Data Mining: Fundamentals

Group 12

Bruno Barbieri, Noemi Dalmasso, Gaia Federica Francesca Ferrara

Contents

Introduction

The aim of this report is to display an analysis carried out on the IMDb dataset, which contains data about movies, TV shows, and other forms of visual entertainment, along with their ratings generated by the internet community. The analysis has been conducted making use of data mining methodologies. After the data understanding and preparation phase, clustering, classification, and pattern mining techniques have been applied.

1. Data Understanding and Preparation

1.1 Data Semantics

The dataset *train.csv* contains 16431 titles of different forms of visual entertainment that have been rated on IMDb, an online database of information related to films, television series etc. Each record is described by 23 attributes, both numerical and non-numerical. All the variables of the dataset are introduced and explained in Table 1.1 and Table 1.2.

Attribute	**Type**	**Description**
originalTitle	Nominal	Title in its original language
rating	Ordinal	IMDB title rating class The range is from (0,1] to (9,10]
titleType	Nominal	The format of the title
canHaveEpisodes	Nominal (Binary)	Whether or not the title can have episodes True: can have episodes; False: cannot have episodes
isRatable	Nominal (Binary)	Whether or not the title can be rated by users True: it can be rated; False: cannot be rated
isAdult	Nominal (Binary)	Whether or not the title is for adults 0: non-adult title; 1: adult title
countryOfOrigin	Nominal	The country(ies) where the title was produced
genres	Nominal	The genre(s) associated with the title

Table 1.1: Description of non-numerical attributes

1.2 Distribution of the variables and statistics

1.2.1 Discrete attributes

In this paragraph, we examine the most informative discrete attributes of the dataset to provide an overview of their statistics and frequencies.
From Fig.1.1(a), we observe that the classes of the *titleType* attribute are unbalanced, with *movie* being the most frequent class (5535 records) and *tvShort* the least frequent (40 records).

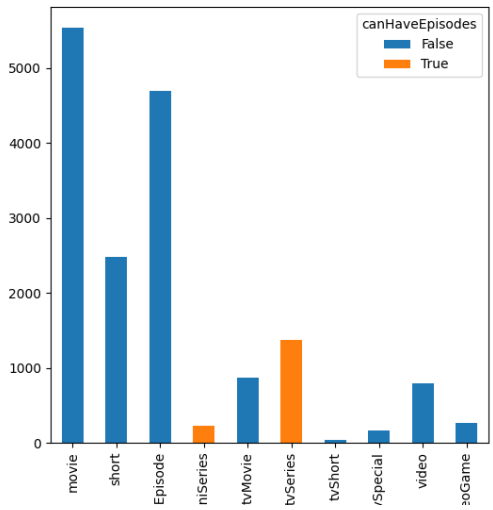
Attribute	Type	Description
runtimeMinutes	Numeric	Runtime of the title expressed in minutes
startYear	Interval	Release/start year of a title
endYear	Interval	TV Series end year
awardWins	Numeric	Number of awards the title won
numVotes	Numeric	Number of votes the title has received
worstRating	Numeric	Worst title rating
bestRating	Numeric	Best title rating
totalImages	Numeric	Number of Images on the IMDb title page
totalVideos	Numeric	Number of Videos on the IMDb title page
totalCredits	Numeric	Number of Credits for the title
criticReviewsTotal	Numeric	Total Number of Critic Reviews
awardNominationsExcludeWins	Numeric	Number of award nominations excluding wins
numRegions	Numeric	The regions number for this version of the title
userReviewsTotal	Numeric	Number of User Reviews
ratingCount	Numeric	The total number of user ratings for the title

Table 1.2: Description of numerical attributes

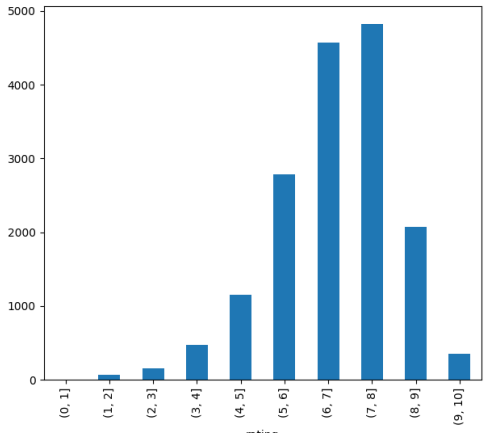
By analyzing the *canHaveEpisodes* attribute within these *titleType* values, we found out that only *tvSeries* and *tvMiniSeries* can have episodes, as expected.
As shown in Fig.1.1(b), the frequency of rating classes is skewed toward higher values. The most frequent rating class is (7, 8], which is the rating of 4822 titles.
Another important aspect is that all 16341 titles are ratable and the vast majority of them (16005) are non-adults contents, as shown in Fig.1.1(c)

1.2.2 Continuous attributes

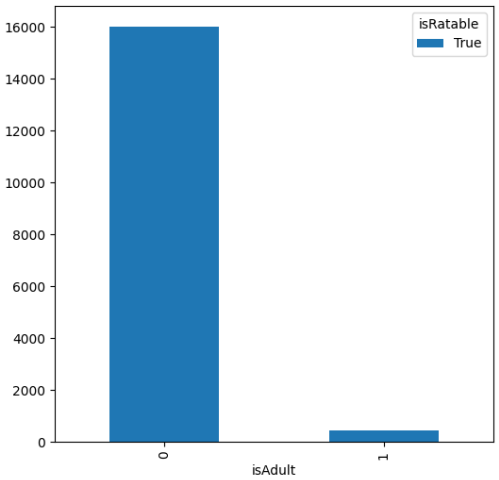
...content... ...content... ...content... ...content... ...content... ...content... ...content... ...content... ...content... ...content...



(a) (a)



(b) (b)



(c) (c)

Figure 1.1: Bar chart of the discrete attributes: (a): counting of the title types frequencies combined with the canHaveEpisodes variable (b): counting of ratings frequencies (c): counting of the adult and non-adult frequencies combined with the isRatable attribute.