PROJECT REPORT

ASSOCIATION RULE MINING

SUBMITTED TO :
POONAM CHAUDHARY

SUBMITTED BY :
SHASHWAT GUPTA

DATE:
20 SEPTEMBER 2020

ROLL N0.
18CSU204

OBJECTIVE

Association Rule Mining, as the name suggests, association rules are simple If/Then statements that help discover relationships between seemingly independent relational databases or other data repositories.
Most machine learning algorithms work with numeric datasets and hence tend to be mathematical. However, association rule mining is suitable for non-numeric, categorical data and requires just a little bit more than simple counting.
Association rule mining is a procedure which aims to observe frequently occurring patterns, correlations, or associations from datasets found in various kinds of databases such as relational databases, transactional databases, and other forms of repositories.

Some of the areas where Association Rule Mining has helped a lot are:
1.Market Basket Analysis (MBA, often used synonymous to Association Rule Mining)
2.Medical Diagnosis
3.Census Data
4.Protein Sequence

DATA-SET DESCRIPTION

During lockdown many of us spend our time Binge-watching various movies and series on various OTT platforms like Netflix, Amazon-Prime, Hotstar etc. Often before getting started with a new movie or series we check for their reviews and ratings. IMDb is one such platform. IMDb is an online database, owned by Amazon, of information related to films, television programs, home videos, video games, and streaming content online – including cast, production crew and personal biographies, plot summaries, trivia, ratings, and fan and critical reviews.

Hence, I chose one such dataset. The data is a set of 1,000 most popular movies on IMDb for the years 2006-2016.The data fields included are:

1.Title: Contains title of the movie

2.Genre: Contains a comma-separated list of genres of genres used to classify the movie

3.Description: Brief one-line movie summary

4.Actors: The name of actors starring in the movie

5.Year: The year which the movie was released

6.Runtime: Duration of the movie in minutes

7.Rating: User rating of the movie ranging from 1-10

8.Votes: Number of votes received by the movie

9.Revenue: The revenue generated by the movie in millions

10.Director: The name of the movie director

Why Choose "Genres" for applying Apriori:

Since the Apriori algorithm is use for finding the frequently occurring itemset in the given dataset. In this dataset I chose to apply the Apriori algorithm on the "Genre" of the movies. I believe it was best suited for applying the apriori algorithm as it was the only attribute in the dataset with least unique values.The data set contains, having a thousand movies, just 20 distinct Genres (Action ,Adventure, Animation, Biography, Comedy, Crime, Drama, Family, Fantasy, History, Horror, Music, Musical, Mystery, Romance, Sci-Fi, Sport, Thriller, War, Western)

Data Preprocessing

Luckily this data set was pre-processed not much of pre-processing was required.

However, it was required for the application of apriori algorithm, that the data-set attribute on which the algorithm is to be applied be in a binary form, i.e. in the form of 0 and 1. For this I have used "str.get_dummies()" function from the Pandas library of Python. This function is used to separate each string in caller series at the passed separator (comma, semi-colon, strip (|)). A data frame is returned with all the possible values after splitting every string. If the text value in original data frame at same index contains string (Column name/ Splited values) then the value at that position is 1 other wise, 0.

RULE MINING PROCESSES

The Association Rule has two parts

A. **an antecedent** (if)

B. **a consequent** (then)

An antecedent is something that's found in a data, and a consequent is an item that is found in combination with antecedent.

Association rules are created by thoroughly analyzing data and looking for frequent if/then patterns. Then, depending on the following two parameters, the important relationships are observed:

1.**Support**: Support indicates how frequently the if/then relationship appears in the database

The support of X with respect to T is defined as the proportion of transactions t in the dataset which contains the itemset X.

2.**Confidence**: Confidence tells about the number of times these relationships have been found to be true

Confidence is defined as: **conf(X ---> Y)=supp(X U Y)/supp (X)**

Other metric used:

Lift : how likely item Y is purchased when item X is purchased, while controlling for how popular item Y is. The lift of a rule is defined as:

lift (X ---> Y) = sup (X U Y) / sup (X) x sup (Y)

The ratio of the observed support to that expected if X and Y were independent.

RESULTING RULES

The final rules to be shared with the client are:

	antecedents	consequents	lift
0	(Adventure)	(Action)	1.975101
1	(Action)	(Adventure)	1.975101
2	(Comedy)	(Action)	0.532311
3	(Action)	(Comedy)	0.532311
4	(Crime)	(Action)	1.188119
5	(Action)	(Crime)	1.188119
6	(Action)	(Drama)	0.463204
7	(Drama)	(Action)	0.463204
8	(Fantasy)	(Action)	1.307061
9	(Action)	(Fantasy)	1.307061

	antecedents	consequents	lift
88	(Mystery, Drama)	(Crime)	2.948718
89	(Crime)	(Mystery, Drama)	2.948718
90	(Mystery)	(Crime, Drama)	2.236919
91	(Drama)	(Crime, Mystery)	1.546011
92	(Thriller, Drama)	(Crime)	2.000000
93	(Thriller, Crime)	(Drama)	1.063264
94	(Crime, Drama)	(Thriller)	1.268834
95	(Thriller)	(Crime, Drama)	1.268834
96	(Drama)	(Thriller, Crime)	1.063264
97	(Crime)	(Thriller, Drama)	2.000000

97 Rules.

lift (X ---> Y) = sup (X U Y) / sup (X) x sup (Y)

The ratio of the observed support to that expected if X and Y were independent.

As we know the greater is the value of lift the more is likeliness of choosing that option. However, if the value of lift is negative there are no chances of those items to be together. In this dataset the lift values varies from 0.5 to 9.3.

RECOMMENDATIONS:

We apply a APRIORI ALGORITHM to a top rated movies dataset. The technique does not provide a recommendation in a fine-grained user level, but it does enable us to investigate an underlying relationship within the movies. We can utilize such findings to construct a new marketing campaign, research customer's behavior, or make a product suggestion.

After applying various metrics it was seen the that the most successful genre was a combination of Crime, Thriller and Drama.

As a movie maker, one could try to explore other than these genres, which are not explored much. The makers can either earn by making more movies of such genres that are preferred by the masses or they could try and bring innovative ideas ,story, screenplay and direction to add the currently not so liked genres to the liking list of the audiences.