Movie Success Prediction using Data Mining

[1]Anantharaman V [2]Ebin G. Job [3]Neha Sam [4]Neha Sam [5]Asst. Prof. Sheryl Maria Sebastian
[1,2,3,4,5]Department of Computer Science and Engineering
[1,2,3,4,5]Albertian Institute Of Science and Technology, Kalamassery, Ernakulam, Kerala

Abstract

In this project, we developed a mathematical model to predict the success and failure of upcoming movies based on several attributes. Some of the attributes in calculating the success of a movie include director, actor, genre, and budget. In this model Data mining process was used to extract patterns and trends which can be beneficial in predicting movies success. The data mining techniques were applied to a movie database, but before the mining techniques could be used, the data went through the cleaning and integration process. Decision tree classifiers are used in this paper to generate decision tree for given inputs containing various attributes. Movie success prediction is also significant for the movie watchers who need to know in advance the quality and success rating of a movie.

Keyword- Data Mining, Cleaning, Decision Tree, Attributes

I. INTRODUCTION

Movies have become popular in last century. It is the best way to disconnect yourself from your world for 2-3 hours. If the movie is good, you end up laughing or if the movie is bad, you end up hating it. So predicting the success of movie will help business significantly. So paper proposes a model that predicts the success of upcoming movies using data mining. Data mining techniques are used for analyzing patterns in previous records of movies. They help in deciding the success or failure of upcoming movies. Data mining approach is important since it can help to identify hidden patterns and relationship among various variables [1]. It could be useful in many scenarios including profit predictions, investment decision, medical purpose and many more.

Movie success prediction is important because it involves significant time and investment. It is very useful for the movie watchers who need to know in advance the success rating and quality of upcoming movies [8]. So it is important to have more accuracy in prediction. It can be achieved using data mining. The data mining techniques are applied to the movie dataset and before the mining technique is applied, the data goes through data pre-processing stages.

In this paper, we have proposed a mathematical model for predicting the success of upcoming movies depending on various attributes such as director, genre, actor, budget etcetera [5]. We make use of historical data of movies to predict the success of upcoming movies [6]. Data mining techniques are used to analyze the pattern in previous records and classification is done based on this. Decision tree classifiers are used in this paper to generate decision tree for given inputs containing various attributes such as director, actor, and genre, budget to classify the input tuple as hit or flop. The goal of this model is to predict the success of upcoming movies with high accuracy.

II. LITERATURE SURVEY

The classical movie attributes such as cast, director, producer, and genre play a crucial role in the movies success. Dan Cocuzzo et al have used Naive Bayes and Support vector machine to predict the movie success. In Naive Bayes algorithm, they represented movie as independent combination of associated personas and attributes which was given by, P (rating | movie) proportional to P(movie | rating) * P(rating) ,where P(movie | rating) is product of individual conditional probabilities for each person.

Sitaram Asur and Bernardo A. Huberman applied data mining tools to generate interesting patterns for predicting box office performance of movies using data collected from multiple social media and web sources including Twitter, YouTube and the IMDb movie database. The prediction is based on decision factors derived from a historical movie database, followers count from Twitter, and sentiment analysis of YouTube viewers' comments. We label the prediction in three classes, Hit, Neutral and Flop, using Wekas K-Means clustering tool. Interesting patterns for prediction are generated by Wekas J48. Since our prediction is for movies yet to be released in summer 2013, the performance of the final results will be validated by a follow-up study.

Machine learning has also been used for predicting movie success by using algorithms like RF and SVM. Although the use of RF and SVM within the movie domain seems to be fairly limited, the two algorithms have been applied and evaluated in many applications for the purpose of regression as well as classification. Within recent study Verikas et al. (2011) have surveyed a number of large as well as small scale comparisons on data mining and machine learning, all of which include the RF algorithm, specifically issuing its prediction performance in comparison to other algorithms as well as the use of the variable importance estimates available from RF. Among the previous applications and algorithm comparisons included by Verikas et al. (2011) are

several large scale studies such as Meyer et al. (2003) and Statnikov et al. (2008), evaluating RF and SVM among other algorithms over a number of 33 and 22 datasets respectively.

III. PROPOSED METHOD

In our work, we have developed a mathematical model which is used to predict the success and failure of upcoming movies depending on certain parameters such as director, actor, actress, date of release, location, genre of movie the movie. Our work provides advantage in that strong correlations were found between different criteria and movie success rating. Unlike the related work discussed, our work can be used to predict movie success even before it is released. Our work makes use of historical data in order to successfully predict the ratings of movies to be released. Data mining techniques are used for analyzing patterns in previous records of movies based on which classification is done.

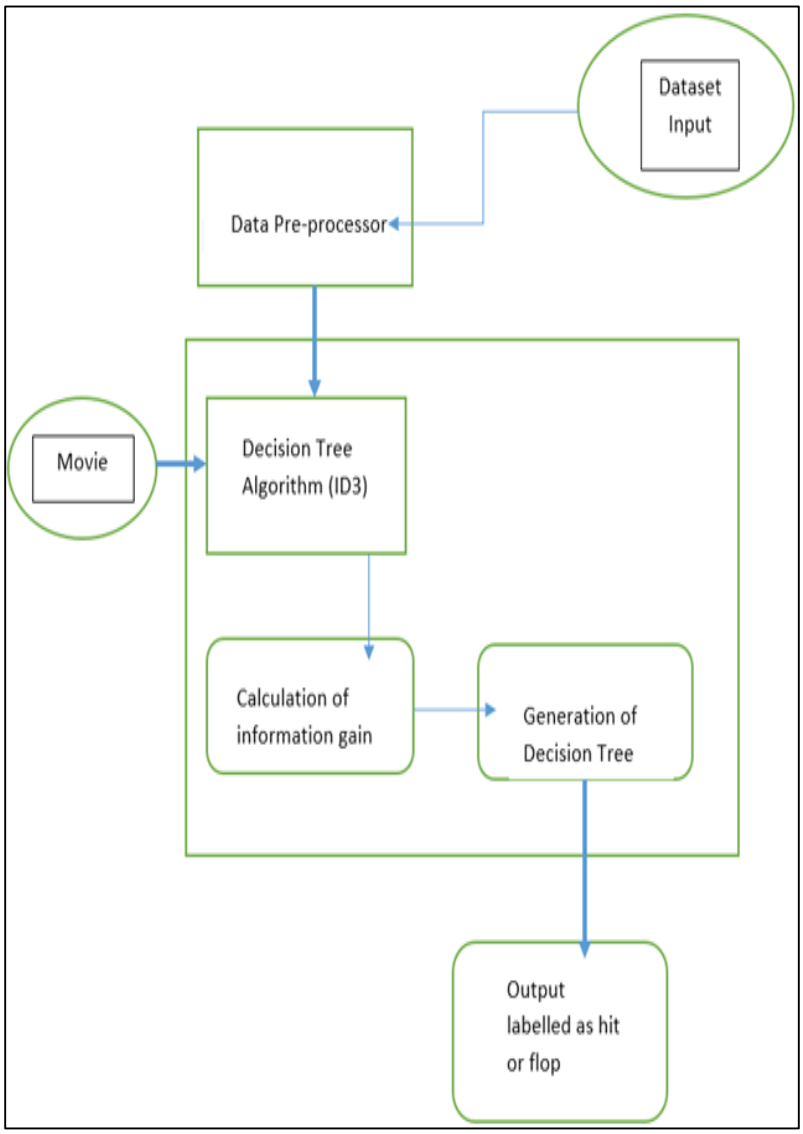

Fig. 1: System Architecture

The system makes use of decision tree classifiers to generate a decision tree for the given input tuple that contains various attribute values like name of the director, main cast (up to 3 actors), the genre of the movie, its budget etcetera to classify the tuple as hit or flop.

We use over 5000 movies data to train and test the classifier, with the dataset being obtained from IMDb. The CSV file is used as the input to the classifier algorithm that generates the output as the class label. Our classifier will have an accuracy of at least 85%, so it is safe to say it will be fairly accurate. The formula is:

$$\text{Accuracy} = \frac{\text{Number of correctly classified tuples}}{\text{Total number of tuples}}$$

The various functional and/or logic stages are described below:

A. Data Pre-processing

Although initially there will be only a single dataset, on further development, different datasets would be of different formats. This needs to be integrated. Also attribute subset selection is done to ensure the best attributes are selected. This is done manually.

B. Decision Tree Algorithm

ID3 algorithm is used to generate decision trees for the given input movie name using the dataset is calculated accordingly. After all calculations, decision tree is generated and the class label is predicted for the entered movie.

Apart from this the administrator will be able to edit the system, as in change the attributes selected, add, edit or remove details into/from the datasets, edit user profiles etc.

The algorithm is as follows:

ID3 (Examples, Target_Attribute, Attributes)
Create a root node for the tree
If all examples are positive, Return the single-node tree Root, with label = +.
If all examples are negative, Return the single-node tree Root, with label = -.
If number of predicting attributes is empty, then Return the single node tree Root,
With label = most common value of the target attribute in the examples.
Otherwise Begin
A ← The Attribute that best classifies examples.
Decision Tree attribute for Root = A.
For each possible value, vi, of A,
Add a new tree branch below Root, corresponding to the test A = vi.
Let Examples (vi) be the subset of examples that have the value vi for A
If Examples (vi) is empty
Then below this new branch add a leaf node with label = most common target value in the examples
Else below this new branch add the subtree ID3 (Examples (vi), Target_Attribute, Attributes – {A})
End
Return Root

IV. CONCLUSION

In this work we have come up with a mathematical model for predicting the success of movies depending on various attributes such as director, genre, actor, budget etcetera. Data mining techniques are used for analyzing patterns in previous records of movies based on which classification is done. Our work and results can be used to predict success or failure of upcoming movies. Our model shows an accuracy of 80% which is higher than other methods.

REFERENCES

[1] Javaria ahmad, Prakash Duraisamy, Amr Yousef, Bill Buckles, "movie success prediction using data mining" july 2017
[2] SitaramAsur, Bernardo A. Huberman, "Predicting the Future with Social Media", March 2010.
[3] Ajay Siva, Santosh Reddy, Pratik Kasat, Box-Office Opening Prediction of Movies based on Hype Analysis through Data Mining in International Journal of Computer Applications, Foundation of Computer Science, New York, USA, vol.56, no. 1, October 2012.
[4] Dan Cocuzzo, Stephen Wu ; Hit or Flop: Box Office Prediction for Feature Films; Stanford University , 2013.
[5] Anand Bhave, Himanshu Kulkarni, Vinay Biramane, "Role of Different Factors in Predicting Movie Success", 2015
[6] Nithin VR, Pranav, Sarath Babu PB, Lijiya A" Predicting Movie Success Based on IMDb Data" ,October 2017
[7] Steven Yoo, Robert Kanter, David Cummings; Predicting Movie Revenue from IMDb Data; Stanford University, 2011
[8] Jeffrey Ericson, Jesse Grodman; A Predictor for Movie Success ";Stanford University, 2013.