IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
Volume 6, PP 66-69

Movie Success Prediction

Komal Gothwal[1], Dhiral Sankhe[2,]Nirav Waghela[3], Mitul Sharma[4], Ramanand Yadav[5]

(Department of Information Technology, Atharva College of Engineering, Mumbai, Maharashtra, India)
(Department of Information Technology, Atharva College of Engineering, Mumbai, Maharashtra, India)
(Department of Information Technology, Atharva College of Engineering, Mumbai, Maharashtra, India)
(Department of Information Technology, Atharva College of Engineering, Mumbai, Maharashtra, India)
(Department of Information Technology, Atharva College of Engineering, Mumbai, Maharashtra, India)

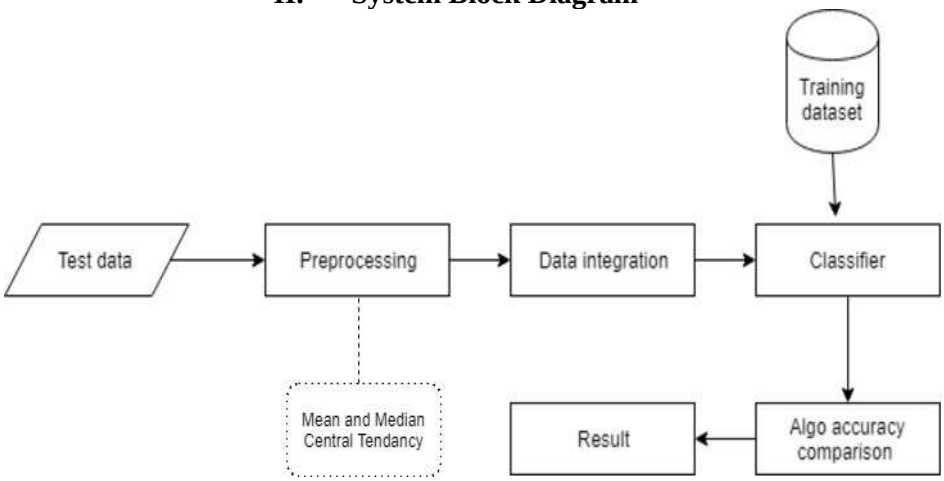
Abstract: *In this system we have developed a mathematical model for predicting the success class such as flop, hit, neutral of the movies. For doing this we have to develop a methodology in which the historical data of each component such as actor, actress, director, music that influences the success or failure of a movie is given is due to weight age and then based on multiple thresholds calculated on the basis of descriptive statistics of dataset of each component it is given class flop, hit, neutral label. Based on the weight age of historical data of each film crew the movie will be labelled as neutral, hit or flop. This system helps to find out whether the movie is super hit, hit, flop on the basis of historical data of actor, actress, music director, writer, director, marketing budget and release date of the new movie. If the movie releases on weekend, new movie will get higher weight age or if the movie releases on week days new movie will get low weight age. The factors such as actor, actress, director, writer, music director and marketing budget historical data of each component are calculated and movie success is predicted. This application helps to find out the review of the new movie.*

Keywords: *Data Mining, SVM, k-NN, Machine Learning, Movies*

I. Introduction

Movies is the most convenient way to entertain yourself. Many movies are produces ny the movie industry in a year. However only few movies get higher success and are ranked high. Given the low success rate, models and mechanisms to predict reliably the ranking and / or box office collections of a movie can help de-risk the business significantly and increase average returns. So if the movie has low success rate it will affect various stakeholders such as actor, financiers, directors etc. so these stakeholders can use these prediction to make more informed decisions. The excellent way to find detailed information about almost every film ever made is through IMDb. Vast amount of data, which contains much valuable information about general trends in films.However, very few movies taste success and are ranked high. Given the low success rate, models and mechanisms to predict reliably the ranking and / or box office collections of a movie can help de-risk the business significantly and increase average returns. Various stakeholders such as actors, financiers, directors etc. can use these predictions to make more informed decisions. The IMDb is an excellent resource to find detailed information about almost any film ever made. It contains a vast amount of data, which undoubtedly contains much valuable information about general trends in films. Data mining techniques enable us to uncover information which will both confirm or disprove common assumptions about movies, and also allow us to predict the success of a future film given select information about the film before its release. The main difficulty in attempting to use data mining to extract useful information from the data source (eg. IMDb, rotten tomatoes, etc) because the format of the source data – it is only available in a number of inconsistently structured text files. The outcome of this research is therefore twofold; it provides tools/techniques to transform the database data into a format suitable for data mining, and provides a selection of information mined from this refined data.

II. System Block Diagram



First the data from the database is extracted into the system. The database source can be any of the freely available dataset online. This is the data gathering step. The pre-processing step starts where the data is cleaned. Each test file will contain best attributes and rebalanced. Data integration is the next step where the data from multiple sources is merged together. The various attributes are unified to single format. After this is data mining classifier is applied, here we are using SVM and KNN algorithm. The results from them is compared and final result is displayed. Thus when user enters the Movie data in the UI the success is predicted.

III. Proposed System

The Proposed methodology consist of following steps.

1. Data Gathering
2. Data Pre-Processing
3. Data Mining
4. Data Interpretation

3.1. DATA GATHERING

The process of data gathering is that involves in collecting all available information about movie. The set of factor should be identified that can affect movie performance and collected from different available data sources. The collected characteristics or risk factors that can influence to movie failure or hit. The initial dataset used will be collected from IMDB. It will consist of movies that were released from 2011 to 2016. Among these movies, we selected the ones that were released in the United States and are in English, we removed movies which don't have any information about Box office details. We got data regarding 200 films. The following data types were removed because they were irrelevant to the task thus contain information: aka-names, aka-titles, cinematographers, complete-cast, complete-crew, costume-designers, editors, german-aka-titles, iso-aka-titles, italian-aka-titles, keywords, laserdisc, miscellaneous, miscellaneous-companies, movie-links, producers, production-designers, sound-mix, and soundtracks, technical. Data which we collected was from IMDb dataset. Initially it was very difficult to analyze the data then using some algorithms data was finally extracted in the required format.

3.2. PRE-PROCESSING

In this stage dataset is prepared for applying data mining technique. Before applying data mining technique, pre-processing methods like cleaning, variable transformation and data partitioning and other techniques for attribute selection must be applied. After pre-processing we have attributes or variables for each movie. Each test file will contain best attributes and rebalanced. . As the data is taken in the raw format from IMDb it is first required to be pre-processed. To overcome missing value scenario central tendency method is used both mean and median and later the duplicate items are removed. Pre-processing is the crucial phase for the project as it mainly focuses on the working of the algorithm. As the data is now pre-processed next comes data integration and transformation in which the alpha numerical data need to be converted to the numerical data as it is required for regression model. The correlation between the features is identified using the greedy backwards method.

3.3. DATA MINING

In this stage Data mining techniques are applied. Data mining refers to the process in which large amounts of data in a database is examined and analyzed in order to generate new information. The data in the database is classified based on which we analyses data patterns in large batched of data using one or more software. Here the data mining techniques are used for classification and regression. The classification is based on best attribute selection from data set. Finally the result with the test file of classification is shown. With the data cleaned, integrated, selected and transformed, the actual data mining could begin. For further analysis we generalized the rating in 3 categories as hit, flop and neutral where each of the category were given values as for flop 0 – 6.5, neutral 6.6 to 7.5 and hit as 7.6 to 10. Our mining is to analyses the two algorithms SVM and kNN with their efficiency and accuracy of the rating which they provide. To reduce the complexity of dealing with the huge load of data we selected few of the attributes only up to 13 so as only the relevant data is used and can be understood by the user.

3.4. INTERPRETATION

In which, the obtained results are analyzed to predict movie hit, flop or neutral. To achieve this, previous test results are taken for comparison. At this stage classification rules are applied for predicting relevant factors and relationships that lead to pass or fail. Classification techniques in data mining such as Support Vector Machine (SVM) and KNN algorithms are used for the prediction of a movie into hit, flop or neutral. Classification divides the data into predefined set of data groups. Through the initial result we can to know about the accuracy of both the algorithms. As far with the limited database of only Hollywood movies the difference between both the algorithms are not so different. Further we are looking for the integration of the Bollywood database which will be more helpful of the end user for predicting the movie.

IV. Algorithms

4.1. SVM

Support Vector Machine (SVM) is part of a group of kernel based methods which are used for pattern classification and regression. A classifier takes an input pattern called feature vector, and determines to which class it belongs to. Let xi, i =1, 2, …, M be feature vectors of a training set X, which belong to either of two classes ω1 and ω2. Using this training data, SVM finds an optimal hyper plane with maximum margin that separates the unknown input patterns into 2 classes as shown in Figure 1. Many hyperplanes separating the feature vectors are possible, SVM finds the one that has maximum margin and better generalization performance for classification.

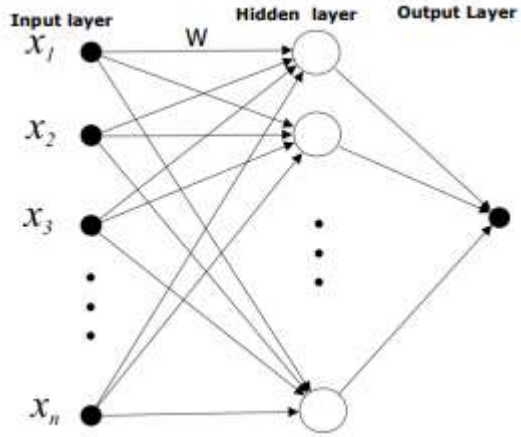


Fig 4.2.1(SVM algorithm)

SVM is basically a linear classifier that classify linearly separable data, but in general, the feature vectors might not be linearly separable. To overcome this issue, kernel trick is used. The original input space is mapped into a high-dimensional feature space using kernel functions where it becomes linearly separable. The performance of an SVM classifier is dependent on the choice of a proper kernel function. Different kernel functions have been employed for different classification tasks. We employ four kernels functions (polynomial, radial basis function, Mahalanobis, and sigmoid) for breast cancer detection and compare their performance.

4.2. k-NN

In pattern recognition, the k-nearest neighbour's algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbour. In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbours.

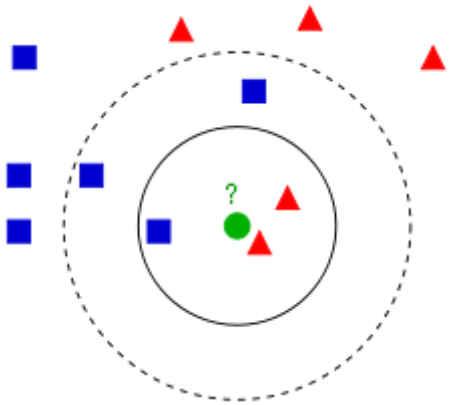


Fig. 4.2.2(kNN Algorithm)

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms. Both for classification and regression, it can be useful to assign weight to the contributions of the neighbors, so that the nearer neighbors contribute more to the average than the more distant ones. For example, a common weighting scheme consists in giving each neighbor a weight of 1/d, where d is the distance to the neighbor. The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required. A peculiarity of the k-NN algorithm is that it is sensitive to the local structure of the data. The algorithm is not to be confused with k-means, another popular machine learning technique.

V. Conclusion

The proposed research aims to predict movies popularity. We have used machine learning approach for our experimentation. Machine learning have powerful classification algorithms like SVM and k-NN for classification and regression. Our research aims to improve previous researches. Performing data mining on IMDB is a hard task because of so many attributes related to a movie and all in different dimensions with lots of noisy data and missing fields. Though the movie can be predicted through the available data from IMDb but if data is not available for certain factors like new hero, or any new genre then prediction cannot be done.

References

Journal Papers:

[1] Krushikanth R. Apala, Merin Jose, Supreme Motnam: "Prediction of Movies Box Office Performance UsingSocial Media", *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*

[2] Muhammad Hassan Latif, Hammad Afzal: "Prediction of Movies popularity Using Machine Learning Techniques", *IJCSNS International Journal of Computer Science and Network Security, VOL.16 No.8, August 2016*

[3] Nithin VR, Pranav M, SarathBabu PB, Lijiya "A Predicting movie success based on IMDB data" *International journal of data mining and techniques , Volume 03, june 2014, pages 365-368*

[4] Darin Im and Minh Thao Nguyen : "PREDICTING BOX-OFFICE SUCCESS OF MOVIES IN THE U.S. MARKET ", *CS 229, Fall 2011*

[5] Andrei Oghina, Mathias Breuss, Manos Tsagkias, and Maarten de Rijke , "Predicting IMDB Movie Ratings Using Social Media", *Advances in Information Retrieval , Volume 7224, 2012, pp 503-507*

[6] Ramesh Sharda , DursunDelen :"Predicting box-office success of motion pictures with neural networks", *Expert Systems with Applications 30 (2006) 243–254*