Prediction of Movies popularity Using Machine Learning Techniques

Muhammad Hassan Latif[†], Hammad Afzal^{††}

National University of Sceinces and technology, H-12, ISB, Pakistan

Summary

Number of movies are released every week. There is a large amount of data related to the movies is available over the internet, because of that much data available, it is an interesting data mining topic. The prediction of movies is complex problem. Every viewer, producer, director's production houses all are curious about the movies that how it will perform in the theatre. Many work has been done relating to movies using social networking, blogs articles but much less has been explored by the data and attributes related to a movie which is continuous and in different dimensions. We have used IMDB for our experimentation. We created dataset and then transformed it and applied machine learning approaches to build efficient models that can predict the movies popularity.

Key words:

machine learning, classification, movies, logistic.

1. Introduction

A movie is also called a film or motion picture, is a combination of still images, when displayed on a screen, behaves as an illusion of moving images, it is because of the phi phenomenon. The process of movie making is both an industry and an art [1]. Movies are a great source of entertainment and people are crazy about movies. Movie Industry produces hundreds of movies every year of different genres such as animation, war, comedy, thriller horror etc. Most of the time, people are not sure about which particular movie to look for so that their spare time is utilized in entertainment. There are many online platforms that keep track of movies like RottenTamatoes, Metacritic and Internet Movie Database (IMDb) which provide information about movies such as actors, directors, budget, as well as user ratings and comments which provide a fair information about the movie.

Internet movie database (IMDB) is the number one consumer site of movies. It contains information about programs, films and television including financial information, biographies, user rating, cast, reviews, crew, actors, directors, summaries etc. It has database of approx. 60 million registered users and 6.6 million personalities with 3.4 million movie and episodes titles [2].

"Hollywood is the land of hunch and the wild guess" [3] [5].Thousands of movies are released every year. According to a study, movie industry in the United States generate revenue up to 10 billion dollars. Each movie cost 100 million approx. but still there is a great deal of uncertainty that the movie will do business or not [3]. Movie industry is a big business, which can give profits or loss up to several millions dollar [4]. It will be a hit or flop this give rise to the movie prediction problem.

A lot of research has been done on prediction of movies. Most of them include user ratings on different movies, whereas, some of them use social media (e.g. YouTube, Twitter etc) for prediction. However, less work has done on using movies attributes such as crew, dates etc. to predict movies. The amount of data available about the movies over the internet makes its serious candidate for data mining, knowledge discovery and also machine learning. Most of the work done on movies relating to its ratings and reviews over the internet. Prediction of a movie is of great importance to industry; movie makers are still never sure about whether there movie will do business or not; when they should release the movie and how to advertise it [3][6].

2. Literature Overview

Many different researches are carried for the prediction of movies by using different approaches by using news, articles, blogs, articles and social media etc. but quite a few have explored through attributes related to a movie. The research has been done on predicting the future of movies in terms of its business (box office revenue) using data from social media and employing sentiment analysis. [7], another similar work has been presented in [8] where social media including twitter and YouTube's comments are used for same purpose. [9]Presents prediction of popularity of a movie by the articles on Wikipedia. The research shows that these articles can be used to get some future outcomes. It also uses financial data of movies from box office mojo by using Pearson's correlation coefficient and linear regression. There is a research that predicts the opening weekend revenue. It takes the movie information like actors, director, genre and released date etc. from metacritic and financial data like budget, opening week gross revenue from the numbers. The pre-released articles about the movies are collected from seven different articles sources. Mean Absolute error, person's correlation

Manuscript received August 5, 2016 Manuscript revised August 20, 2016

coefficient and linear regression are employed [10]. [11] Presents predicting the gross of movies using the news articles as well. [12] Uses IMDb data and data from 'boxofficemojo'. They applied PART and C4.5 and used correlation coefficient as a measure, they have created 2 dataset of Pre-release movies and post release movies and applied experimentation on it. [6] Uses IMDb data, rotten tomatoes and Wikipedia data about the movie and machine learning algorithms are applied on it like linear regression, SVM regression and logistics regression.

3. Methodology

The Proposed methodology is illustrated in Figure 1. It contains following steps

- Data Extraction
- Data Preprocessing
- Data Integration and Transformation
- Feature Selection
- Classification





3.1 Data Extraction

The dataset to be used is collected from internet movie database (IMDB). We targeted the movies that were released from year 2004 to 2014 to get the latest trend of market. Only those movies are selected that are listed on Wikipedia film of the year pages and are English movies released in United States all other movies are excluded. We then write the script in python and C# to pull the data from IMDB and stored in MySQL. Our dataset contains about 2000 data points.

3.2 Data Preprocessing

The data we extracted from IMDB need to be cleaned as the data is obtained from multiple sources mainly IMDB and Wikipedia. The data is inconsistent, missing and very noisy as well. To cater missing fields issue we have used central tendency as a standard for filling missing values for attributes [6]

3.3 Data Integration and Transformation

The data extracted from IMDB need to be integrated and transformed so that it can be used for analysis and classification purposes

3.3.1 Rating

Each movie in IMDB has a rating from 0 to 10. Every user can rate the movie which is a vote and average of voting is the rating for the movie. As the values will be continuous we need to generalize the rating for analysis and classification. This will be our class in prediction. We have used the same approach as used by [12] [13] in there research.

Tal	ble 1. Consideration for classical				
	Class	Rating			
	Terrible	0.0-2.4			
	Poor	2.4-5.0			
	Average	5.1-7.4			
	Excellent				

^{3.3.2} MPAA Rating

Motion picture Association of America (mpaa) is a body that assign rating to the movies. These ratings represent violence, sexual content, and language in a movie. There are 5 categories for each of the movies mainly R, PG, PG13, G and NR. We have represented mpaa rating for a movie with 5 binary values each representing 5 categories of mpaa rating [5] [14].

3.3.3 Genre

Genre represent the type of content present in a movie. A movie can have a single genre or many for example Fish Tank a movie released in 2009 has a single genre drama where as a movie Exodus: Gods and kings released in 2014 has 3 genre that are action, adventure and biography. Our dataset includes following genre for each of the movie that are action, adventure, thriller, biography, crime, drama, horror, comedy, fantasy, animation, mystery, music, war, documentary, romance , Scifi, western, family, sport, and short. To represent genre for each of the movie we have assign 20 binary variables to each of a movie [5] [14]. 3.3.4 *Awards*

Oscars and golden globe awards are the most prestigious awards for the movies. Many movies win awards and many get nominated. Even getting nominated is a big achievement. We have assigned 4 binary value to represent awards 2 for Oscar won and nominated and 2 for golden globe wining and nominee movie [14].

3.3.5 Number of Screens

When a movie is released it is displayed in cinemas and screen. Before the release of a movie it is already decided that how many the number of screens will be used to screen the movie [14]. We have used this attribute for our dataset since it tells us about the number of screens for the opening week of a movie [11].

3.3.6 *Opening week business*

Each movie revenue is calculated mostly in weeks. The revenue generated by a movie to the first week of its release is used as an attribute name opening weekend business [15]. The revenue for each movie is a continuous attribute with values ranging from several to many thousands dollars, we discretize this attribute into 9 buckets using existing approaches [3][5].

3.3.7 Budget

Budget is the amount of resources that is used in the making of a movie. It is the total amount of money that is used in the whole making. Budget can range from several thousand dollars up to million dollars [11] [16]. Budget is a continuous attribute so we have discretize this attribute by using the existing approaches [3] [5].

3.3.8 Metascore

Metacritic is a website that keep track of movies, episodes, albums, books games etc. Metacritic assign a score to each of the product for example a movie. Each movie is assigned score based on the type of review and score given by a user and average of score is set for the product and that is metascore [20]. IMDB beside its rating and score also uses metascore so we have used this as our attribute in dataset

3.3.9 *Number of votes*

Imdb user can log in to IMDB and explore the movies of their choices and get fair amount of information about a movie. Each user can also rate the movie from a scale of 0 upto10. For each movie, number of votes are also calculated that rated the movie [2].

Table 2 .Summary of Independent variable	Table 2	.Summary	of Inde	pendent	variable
--	---------	----------	---------	---------	----------

Name of Variable	No. of Values	Values
Rating	4	Terrible, Poor, Average, Excellent
MPAA Rating	5	R, PG, PG-13, G, NR
	20	Action, adventure, thriller, biography, crime, drama, horror, comedy, fantasy, animation, mystery, music, war ,documentary, romance, Sci-fi, Westren, Family, sport,short
Awards	4	Oscar Won, Oscar Nominee, Golden Globe won, Golden globe nominee
Screens	1	Positive integers
Opening weekend	9	1,2,3,4,5,6,7,8,9
Metascore	1	Positive integers
Number of votes	1	Positive integers
Budget	9	1,2,3,4,5,6,7,8,9

Table 2 illustrate the values and number of attributes that are used in our experimentation

3.4 Feature Selection

There are many ways for feature selection process and its identification. Feature selection addresses the issue of most important attribute from the dataset. It eliminates those attributes that do not add information to the analysis process. It finds out the weight of mot contributed attribute and the attribute that contribute the least. There are many approaches for feature selection we will use Information Gain for feature section process [17]. The information gain of an attribute can be calculated by formula [19] where m is no. of classes and probability of class i is represented by p_i

$$Entropy = \sum_{i=1}^{m} p_i \, \log_2 p_i \quad (1)$$

Info_gain = entropy (parent)[avg.entropy (children)]

The information gain depends on the value of entropy. Information gain will be high if the entropy of an attribute is maximum. Entropy calculates on the basis of instance division into subsets. It calculates dependency in dataset. If it is completely dependent then entropy will be zero [20].

3.5 Classification

There are many data mining tools available. We have used WEKA for our experimentation. WEKA is well suited for data mining tasks with collection of machine learning algorithms. It can perform classification, data preprocessing, clustering, regression, visualization and association rules [12][18]. The machine used for experimentation has Intel core I3 processor with 4GB ram. As we are doing supervised learning following some linear and non-linear classifier are used for classification, implemented and tested. Following classifiers are selected for experimentation are shown in the table 3.

Table 3. List of Classifier
Classifier
Logistic regression
Simple Logistic
Multilayer Perceptron
J48
Naive Bayes
PART

We have applied the following classifier on our dataset. Then each of the classified results are tested against k-fold cross validation where the value of X = 10.

4. Results

The results obtained from each of the classifier is shown in figure 2. These result shows the percentage of time we are able to correctly predict the instances. We are able to achieve highest accuracy with simple logistic and logistic regression 84.34 % and 84.15% respectively. We also implemented some other classifier, neural network classifier that is multilayer perceptron produces accuracy pf 79.07%, and decision tree j48 results are 82.42%.



Figure 2. Classification Results

Receiver operating characteristic curve also known as ROC curve is a depiction of true positive in opposition to false positive for possible different points in a test. It demonstrate that compromise between the two specificity and sensitivity when the sensitivity increases specificity decreases. The accuracy of test can be seen if the curve is closer to the top and left of graph, the more it is closer the result are better same way if the curve comes near 45 degree angle the results would not be accurate. If the ROC curve value is above 0.9 the results are excellent, and are good at 0.8-0.9 and fair at 0.7-0.8 and poor below 0.6 [21][22].

ROC Area is shown in figure 3 for each of the algorithm which correctly predicted the movie to be excellent. Again we have got best results with simple logistic and logistic regression with ROC area of 0.932 and 0.922 respectively. Naive Bayes algorithm also produces a fine curve with 0.923. Other algorithm produces ROC area of 0.859 in multilayer perceptron, 0.825 in J48 and 0.829 in PART.



Figure 3. Roc Area

According to a study trying to do prediction and analysis for movies, they applied locally weighted linear regression and support vector machine, they were able to predict ratings for IMDB 70% of the time [13] whereas our result shown in the figure 2 are much better, we achieved the least accuracy with neural network multilayer perceptron 79.07% and highest with simple logistic of 84.34%. Another study applied logistic regression, linear regression and support vector machine, there success rate for SVM regression 39%, for logistic regression were 42.2% and for linear regression 50.7%. Whereas we have achieved much higher accuracy of 84.15% for logistic regression. In a research of predicting movie popularity using C4.5 and PART for movie popularity and they were able to correctly classify instance 77.3562% of the time with c4.5 and 77.7234% with PART classifier. In our experimentation, our classifier correctly predict the instance 79.52% with PART classifier.

5. Conclusion and Future works

The proposed research aims to predict movies popularity. We have used machine learning approach for our experimentation. Machine learning have powerful classification algorithms for classification. Our research aims to improve previous researches. Performing data mining on IMDB is a hard task because of so many attributes related to a movie and all in different dimensions with lots of noisy data and missing fields. After performing classification, we have found out that our best results are achieved through simple logistic and logistic regression at around 84 %. The attributes that contributed the most to information are metascore and number of votes for each movie, Oscar awards won by the movies and the number of screens the movie is going to be screened.

References

- [1] https://en.wikipedia.org/wiki/Film , Accessed on August 1st, 2015
- [2] https://en.wikipedia.org/wiki/Internet_Movie_Database Accessed on August 1st, 2015
- [3] Darin Im and Minh Thao Nguyen : "PREDICTING BOX-OFFICE SUCCESS OF MOVIES IN THE U.S. MARKET ", CS 229, Fall 2011
- [4] Jeffrey S. Simonoff and liana R.Sparrow : "Predicting Movie Grosses : Winners and Losers, Blockbusters and Sleepers". Chance, vol. 13(3), pp. 15–24, 2000.
- [5] Ramesh Sharda , Dursun Delen :"Predicting box-office success of motion pictures with neural networks", Expert Systems with Applications 30 (2006) 243–254
- [6] Nithin VR, Pranav M, Sarath Babu PB, Lijiya "A Predicting movie success based on IMDB data" International journal of data mining and techniques, Volume 03, june 2014, pages 365-368
- [7] Sitaram Asur, Bernardo A. Huberman "Predicting the Future With Social Media", Hp Labs
- [8] Andrei Oghina, Mathias Breuss, Manos Tsagkias, and Maarten de Rijke, "Predicting IMDB Movie Ratings Using Social Media", Advances in Information Retrieval, Volume 7224, 2012, pp 503-507
- [9] Mestya'n M, Yasseri T, Kerte'sz J (2013): "Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data". PLoS ONE 8(8): e71226.doi:10.1371/journal.pone.0071226
- [10] Mahesh Joshi Dipanjan Das Kevin Gimpel Noah A. Smith: "Movie Reviews and Revenues: An Experiment in Text Regression", The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Pages 293-296
- [11] Wenbin Zhang ,Steven Skiena :"Improving Movie Gross Prediction Through News Analysis", Department of computer science stony brook university, 2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology – Workshops, Pages 301-304
- [12] Khalid Ibnal Asad, Tanvir Ahmed, Md. Saiedur Rahman: "Movie Popularity Classification based on Inherent Movie Attributes using C4.5,PART and Correlation Coefficient", IEEE/OSA/IAPR International Conference on Infonnatics, Electronics & Vision, Pages 747 - 752
- [13] Saraee, M, White, S and Eccleston, J 2004, "A data mining approach to analysis and prediction of movie ratings", The Fifth International Conference on Data Mining, Text Mining and their Business Applications, 15-17 September 2004, Malaga, Spain
- [14] Jeffrey Ericson & Jesse Grodman : "A Predictor for Movie Success" CS229, Stanford University
- [15] Nikhil Apte, Mats Forssell, Anahita Sidhwa :"Predicting Movie Revenue", CS229, Stanford University ,December 16,2011
- [16] Steven Yoo, Robert Kanter, David Cummings : "Predicting Movie Revenue from IMDb Data"
- [17] Jiawei Han, Micheline Kamber, Jian Pei : "Data mining concepts & techniques", third edition, 2011
- [18] http://www.cs.waikato.ac.nz/ml/weka , Accessed on August 1,2015
- [19] Saba Bashir, Usman Qamar, Farhan Hassan Khan, M.Younus Javed : "An Efficient Rule-based Classification

of Diabetes Using ID3, C4.5 & CART Ensembles", 12th International Conference on Frontiers of Information Technology, Pages 226 - 231

- [20] http://www.metacritic.com/ , Accessed on August 1,2015
- [21] http://gim.unmc.edu/dxtests/ROC1.htm , Accessed on August 1,2015 Charles E. Metz , "Basic principles of ROC analysis", volume 8, issue 4, pages 283-298, Deparmtnet of radiology, university of chicago



Hammad Afzal is working at National University of Sciences and Technology, Pakistan as an Assistant Professor since May, 2010. He has previously done his PhD, working under the supervision of Dr. Goran Nenadic and Dr. Robert Stevens, based in School of Computer Science, University of Manchester UK. Dr. Hammad Afzal has also worked as a Research Fellow at Digital

Enterprise Research Institute (DERI), National University of Ireland, where he was working with Dr. Paul Buitelaar, as part of the Unit of Natural Language Processing. (Worked after PhD). He is currently involved in a teaching and a number of UG and PG projects, mainly in the fields of Natural Language Processing, Semantic Web, Linked Data and Smart Phone Applications (UG)



Muhammad Hassan Latif He has previously done his Undergraduate in Computer Software Engineering, and is doing his Master's degree in Computer Software engineering s well and working under the supervision of Dr. Hammad Afzal. He is a senior .NET web developer at a software company based in Islamabad Pakistan.