Data Mining: Foundations

Group 12

Barbieri, Dalmasso, Ferrara

Indice

Elenco delle figure

Introduction

1. Data Understanding and Preparation

1.1 Data Introduction

The dataset *complete_df.csv*, which is the merge of the original *train.csv* and *test.csv* datasets, contains 21909 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 Range: from (0,1] to (9,10], converted into an integer interval: [1, 10]
titleType	Nominal	The type of media product
canHaveEpisodes	Binary	Whether or not the title can have episodes True: can have episodes; False: cannot have episodes
isRatable	Binary	Whether or not the title can be rated by users True: it can be rated; False: cannot be rated The training set, as well as the test set, only contain True values; hence, this attribute will be discarded
isAdult	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

Tabella 1.1: Description of non-numerical attributes

Attribute	Type	Description
runtimeMinutes	Integer	Runtime of the title expressed in minutes
startYear	Year	Release/start year of a title
endYear	Year	TV Series' end year
awardWins	Integer	Number of awards the title won
numVotes	Integer	Number of votes the title has received
worstRating	Integer	Worst title rating Range: [1, 10] Always equal to 1, so the column was discarded
bestRating	Integer	Best title rating Range: [1, 10] Always equal to 10, so the column was discarded
totalImages	Integer	Number of Images on the IMDb title page
totalVideos	Integer	Number of Videos on the IMDb title page
totalCredits	Integer	Number of Credits for the title
criticReviewsTotal	Integer	Total Number of Critic Reviews
awardNominationsExcludeWins	Integer	Number of award nominations excluding wins
numRegions	Integer	The number of regions where this version of the title is available
userReviewsTotal	Integer	Number of User Reviews
ratingCount	Integer	The total number of user ratings for the title

Tabella 1.2: Description of numerical attributes

1.2 Distribution of the variables and statistics

1.2.1 Discrete attributes

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

1.2.2 Continuous attributes

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

2. Clustering

2.1 Centroid-based methods

2.2 Density-based methods

2.3 Analysis by hierarchical clustering

2.4 General considerations

3. Classification

Key findings and conclusions

Bibliografia