MOVIE IMDB RATING PREDICTOR - STATUS REPORT

1. Group Member

Zhilin Chen(zcj456)

2. Task

I. Goal

Out task is to predict the IMDb rating of a movie based on the people who are involved in making this movie. (For more information about IMDb, please refer to http://www.imdb.com/ and https://en.wikipedia.org/wiki/IMDb)

II. Input & Output

Given: Genres (at most 3 genres), Director, Writers (at most 3 writers), 4 main Actors, Language and Country of a movie; Returns/Predict: The IMDb rating of a movie.

III. Reasons

Each movie is jointly produced by Director, Writers, Actors and so on. These people (with their professional skills, specific understanding of movies and their audiences or any chemistry among them) and the content (some basic features such as Genres, Language and Country) could mainly decide the quality of a movie (in terms of IMDb rating, which could reflect how audience like this movie). In this task, I try to use statistical method (or Machine Learning techniques) to analyze and predict how these features affect the IMDb rating of a movie.

In general, this task would help: 1) provide businessmen who want to make profitable movies a guideline to choose the staffs. 2) provide us an insight into how the taste of audiences changes among years.

3. Data

I. Number of examples: 13428, collected via OMDb API (http://www.omdbapi.com/) and a self-written python script according to movies list (https://en.wikipedia.org/wiki/Lists_of_films)

II. Features: Title, Year, Genre_1, Genre_2, Genre_3, Director, Writer_1, Writer_2, Writer_3, Actors_1, Actors_2, Actors_3, Actors_4, Metascore, imdbVotes, Language, Country, imdbRating. Actually Title, Year, imdbVotes, Language and Country won't participate in the actual analysis, it could help us filter the data set before analysis.

III. Data Partition: So far, I didn't explicitly split the data set into training, validation and testing. However, I use 10-fold validation to test if a model fit in my task.

4. Preliminary Result

I. Preprocessing: In order to facilitate the analysis, I replaced the values of each genre, directors, writers and actors with a number corresponding to the average or imdbRating of the movies these people have involved in. I also classified the imdbRating into 5 classes: $S(>8.5)$, $A(7.5-8.5)$, $B(6.5-7.5)$, $C(5.5-6.5)$ and $F(<6.5)$. Actually, this preprocessing also improves the accuracy of my predictor.

II. Accuracy:

I used Weka and 10-folds cross-validation to choose a suitable model for my task:

Classifier	ZeroR	J48	IB1	IB5	BayesNet	Naïve Bayes	KStar
Accuracy	38.62%	69.69%	63.53%	65.67%	82.73%	79.77%	73.38%

According to the table above, I would like to use BayesNet methods for my predicator.

5. Future Plans

I. Data Expansion: ~10000 examples are not enough for generating a comprehensive IMDb predictor. As my data set includes only a small part of all the movies, it could make the result deviate from the facts.

II. More Specific Classifier/Program: Though Weka is easy to use and it helps me choose a suitable classification model for my task, it cannot provide me more than that. In order to generate a Learning Curve, how the taste of audiences change among years, who are the top K directors, writers and actors, and other relevant tasks, I would implement a specific python program with scikit (http://scikit-learn.org/stable/user_guide.html), an open source machine learning module for python.