Data Mining Approach to Analysis and Prediction of Movie Success

Upeksha P. Kudagamage
Department of Computing & Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka.
upekshakg@gmail.com

Banage T.G.S. Kumara
Department of Computing & Information Systems
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka.
btgsk2000@gmail.com

Chaminda H. Baduraliya
Department of Physical Sciences & Technology
Sabaragamuwa University of Sri Lanka
Belihuloya, Sri Lanka.
chamihb@gmail.com

Abstract— **Data mining is a very efficient approach to uncover information which will both confirm or disprove common assumptions about movies, and it also allows us to predict the success or failure of a future movie using the known information about the particular movie before its release. The main aim of this study is to analyze data mining approaches to explore the attributes affecting the success or failure of a movie. Each and every data mining algorithm provides separate prediction accuracy details. This study integrates four data mining algorithms (Decision Trees, Naïve Bayes, Support Vector Machine, Neural Networks) and an Ensemble approach in order to address the intriguing problem of the movie success prediction and also demonstrates the correlation between success or failure of a movie and different attributes of movies like Opening weekend Gross, Sequel, Theaters, Budget, Genre, Distributors, Country, IMDB Rating, MPAA Rating, Run Time etc. The prediction performance of these models has been evaluated using Accuracy, Precision, Recall, F-Measure, MCC, ROC Area, PRC Area, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) etc. Further, a spatial clustering technique called the Associated Keyword Space (ASKS) was applied for this study, which is effective for noisy data and projected clustering result from a three-dimensional(3D) sphere to a two dimensional(2D) spherical surface for 2D visualization. Similarities between movies were calculated using the Cosine Similarity and these affinity values were used for this clustering model. Movies were categorized under the success or failure of movies by clustering them into four clusters as Most Successful Movies, Successful Movies, Unsuccessful Movies and Least Successful movies. Experimental results show the most effective attributes towards the success or failure of a movie out of these movie attributes considered in this study. Moviemakers can use these results to identify which movie attributes are the most effective and can consider them for the success of their future movie productions. Also, using the Correlation Coefficient, a mathematical model that can be used to predict the movie's success or failure is proposed in this study.**

Keywords*— *decision tree, naïve Bayes, neural networks, support vector machines, ensemble, spatial clustering

I. INTRODUCTION

The process of movie making is both an industry and an art. Movies are a great source of entertainment and people are crazy about movies. Every year movie industry produces thousands of movies of different genres [1]. There are many online platforms that keep track of movies like Box Office Mojo (BOM), Rotten Tomatoes, and Internet Movie Database (IMDB), which provide information about movies such as actors, directors, budget, as well as user ratings and comments, which provide fair information about movies. There is a great deal of uncertainty that the movie will do business or not [1]. The movie industry is a big business, which can give profits or loss up to several million dollars [2]. Moviemakers are still never sure about whether their movie will do business or not; when they should release the movie and how to advertise it [1] [3]. Most of the time, people are not sure about which particular movie to look for so that their spare time is utilized in entertainment.

Being able to predict into the future is of great importance in decision-making, playing a key role in many areas of science, medicine, finance and industry [4]. If there is a way to predict the percentage of success of movie which is yet to be released and predict a movie being a hit or flop in terms of various parameters such as language, country, budget etc., it is very important for everyone involved in the movie such as cast, directors, producers, investors, other artists and also the audience.

Data mining algorithms are generally based upon the notion of finding trends, patterns or anomalies in a set of data, and could be beneficially used for identifying the most contributed attributes towards the success or failure of a movie and predicting the success or a failure of a movie.

II. LITERATURE REVIEW

For the prediction of movies, different types of researches have been carried out by using different approaches by using news, articles, blogs, and social media etc. But very few researchers have explored through attributes related to a movie.

S. Asur and B. Huberman [5] used data from social media and employing sentiment analysis to predict the future of movies in terms of its business (box office revenue). Another research [6] presents the prediction of the popularity of a movie by the articles on Wikipedia. Another research [7] presents predicting the gross of movies using the news articles as well. Their results suggest that they can achieve better performance by using the combination of IMDB data and news data. Reference [3] used movie data from IMDB, Rotten Tomatoes and Wikipedia and applied machine learning algorithms like linear regression, SVM regression and logistics regression.

Some [2] have studied attributes like MPAA ratings, sequel and genre that affect the box office revenue of movies, but however, they did not get sufficiently accurate results. And also, they failed to consider social media as one of the parameters in their forecasting model. Another research [8] is conducted to predicting the numerical user ratings of a movie using pre-release attributes such as its cast, directors, budget and movie genres. Their aim is to evaluate the prediction performance of random forests in comparison to support vector machines. Both algorithms show great similarities in terms of their prediction performance, making it hard to draw any general conclusions on which algorithm yield the most accurate movie predictions.

Yoo et al. [9] predict the movie revenue from movie data collected from IMDB. This study fails to give sufficiently precise results to be used in practice. There is a research [10] that predicts revenue of movies using reviews on blogs and critical reviews by practical critics but not using social media.

In this paper, this study differs from others as follows. First, there is no reported study that compares the performance between various data mining algorithms to find out what provides the best accuracy for predicting the success or failure of a movie and correlation of the attributes that affects the movie's success/failure. This study seems to be the first modern attempt of its kind in this problem domain that compares the data mining classifiers using the accuracy, correlation, precision, recall etc. This research investigates the application of data mining algorithms such as Decision Trees, Naïve Bayes, SVM, Neural networks, Ensemble etc. to predict the success of movies prior to release and to introduce a simple solution for predicting success of movies in terms of various parameters such as opening weekend gross, distributor, released month, genre, runtime, MPAA rating, budget, theaters (no of screens) and sequel etc. The predictions are based on historical data collected from Box Office Mojo, IMDB, Rotten Tomatoes, and Wikipedia. The predictions are then evaluated and compared in order to find those which provide the best and most accurate performances. Here this study intends to find the issues within these classifiers and introduce a new, more accurate algorithm that overcomes those issues to make the most accurate predictions and a mathematical model for predicting the success/failure of a movie. And also this study hopes to conduct a study on the movies used here, using Euclidean Distance, Similarity and Correlation etc. Movie clustering is used here to discover movies efficiently by providing a visualization of movie data. This is used to compare and evaluate the movie in terms of their similarities, distances and correlations between movies. And also a mathematical model is introduced here for predicting a movie being a success or a failure.

III. Methods

The objective of the project is to analyze different attributes that affect the success or failure of movies from various sources available on the Internet and to predict the success or failure of the movies based on various criteria using different data mining algorithms (classifiers) and to study the similarities and distances between movies. Fig. 1 shows the flow of various processes for analysis.

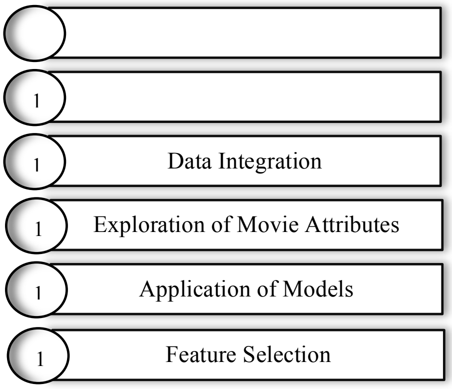


Fig. 1. Process Flow Diagram

A. Data Collection

For this study, there was no public dataset available which had all of the attributes that are deemed to be necessary for predicting movie successes. In order to properly evaluate the relevant data mining algorithms on movie success or failure predictions, a number of relevant datasets have to be considered for the experiment. Therefore, different options for acquiring the data needed were considered. Main sources of data were IMDB, Rotten Tomatoes, Box Office Mojo (BOM) and Wikipedia. About 200 movies that were released in 2017 were considered.

B. Data Preprocessing

The data extracted from the various sources need to be cleaned as the data was obtained from multiple sources. The data was inconsistent, missing and very noisy as well. The Central Tendency was used as a standard for filling missing values for attributes, to cater to missing fields' issue [3]. The collected data were stored in .csv files, with all the attributes and information that are linked to the movies' success or failure for the study.

C. Data Integration

The data extracted from Box Office Mojo, IMDB and other sources need to be integrated and transformed so that it can be used for analysis and classification purposes. Using the .csv files as my source, the selected attributes were transformed into a format to facilitate mining, and produce database queries to select the data to be mined. Here, the data were grouped into a separate .csv file with only the required attributes that have an impact on the success or failure of a movie.

D. *Exploration of Movie Attributes*

Following attributes given in TABLE I and TABLE II, which affect the success or failure of movies, were considered for this study. Here values were categorized into groups for ease of analysis.

TABLE I
SUMMARY OF INDEPENDENT ATTRIBUTES

Name of Attribute	**No. of Values**	**Values**
Opening Weekend Gross	9	Numeric values(1,2,3,4,5,6,7,8,9)
Total Worldwide Gross	9	Numeric values(1,2,3,4,5,6,7,8,9)
Total Profit	9	Numeric values(1,2,3,4,5,6,7,8,9)
Country	21	USA, Russia, India, China, Germany, Thailand, Norway, France, Turkey, Iran, Nigeria, South Korea, Netherland, Australia, Israel, Belgium, Mexico, Brazil, Japan, Vietnam, Indonesia
Genre	20	Action, Adventure, Thriller, Horror, Biography, Crime, Drama, Horror, Comedy, Fantasy, Animation, Mystery, Musical, War, Documentary, Romance, Sci-Fi, Family, Sport, Short
Sequel	2	Sequel, Not Sequel
Budget	9	Numeric values(1,2,3,4,5,6,7,8,9)
Theaters (No of Screens)	3	Numeric values(1,2,3,4,5)
IMDB Rating	4	Terrible, Poor, Average, Excellent
MPAA Rating	5	R, PG, PG-13, G, NR
Release Month	12	Jan., Feb., March, April, May, June, July, Aug., Sept., Oct., Nov., Dec.
Runtime	4	Too Long(TL), Long(L), Short(S), Too Short(TS)
Distributors	44	BV, Sony, WB, WB(NL), Uni., Fox, Par., LGF, TriS, LG/S, STX, Focus, ENTMP, A24, Wein, FoxS, PNT, Aviron, SGem, ORF, BST

TABLE II
SUMMARY OF THE CLASS ATTRIBUTE

Name of Attribute	**No. of Values**	**Values**
Success/Failure	2	Yes(Y), No(N)

E. *Application of Models*

There are many data mining tools available. WEKA (Waikato Environment for Knowledge Analysis) and MATLAB were used for the experimentation. This paper mainly focused on comparing the performance of each data mining algorithm (classifier) to determine whether each movie would be a success or a failure and find out the correlation of the attributes towards the success or failure of movies. Following classifiers shown in TABLE III were selected for the experimentation.

TABLE III
LIST OF CLASSIFIERS

Classifiers
Decision Tree
Naïve Bayes
Support Vector Machine (SVM)
Neural Networks
Ensemble (Voting)
Spatial Clustering (Associated Keyword Space – ASKS)

1) Data Mining Algorithm Comparison: Decision Tree (J48), Naïve Bayes, Support Vector Machine, Neural Networks were evaluated and compared to find out the classifier with the best prediction performance based on various prediction criteria such as Prediction Accuracy, Precision. Recall, F Measure, MCC, ROC Area, PRC Area, RMSE and MAE values etc.

2) Ensemble Approach: Ensemble approach that combined the predictions from multiple models was used to increase the prediction accuracy. Different ensemble techniques were used such as Bagging, Voting, Boosting, and Stacking.

3) Clustering Approach: A spatial clustering technique called the Associated Keyword Space (ASKS) was used to plot similar movies into the same area allowing us to search movies by visualization of the movie data on a spherical surface. The movie similarities were calculated using Cosine Similarity method to get the affinity matrix. Following Equation (1) is used to integrate the features of movies to compute the final similarity value Similarity ($M1_i$, $M2_i$) between movies M1 and M2. Given two vectors of attributes, M1 and M2, the cosine similarity, *cos θ*, is represented using a dot product and magnitude as,

$$Similarity = \cos\theta = \frac{M1 \, . \, M2}{||M1|| \, ||M2||} = \frac{\sum_{i=1}^{n} M1_i M2_i}{\sqrt{\sum_{i=1}^{n} M1^2} \, \sqrt{\sum_{i=1}^{n} M2^2}} \tag{1}$$

F. *Feature Selection*

To do the attribute ranking according to their contribution towards the success or failure of a movie, Feature Selection Process was done using various feature selection techniques such as InfoGainAttributeEval, GainRatioAttributeEval, CorrelationAttributeEval.

IV. RESULTS AND DISCUSSION

This section discusses the results obtained from the analysis. The experiments were conducted on Microsoft Windows 7, Intel Core i3-3110M, 2.40 GHz and 4GB RAM. Data were gathered from Box Office Mojo, IMDB, Rotten Tomatoes, and Wikipedia. The manual classification was performed in order to categorize the movie data set to compare the results. MS Excel was used for data storage and classification. WEKA and MATLAB were used for the analysis.

A. *Result 01— Classifier Evaluation & Comparison*

Following TABLE IV shows how each algorithm performance according to their accuracy, precision, recall, F-Measure, MCC, ROC Area, and PRC Area, RMSE, MAE etc.

TABLE IV
OVERALL ALGORITHM CLASSIFICATION RESULTS

Classifier Measure	Decision Tree	Naive Bayes	SVM	Neural Networks
Accuracy	71.42 %	76.62 %	89.61%	86.66 %
Precision	0.733	0.775	0.898	0.867
Recall	0.714	0.766	0.896	0.867
F Measure	0.716	0.768	0.896	0.867
MCC	0.443	0.535	0.790	0.700
ROC Area	0.724	0.870	0.898	0.760
PRC Area	0.669	0.878	0.858	0.782
RMSE	0.4429	0.388	0.3223	0.3643
MAE	0.3923	0.2679	0.1039	0.1826

The results obtained from each of the classifiers are shown above in TABLE IV. These results show the percentage of time that the experiment was able to correctly predict the instances. Highest accuracy with SVM which is 89.61 % was achieved. Decision Tree, Naïve Bayes, and Neural Networks produced an accuracy of 71.42 %, 76.62 %, and 86.66 % respectively. Ensemble approach that was used to increase the movie prediction accuracy produced an accuracy of 92.85 %.

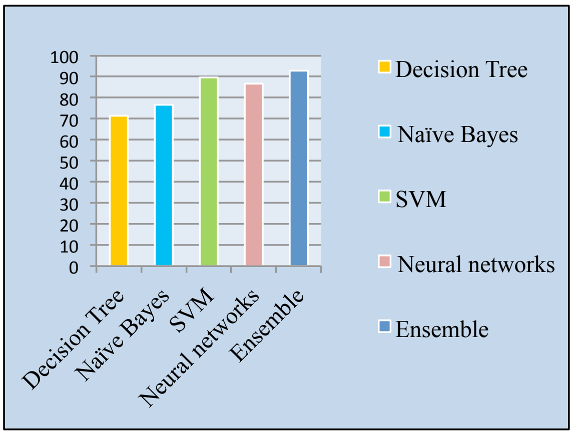


Fig. 2. Overall Comparisons of Data Mining Algorithms

B. Result 02– Clustering Approach

A program that returns similar movie list for a given input movie was implemented. Benchmark of 200 movies released in 2017 was selected as the test data set. Fig. 3 shows the result of spatial clustering approach.

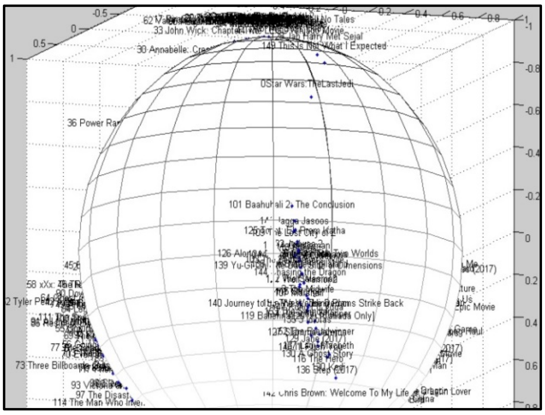
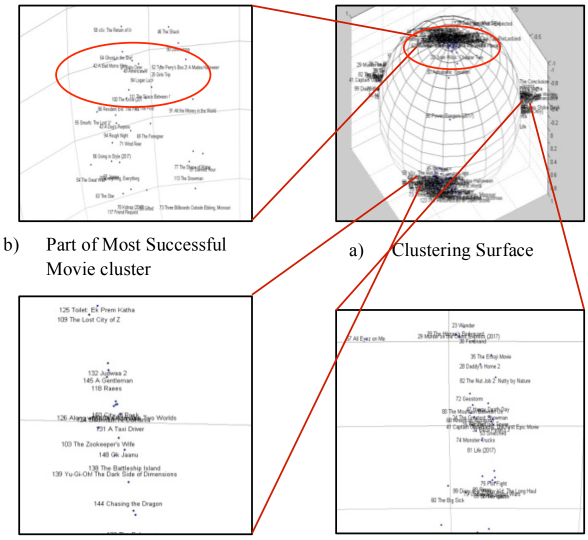


Fig. 3. Result of spatial clustering and visualization

ASKS plotted similar movies into the same area allowing us to search movies by visualization of the movie data on a spherical surface.

On the spherical surface, the movies are distributed according to their similarity. When analyzing the spherical surface, four main regions were identified; Most Successful Movies, Successful Movies, Unsuccessful Movies, Most Unsuccessful movies, where movies are placed and the results show that similar movies in the same domain are placed into one region. Clear separation of regions and density variation of movies within the region that can be considered these regions as movie clusters were observed.

Highlighted areas in Fig. 4 (a) show some similar movies. Fig. 4 (b), Fig. 4 (c) and Figure 4.7 (d) show parts of Most Successful Movies, Successful Movies and Unsuccessful Movies clusters respectively.



b) Part of Most Successful Movie cluster

a) Clustering Surface

c) Part of Successful Movie cluster

d) Part of Unsuccessful Movie cluster

Fig. 4. Result of spatial clustering and visualization

Moreover highlighted area in Fig. 4 (b) shows that more similar movies are placed in the same area within the cluster. For example "Star Wars: The Last Jedi" movie is placed inside the cluster more closely to the “Beauty and the Beast (2017)” movie than “Power Rangers (2017)” movie.

As the next step of the evaluation procedure to evaluate the performance of the affinity calculation method which uses Cosine Similarity, affinity values were calculated using these similarity values. Here also four main regions were observed, but when further analyzing the clustering region, observation shows that some correctly clustered movies in cosine similarity method are incorrectly placed within the clustering surface.

C. *Result 03—Mathematical Approach*

The attributes were given weightages according to the results of the analysis. Attributes were ranked, according to the Feature Selection Process**.** The weightages were calculated and assigned to each attribute according to their contribution towards the movie success prediction. The prioritized values were given to each group of each attribute according to their contribution towards the success or failure of movies.

TABLE V
WEIGHTAGES & PRIORITIZED VALUES GIVEN TO ATTRIBUTES

Attributes	Rounded Weightage	Prioritized Values
Opening Weekend Gross	0.23	0,1,2,3,4,6,7,8
No of Theaters	0.23	0,1,2,3,4
Sequel	0.20	0,1
Country	0.10	0,1,2,3,4,5,6,7,8, 9,10,11,12,13,14, 15,16,17,18,19,20
Distributor	0.56	0,1,2,3,4,5,6,7,8, 9,10,11,12,13,14, 15,16,17,18,19,20
Budget	0.05	0,1,2,3,4,5,6,7,8
Genre	0.04	0,1,2,3,4,5,6,7,8,9,10, 11,12,13,14,15,16
IMDB Rating	0.03	0,1,2,3
Release Month	0.03	0,1,2,3,4,5,6,7,8,9,10,11
MPAA Rating	0.02	0,1,2,3
Run Time	0.01	0,1,2,3

A new mathematical way to calculate the percentage of the success of a movie in terms of selected attributes of movies was introduced here. This is a linear mathematical equation.

$$y = \sum_{i=1}^{11} x_i \times w_i \tag{2}$$

$$y = x_1 \times w_1 + x_2 \times w_2 + x_3 \times w_3 + x_4 \times w_4 + x_5 \times \quad w_5 + x_6 \times w_6 + x_7 \times w_7 + x_8 \times w_8 + x_9 \times w_9 + \quad x_{10} \times w_{10} + x_{11} \times w_{11} \tag{3}$$

where x_i (i = 1,2,3,…) is the prioritized value of each attribute and w_i (i = 1,2,3,…) is the weightage given to each attribute.

If $(\frac{y}{y_{max}} \times 100) > 50\%$, then the movie is predicted as a success and otherwise, it is predicted as a failure.

This mathematical model provides a prediction accuracy of 85 % – 90 %. This model is quite simple but still powerful enough to make good predictions.

V. CONCLUSION

This study using a Data Mining approach aims to introduce a new model for predicting a movie’s success or failure considering a number of attributes related to that movie. Data mining algorithms were compared and evaluated. SVM produces the highest movie prediction accuracy of 89.62 % and as expected the Ensemble approach increases the accuracy up to 92.85 %. Opening Weekend Gross, Sequel and No of Theaters were identified as the most contributed attributes towards movie success prediction. It is clear that a movie’s success is determined by much more than obvious attributes such as genre. Clustering approach provided a 3D visualization of movies, useful in comparison of movies. The proposed mathematical model for predicting the success or failure of movies is powerful enough to make good predictions with a prediction accuracy of 85 % – 90 %.

Movie makers can use these results to identify which movie attributes are the most effective and can consider them for the success of their future movie productions. All the people who are involved in a movie can use this predictor to do the predictions prior to its release.

VI. RECOMMENDATION

This research shows promise for further development in this area. The prediction performance of data mining algorithms and the performance of the clustering approach can be improved by considering new attributes that may affect the movie success or failure. More data mining classifiers can be tested against the results of the used classifiers in this study. Experiments will be carried out to increase the performance of movie discovery and comparison by using visualized movie clusters. Comparison of the movies can be improved using more clustering approaches as well. Prediction performance of the mathematical model can be improved considering the attributes that were not discussed.

ACKNOWLEDGMENT

This paper is based upon the research report submitted in partial fulfillment of the requirement for B.Sc. Special Degree in Computing & Information Systems supported and guided by Dr. B.T.G.S. Kumara and Dr. C.H. Baduraliya.

REFERENCES

[1] D. Im and M. T. Nguyen, "Predicting box-office success of movies in the US market," CS 229, 2011.

[2] J. S. Simonoff and I.R. Sparrow, "Predicting movie grosses : winners and losers, blockbusters and sleepers," Chance, vol. 13(3), pp. 15–24, Sept. 2012.

[3] V.R. Nithin, M. Pranav, P.B.S. Babu, and Lijiya, "A Predicting movie success based on IMDB data," International journal of data mining and techniques, vol. 3, pp. 365-368, June 2014.

[4] T. Hastie, T. Tibshirani, and J. Friedman, The elements of statistical learning: data mining, inference, and prediction, 2nd ed., 2009.

[5] S. Asur and B.A. Huberman, "Predicting the future with social media," March 2010.

[6] N.M. Mestyan, T. Yasseri, and J. Kerte´sz, "Early prediction of movie box office success based on wikipedia activity big data," August 2013.

[7] W. Zhang and S. Skiena, "Improving Movie Gross Prediction Through News Analysis," in WI-IAT '09 Proceedings of the 2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology – Workshops, vol. 1, Washington, 2009, pp. 301-304.

[8] K.Persson, "Predicting movie ratings: a comparative study on random forests," Computer Sciences, University of Skovde, School of Informatics, 2015.

[9] S. Yoo, R. Kanter, and D. Cummings, "Predicting movie revenue from IMDB data".

[10] A.Kennedy, "Predicting box office success: do critical reviews really matter?," 2008.