商务智能第四次作业 关联分析apriori实战

数据集来源：https://www.kaggle.com/datasets/rounakba movies-dataset?select=movies_metadata.csv

2108080217 余睿

代码

In [ ]:
```
import pandas as pd
import json
import gc
from mlxtend.frequent_patterns import apriori
from mlxtend.frequent_patterns import association_rules
```

In [ ]:
```
pd.options.display.max_columns=100
```

1.读取数据

In [ ]:
```
# 读入元数据
movies_metadata = pd.read_csv("../data/movies_metadata.csv")
```

```
d:\OTHER\software\Anaconda3\envs\doog\lib\site-packages\IPython\core\interactives
hell.py:3258: DtypeWarning: Columns (10) have mixed types.Specify dtype option on
import or set low_memory=False.
  interactivity=interactivity, compiler=compiler, result=result)
```

In [ ]:
```
# 只要 id 标题 题材（原始数据）
movies = movies_metadata[{'id', 'title', 'genres'}]

# 回收metadata
del movies_metadata
gc.collect()

movies
```

Out[ ]:

	title	id	genres
0	Toy Story	862	[{'id': 16, 'name': 'Animation'}, {'id': 35, '...
1	Jumanji	8844	[{'id': 12, 'name': 'Adventure'}, {'id': 14, '...
2	Grumpier Old Men	15602	[{'id': 10749, 'name': 'Romance'}, {'id': 35, ...
3	Waiting to Exhale	31357	[{'id': 35, 'name': 'Comedy'}, {'id': 18, 'nam...
4	Father of the Bride Part II	11862	[{'id': 35, 'name': 'Comedy'}]
...	...	...	...
45461	Subdue	439050	[{'id': 18, 'name': 'Drama'}, {'id': 10751, 'n...
45462	Century of Birthing	111109	[{'id': 18, 'name': 'Drama'}]
45463	Betrayal	67758	[{'id': 28, 'name': 'Action'}, {'id': 18, 'nam...
45464	Satan Triumphant	227506	[]
45465	Queerama	461257	[]

45466 rows × 3 columns

制作数据集

In [ ]:

```
# gpt-4编写的字符串处理函数
# 转换体裁

def genres2genre(str):
    # Since the input string uses single quotes, we need to replace them with do
    json_string = str.replace("'", '"')

    # Load the string as a JSON object (list of dictionaries)
    data = json.loads(json_string)

    # Extract the 'name' key from each dictionary and join them with '|'
    result = '|'.join(d['name'] for d in data)
    return result
```

In [ ]:

```
# 将genres转换成容易处理的形式

movies['genre'] = movies['genres'].apply(genres2genre)
movies.drop(columns='genres', inplace=True)
movies
```

Out[ ]:

	title	id	genre
0	Toy Story	862	Animation|Comedy|Family
1	Jumanji	8844	Adventure|Fantasy|Family
2	Grumpier Old Men	15602	Romance|Comedy
3	Waiting to Exhale	31357	Comedy|Drama|Romance
4	Father of the Bride Part II	11862	Comedy
...	...	...	...
45461	Subdue	439050	Drama|Family
45462	Century of Birthing	111109	Drama
45463	Betrayal	67758	Action|Drama|Thriller
45464	Satan Triumphant	227506	
45465	Queerama	461257	

45466 rows × 3 columns

In [ ]:

```
# 队电影题材进行ont-hot编码
movies = movies.join(movies.genre.str.get_dummies())
movies.drop(columns='genre', inplace=True)
movies
```

Out[ ]:

	title	id	Action	Adventure	Animation	Aniplex	BROSTA TV	Carou Productic
0	Toy Story	862	0	0	1	0	0	
1	Jumanji	8844	0	1	0	0	0	
2	Grumpier Old Men	15602	0	0	0	0	0	
3	Waiting to Exhale	31357	0	0	0	0	0	
4	Father of the Bride Part II	11862	0	0	0	0	0	
...	...	...	...	...	...	...	...	...
45461	Subdue	439050	0	0	0	0	0	
45462	Century of Birthing	111109	0	0	0	0	0	
45463	Betrayal	67758	1	0	0	0	0	
45464	Satan Triumphant	227506	0	0	0	0	0	
45465	Queerama	461257	0	0	0	0	0	

45466 rows × 34 columns

关联分析

In [ ]:

```
# 获取频繁项集
frequent_itemsets_movies = apriori(movies.drop(columns={'title', 'id'}), use_col
```

In [ ]:

```
frequent_itemsets_movies
```

Out[ ]:

	support	itemsets
0	0.145075	(Action)
1	0.076893	(Adventure)
2	0.042559	(Animation)
3	0.289931	(Comedy)
4	0.094730	(Crime)
...	...	...
70	0.016870	(Action, Crime, Thriller)
71	0.019157	(Action, Drama, Thriller)
72	0.030836	(Drama, Romance, Comedy)
73	0.025821	(Drama, Crime, Thriller)
74	0.015594	(Drama, Thriller, Mystery)

75 rows × 2 columns

In [ ]:

```
# 获取规则
rules_movies = association_rules(frequent_itemsets_movies, metric='lift', min_th
```

In [ ]:

```
rules_movies
```

Out[ ]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(Action)	(Adventure)	0.145075	0.076893	0.038116	0.262735	3.416908
1	(Adventure)	(Action)	0.076893	0.145075	0.038116	0.495709	3.416908
2	(Action)	(Crime)	0.145075	0.094730	0.030088	0.207398	2.189361
3	(Crime)	(Action)	0.094730	0.145075	0.030088	0.317622	2.189361
4	(Action)	(Fantasy)	0.145075	0.050873	0.011019	0.075955	1.493029
...	...	...	...	...	...	...	...
77	(Thriller)	(Drama, Crime)	0.167686	0.055536	0.025821	0.153987	2.772749
78	(Drama, Thriller)	(Mystery)	0.075375	0.054260	0.015594	0.206886	3.812850
79	(Drama, Mystery)	(Thriller)	0.025887	0.167686	0.015594	0.602379	3.592309
80	(Thriller)	(Drama, Mystery)	0.167686	0.025887	0.015594	0.092996	3.592309
81	(Mystery)	(Drama, Thriller)	0.054260	0.075375	0.015594	0.287394	3.812850

82 rows × 10 columns

In [ ]:

```
# 选取提升都大于3的电影
rules_movies_lift3 = rules_movies[rules_movies['lift'] > 3].sort_values('lift',
rules_movies_lift3
```

Out[ ]:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
19	(Family)	(Animation)	0.060925	0.042559	0.018849	0.309386	7.269538
18	(Animation)	(Family)	0.042559	0.060925	0.018849	0.442894	7.269538
38	(Fantasy)	(Family)	0.050873	0.060925	0.013483	0.265024	4.350026
39	(Family)	(Fantasy)	0.060925	0.050873	0.013483	0.221300	4.350026
15	(Fantasy)	(Adventure)	0.050873	0.076893	0.015000	0.294855	3.834635
14	(Adventure)	(Fantasy)	0.076893	0.050873	0.015000	0.195080	3.834635
81	(Mystery)	(Drama, Thriller)	0.054260	0.075375	0.015594	0.287394	3.812850
78	(Drama, Thriller)	(Mystery)	0.075375	0.054260	0.015594	0.206886	3.812850
12	(Adventure)	(Family)	0.076893	0.060925	0.017244	0.224256	3.680880
13	(Family)	(Adventure)	0.060925	0.076893	0.017244	0.283032	3.680880
73	(Drama, Thriller)	(Crime)	0.075375	0.094730	0.025821	0.342574	3.616312
76	(Crime)	(Drama, Thriller)	0.094730	0.075375	0.025821	0.272580	3.616312
48	(Thriller)	(Mystery)	0.167686	0.054260	0.032882	0.196091	3.613898
49	(Mystery)	(Thriller)	0.054260	0.167686	0.032882	0.605999	3.613898
79	(Drama, Mystery)	(Thriller)	0.025887	0.167686	0.015594	0.602379	3.592309
80	(Thriller)	(Drama, Mystery)	0.167686	0.025887	0.015594	0.092996	3.592309
53	(Drama, Adventure)	(Action)	0.022940	0.145075	0.011481	0.500479	3.449787
54	(Action)	(Drama, Adventure)	0.145075	0.022940	0.011481	0.079139	3.449787
1	(Adventure)	(Action)	0.076893	0.145075	0.038116	0.495709	3.416908
0	(Action)	(Adventure)	0.145075	0.076893	0.038116	0.262735	3.416908
61	(Action, Thriller)	(Crime)	0.052127	0.094730	0.016870	0.323629	3.416323
64	(Crime)	(Action, Thriller)	0.094730	0.052127	0.016870	0.178082	3.416323
60	(Action, Crime)	(Thriller)	0.030088	0.167686	0.016870	0.560673	3.343591
65	(Thriller)	(Action, Crime)	0.167686	0.030088	0.016870	0.100603	3.343591
41	(Science Fiction)	(Fantasy)	0.067061	0.050873	0.011393	0.169892	3.339515

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
40	(Fantasy)	(Science Fiction)	0.050873	0.067061	0.011393	0.223952	3.339515
11	(Adventure)	(Animation)	0.076893	0.042559	0.010755	0.139874	3.286572

保存数据

In [ ]:

```
frequent_itemsets_movies.to_csv('../data/frequent_itemsets_movies.csv', index=Fa
rules_movies_lift3.to_csv('../data/rules_movies_lift3.csv', index=False)
```

In [ ]:

```
```