IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

A Compendium for Prediction of Success of a Movie Based Upon Different Factors

Shraddha Mehta[1], Hitarthi Bhatt[1], Prof. Darshana Desai[1]

Dwarkadas J. Sanghvi College of Engineering, Mumbai, India[1]

Abstract: The success of a movie is uncertain but it is no secret that it is dependent to a large extent upon the level of promotion and also upon the ratings received by the major movie critics. Time and money are valuable to the general audience and hence, they refer to the leading critics when making a decision about whether to watch a particular movie or not. Over 1000 movies on an average are produced per year. Therefore, in order to make the movie profitable, it becomes a matter of concern that the movie succeeds. Due to the low success rate, models and mechanisms to predict reliably the ranking and or box office collections of a movie can risk the business significantly. The current predictive models that are available are based on various factors for assessment of the movie. These include the typical factors such as cast, producer, director etc. or the social factors in form of response of the society on various online platforms. Various stakeholders such as actors, financiers, directors etc. can use these predictions to make more informed decisions.

Keywords: Movie success; Decision tree; Weka; Box office prediction; ID3 Algorithm.

I. INTRODUCTION

The movie industry worldwide produces a large number of movies per year. However, very few movies are a success and are ranked high. Given the low success rate, models to predict reliably the box office collections of a movie can help by improving the business significantly and increase average returns.[2] Looking at this situation from the movie industry's perspective, if there is a link between critical reviews and getting people out to see a movie, this could help with distribution decision making. If a movie does well in test screenings or if they anticipate good reviews from the critics then they can decide to release it on opening weekend in more theatres in hopes of bringing in more revenue [1].

Various stakeholders such as actors, financiers, directors etc. can use these predictions to make more informed decisions. Predictive analytics encompasses a variety of statistical techniques from modelling, machine learning and data mining that analyze current and historical facts to make predictions about future or otherwise unknown events. In business, predictive models exploit patterns founding historical and transactional data to identify risks and opportunities. Models, actual relationships among many factors, allow assessment of risk or potential associated with particular set of conditions guiding decision making for candidate transactions.[6] The current predictive models available are based on various factors for assessment of the movie such as the classical factors such as cast, producer, director etc. or the social factors in form of response of the society on various online platforms.

The paper suggests that the integration of both the classical and the social factors to generate the result and the study of interrelation among the classical factors will lead to more accuracy. To achieve this, collecting the data scattered across internet is necessary and thus data on various platforms such as YouTube, Twitter, and Wikipedia etc. is taken into account along with the classical factors resulting in effective integration.

II. LITERATURE SURVEY

A. ID3 Algorithm

In decision tree learning, ID3 (Iterative Dichotomiser 3) is an algorithm invented by Ross Quinlan used to generate a decision tree from the dataset.[3] To model the classification process, a tree is constructed using the decision tree technique. Once a tree is built, it is applied to each tuple in the database and this results in classification for that tuple.

The following issues are faced by most decision tree algorithms[4]:

- To choose splitting attributes
- Order of splitting attributes
- Number of splits to be taken
- Balance of tree structure and pruning
- The stopping criteria

The decision tree algorithm is based on Entropy, its main idea is to map all examples to different categories based upon different values of the condition attribute set; its core is to determine the best classification attribute from condition attribute sets. The algorithm chooses information gain as attribute selection criteria; usually the attribute that has the highest information gain is selected as the splitting attribute of the current node.[9] Branches can be established based on different values of the attributes and the process above is recursively called on each branch to create other nodes and branches until all the samples in a branch belong to the same category. To select the splitting attributes, the concepts of Entropy and Information Gain and Gain Ratio are used.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Entropy

It is a measure in the information theory, which characterizes the impurity of an arbitrary collection of examples. If the target attribute takes on 'c' different values, then the entropy S relative to this c-wise classification is defined as

$$\text{Entropy}(s) = \sum - \text{Pi} \log 2 \text{Pi}$$

Where Pi is the probability of S belonging to class i. Logarithm is base 2 because entropy is a measure of the expected encoding length measured in bits. For e.g. if training data has 14 instances with 5 positive and 9 negative instances, the entropy is calculated as

$$\text{Entropy}([5+,9-]) = -\frac{5}{14}\log_2\frac{5}{14} - \frac{9}{14}\log_2\frac{9}{14}$$

A key point to note here is that the more uniform the probability distribution, the greater is its entropy. If the entropy of the training set is close to one, it has more distributed data and hence, considered as a good training set.

Information Gain

The decision tree is built in a top-down fashion. ID3 chooses the splitting attribute with the highest gain in information, where gain is defined as difference between how much information is needed after the split. This is calculated by determining the differences between the entropies of the original dataset and the weighted sum of the entropies from each of the subdivided datasets. The motive is to find the feature that best splits the target class into the purest possible children nodes - pure nodes with only one class.[10]

This measure of purity is called information. It represents the expected amount of information that would be needed to specify how a new instance of an attribute should be classified. The formula used for this purpose is:

$$G(D, S) = H(D) - \sum P(Di)H(Di)$$

The attribute with highest value of information gain is used as the splitting node thereby constructing the tree in top down fashion.

B. CART

CART stands for Classification and Regression Trees (Breiman et al., 1984). It is characterized by the fact that it constructs binary trees, namely each internal node has exactly two outgoing edges. The splits are selected using the towing criteria and the obtained tree is pruned by cost–complexity Pruning. When provided, CART can consider misclassification costs in the tree induction. It also enables users to provide prior probability distribution. [11]

C. C4.5

C4.5 is a well-known algorithm used to generate decision trees. The decision trees generated by the C4.5 algorithm can be used for classification, and for this reason, C4.5 is also referred to as a statistical classifier.

At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits data set of samples S into subsets that can be in one class or the other [12]. It is the normalized information gain that results from choosing an attribute for splitting the data. The attribute factor with the highest normalized information gain is considered to make the decision. The C4.5 algorithm then continues on the smaller sub-lists having next highest normalized information gain.

D. CHAID

CHAID is a type of decision tree technique, based upon adjusted significance testing (Bonferroni testing). CHAID can be used for prediction as well as classification, and for detection of interaction between variables. In practice, CHAID is often used in the context of direct marketing to select groups of consumers and predict how their responses to some variables affect other variables, although other early applications were in the field of medical and psychiatric research.

Reasons to choose ID3

- Understandable prediction rules are created from the training data.
- Builds the fastest tree.
- Builds a short tree.
- Only need to test enough attributes until all data is classified.
- Finding leaf nodes enables test data to be pruned, reducing number of tests.
- Whole dataset is searched to create tree.
- Whole hypothesis space is searched.

Weka

Weka (Waikato Environment for Knowledge Analysis) is free software available under the GNU General Public License that implements a large collection of machine learning algorithms and is widely used in data mining applications [8]. Weka supports several standard data mining tasks, more specifically, data pre-processing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as a single flat file or relation, where each data point is described by a fixed number of attributes. [13]

III.METHODOLOGY

The movie industry is high influenced by critics rating, the cast and crew and box office collection. Therefore we take into consideration these attributes in order to predict whether a movie is a hit or a flop. Depending on these parameters, manual classification is done whether the movie is hit or flop. A decision tree is then implemented to determine the probable outcome of a movie being successful.

The following set of data is considered as the base set for the proposed system. The data set comprises of different aspects of 20 movies. The attributes such as rating of the movie, Profit in Crores, Genre, Name of the movie, success, etc. have been taken into consideration.

With these attributes, a decision tree using Weka tool is obtained. The J48 classification algorithm which is an extension of ID3 algorithm is used to generate the decision tree.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 12, December 2015

Name	Genre	Rating	Profit(in Cr)	Success
Ramleela	Drama	6.2	110	Yes
Hasee toh Phasee	Comedy	7	37	No
Humpty Sharma ki Dulhaniya	Romance	6	56.81	No
2 States	Drama	7.1	117	Yes
Vicky Donor	Comedy	7.8	60	No
Raaz 3	Horror	3.9	74	No
Ghanchakkar	Comedy	5.7	8	No
Tanu Weds Manu	Comedy	6.6	41	No
Ek Villain	Action	6.4	117	Yes
Drishyam	Drama	8.9	12	No
Kahani	Thriller	8.2	98	Yes
Rang de Basanti	Drama	8.4	67	Yes
Shandaar	Romance	3.5	33	No
Kill Dill	Action	4.5	1	No
3 Idiots	Comedy	8.5	498	Yes
A Wednesday	Drama	8.4	25	No
Madras Café	Action	7.8	16	Yes
Kai Po Che	Drama	7.7	32	No
Ek Tha Tiger	Action	5.4	245	Yes
Dabang	Action	6.3	170	Yes

Fig. 1 A sample collection of movie database

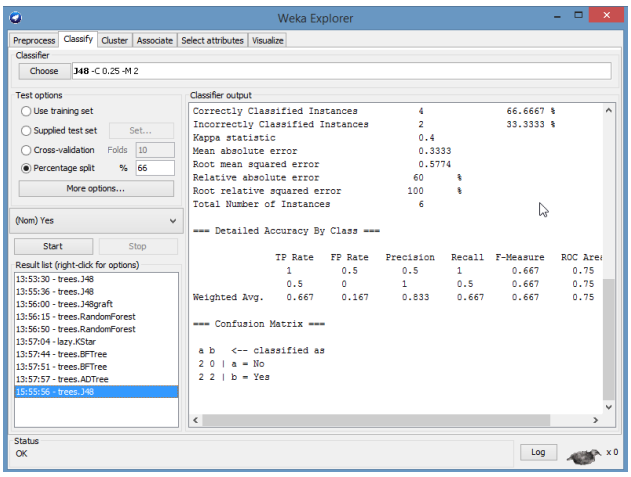

Fig. 2. Weka Explorer Window

The algorithm gives a pruned decision tree with leaves as the decision that is placed or not placed. The primary node consists of IMDb Rating which can range from 0 to 10. If the ratings are Poor, the movie is not a hit. Furthermore, if the ratings are good, the movie may be successful based on the credentials such as budget, genre, profit, etc.. Also, if the movie has average ratings, it may still be a hit based on other attributes like actors, production companies involved, total collections worldwide, etc.[7] The leaf nodes hold the value whether the movie is a hit or not.

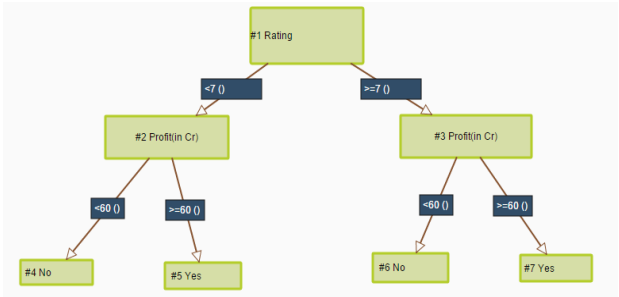

Fig. 3. Decision Tree

The root node chosen here is ratings. Further classification is done by calculating information gain and entropy for each attribute.

Advantages of Decision Tree[5]

1. Variable screening or feature selection is implicitly performed by decision trees. Upon fitting a decision tree to a training set, within the dataset, the most essential variables are the top few nodes based on which the tree is split.

2. For data preparation, decision trees need less effort from users. To overcome scale differences between parameters - for example if there is a dataset which measures revenue in millions and loan age in years, say; this will require some form of normalization or scaling before it can fit a regression model and interpret the coefficients. Such variable transformations are not required with decision trees because the tree structure will remain the same with or without the transformation. Decision trees are also not sensitive to outliers since the splitting happens based on proportion of samples within the split ranges and not on absolute values.

3. Nonlinear relationships between parameters do not affect tree performance. As described here, highly nonlinear relationships between variables will result in failing checks for simple regression models and thus make such models invalid. However, decision trees do not require any assumptions of linearity in the data. Thus, they can be used in scenarios where known parameters are nonlinearly related.

4. The best feature of using trees for analytics is that they are easy to interpret and explain. Decision trees are very intuitive and easy to explain.

IV. CONCLUSION

In this paper ID3 classification algorithm is used to generate decision rule. The generated decision rule can be used to predict the success of a movie. The result of this algorithm can be used by the actors, directors, producers, etc. to identify those set of movie that are likely to underperform at the box office. The classification model can play an important role in increasing the success statistics. It can be concluded that classification algorithms can be used successfully in order to predict the success of a movie. Further the implementation can be done in development and application of novel computational techniques for the analysis of large datasets.

ACKNOWLEDGMENT

We express our sincere gratitude towards our guide **Prof. Darshana Desai** who assisted us throughout our work. We thank her for directing us to the right tract and for the motivation and prompt guidance she has provided whenever we needed it.

REFERENCES

[1] Alec Kennedy; “Predicting box office success: Do critical reviews really matter?”; UC,Berkeley

[2] Dan Cocuzzo, Stephen Wu ; “Hit or Flop: Box Office Prediction for Feature Films”; Stanford University , 2013.

[3] Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao, " Predicting Students' Performance using ID3 and C4.5 classification algorithm", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.3, No.5, September 2013

[4] Dunham, M.H., (2003) Data Mining: Introductory and Advanced Topics, Pearson Education Inc.

[5] Key advantages of using decision trees for predictive analyticsSimafore [Online]

http://www.simafore.com/blog/bid/62333/4-key-advantages-ofusing-decision-trees-for-predictive-analytics

[6] Nikhil Apte, Mats Forssell, Anahita Sidhwa; "Predicting Movie Revenue" ; Stanford University.

[7] Marton Mestyan, Taha Yasseri, Janos Kertesz; "Early Prediction of Movie Box Office Success Based onWikipedia Activity Big Data",2013.

[8] [Online]. https://en.wikipedia.org/wiki/Weka_%28machine_learning%29.

[9] Lyric Doshi; "Using Sentiment and Social Network Analysis to Predict Opening-Movie Box-Office Success."; Massachusetts Institute of Technology, 2010.

[10] Ramesh Sharda, Dursun Delen; "Predicting box office success of motion pictures with neural networks"; Elsevier (2005).

[11] X. Amatriain, A. Jaimes, N. Oliver and J. Pujol, "Data Mining Methods for Recommender Systems," in Recommender Systems Handbook, F. Ricci, Ed., Springer Science Business Media, 2011, pp. 39-71.

[12] Xiaoliang, Z., Jian, W., Hongcan Y., and Shangzhuo, W., (2009) "Research and Application of the improved Algorithm C4.5 on Decision Tree", International Conference on Test and Measurement (ICTM), Vol. 2, pp184-187.

[13] M. Saraee, S. White and J. Eccleston, "A data mining approach to analysis and prediction of movie ratings," in Data Mining V, WIT Press, 2004, pp. 343-352.

DOI 10.17148/IJARCCE.2015.41268