MOVIE RATING PREDICTOR AND RECOMMENDER SYSTEM

Piyush Gupta Priyanka

Thapar University Thapar University

piyush.knl@hotmail.com
priyankabhanot5@gmail.com

ABSTRACT

Rating predictors and recommender systems in entertainment media are one of the hot topics of research area as accuracy and computational power remains the challenge till present. This system predicts the rating of a movie, based on user information and genres of the movie. The open-source IMDb dataset was mined and user information, movie information and user rating information were integrated to form a single set. Using classification on this set, movie is labelled as hit or flop. The proposed system considers both genres of movies and user relationships on open IMDB dataset for predicting movie ratings and uses association to find correlations between different attributes. There are 22 attributes including movie genres, user's age, occupation and residential location after data preparation, cleaning and analyzing importance of each attribute. Data has been discretized using equal frequency binning. The system applies J48 algorithm for classification with 21 input attributes and 'Rating' as the class attribute. More than 62 percent accuracy is obtained with 480 leaves and 532 tree size. To avoid over fitting, tree has been pruned to get 120 leaves and 132 tree size.

1.0 INTRODUCTION

IMDb is the world's most popular and authentic source for movie and TV content. The IMDb has more than 200 million unique monthly visitors combining its web and mobile audience. IMDb offers a searchable database of more than 180 million data items including more than 3 million movies, TV and entertainment programs and more than 6 million cast and crew members.

The proposed system focusses on building a predictor of movie ratings based on user information like occupation, age, sex and movie genres. The database was reconstructed by integratin user information, occupation, age and sex, and movie information that is, movie id, movie name, genres like comedy, action, romantic, etc. with the ratings provided for each movie by the users. Using classification, the system would be able to assign rating to movies as hit or flop. And using association rule mining, interesting relationships can be found between movie rating and user information or movie genres.

1.1 NEED OF SYSTEM

The need of the system is felt due to primarily two reasons. First is to provide an accurate prediction of rating of a movie of different genres by a user belonging to a particular age group and with a particular occupation. Second, to recommend movies to a new user based on past results and user information.

Such information may be useful for the movie makers to predict which age group and job sector to focus on and target that group by better publicity to reach more people belonging to that group, and increase profits. It also gives an idea of keen interest of a particular age group or occupation. It may be useful to analyze which genres are the most popular, that is, which genre increases the probability of a movie being a hit.

1.2 APPLICATIONS

The system can be applied by recommender websites to recommend movie, TV shows to daily users based on their information and past experience. It can be used for new users also, to recommend movies and other media content based on their age, gender, profession and location.

Producers can also be benefitted from this model as they can analyze which movie genre targets largest number of audience. It can also be used for the promotion strategy for online and offline media to target different groups of users, emphasizing different genres of movie.

1.3 CHALLENGES IN DEVELOPMENT

The first challenge in the project was to choose a good and reliable dataset that was diverse enough to be applied anywhere across the globe. Raw data from IMDb open-source datasets was obtained, spread across multiple Excel sheets, one containing user information, that is, user id, age, sex, occupation and location, second containing a list of different occupations, third containing movie information, that is, movie id, movie name, release date, and different genres, and finally, the one with rating information user id, movie id and rating. The most challenging task was to clean, collaborate and integrate huge dataset with rows greater than 65500 and more than 25 attributes into a single excel sheet and choosing only the relevant attributes.

2.0 EXISTING WORK

Other proposed systems that had conducted various analyses on the IMDb dataset to predict movie rating were reviewed.

Movie Popularity Classification based on Inherent Movie Attributes proposed by Khalid Ibnal Asad Tanvir Ahmed Md. Saiedur Rahman utilized C4.5, PART and Correlation Coefficient to provide a suitable approach for developing pre-release and post release movie datasets using IMDB data.

A Movie Rating Approach and Application Based on Data Mining proposed by S. Kabinsingha, S. Chindasorn, C. Chantrapornchai used data mining to the movie classification. Prototype was built based on the decision tree (J48) using Weka in which the movies were rated into PG, PG-13 and R. 240 prototype movies were used from IMDB. The selected attributes depended mainly on the genres of the movies and the words used in the movies.

The success of these two models proves that IMDB data is consistent. Though both these proposed systems focus on movie data that is, the actors, directors and budget of movies, in this system, focus was put more on user age group, occupation and location, and the genre of the movie. Prediction based on movie data helps in having an idea of how the movie would be rated based on movie characteristics but using user information as the basis, an idea of movie rating can be made based on the age and job of the user. It helps in predicting how a new user with given attributes rate a movie.

3.0 WORKING OF THE PROPOSED SYSTEM

This system aims to predict user rating of movies as either hit or flop and finding correlations amongst attributes using machine learning techniques. Machine learning is the field of study that gives computers the ability to learn without being explicitly programmed. The two types of machine learning are Supervised and Unsupervised learning.

Supervised learning deals with learning a function from available training data which can be used for mapping new examples. Unsupervised learning makes sense of unlabeled data without having any predefined dataset for its training. It involves analyzing available data and looking for patterns and trends.

This system is built on two of the algorithms of machine learning- Classification and Association rule mining. Classification involves generalizing known structure to apply to new data. In this case, possible outcomes are known, and based on algorithm developed from training dataset, possible outcome for a new observation is predicted. The different approaches under classification are J48 (based on decision trees and information gain), Naive Bayes Algorithm (based on conditional probability).

Association rule mining involves searching for relationships between variables. Market Basket Analysis is an important example of association, used to find correlations between different input variables. Some of the algorithms of association rule mining are Apriori (based on Confidence), Predictive Apriori (based on Support and Confidence i.e. Predictive accuracy) FP-Growth (based on Support, Confidence and Lift).

But before using any of these algorithms, a little data pre-processing is required. Data was collected, integrated and cleaned using Microsoft Excel and Microsoft Query as shown in Fig 1 and Fig 2. Then, relevant attributes were identified using InfoGainAttributeEval algorithm and unnecessary attributes were removed (Ref. Fig 1 Preprocessing block and Fig 2 Select Attribute subroutine). In order to convert numerical data to nominal, equal frequency discretization was used as shown in Fig 2. This converted 'Age' attribute which was in numerical format ranging from 7 to 73 into 10 bins with equal frequency. Then, J48 algorithm for training and testing is used to predict user rating as hit or flop. J48 is a tree based algorithm which builds a decision tree based on Information Gain. Apriori algorithm is also used for association rule mining to identify interesting relationships between different attributes. This way, which attributes play major role in deciding whether a movie will be rated as hit or flop by a user are analyzed. The block diagram and flow chart of the system are shown in figure 1 and figure 2 respectively.

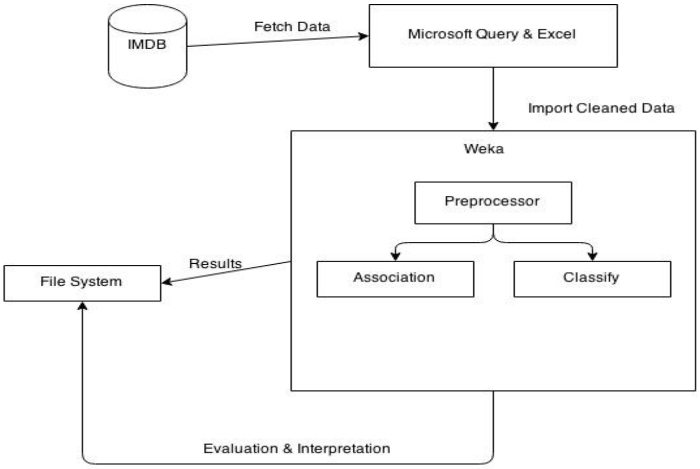


Fig 1 Block Diagram

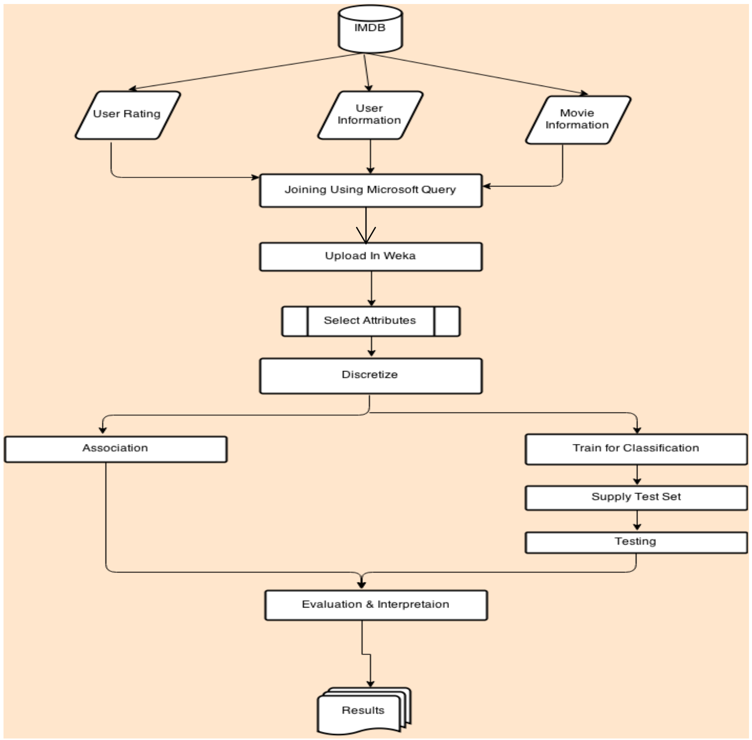


Fig 2 Flow Chart

4.0 DATA COLLECTION AND DATA PREPARATION

Data of IMDB was taken from *http://grouplens.org/datasets/movielens/*. GroupLens Research has collected and made available rating data sets from the MovieLens web site (http://movielens.org) which is indeed a collection of IMDB set. The data sets were collected over various periods of time, depending on the size of the set. MovieLens 100k dataset has been taken. It is a stable benchmark dataset. It has 100,000 ratings from 1000 users on 1700 movies which was Released in 4/1998. It consisted of 3 excel files.

First contained full dataset, 100000 ratings by 943 users on 1682 items. Each user has rated at least 20 movies. The data was randomly ordered. It consisted of 4 attributes, namely user id, item id, rating and timestamp. Second contained information about the items (movies). The attributes were movie id, movie title, release date, video release date, IMDb URL, unknown, Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western. Each of these is described in Table 1. The last 19 fields are the genres. Third contained demographic information about the users. The attributes were namely user id, age, gender, occupation and zip code.

From inspection it is interpreted that userId, movieId, movie title and IMDB url should not determine classification, hence removed. Using Information gain, attributes namely Sex, Release date and timestamp have been removed. Dataset is dicretized to convert age and zip to nominal values using equal frequency binning since, age and zip values are concentrated in a small range.

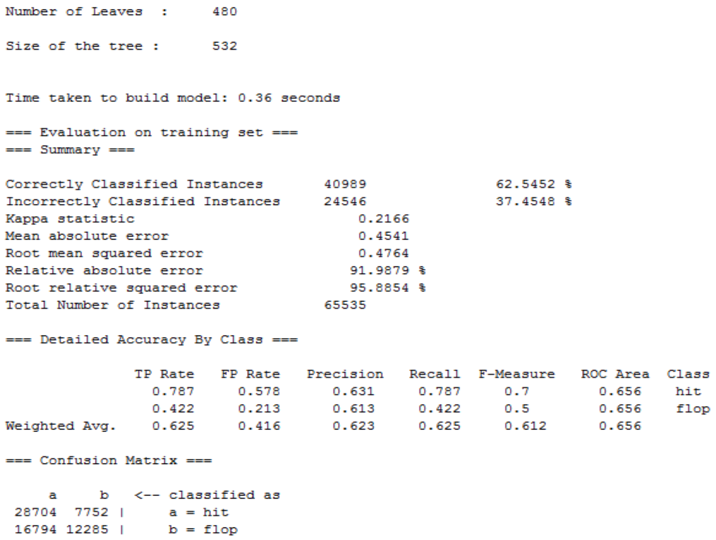
Table 1 Attributes, Types and Their Descriptions

Attribute	Type	Description
Age	Nominal	Age of the user
Occupation	Nominal	Domain-Administrator, artist, doctor, educator, engineer, entertainment, executive, healthcare, homemaker, lawyer, librarian, marketing, none, other, programmer, retired, salesman, scientist, student, technician, writer
Zip	Nominal	

Action, adventure, animation, children, comedy, crime, documentary, drama, fantasy, filmnoir, horror, musical, mystery, romance, scifi, thriller, war, western	Binary	
Rating	Class	Domain- Hit, Flop

5.0 TRAINING OF MODEL

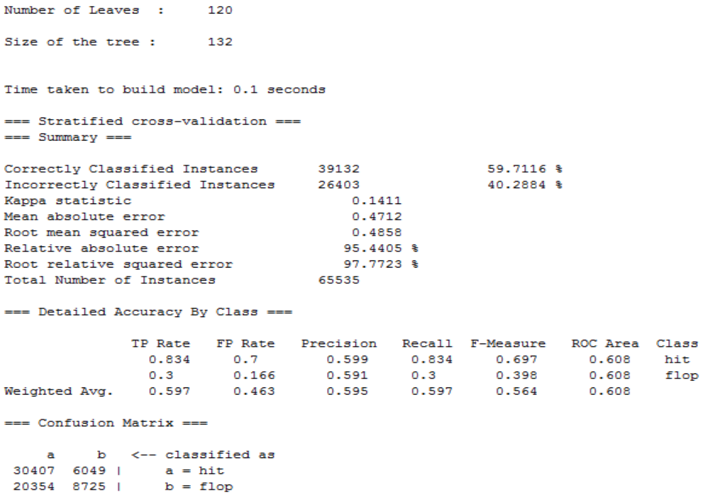
After uploading the clean data in Weka and discretizing it under the Preprocessor tab, J48 algorithm is chosen under Classify tab. Cross validation is chosen with 10 folds as training option. Results are shown in figure 3.



```
Number of Leaves  :     480

Size of the tree :      532


Time taken to build model: 0.36 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances       40989               62.5452 %
Incorrectly Classified Instances     24546               37.4548 %
Kappa statistic                          0.2166
Mean absolute error                      0.4541
Root mean squared error                  0.4764
Relative absolute error                 91.9879 %
Root relative squared error             95.8854 %
Total Number of Instances            65535

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.787     0.578      0.631     0.787     0.7        0.656    hit
                 0.422     0.213      0.613     0.422     0.5        0.656    flop
Weighted Avg.    0.625     0.416      0.623     0.625     0.612      0.656

=== Confusion Matrix ===

     a     b   <-- classified as
 28704  7752 |     a = hit
 16794 12285 |     b = flop
```

Fig 3 Classifier Output with minNumObject=2

The tree obtained from Weka can be pruned by increasing the value of minNumObject in the J48 options to reduce its complexity. A simplified tree is preferred even if accuracy decreases a little in order to avoid over-fitting problem. The results after pruning are as in Fig 4

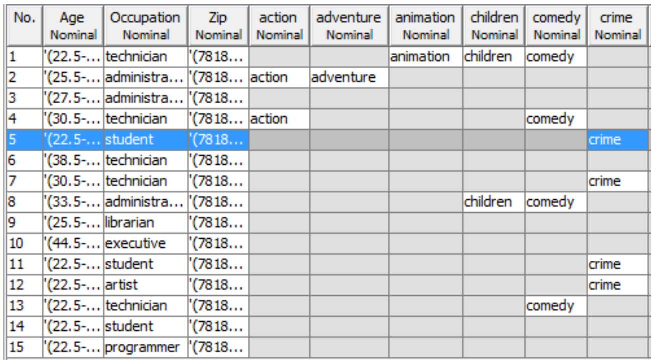


```
Number of Leaves  :     120

Size of the tree :      132


Time taken to build model: 0.1 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       39132               59.7116 %
Incorrectly Classified Instances     26403               40.2884 %
Kappa statistic                          0.1411
Mean absolute error                      0.4712
Root mean squared error                  0.4858
Relative absolute error                 95.4405 %
Root relative squared error             97.7723 %
Total Number of Instances            65535

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.834     0.7        0.599     0.834     0.697      0.608    hit
                 0.3       0.166      0.591     0.3       0.398      0.608    flop
Weighted Avg.    0.597     0.463      0.595     0.597     0.564      0.608

=== Confusion Matrix ===

     a     b   <-- classified as
 30407  6049 |     a = hit
 20354  8725 |     b = flop
```

Fig 4 Classifier output with minNumObject=333

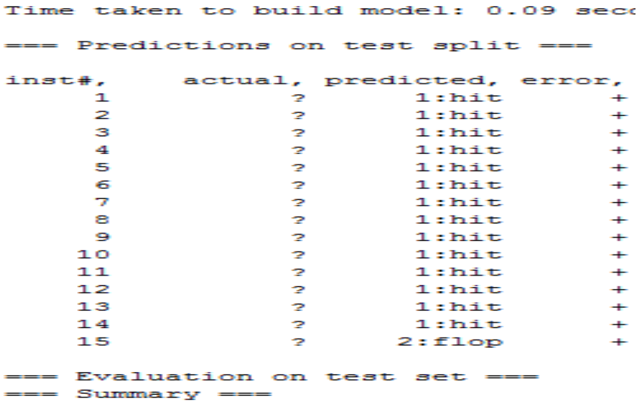
6.0 TESTING OF MODEL

Supplied dataset is created using arff viewer of Weka (Fig. 5). Then, J48 algorithm is applied using the Supplied Dataset as the training option. The results are shown in Fig 6.



No.	Age Nominal	Occupation Nominal	Zip Nominal	action Nominal	adventure Nominal	animation Nominal	children Nominal	comedy Nominal	crime Nominal
1	'(22.5-...	technician	'(7818...			animation	children	comedy	
2	'(25.5-...	administra...	'(7818...	action	adventure				
3	'(27.5-...	administra...	'(7818...						
4	'(30.5-...	technician	'(7818...	action				comedy	
5	'(22.5-...	student	'(7818...						crime
6	'(38.5-...	technician	'(7818...						
7	'(30.5-...	technician	'(7818...						crime
8	'(33.5-...	administra...	'(7818...				children	comedy	
9	'(25.5-...	librarian	'(7818...						
10	'(44.5-...	executive	'(7818...						
11	'(22.5-...	student	'(7818...						crime
12	'(22.5-...	artist	'(7818...						crime
13	'(22.5-...	technician	'(7818...					comedy	
14	'(22.5-...	student	'(7818...						
15	'(22.5-...	programmer	'(7818...						

Fig 5 Supplied Dataset



```
Time taken to build model: 0.09 seco

=== Predictions on test split ===

inst#,     actual, predicted, error,
     1          ?      1:hit       +
     2          ?      1:hit       +
     3          ?      1:hit       +
     4          ?      1:hit       +
     5          ?      1:hit       +
     6          ?      1:hit       +
     7          ?      1:hit       +
     8          ?      1:hit       +
     9          ?      1:hit       +
    10          ?      1:hit       +
    11          ?      1:hit       +
    12          ?      1:hit       +
    13          ?      1:hit       +
    14          ?      1:hit       +
    15          ?     2:flop       +

=== Evaluation on test set ===
=== Summary ===
```

Fig 6 Predicted output

7.0 RESULTS AND DISCUSSIONS

Classification is applied followed by association on the dataset. Each of this is discussed in the following section.

7.1 CLASSIFICATION

J48 algorithm is applied for classification with 21 input attributes and 'Rating' as class attribute. The results as shown in the previous section depict greater than 62 percent accuracy with 480 leaves and 532 tree size. The confusion matrix (Fig.3) reveals that 28,704 instances are correctly classified as hits and 12,285 instances are correctly classified as flops. This implies that around 41,000 instances were correctly identified out of 65,535 instances. To avoid over-fitting, tree is pruned to get 120 leaves and 132 tree size. In this case, though the accuracy gets reduced slightly but the tree obtained is simplified and more general. The confusion matrix (Fig. 4) shows that 30,407 instances have been correctly classified as hits and 8,725 as flops. Even though the overall accuracy gets reduced by one percent, accuracy of classifying hits is improved by 0.5 percent, approximately.

.Table 2 Before and After pruning

METHOD (PRUNING)	**ACCURACY**	**MEAN ABSOLUTE ERROR**	**RMS ERROR**	**RELATIVE ABSOLUTE ERROR**
Before	62.5452	0.4541	0.4764	91.9879%
After	59.7116	0.4712	0.4656	95.4405%

On analyzing the classification model, it is found that the most important attribute for prediction is Occupation, followed by Zip and Age. These attributes mainly decide the verdict of the movie. Genre of the movie does not appear in the decision making process.

7.2 ASSOCIATION

Apriori algorithm is applied for association with 5 percent support and 60 percent confidence. The following association rules are obtained (Table 3).

Table 3 Association rules

Determinant (instances)	**Consequence (instances)**	**Confidence**
Age=’-inf-19.5’(5917)	Occupation=’student’(5428)	0.92
Adventure=’Adventure’(8831)	Action=’Action’(6599)	0.75
War=’War’(6066)	Rating=’Hit’(4071)	0.74
Occupation=’Student’, Drama=’Drama’	Rating=’Hit’(3319)	0.64
Drama=’Drama’, Romance=’Romance’(5189)	Rating=’Hit’(3283)	0.63
Drama=’Drama’(26193)	Rating=’Hit’(16302)	0.62
Scifi=’Scifi’(8163)	Action=’Action’(5076)	0.62
Age=’50.5-inf’(6690)	Rating=’Hit’(4104)	0.61
Occupation=’educator’(6723)	Rating=’Hit’(4109)	0.61

These results show which movie genres are most popular amongst the IMDB users. If the genre of movie is war then the probability of it being rated as hit is 67 percent. Further, it can also be interpreted that users from which age group and occupation background prefer to watch which genre(s). If age group is below 19.5 years, then probable occupation is student, probability being 92 percent. Students are likely to prefer a movie with genre drama with probability 64 percent. Additionally, associations between age groups and occupation is obtained along with associations among different genres. Adventure and action occur together with 75 percent probability.

From association between genre and rating, one can predict if putting a movie belonging to specific genre online will grab more views or not. One can also predict which genres are more popular amongst which age groups and/or occupation background.

8.0 CONCLUSION AND FUTURE SCOPE

Using Weka, it was interpreted that the proposed system for the movie’s rating based on user’s information and for the genre’s popularity based on the movies was fairly accurate. However, the system displayed a slight positive bias and tended to assign movies as hit rather than flop. This can be attributed to the form of the data, in which people tend to be positively biased in giving ratings.

There are often very few low ratings, thus, our predictor, when trained on this data, has a higher tendency to assign movies as hits than the true rating when analyzing a movie with a low true rating. In the course of building the proposed system, it was noted that one main limitation of the dataset was that it had rating information limited to a continent.

This project can be expanded and improvised further to provide better accuracy and promising results by using one metric rating rather than binary rating as hit or flop. Second improvement that can be incorporated is, besides relying only on the IMDB dataset, data can be fetched from social networking websites using sentimental analysis to predict the true rating of the concerned movies. As of now, analyzing is done on the basis of genres and user information but this can further be extended to movie information like Directors, Actors and Producers along with genre.

9.0 REFERENCES

[1] Weka Tutorials - technologyforge.net/ WekaTutorials/

[2] Datasets - http://grouplens.org/datasets /movielens/

[3] Introduction to Data Mining with Case Studies 2nd Edition by GK Gupta

[4]https://sites.google.com/site/parteekbhatia /home

[5]http://arxiv.org/ftp/arxiv/papers/1209/1209 6070.pdf

[6]http://www.ijeit.com/vol%202/Issue%201/ IJEIT1412201207_14.pdf