Proceedings of the Fourth International Conference on Computing Methodologies and Communication (ICCMC 2020)
IEEE Xplore Part Number:CFP20K25-ART; ISBN:978-1-7281-4889-2

Decision Tree Algorithms for Accurate Prediction of Movie Rating

KavyaPradeep
Student, Department of CS
Amrita school of arts and science
Kochi, Kerala
Amrita vishwavidyapeetham, India
kavyathachu@gmail.com

TintuRosmin C R
Student, Computer Science department
Amrita School of arts and science
Kochi, Kerala
Amrita Vishwavidyapeetham, India
rosemol148@gmail.com

Sherly Susana Durom
Student,Computer Science department
Amrita school of arts and science
Kochi, Kerala
Amritavishwavidyapeetham,India
sherlysusana16@gmail.com

G S Anisha
Faculty Associate,Department of CS & IT
Amrita School of Arts and Sciences
Kochi, Kerala
Amritavishwavidyapeetham,India
gs.anisha21@gmail.com

1. ABSTRACT

In this paper, we aim to find an accurate algorithm for implementing information mining systems utilizing Weka tool to foresee the achievement or disappointment of motion pictures dependent on a few important qualities with respect to system by systematically analyzing decision tree algorithms. The weightage for each attribute is calculated in the first stage, and then weightage of that movie is calculated by combined calculation of attributes using decision tree algorithms. Here we are trying to implement this idea on mainly three decision tree algorithms. J48 algorithm, Random Forest, Hoeffiding tree. Here, we try to find key factors for a profitable movie. This model assists with discovering the rating of the upcoming motion picture through qualities or attributes of that movie.

(Key words: Data mining, Weka Tool, J48 algorithm, Random forest algorithm, Hoeffding Tree)

2. INTRODUCTION

Movies are only a suitable way to preoccupy the hearts of viewers. Silver screens can also be used for conveying the messages to society. Unfortunately only a few movies get respect from society. They will get higher appreciation. Film industry gives rise to a large number of movies in a year. The revenue from movies are based upon many elements for example, cast, acting, spending plan for making the movie, review of moviecensors, score for the film, releasing time of the film, etc.

As you see above, there are many components that determine the fortune of the movie, Because of that there is no particular formula for analyzing and calculating the fortune and net gross it could earn. However by analyzing revenues from movies, a prototype can be built that could be given a hand in predicting the estimated income for a specific film.

Today a large amount of movies are created for many purposes. If some are created to entertain society, some are doing it to convey messages to society, and some are doing it for reaping money. In order to expand the creation of movies as a business we need the help of technology. Technological use will help the business significantly. Through the prediction of the movie, the businessman is able to know whether fortune becomes positive or negative. If the fortune is prophesying negative, they are able to change the genre or director or actor accordingly preferable to a positive outcome.

Indian Hindi Film business which is also identified as Bollywood has got in peak of commercial, movies produced and its reach. With so many stakes the failure of the movie is not a simple thing. For competing with that, the prediction of the movie can be used as criteria for success.

RELATED WORKS

1. Movie Success Prediction Using Data Mining

In 2017 Ahmad J, Duraisamy P, Yousef A and Buckles B had done a project on motion picture achievement forecast utilizing information mining. In that project they built up a scientific model which is utilized to anticipate the achievement and disappointment of up and coming motion pictures relying upon specific criteria. Their work gives favorable position of solid connections that were found between various criteria and motion picture achievement rating. Their work can be utilized to foresee film achievement even before it is discharged. Their work utilizes chronicled information so as to effectively anticipate the appraisals of motion pictures to be discharged. They projected the X^2 prototypical to classify the data. X^2 is a widely accepted classification technique for classifying data. In this method, X^2 between two attributes are calculated. They find X^2studyamongstkinds and rating of the cinemas, actors and movies ratings and actors and movie genres. Finally they identified a robust connection amongst artists and kinds. X^2 value of these attributes is 20.6. Using this correlation, the data is predicted.

2. A Data mining Technique for Analyzing and Predicting the success of Movie

In 2018 K Meenakshi, G Maragatham, Neha Agarwal and IshithaGhosh had a project on Information digging procedure for examining and foreseeing the achievement of films. For that they used 4 major components such as Data collection, Data Cleaning, Data Transfer, Data Analysis and Prediction. They took data from IMDB. The problem they faced was to find the structural database or a pivotal vault. IMDB rated movie according to the Bayesian principal. They used K-Mean clustering for the classification of data. Through this classification, they classify the movie into two different categories. Here Euclidean distance is used to cluster the data into a given algorithm. The problem they faced here was data from IMDB needs proper cleaning, integration which consumes a large amount of time. Textural information other than numerical arrangement makes mining progressively troublesome.

3. A data mining approach to analysis and prediction of movie ratings

In 2004 M Saraee, S White, and J Eccleston had done a paper referring to the production detail of about 390,000 movies from IMDB. They gathered a series of interesting factors from video games, box office taking, television series etc. for analyzing. They found that the budget of the film indicates no idea on the rating of that film. There is a descending pattern in the nature of the motion pictures step by step. They found a difficulty in extracting details from IMDB as it has no definite structure. They first classified the whole sole data in basis of Budget, which points out that there is no relation between budget of the movie and the rating. They couldn't arrange the information with 100% sureness significantly after 10 levels in decision tree. They found out that the actress or actor appearing in that movie is 90% relevant each. Director is 55% relevant and Budget plays only 28% relevance in the prediction.

3. METHODOLOGY

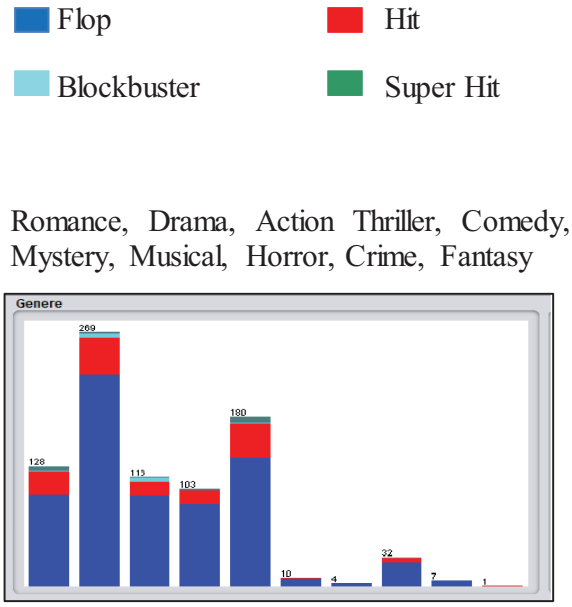
Data mining seems to be the youngest technology in the field of Computer Science for identifying the pattern in the dataset for prediction. In this proposed system, there are three different algorithms. By comparing these algorithms we identify the best algorithm that shows the highest accuracy and with the help of this algorithm we can predict the success of upcoming movies. For finding the accurate algorithm, we downloaded dataset from Kaggle.com namely "Bollywood movies". It needed a high level cleaning. We included rating, budget and net-gross for each movie manually with the help of IMDB and Box-Office India websites. We then processed this raw data in order to classify those using decision algorithms. 400 movies from the considered movie set had taken for training set. This training set ispre-processed and loaded in the first stage. After that using this remaining data is classified considering the classified training set.

J48 Algorithm:
J48 classification is the process of building a prototype of classes from a set of data (either raw or processed) containing class labels or attributes. It is an implementation of ID3 algorithm. J48 algorithm gives importance to the attributes.J48 algorithms will first find the label and classifies the data with respect to that algorithm. This process will be iterated until every set of data is classified according to the label decision tree algorithm is to discover the way the qualities vector responds for various examples joining with different traits.

Random Forest Algorithm:
Random Forests is a significant adjustment of bootstrap aggregation that fabricates an enormous numbers of non-connected trees, and afterward midpoints them for one single outcome. On many problems the performance of random forests can be closely related to boosting and they can be trained and tuned without any complication.

Hoeffding tree:
Hoeffding bound gives a specific degree of certainty on the best credit to part the tree, subsequently we can manufacture a model dependent on certain number of occasions that we have seen.

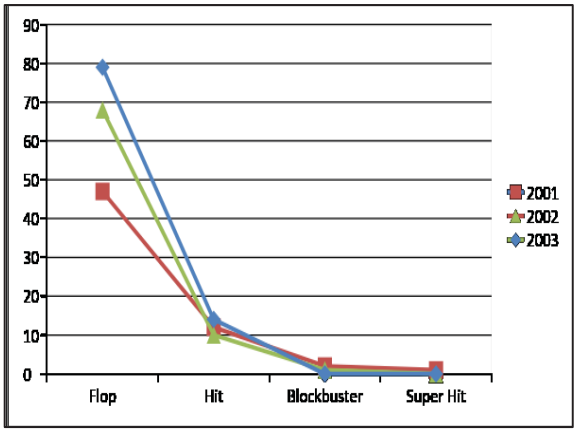
4. IMPLEMENTATION



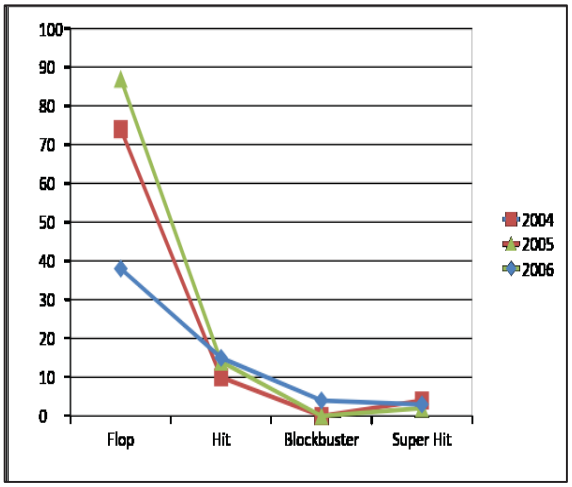
(Bardiagram-1
Representing the total number of flop, blockbusters, hits, superhits that has been released in the period of 2001-2010 based on the category of the movie.)

From this diagram it's been clear that in the 10 years we analyzed drama movies are released more than any genre. Romance, drama and action movies got blockbusters also. Comedy movies have more hits and others are just flops. For more accuracy we analyzed the basis of the year too.
From that it becomes clear that year doesn't matter in the case of movie release and ratings. But from these analyzes it became

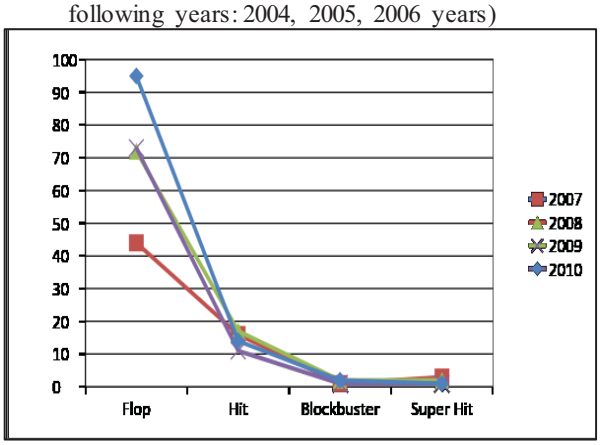
clear that more movies are releasing year by year comparing to the previous year.



(Line diagram 1 representing the movies released in following years: 2001, 2002, 2003 years)



(Line diagram 1 representing the movies released in following years: 2004, 2005, 2006 years)



(Line diagram 1 representing the movies released in following yearsin 2007, 2008, 2009, 2010 years)

We took about 850 movies that were released in year from 2001 to 2010. Most of them were flop movies, According to our research based on movies; we found many things that were new to us. Many movies are releasing in a year in Bollywood itself. Most of them are not accepted by viewers. These viewers are the one who decides which should be accepted, which should be rejected. Most of the movies that were blockbuster was drama movies. Romantic films were there but not as much as drama. Secondly, comedy films are getting appreciation from the viewers. Unfortunately, all the musical films in this range were flop.

We tried many algorithms in this movie data for classifying it accurately. Some of them are:

1) J48: When we classified the data using the algorithm J48, data is getting classified with 85.75 percent accurately. Which means, on the total of 685 flop movies, 637 is identified as Flop itself. Remaining 48 were considered as Hit. On total of 135 hit movies, 84 were identified as hit movies as well as 44 were identified as flop movies and 6 of them were identified as blockbuster and 1 of them were identified as Super hit. Of 13 blockbuster movies, 9 of them were identified as blockbuster itself. 4 of them were identified as Hit movies. Unfortunately in the remaining results, none of them were identified as super hit movies. If considered total number of instances, 730 movies of 850 movies were classified accurately.

   J48 uses iterative classification of algorithm. Initially it classifies the data using the prior attribute. Then the classified data is again classified using the second prior attributes. Thus

a virtual tree is made with number of leaf nodes same as the number of data taken for consideration.

2) HoeffdingTree: When we classified the movie data using the Hoeffding Tree, data is getting classified with 79.29 percent accurately. Which means, on total of 685 Flop movies, 601 were classified correctly and remaining were classified in remaining three categories. On total of 135 movies, 64 were identified as Flop as well as 67 were identified as Hit movies and 1 of them were identified as super hit and 3 of them were identified as Blockbuster. In 13 movies 3 of them were identified as Flop movies as well as 3 movies were identified as Hit and 6 of them were labeled as Blockbuster and 1 of them as Super Hit. On total of 17 Super Hit movies, 2 of them are classified as Flop as well as 14 of them are classified as Hit and remaining 1 as Blockbuster and unfortunately none of them are Super Hit.
3) Random Forest: When we classified the collected movie data using the Random Forest algorithm the data is classified with the accurate value of 80.47 percentages. That shows on total 685 flop movies, 679 movies are identified into flop movies and remaining 6 as Hit movies. On total of 135 hit movies 133 movies are labeled as flop and remaining 2 as Hit movies. On total of 13 Blockbuster movies, 13 of them are classified as Flop movies. Unfortunately none of them are classified as Hit, Blockbuster and Super Hit. Of the 17 super hit movies, 3 of them are identified as super hit itself.13 of them are labeled as Flop and 1 of them as Hit.

5. CONCLUSION

From our research based on movie prediction using data mining, we come up with J48 Algorithm to discover the achievement of films dependent on specific elements. Here we use J48, Random Forest and Hoeffding Tree algorithms and compare these algorithms to know which will give best accuracy. We find that J48 Algorithm has best accuracy. On the basis of this algorithm we foresee the victory or letdown of a movie.

6. FUTURE WORKS

There is wide scope for the future I this project. We had considered Bollywood movies from the period 2001 to 2010. There were only 850 data for us. This could be extended by taking more movies from wider range. Movies can also be taken from Mollywood, Kollywood, Tollywood, Hollywood etc. Through extending historical data, accuracy of the prediction can be increased. And also here we had taken only 6 attributes of the move. Considering more attributes of the movie could result in the more accurate prediction of the rating. Not only like that here we are only taking Decision tree algorithms, such a way naïve base algorithms, function algorithms, artificial intelligence algorithms can also be applied to the wide range of historical data in order to get more accuracy on the prediction. And also other than using Weka tool, more efficient program can be written with many languages. Like that also more accuracy can be achieved in predicting.

7. REFERENCES

1. Ahmad J, Duraisamy P, Yousef A, & Bill Buckles (2017). "*MovieSuccess Prediction Using Data Mining*" 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT).

2.K Meenakshi, G Maragatham, NehaAgarwal and IshithaGhosh, (2018). "*A Data mining Technique for Analyzing and Predicting the Success of Movie*". Department of Information Technology, SRM Institute of Science and Technology,2018.

3. M Saraee, S White & J Eccleston University of Salford, England, (2004). "*A data mining approach to analysis and prediction of movie ratings*" 2004 Fifth International Conference on Data Mining, Text Mining and their Business Applications.

4.Apala K R, Jose M, Motnam S, Chan C C, Liszka K J, & de Gregorio, (2013). "*Prediction of Movies Box Office Performance Using Social Media*" Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining - ASONA'13.

5.GaganjotKaurAmitChhabra (2014). "*Improved J48 Classification Algorithm for the Prediction of Diabetes*" Department of Computer Science and Engineering GNDU, Amritsar (Pb.), India. International Journal of Computer Applications (0975 – 8887) Volume 98 – No.22, July 2014

6.Uzair Bashir, Mewar University, Chittorgarh, Rajasthan, "*Performance evaluation of J48 and Bayes algorithms for intrusion detection system*" India &ManzoorChachoo University of Kashmir, Srinagar, India.

7.Saurabh Kumar, AvinayMetha, Joy Pal. "*Movie Success Prediction using Data Mining*". Data Mining and Business Intelligence(ITA5007) of Master of Computer Application, School Of Information Technology and Engineering April, 2019.

8.AntaraUpadhyay, NiveditaKamath, ShalinShanghavi, TanishaMandvikar, PranaliWagh, Asst. Professor. "*Movie Success Prediction Using Data Mining*". Department of Information Technology, Shah & Anchor Kutchhi Engineering College, Mumbai, India. © IJEDR 2018 | Volume 6, Issue 4 | ISSN: 2321-9939

9.Ms. Kenvi Shah, Mr. JigeshKapadia, Mr. YashSamel, Mr. SarveshSaple, Mrs. PallaviDeshmane. "*Movie Success Prediction using Data Mining and Social Media*". International Research Journal of Engineering and Technology (IRJET),March 2019.

10. Upeksha P. Kudagamage ,Banage T.G.S. Kumara,Chaminda H. Baduraliya."*Data Mining Approach to Analysis and Prediction of Movie Success*".Department of Computing & Information Systems Sabaragamuwa University of Sri Lanka ,Belihuloya, Sri Lanka.2018 International Conference On Business Innovation(ICOBI),25 - 26 August 2018.

11. Md Shamsur Rahim, A Z M Ehtesham Chowdhury, Md. Asiful Islam, Mir Riyanul Islam. *"Mining Trailers Data from YouTube for Predicting Gross Income of Movies"*,2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC),PP 551-554

12. Javaria Ahmad, Prakash Duraisamy, Amr Yousef, Bill Buckles. " *Movie Success Prediction Using Data Mining"*, 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), PP 1- 4.

13.Anantharaman V, Ebin G Job, Neha Sam, Asst.Prof. Sheryl Mariya Sebastian. *"Movie Success Prediction Using data Mining"*, Department of Computer Science and Engineering, Albertian Institute of Science and Technology, Kalamassery, Ernakulam, Kerala. GRD Journals ,CONFERENCE,May 2019.

14.Gaganjot Kaur, Amit Chhabra "*Improved J48 Classification Algorithm for the Prediction of Diabetes"*Department of Computer Science and Engineering GNDU, Amritsar (Pb.), India

15. A. Kuhn, T. Senst, I. Keller, T. Sikora, and H. Theisel*"Comparison of Four Text Classifiers on Movie reviews"*2012 pp. 387-392.

16. Jay Gholap, "*Performance tuning of J48 algorithm for prediction of soil fertility"* Department Of Computer Engineering , College Of Engineering, Pune,
Maharashtra, India