CSP 571 Data Preparation and Analysis
Project Outline & Proposal

Group Member:

Gokul Monday Ramesh	A20401522	gmondayramesh@hawk.iit.edu
Ranganathan Karpagam Arumuga	A20403080	rkarpagamarumuga@hawk.iit.edu
Yudong Wu	A20405374	ywu135@hawk.iit.edu

Problem definition and planning:

Description

Number of movies are released every week. There is a large amount of data related to the movies is available over the internet, because of that much data available, it is an interesting data mining topic. We have used Movie Lens dataset for our experimentation. We created dataset and then transformed it and applied data mining approaches to build efficient models that can predict the movies popularity.

Questions to be addressed

A lot of research has been done on prediction of movies. Most of them include user ratings on different movies, whereas, some of them use social media (e.g. YouTube, Twitter etc.) for prediction. However, less work has done on using movies attributes such as crew, dates etc. to predict movies. The amount of data available about the movies over the internet makes its serious candidate for data mining, knowledge discovery and machine learning. Prediction of a movie is of great importance to industry; movie makers are still never sure about whether their movie will do business or not; when they should release the movie and how to advertise it.

In this kind of model and classification, it will be interesting to study some interesting insights in movie business such as what were the best movies of every decade, what were the best years for a genre, how many movies were produced every year or what cast is the ultimate movie cast.

Proposed methodology/approach

We will be using Jaccard similarity, cosine similarity and Pearson similarity. Also, k-cluster, Content based algorithms, User-based collaborative Filtering approach to predict the popularity. We will be using classifier to make recommendation and other functions in R studio lib to provide a data visualization.

Metric for measure analysis result:

The Jaccard similarity between recommend result & actual result.

Project Outline:

Data source and Reference Data

Dataset

MovieLens 1M Dataset

URL: https://grouplens.org/datasets/movielens/1m/

README file

These files contain 1,000,209 anonymous ratings of approximately 3,900 movies made by 6,040 MovieLens users who joined MovieLens in 2000.

URL: http://files.grouplens.org/datasets/movielens/ml-1m-README.txt

Data Overview

There are three files, altogether forming movie lens dataset. They are Users, Movies and Ratings

Rating File Description

- UserIDs range between 1 and 6040
- MovieIDs range between 1 and 3952
- Ratings are made on a 5-star scale (whole-star ratings only)
- Timestamp is represented in seconds since the epoch as returned by time(2)
- Each user has at least 20 ratings

User File Description

- Gender is denoted by a "M" for male and "F" for female
- Age is chosen from the following ranges:
  - 1: "Under 18"
  - 18: "18-24"
  - 25: "25-34"
  - 35: "35-44"
  - 45: "45-49"
  - 50: "50-55"
  - 56: "56+"
- Occupation is chosen from the following choices:
  - 0: "other" or not specified
  - 1: "academic/educator"
  - 2: "artist"
  - 3: "clerical/admin"
  - 4: "college/grad student"
  - 5: "customer service"
  - 6: "doctor/health care"
  - 7: "executive/managerial"
  - 8: "farmer"
  - 9: "homemaker"
  - 10: "K-12 student"
  - 11: "lawyer"
  - 12: "programmer"
  - 13: "retired"
  - 14: "sales/marketing"
  - 15: "scientist"
  - 16: "self-employed"
  - 17: "technician/engineer"
  - 18: "tradesman/craftsman"
  - 19: "unemployed"
  - 20: "writer"

Movie File Description

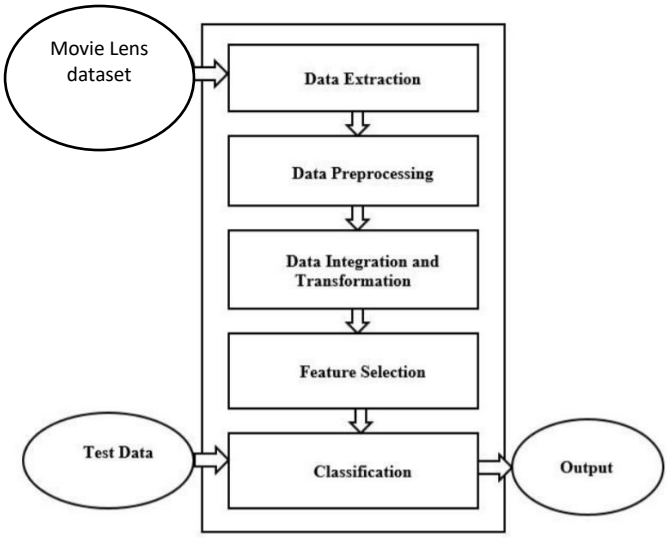
- Titles are identical to titles provided by the IMDB (including year of release)
- Genres are pipe-separated and are selected from the following genres:
  - Action

Adventure
Animation
Children's
Comedy
Crime
Documentary
Drama
Fantasy
Film-Noir
Horror
Musical
Mystery
Romance
Sci-Fi
Thriller
War
Western

- Some MovieIDs do not correspond to a movie due to accidental duplicate entries and/or test entries
- Movies are mostly entered by hand, so errors and inconsistencies may exist

Data processing

The data we extracted from IMDB need to be cleaned as the data is obtained from multiple sources mainly Movie lens and Wikipedia. The data is inconsistent, missing and very noisy as well. To cater missing fields issue we have used mean values as a standard for filling missing values for attributes. The data extracted from Movie lens need to be integrated and transformed so that it can be used for analysis and classification purposes.



Model selection

The purpose of this project is to develop a multiple linear regression model to understand what attributes make a movie popular. Our goal is to analyze the movie's popularity by measuring the audience score, related to the type of the movie, genre, runtime, user rating, critics score etc. Software packages

Software packages, libraries used

Software: R studio

Cran packages: ggplot2, tidyverse, ggmap

References:

[1] Darin Im and Minh Thao Nguyen : "PREDICTING BOXOFFICE SUCCESS OF MOVIES IN THE U.S. MARKET ", CS 229, Fall

[2] 2011https://en.wikipedia.org/wiki/Film , Accessed on August 1st, 2015

[3] https://en.wikipedia.org/wiki/Internet_Movie_Database , Accessed on August 1st, 2015

[4] 'Prediction of Movies popularity Using Machine Learning Techniques', Muhammad Hassan Latif†, Hammad Afzal. National University of Sciences and technology, H-12,ISB.