScore, RuntimeScore
Genres, Directors, Actors, Writers, Country, Locations,
Runtime
weight
OUTPUT:
BestWeight
BestError
PROGRAM:
BestError = infinite
For i = 0 To weight.count-1
 For Each m In movie
     Error = Abs(m.Score – (GenresScore*weight(i,1) + Di-
rectorsScore*weight(i,2) +
     ActorsScore*weight(i,3) + WritersScore*weight(i,4) +
CountryScore*weight(i,5) +
     LocationsScore*weight(i,6)) +
RuntimeScore*weight(i,7)))
     If  Error < BestError Then
        BestError = Error
  BestWeight = i
     End If
   End For
End For
```

$$\text{Predicted user rating} = 0.05 \times \text{Score}_{\text{genres}} + 0.05 \times \text{Score}_{\text{directs}} + 0.15 \times \text{Score}_{\text{actors}} + 0.75 \times \text{Score}_{\text{writers}}$$

4.2 Multiple Linear Regression Model

In this section, we use multiple linear regression analysis to yield another user rating prediction model. User rating is dependent variable and other attribute (predictor variables) are independent variables. We applied stepwise regression technique to select predictor variables. In each step, we included a significant variable (at the 5% level) that brought the highest increase in adjusted R^2. After each variable inclusion step, we removed any previously included variable if the variable is no longer significant (at the 10% level). We stopped adding variables when the adjusted R^2 did not increase when additional variables were no longer significant. In this study, all variables were included in the regression model. We listed the result of the 7^{th} step in stepwise regression procedure in table 2.

Table 2. Results of stepwise regression

Step		Coefficient		Standard Coefficient	t-value	p-value
		Beta	Standard Error	Beta		
7	constant	-0.632	0.054		-11.804	.000
	writers	0.409	0.008	0.384	49.113	.000
	actors	0.556	0.009	0.432	61.934	.000
	directors	0.192	0.008	0.181	25.060	.000
	runtime	0.028	0.005	0.014	5.894	.000
	country	-0.031	0.006	-0.012	-4.944	.000
	genres	-0.028	0.006	-0.012	-4.768	.000
	film_ location	-0.012	0.003	-0.009	-3.929	.000

The prediction model is as follows:

$$\text{Predicted user rating}=-0.632+0.409 \times \text{Score}_{\text{writers}} + 0.556 \times \text{Score}_{\text{actors}} + 0.192 \times \text{Score}_{\text{directors}} + 0.028 \times \text{Score}_{\text{runtime}} - 0.031 \times \text{Score}_{\text{country}} - 0.028 \times \text{Score}_{\text{genres}} - 0.012 \times \text{Score}_{\text{file\_location}} + \varepsilon$$

4.3 Neural Networks Model

Neural networks is a massive parallel distributed processor made up of simple processing units[23]. Neural networks is composed of several interconnected nodes and links. It modifies its interconnection weights by apply a set of training data. The attribute scores is the input vector and the corresponding output is actual user rating. The prediction model was developed using a commercial software product called SQL Server Business Intelligence Development Studio.

It is difficult to interpret the meaning behind the interconnection weights and hidden layer in the networks [21]. Due to the poor interpretability, the result of neural networks was used to illustrate how effective these attributes are.

5 Conclusion and Future Works

5.1 Forecast Accuracy

In order to test the forecast accuracy of the prediction models, we use 1,462 movies to be testing data which we obtained from IMDb. For testing the forecast accuracy, we used the testing data to calculate the predicted user rating from three prediction models which we developed in chapter 4. Then, we calculate the difference (Prediction absolute error; PAE) between the forecast value and actual user rating. The smaller value of PAE is, the better forecast accuracy is. We calculate the percentage of the

appearance frequency in the different PAE area accounting for all testing data, to be used to compare the forecast accuracy of the three kinds of method.

$$\text{PAE} = |\text{Predicted user rating} - \text{Actual user rating}|$$

$$\text{Average PAE} = \frac{\sum_1^n \text{PAE}_n}{n}$$

$$\text{Percent of PAE between a and b} = \frac{\text{a} \le \text{number of PAE} < b}{\text{total number of testing data}} \quad 0 \le a < b \le 10$$

Table 3. Comparison predicted absolute error between the linear combination method, multiple linear regression and neural networks prediction models

Prediction model	Average PAE	PAE			
		$0 \le \text{PAE} < 1$	$1 \le \text{PAE} < 2$	$2 \le \text{PAE} < 3$	$3 \le \text{PAE} < 4$
Linear combination	0.7347	72.73%	24.45%	2.19%	0.31%
Multiple linear regression	0.8186	67.08%	28.21%	4.08%	0.31%
Neural networks	0.6973	76.8%	18.5%	4.39%	0.31%

As shown in Table 3, the average PAE of the linear combination is 0.7347, lower than the average PAE 0.8186 of multiple linear regression, while the average PAE of neural networks is only 0.6973, as the method with the lowest average PAE. The PAE percentage of the linear combination lower than 1 is 72.73%, higher than the 67.08% of multiple linear regression by 5.65%, while the PAE of neural networks lower than 1 is 76.8%, higher than 72.73% of the linear combination by 4.07%. The results of paired t-test were also indicated that there is a significant difference between the PAE of neural networks and the PAE of multiple linear regression. However, there is no significant difference between the PAE of neural networks and the PAE of linear combination. As mentioned above, it can be seen that the forecast performance of using neural networks prediction model to be greater than or equal to the linear combination prediction model, while the forecast performance of the neutral networks prediction model is better than that of multiple linear regression model.

5.2 Conclusion and Future Work

A proper weight combination forecast equation is obtained in this study to solve the linear combination (convex combination) by enumerating value systematically; meanwhile, multiple linear regression and neural networks applied to the development of forecast models. The result indicated that using neural networks is superior to the optimal weight combination provided in this study, while the forecast performance of the linear combination is better than multiple linear regression. These findings are in line with previous studies [17]. It is noteworthy that if we only focus on the PAE lower than 2, the linear combination model is the great ratio 97.18% (72.73%+ 24.45%), that is, 97.18% testing data predicted user rating error lower than 2 when we used linear combination prediction model. The PAE of neutral networks model is 95.3% in the same condition.

In Table 4, we listed the weights of linear combination and standard coefficient of multiple linear regression. Writers, actors and directors profoundly affect user rating. A writer is in charge of such core elements as the scheme, characters, scene, and structure of the whole movie; a good screen scripts can find an echo in everyone's heart. On the contrary, a poor screen script hardly gains the favor even under sufficient resources of various aspects. The next important factor is actors. The actor selection of a movie production company is extremely important. Most studies considered star as one of the covariates with box office performance. However, directors and genres also account for considerable influence on user rating. Therefore, before investors and a movie production company prepare to shoot a movie, they may well consider the favorable portfolios of the audience from such aspects of writers, actors, directors and genres to acquire a higher anticipated user rating. Once the anticipated user rating is reached, the increase of movie revenue will take place.

Table 4. Comparison between weights of the linear combination and standard coefficient of multiple linear regression

	Weight/ Standard Coefficient						
	Genres	Directors	Actors	Writers	Country	Film_ locations	Runtime
Linear combination	0.05	0.05	0.15	0.75	0	0	0
Multiple linear regression	-0.012	0.181	0.432	0.384	-0.012	-0.09	0.014

While our results are encouraging, there are still many improvements to be made. We consider that there are many factors with impact on user rating which are not explored. Some potential endogenous relationships exist among the factors [4], it is recommended that more studies of these questions could be performed.

This study uses enumerating value systematically to find out the proper weight combination; such kind of method highly consumes time and computer resources. The time performance of linear combination method is about 450 minutes, on the contrary, multiple linear regression is about 4 seconds and neural networks is about 8 seconds. Further research might adopt the other algorithms of solving a convex combination to reduce the calculation time and resources.

References

1. Elberse, A., Eliashberg, J.: Demand and supply dynamics for sequentially released products in International markets: The case of motion pictures. Marketing Science 22(3), 329–354 (2003)
2. Reinstein, D.A., Snyder, C.M.: The influence of expert reviews on consumer demand for experience goods: A case study ofmovie critics. The Journal of Industrial Economics 53(1), 27–51 (2005)

3. Eliashberg, J., Shugan, S.M.: Film critics: Influencers or predictors? Journal of Marketing 61(2), 68–78 (1997)
4. Basuroy, S., Chatterjee, S., Ravid, S.A.: How critical are critical reviews? The box office effects of film critics, star power, and budgets. Journal of Marketing 67(4), 103–117 (2003)
5. Motion Picture Association of America, I., Theatrical Market Statistics Report, Motion Picture Association of America, Inc. p. 31 (2013)
6. Motion Picture Association of America, I., MPAA Economic Review (2004)
7. Eliashberg, J., Hui, S.K., Zhang, Z.J.: From story line to box office: A new approach for green-lighting movie scripts. Management Science 53(6), 881–893 (2007)
8. Eliashberg, J., Elberse, A., Leenders, M.A.A.M.: The motion picture industry: Critical issues in practice, current research, and new research directions. Marketing Science 25(6), 638–661 (2006)
9. Jones, J.M., Ritz, C.J.: Incorporating distribution into new product diffusion models. International Journal of Research in Marketing 8(2), 91–112 (1991)
10. Krider, R.E., Weinberg, C.B.: Competitive dynamics and the introduction of new products: The motion picture timing game. Journal of Marketing Research 35(1), 1–15 (1998)
11. Ainslie, A., Drèze, X., Zufryden, F.: Modeling movie life cycles and market share. Marketing Science 24(3), 508–517 (2005)
12. Zufryden, F.S.: Linking advertising to box office performance of new film releases: A marketing planning model. Journal of Advertising Research 36, 29–42 (1996)
13. Ravid, S.A.: Information, blockbusters, and stars: A study of the film industry. The Journal of Business 72(4), 463–492 (1999)
14. Hennig-Thurau, T., Houston, M.B., Sridhar, S.: Can good marketing carry a bad product? Evidence from the motion picture industry. Marketing Letters 17(3), 205–219 (2006)
15. Litman, B.R., Kohl, L.S.: Predicting financial success of motion pictures: The '80s experience. Journal of Media Economics 2(2), 35–50 (1989)
16. Simonton, D.K.: Cinematic success criteria and their predictors: The art and business of the film industry. Psychology & Marketing 26(5), 400–420 (2009)
17. Sharda, R., Delen, D.: Predicting box-office success of motion pictures with neural networks. Expert Systems with Applications 30(2), 243–254 (2006)
18. Malczewski, J.: On the use of weighted linear combination method in GIS: Common and best practice approaches. Transactions in GIS 4(1), 5–22 (2000)
19. Dawes, R.M.: The robust beauty of improper linear models in decision making. American Psychologist 34(7), 571–582 (1979)
20. Johnson, J.W., LeBreton, J.M.: History and use of relative importance indices in organizational research. Organizational Research Methods 7(3), 238–257 (2004)
21. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn. Morgan Kaufmann, San Francisco (2006)
22. Lee, F.L.F.: Cultural discount and cross-culture predictability: Examining the box office performance of American movies in Hong Kong. Journal of Media Economics 19(4), 259–278 (2006)
23. Kantardzic, M.: Data Mining: Concepts, Models, Methods, and Algorithms. IEEE Press, Piscataway (2003)