Predicting Box-Office Success Of Movies

Gorang Maniar
Deparment of Computer
Science and Engineering
Email: gorang.maniar@research.iiit.ac.in

Ankush Khandelwal
Deparment of Computer
Science and Engineering
Email: ankush.k@research.iiit.ac.in

Ishaan Arora
Deparment of Computer
Science and Engineering
Email: ishaan.arora@students.iiit.ac.in

***Abstract*—Cinema industry popularly known has reached staggering proportions in terms of volume of business (18400 billion), manpower employment (over 10 million workers), movies produced (more than 100 in a year) and its reach (more than 100 countries worldwide). With so much at stake and highly uncertain nature of returns, it is of commercial interest to develop a model which can predict success of a movie. This however, is not an easy work, since movies have been described as experience goods with very less self life; it is difficult to forecast demand for a movie. There are number of parameters that may influence success of a movie like time of its release, marketing gimmicks, lead actor, lead actress, director, producer, writer, music director being some of the factors. The present study aims to develop a model based upon multiple pattern recognition techniques that may help in predicting the success of a movie in advance thereby reducing certain level of uncertainty.**

Keywords*—Multilayer neural network, Support vector machine, Logistic regression

I. Introduction

Today, the trouble is that the more things change, the more they stay in the same horizons. However, this may not be time for the movie industry, as it can break completely free of the cycles which had marked its history for hundreds of years , and it will be in fact , a departure from reality , Its not predicting the future success of movie is problematical, its the realization that you have to relive the past again and again and still make highly intelligent guess about the success and failure of the movie. Predicting the outcome of events and the success of products is a fundamental problem in pattern classification and predictive analytics. A variety of techniques have been proposed to address real world prediction problems arising in different domains. In this work, we address the problem of predicting movie success based on the box office income i.e. the gross revenue of the movie combined for all theatres showing the movie. Movie success prediction has a lot of use for companies to plan their resources. For example, a Hollywood studio, that expects its newest movie to be highly successful will rent more theatre rooms in advance, increasing revenue if the prediction turns out to be true. If it rent less theatre rooms, not all viewers might have been able to watch the movie in its opening weekend. Before the advent of the internet, critics published their opinions on the quality of a movie, so a movie would be broadly rated on the opinions of a small elite educated audience or polling the audience after the movie. The internet not only changed the way we consume movies but also how movie quality is determined. Thanks to the IMDb (Internet Movie Database) [IMDb, 2014], one of the first online movie databases and the biggest today, it is possible for everyone to rate movies, the positive effect being it is more or less anonymous and statistically more significant, because of the bigger sample size (currently 45 million registered users). None of the studies thus far has succeeded in suggesting a model good enough to be used in the industry.In this study, we attempt to use critics rating in terms of IMDb data, rotten tomatoes, metacritics and features including actors, directors, writers, production house to predict the gross revenue of the movies.

The outline of the paper is as follows. First we consider some related work done in solving problems with similar issues. In section 3, we formalize the task of evaluations described above. In section 4, we describe in detail the learning framework and how it is used to make the predictions. In section 5, we describe the implementation details of the system for this purpose. Finally, in section 7 we evaluate this approach using our training and test dataset.

II. Related Work

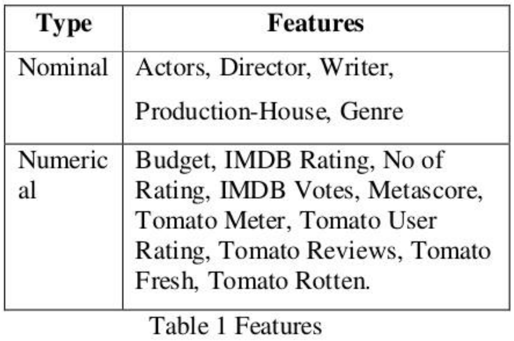
Neural Networks have been extensively used in forecasting and prediction studies, and so it has been also employed for predicting the success and failure the movies also. . Forecasting box office revenue of a movie before its theatrical release is a difficult and challenging problem. In a study conducted by Zhang et al, a multi-layer BP neural network (MLBP) with multi -input and multi output was employed to build the prediction model. All the movies were divided into six categories ranged from blob to bomb according to their box office incomes, and the purpose is to predict a film into the right class. The selections of the input variables were based on market survey and their weight values were determined by using statistical method.

Predicting box-office success of motion pictures with neural networks by Ramesh Sharda and Durus Delen was the paper that we took as a reference. It uses neural networks to predict box-office receipts of new motion pictures. We implemented multilayer neural network, support vector machine and logistic regression to predict the same and compare the three approaches.

III. Dataset

The dataset that we used contains movies released worldwide between year 2000 - 2010. This data was extracted from a wide variety of sources. The website www.omdbapi.com was used to extract the IMDB data of the movies, which includes movie name, year of realese, imdb rating, actors, director(s), production house, writer(s), genre, no. of users, budget, etc. Critics rating from websites like www.rottentomatoes.com, www.metacritic.com were also extracted.

For nominal values every numerical values were assigned by taking into account the past performance of that particular value. Numerical value was calculated from the INCOME/BUDGET ratio of all the movies which that element was a part of in the test data. These numerical values were normalised to 0-1. Following table shows all the features that were extracted:



Type	Features
Nominal	Actors, Director, Writer, Production-House, Genre
Numeric al	Budget, IMDB Rating, No of Rating, IMDB Votes, Metascore, Tomato Meter, Tomato User Rating, Tomato Reviews, Tomato Fresh, Tomato Rotten.

Table 1 Features

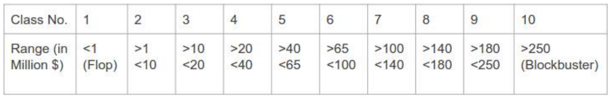
Fig. 1. Feature Table

IV. Method

A. Multilayer Neural Network

Multi layer perceptron (MLP) neural network architecture is known to be a strong function approximator for prediction and classification problems. It has been shown that given the right size and structure, MLP is capable of learning arbitrarily complex nonlinear functions to an arbitrary accuracy level.

The variable of interest in our study is the ratio of box-office gross revenues to the budget of that movie. It does not include auxiliary revenues such as video rentals, international market revenues, toy and soundtrack sales, etc. Another important difference between our study and previous efforts is that the forecasting problem is converted into a classification problem. As mentioned earlier, a movie based on its box-office receipts is classified in one of nine categories, ranging from a flop to a blockbuster. This process of converting a continuous variable in a limited number of classes is commonly called discretization in neural network literature.



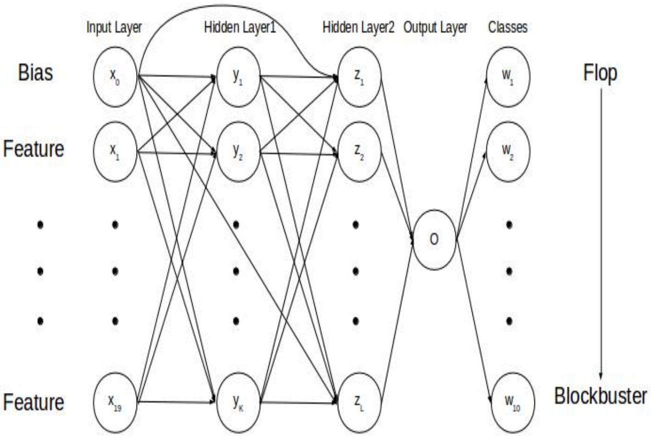
Class No.	1	2	3	4	5	6	7	8	9	10
Range (in Million $)	<1 (Flop)	>1 <10	>10 <20	>20 <40	>40 <65	>65 <100	>100 <140	>140 <180	>180 <250	>250 (Blockbuster)

Fig. 2. Classification

Usually neural networks are composed of multiple neurons. One neuron has multiple inputs which represent the features and one output representing the target value (e.g. class / predicted value). The task of the neuron now is, to find a weight w for each input p, so the output is produced. More precisely this is done by multiplying all weights w with all inputs p and summing up these and the bias, a static term (often 1). The transfer function takes this sum and produces the output. The transfer function is used to obtain a normalized output, normally between -1 and 1, although many different transfer functions exist. In order for the neurons to gain their weights, a neural network needs to be trained, so each neuron can obtain its weights. There are multiple algorithms for training neural networks, the backpropagation algorithm being the most popular. To model an even more complex behaviour, neural networks can also consist of multiple layers, so that the outputs of one layer of neurons are the input for the next, higher level of neurons.

$$bias + \sum_{0<i<n} w(i) * p(i)$$

Best case is obtained when number of hidden layer is 2 and number of neuron in each hidden layer is 10. The graphs for corresponding accuracies are as shown:



Accuracy was calculated by k cross validation where k was varied from 2 to 10.

B. Support Vector Machine (SVM)

The second machine learning technique we applied was SVM. With SVM, we could use all of our input vectors and then pare down the space by use of filter feature selection, which uses the forward search paradigm to choose a subset of features with which to make predictions. As with locally weighted linear regression, we predicted whether a movie performed better or worse than the median found across our entire dataset for each output feature. We implemented three types of kernel function, namely: Linear, radial basis and polynomial functions. We used RBF kernel function to map the data into a high dimensional feature space where linear regression is performed. Hyper parameter C and gamma optimization was done using grid search. Grid search is exhaustive search through a manually specified subset of the hyper parameter space.

C. Logistic Regression

We chose Logistic regression as our third method mainly because it generates a multi- class model with linear weights. Logistic regression measures the relationship between the categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function, which is the cumulative logistic distribution. Thus, it treats the same set of problems as probit regression using similar techniques, with the latter using a cumulative normal distribution curve instead. Equivalently, in the latent variable interpretations of these two methods, the first assumes a standard logistic distribution of errors and the second a standard normal distribution of errors.

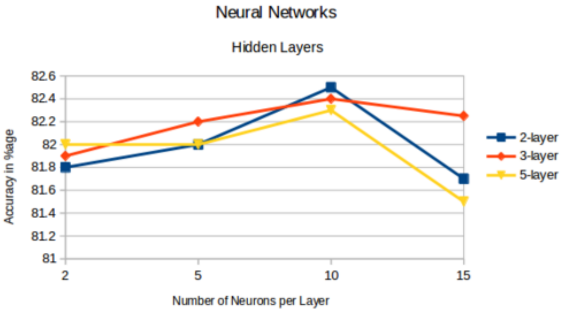
V. Results and Conclusion

Logistic Regression	SVM	Neural Network
61.2%	72%	80%

Logistic Regression Model showed an increase in accuracy as the number of training samples increased. When number of samples were large, the accuracy observed was 61.2percent.

SVM was implemented using 3 different kernel functions, linear, polynomial and radial basis functions . Maximum accuracy was observed in RBF kernel. Average accuracy observed was 72 percent.

Neural Networks exhibited the best accuracy among the techniques used. It showed an accuracy of 82 percent after validating with K-fold Cross Validation.



VI. Future Work

Including Google Search Popularity Index as a feature in our dataset would act as a user based reaction and thus is expected to perform better.

Sentiment analysis of data from social media sites and trending hashtags could improve the performance.

Newly developed hybrid techniques using genetic and other intelligent algorithms can also be used as the results.

VII. References

1. https://www.researchgate.net/publication/222530390

2. Predicting box-office success of motion pictures
with neural networks

3. K. Hornik, M. Stinchcombe H. White,(1990).
Universal approximation of an unknown mapping
and its derivatives using multilayer
feedforward network.

4. http://www.inf.ufrgs.br/ engel/data/
media/file/cmp121/univ$_{a}pprox.pdf$

5. Deniz Demir, Olga Kapralova Hongze Lai,
Predicting IMDB movie ratings using Google Trends

6. http://cs229.stanford.edu/proj2012/
DemirKapralovaLai-PredictingImdbMovie
RatingsUsingGoogleTrends.pdf

7. Sharad Goel, Jake M. Hofman, Sbastien
Lahaie, David M. Pennock and Duncan J.
Watts, Predicting consumer behavior with Web search

8. http://www.pnas.org/content/107/41/17486.full