Movie Genre and Release Year prediction given director's Name using Machine Learning

Assignment report submitted

By

Shoumik Dutta (M.Sc. Computer Science)

NATIONAL INSTITUTE OF TECHNOLOGY, TIRUCHIRAPPALLI

Table of Contents

1. Introduction and Problem Statement

In every year, a numerous number of movies are directed by directors and each movie can be of different genres, having different plots, resulting in a popularity which depends on box office collections, critic reviews, social media posts and likes, etc. In the dataset suggested for the assignment consists of these various details of movies such as the movie title, director name, facebook likes, release year, name of actor and actresses, number of critic reviews, imdb rating, plot keywords,etc.

Based on the dataset, we need to predict how on the basis of the name of director given, the release year with the movie’s probable genres can be predicted using various machine learning techniques.

2. Data Preprocessing

The following steps were conducted for data preprocessing:-

i. At first, the given dataset is read using pandas function pd.read_csv and then we inspect for null values that are present in the dataset. Also using df.describe() we try to gain info about the data distributions for all data in the dataset. Also we count the number of unique distinct type of values each feature takes.

ii. A new feature named total facebook likes is created which is the sum of all the facebook likes attained by the casts, directors, actors and movies and hence instead of using separate columns for each type of facebook like a new column with all the facebook likes are created.

iii. As per the problem statement, the correlation of title year of the movie with all the other features are significant or not needs to be calculated and henceforth a correlation matrix is visualized using all the features of the dataset which consists of the the added new feature combining 4 features. On the basis of this correlation matrix it is observed that all the features are not that much correlated with the feature title year.

iv. Hence a new dataset is created by the name new_df which only consists of the 3 columns namely, director name, title year or the release year of the movie and the movie genres. We remove the rows having null values for movie genres, title year or director names as they are insignificant to our calculations for predicting movie genre or release year of the movie based on director name.

v. We count the number of genres present in the new dataset and count the number times each genre got repeated in the dataset according to the movie, director name or the title year. The genre ‘drama’ has been used most times by most of the directors in the dataset.

vi. In the dataset the genres are written in the format ‘x|y|z’. We convert the genres in the form of [‘x’, ‘y’, ‘z’] . This conversion is important for approaching multilabel classification. For example if a genre is of the format ‘Action|Adventure|Fantasy|Sci-Fi’ then it is converted into ‘[Action, Adventure, Fantasy, Sci-Fi]’.

vii. We use LabelEncoder from sklearn library to encode the director names with numerical values. On the other hand we use multilabel binarizer to encode the genre column to proceed with multilabel classification.

viii. Now we proceed with splitting of dataset into 80 percent training and 20 percent testing dataset. The training dataset consists of 3948 rows and the length of testing dataset is 987.

3. Working Methodology in prediction

Main ideology behind the problem statement is to predict year of release and probable genres of the movie given the director name. So here the director name is ofcourse the input feature and our motive is to predict the genre and the director name.

So the prediction problem is divided into 2 sub parts. In the first part the encoded director names in the training set acts as an input feature and the multi label encoded movie genres as the output feature. The problem in the first subpart is a multi label classification problem. Here 4 machine learning classifiers namely Multinomial Naïve Bayes, K- nearest neighbours with number of neighbours=5, logistic regression and support vector machine (SVM) is used with One-versus-Rest multi label classification criteria. The models are evaluated on the basis of 2 performance metrics F1 score with average set to ‘micro’ and accuracy based on the number of correctly labeled predictions per array.

The evaluation of accuracy can be explained with the following example. Suppose true labels = [[1, 0, 1], [0, 1, 0], [1, 1, 1]] and predicted labels= [[1, 0, 1], [0, 1, 1], [1, 0, 0]], then we calculate accuracy for each sub-array within the array by comparing predicted labels[i] with true labels[i] and computing accuracy score.

Further to inspect how much correctly the genres are predicted we calculate the number of correctly predicted genres per row in the test dataset. For example if the genre is ('Action', 'Crime', 'Drama', 'Thriller') and the predicted genre is ('Action', 'Comedy', 'Thriller') for a particular row then the number of correctly predicted genres are ‘Action’ and ‘Thriller’ which is equal to 2. We compute the total number of such correctly labeled genres over the test dataset.

Now the use of collaborative filtering using SVD or singular value decomposition is used to predict the approximate title year or the year of release of the next movie. In the above approach it uses director name as user id (uid) and genre as the item id (iid) for our prediction purpose.

In order to combine the 2 subparts of the problem together we use iid as the predicted genre for the given director name using the classifier. Hence given the director name, the classifier is used to predict the probable genre and then the singular value decomposition technique is used with the director name as uid and the probable genre as iid to predict the year of release.

This is the procedure adopted to predict the release year of the movie with its probable genres given the name of director as the input feature.

4. EXPERIMENTAL RESULTS

MODEL PERFORMANCES AND PREDICTION BASED ON PROPOSED APPROACH(Table 4.1)

Model_for_Genre_prediction	Model_for_Year_prediction	Model_accuracies_in_multilabel_classification	F1_score	Genres correctly predicted per row for a total of 2276 genres accross test_dataset	Release_year prediction for director name= Mark L. Lester	Genres_prediction for director name= Mark L. Lester
SVM	SVD (RMSE=12.1474)	0.91	0.24	352	1998	(Drama,)
Naïve_Bayes	SVD (RMSE=12.1474)	0.91	0.33	531	1998	(Drama,)
Logistic Regression	SVD (RMSE=12.1474)	0.91	0.33	531	1998	(Drama,)
K nearest neighbours(n_neighbours=5)	SVD (RMSE=12.1474)	0.88	0.32	606	2000	(Action, Comedy, Thriller)

5. CONCLUSION

Based on our study we can reach at following conclusions:-

i. Most of the directors in the dataset aim to include ‘drama’ as a genre in most of the movies.

ii. Considering accuracies of the multi label classifiers used SVM, Naïve Bayes and logistic regression neighbours have high accuracy but considering F1 score SVM does not perform that well as compared to the other 3 models.

iii. Even though evaluating both by proposed accuracy measure and the F1 score with micro averaging technique the Logistic Regression or Naïve Bayes model appears to perform better but considering the number of predicted genre matches per row it is evident that k nearest neighbours with number of neighours =5 performs better with 606 matches across the test dataset.

iv. The low value of RMSE or root mean square error in SVD suggests that the model’s performance is good for year predictions.

v. Even though director’s name can be a useful feature suggesting the release year of movie with its probable genres some other factors may also depend on genres such as plot of the movie, climax sequences, etc. which factors were avoided evaluating focusing on problem statement. This can be suggested as a future work.

THE CODE FOR THE EXECUTION OF THE PROBLEM INCLUDING DATA PREPROCESSING WITH VISUALISATIONS, WORKING METHODOLOGY AND EXPERIMENTAL RESULTS CAN BE FOUND IN THE CODE.

https://colab.research.google.com/drive/1l-lcBEGdr0wUfkUkBLLRi22AhHWx8sRJ?usp=sharing