**频繁模式与关联规则挖掘**

数据集处理

本次实验使用数据集为Trending YouTube Video Statistics，包括16个属性：

•video\_id 视频id，唯一属性

•trending\_date 流行时间

•title 视频标题

•channel\_title 发布频道

•category\_id 类别id（每个国家对应的类别不同）

•publish\_time 发布时间

•tags 视频标签

•views 观看数量

•likes 喜欢数量

•dislikes 不喜欢数量

•comment\_count 评论数量

•thumbnail\_link 视频链接，唯一属性

•comments\_disabled 是否禁止评论

•ratings\_disabled 是否禁止打分

•video\_error\_or\_removed 是否错误或删除

•description 视频描述

在互评实验一中，对此数据集缺失值进行了处理，为了对数据进行分析完全，此次试验对数值属性进行离散化处理，以其最大最小值和四分位数为参考，将数据划分合理区间。

(1)播放量views：'<24000'，'[20000,1800000]'，'[1800000,2400000]'，'>2400000'

(2)喜欢数量likes：'<5000'，'[5000,55000]'，'[55000,75000]'，'>75000'

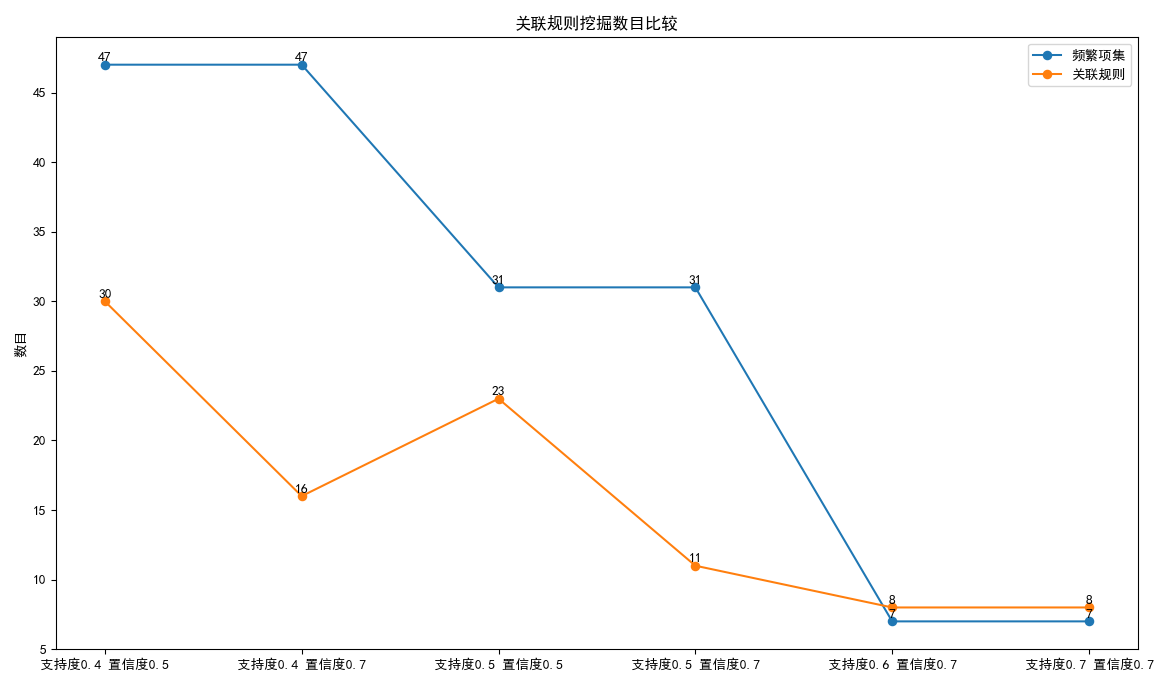
(3)不喜欢数量dislikes：'<200'，'[200,2000]'，'[2000,3700]'，'>3700'

(4)评论数量comment\_count：'<600'，'[600,5700]'，'[5700,8400]'，'>8400'

Apriori算法运行结果展示

一、总体生成频繁项集及关联规则情况

本次实验采用Apriori算法，在七组不同的最小支持度和置信度阈值［0.4,0.5］、［0.4,0.7］、[0.5,0.5]、［0.5,0.7］、［0.6,0.7］、［0.7,0.7］的约束下，对相同的美国Youtube数据集进行了关联规则挖掘，并将所得结果进行比较分析，用Python的2D绘图库Matplotlib可视化出来。



当最小支持度阈值不变时，挖掘得到的频繁项集数目也是不变的，说明挖掘频繁项集的过程只受最小支持度阈值的约束，并且频繁项集的数目随着最小支持度阈值的增大而减小，然后固定最小支持度阈值来讨论最小置信度阈值，可以发现关联规则的数目也会随着最小置信度阈值的增大而减小，当然，由于关联规则是通过频繁项集发现的，所以发现关联规则的过程，也可以看做整个关联规则挖掘的过程，会同时受到这两种参数的约束。

二、支持度0.5，置信度0.5

本文对使用Apriori算法挖掘整组数据集得到的结果进行分析，设置最小支持度阈值为0.5，最小置信度阈值为0.5,使用lift和kulc方法对结果进行评价，其结果为：

（1）32个频繁项集

{‘comments\_disabled=False’}，支持度：0.985

{‘ratings\_disabled=False’}，支持度：0.996

{‘video\_error\_or\_removed=False’}，支持度：0.999

{‘comment\_count=[600,5700]’}，支持度：0.502

{‘dislikes=[200,2000]’}，支持度：0.507

{‘likes=[5000,55000]’}，支持度：0.512

{‘video\_error\_or\_removed=False’,’ratings\_disabled=False’}，支持度：0.995

{‘comments\_disabled=False’,’video\_error\_or\_removed=False’}，支持度：0.984

{‘comments\_disabled=False’,’ratings\_disabled=False’}，支持度：0.983

{‘video\_error\_or\_removed=False’,’likes=[5000,55000]’}，支持度：0.512

{‘ratings\_disabled=False’,’likes=[5000,55000]’}，支持度：0.512

{‘comments\_disabled=False’,’likes=[5000,55000]’}，支持度：0.507

{‘dislikes=[200,2000]’,’video\_error\_or\_removed=False’}，支持度：0.507

{‘dislikes=[200,2000]’,’ratings\_disabled=False’}，支持度：0.507

{‘comments\_disabled=False’,’dislikes=[200,2000]’}，支持度：0.502

{‘comment\_count=[600,5700]’,’video\_error\_or\_removed=False’}，支持度：0.502

{‘comment\_count=[600,5700]’,’ratings\_disabled=False’}，支持度：0.501

{‘comments\_disabled=False’,’comment\_count=[600,5700]’}，支持度：0.502

{‘comments\_disabled=False’,’video\_error\_or\_removed=False’,’ratings\_disabled=False’}，支持度：0.982

{‘comments\_disabled=False’,’comment\_count=[600,5700]’,’ratings\_disabled=False’}，支持度：0.501

{‘comments\_disabled=False’,’comment\_count=[600,5700]’,’video\_error\_or\_removed=False’}，支持度：0.502

{‘comment\_count=[600,5700]’,’video\_error\_or\_removed=False’,’ratings\_disabled=False’}，支持度：0.501

{‘comments\_disabled=False’,’dislikes=[200,2000]’,’ratings\_disabled=False’}，支持度：0.502

{‘comments\_disabled=False’,’dislikes=[200,2000]’,’video\_error\_or\_removed=False’}，支持度：0.502

{‘dislikes=[200,2000]’,’ratings\_disabled=False’,’video\_error\_or\_removed=False’}，支持度：0.507

{‘comments\_disabled=False’,’likes=[5000,55000]’,’ratings\_disabled=False’}，支持度：0.507

{‘comments\_disabled=False’,’video\_error\_or\_removed=False’,’likes=[5000,55000]’}，支持度：0.506

{‘ratings\_disabled=False’,’video\_error\_or\_removed=False’,’likes=[5000,55000]’}，支持度：0.512

{‘ratings\_disabled=False’,’comments\_disabled=False’,’video\_error\_or\_removed=False’,’likes=[5000,55000]’}，支持度：0.506

{‘dislikes=[200,2000]’,’comments\_disabled=False’,’video\_error\_or\_removed=False’,’ratings\_disabled=False’}，支持度：0.502

{‘comments\_disabled=False’,’comment\_count=[600,5700]’,’video\_error\_or\_removed=False’,’ratings\_disabled=False’}，支持度：0.501

（2）23个关联规则

(1){'comments\_disabled=False’}—> {'comment\_count=[600,5700]’}

支持度：0.502，置信度：0.51，lift评价：1.016，kulc评价：0.755

(2){'comment\_count=[600,5700]’}—>{'comments\_disabled=False’}

支持度：0.502，置信度：1.0，lift评价：1.015，kulc评价：0.755

(3){'comment\_count=[600,5700]’}—>{'ratings\_disabled=False’}

支持度：0.501，置信度：0.998，lift评价：1.002，kulc评价：0.751

(4){'ratings\_disabled=False’}—>{'comment\_count=[600,5700]’}

支持度：0.501，置信度：0.503，lift评价：1.002，kulc评价：0.751

(5){'dislikes=[200,2000]’}—>{'comments\_disabled=False’}

支持度：0.502，置信度：0.99，lift评价：1.005，kulc评价：0.75

(6){'ratings\_disabled=False’}—>{'dislikes=[200,2000]’}

支持度：0.507，置信度：0.509，lift评价：1.004，kulc评价：0.755

(7){'video\_error\_or\_removed=False’}—>{'dislikes=[200,2000]’}

支持度：0.507，置信度：0.508，lift评价：1.002，kulc评价：0.754

(8){'comments\_disabled=False’}—>{'likes=[5000,55000]’}

支持度：0.507，置信度：0.515，lift评价：1.006，kulc评价：0.752

(9){'ratings\_disabled=False’}—>{'likes=[5000,55000]’}

支持度：0.512，置信度：0.514，lift评价：1.004，kulc评价：0.757

(10){'video\_error\_or\_removed=False’}—>{'likes=[5000,55000]’}

支持度：0.512，置信度：0.513，lift评价：1.002，kulc评价：0.756

(11){'ratings\_disabled=False’}—>{'comments\_disabled=False’}

支持度：0.983，置信度：0.987，lift评价：1.002，kulc评价：0.992

(12){'video\_error\_or\_removed=False’}—>{'comments\_disabled=False’}

支持度：0.984，置信度：0.985，lift评价：1.0，kulc评价：0.992

(13){'comments\_disabled=False’}—>{'video\_error\_or\_removed=False’}

支持度：0.984，置信度：0.999，lift评价：1.0，kulc评价：0.992

(14){'video\_error\_or\_removed=False’}—>{'ratings\_disabled=False’}

支持度：0.995，置信度：0.996，lift评价：1.0，kulc评价：0.997

(15){'likes=[5000,55000]’}—>{'comments\_disabled=False', 'video\_error\_or\_removed=False’}

支持度：0.506，置信度：0.988，lift评价：1.004，kulc评价：0.751

(16){'video\_error\_or\_removed=False’}—>{'comments\_disabled=False', 'likes=[5000,55000]’}

支持度：0.506，置信度：0.507，lift评价：1.0，kulc评价：0.752

(17){'ratings\_disabled=False’}—>{'dislikes=[200,2000]', 'comments\_disabled=False’}

支持度：0.502，置信度：0.504，lift评价：1.004，kulc评价：0.752

(18){'video\_error\_or\_removed=False’}—>{'ratings\_disabled=False', 'comment\_count=[600,5700]’}

支持度：0.501，置信度：0.502，lift评价：1.002，kulc评价：0.751

(19){'comments\_disabled=False’}—>{'ratings\_disabled=False', 'video\_error\_or\_removed=False’}

支持度：0.982，置信度：0.997，lift评价：1.002，kulc评价：0.992

(20){'video\_error\_or\_removed=False’}—>{'ratings\_disabled=False', 'comments\_disabled=False’}

支持度：0.982，置信度：0.983，lift评价：1.0，kulc评价：0.991

(21){'ratings\_disabled=False’}—>{'comments\_disabled=False', 'video\_error\_or\_removed=False’}

支持度：0.982，置信度：0.986，lift评价：1.002，kulc评价：0.992

(22){'ratings\_disabled=False', 'comments\_disabled=False’}—> {'dislikes=[200,2000]', 'video\_error\_or\_removed=False’}

支持度：0.502，置信度：0.511，lift评价：1.008，kulc评价：0.75

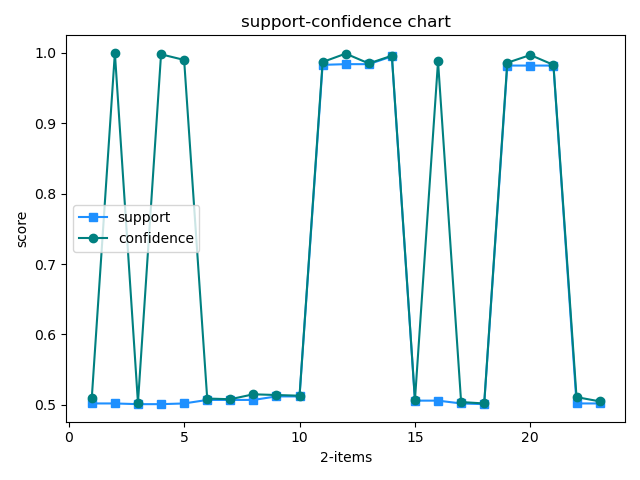
(23){'ratings\_disabled=False', 'video\_error\_or\_removed=False’}—> {'dislikes=[200,2000]', 'comments\_disabled=False’}

支持度：0.502，置信度：0.505，lift评价：1.006，kulc评价：0.752

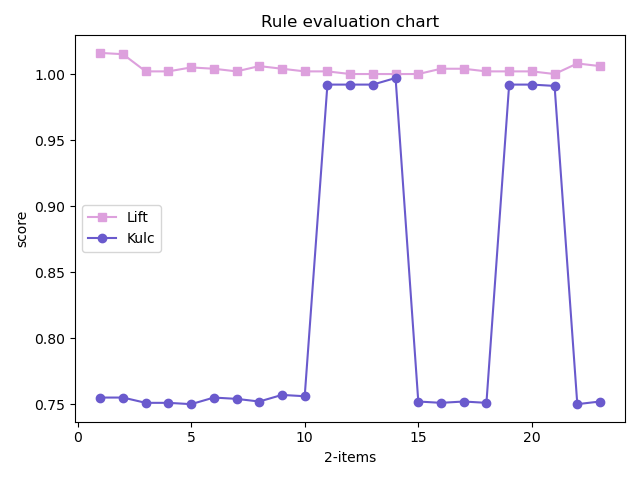
分析可知，对于约50%的评论在[600,5700]以下的视频，其视频都是允许评论和打分的。在50%左右的喜欢人数在[5000,55000]区间的视频无异常且可以评论。在可另外在98.2%的情况下，视频都是无异常且可打分评论的。

（3）可视化

将其23条关联规则的支持度与置信度用折线图进行表示：



将其23条关联规则的使用lift评论和kulc评价的结果用折线图进行表示：



三、支持度0.4，置信度0.7

本文对使用Apriori算法挖掘整组数据集得到的结果进行分析，设置最小支持度阈值为0.4，最小置信度阈值为0.7,其结果为：

（1）47个频繁项集

{'comments\_disabled=False'}，支持度：0.985

{'ratings\_disabled=False'}，支持度：0.996

{'video\_error\_or\_removed=False'}，支持度：0.999

{'comment\_count=[600,5700]'}，支持度：0.502

{'dislikes=[200,2000]'}，支持度：0.507

{'likes=[5000,55000]'}，支持度：0.512

{'views=[20000,1800000]'}，支持度：0.499

{'ratings\_disabled=False', 'video\_error\_or\_removed=False'}，支持度：0.995

{'video\_error\_or\_removed=False', 'comments\_disabled=False'}，支持度：0.984

{'ratings\_disabled=False', 'comments\_disabled=False'}，支持度：0.983

{'video\_error\_or\_removed=False', 'views=[20000,1800000]'}，支持度：0.499

{'ratings\_disabled=False', 'views=[20000,1800000]'}，支持度：0.497

{'comments\_disabled=False', 'views=[20000,1800000]'}，支持度：0.493

{'comment\_count=[600,5700]', 'likes=[5000,55000]'}，支持度：0.405

{'video\_error\_or\_removed=False', 'likes=[5000,55000]'}，支持度：0.512

{'ratings\_disabled=False', 'likes=[5000,55000]'}，支持度：0.512

{'comments\_disabled=False', 'likes=[5000,55000]'}，支持度：0.507

{'dislikes=[200,2000]', 'video\_error\_or\_removed=False'}，支持度：0.507

{'dislikes=[200,2000]', 'ratings\_disabled=False'}，支持度：0.507

{'dislikes=[200,2000]', 'comments\_disabled=False'}，支持度：0.502

{'video\_error\_or\_removed=False', 'comment\_count=[600,5700]'}，支持度：0.502

{'ratings\_disabled=False', 'comment\_count=[600,5700]'}，支持度：0.501

{'comments\_disabled=False', 'comment\_count=[600,5700]'}，支持度：0.502

{'ratings\_disabled=False', 'comments\_disabled=False', 'video\_error\_or\_removed=False'}，支持度：0.982

{'comments\_disabled=False', 'comment\_count=[600,5700]', 'likes=[5000,55000]'}，支持度：0.405

{'ratings\_disabled=False', 'comment\_count=[600,5700]', 'likes=[5000,55000]'}，支持度：0.405

{'ratings\_disabled=False', 'comments\_disabled=False', 'comment\_count=[600,5700]'}，支持度：0.501

{'ratings\_disabled=False', 'video\_error\_or\_removed=False', 'comment\_count=[600,5700]'}，支持度：0.501

{'video\_error\_or\_removed=False', 'comment\_count=[600,5700]', 'likes=[5000,55000]'}，支持度：0.405

{'video\_error\_or\_removed=False', 'comments\_disabled=False', 'comment\_count=[600,5700]'}，支持度：0.502

{'dislikes=[200,2000]', 'ratings\_disabled=False', 'comments\_disabled=False'}，支持度：0.502

{'dislikes=[200,2000]', 'video\_error\_or\_removed=False', 'comments\_disabled=False'}，支持度：0.502

{'dislikes=[200,2000]', 'ratings\_disabled=False', 'video\_error\_or\_removed=False'}，支持度：0.507

{'ratings\_disabled=False', 'comments\_disabled=False', 'likes=[5000,55000]'}，支持度：0.507

{'ratings\_disabled=False', 'video\_error\_or\_removed=False', 'likes=[5000,55000]'}，支持度：0.512

{'video\_error\_or\_removed=False', 'comments\_disabled=False', 'likes=[5000,55000]'}，支持度：0.506

{'ratings\_disabled=False', 'comments\_disabled=False', 'views=[20000,1800000]'}，支持度：0.493

{'ratings\_disabled=False', 'video\_error\_or\_removed=False', 'views=[20000,1800000]'}，支持度：0.497

{'video\_error\_or\_removed=False', 'comments\_disabled=False', 'views=[20000,1800000]'}，支持度：0.493

{'comments\_disabled=False', 'views=[20000,1800000]', 'ratings\_disabled=False', 'video\_error\_or\_removed=False'}，支持度：0.492

{'comments\_disabled=False', 'comment\_count=[600,5700]', 'likes=[5000,55000]', 'video\_error\_or\_removed=False'}，支持度：0.405

{'comment\_count=[600,5700]', 'likes=[5000,55000]', 'ratings\_disabled=False', 'video\_error\_or\_removed=False'}，支持度：0.405

{'comments\_disabled=False', 'comment\_count=[600,5700]', 'likes=[5000,55000]', 'ratings\_disabled=False'}，支持度：0.405

{'comments\_disabled=False', 'likes=[5000,55000]', 'ratings\_disabled=False', 'video\_error\_or\_removed=False'}，支持度：0.506

{'dislikes=[200,2000]', 'comments\_disabled=False', 'ratings\_disabled=False', 'video\_error\_or\_removed=False'}，支持度：0.502

{'comments\_disabled=False', 'comment\_count=[600,5700]', 'ratings\_disabled=False', 'video\_error\_or\_removed=False'}，支持度：0.501

{'ratings\_disabled=False', 'comments\_disabled=False', 'comment\_count=[600,5700]', 'likes=[5000,55000]', 'video\_error\_or\_removed=False'}，支持度：0.405

（2）16个关联规则

(1){'comment\_count=[600,5700]'}—>{'comments\_disabled=False'}

支持度：0.502，置信度：1.0，lift评价：1.015，kulc评价：0.755

(2){'comment\_count=[600,5700]'}—>{'ratings\_disabled=False'}

支持度：0.501，置信度：0.998，lift评价：1.002，kulc评价：0.751

(3){'dislikes=[200,2000]'}—>{'comments\_disabled=False'}

支持度：0.502，置信度：0.99，lift评价：1.005，kulc评价：0.75

(4){'likes=[5000,55000]'}—>{'comment\_count=[600,5700]'}

支持度：0.405，置信度：0.791，lift评价：1.576，kulc评价：0.799

(5){'comment\_count=[600,5700]'}—>{'likes=[5000,55000]'}

支持度：0.405，置信度：0.807，lift评价：1.576，kulc评价：0.799

(6){'views=[20000,1800000]'}—>{'comments\_disabled=False'}

支持度：0.493，置信度：0.988，lift评价：1.003，kulc评价：0.744

(7){'views=[20000,1800000]'}—>{'ratings\_disabled=False'}

支持度：0.497，置信度：0.996，lift评价：1.0，kulc评价：0.747

(8){'ratings\_disabled=False'}—>{'comments\_disabled=False'}

支持度：0.983，置信度：0.987，lift评价：1.002，kulc评价：0.992

(9){'comments\_disabled=False'}—>{'video\_error\_or\_removed=False'}

支持度：0.984，置信度：0.999，lift评价：1.0，kulc评价：0.992

(10){'video\_error\_or\_removed=False'}—>{'comments\_disabled=False'}

支持度：0.984，置信度：0.985，lift评价：1.0，kulc评价：0.992

(11){'video\_error\_or\_removed=False'}—>{'ratings\_disabled=False', 'comments\_disabled=False'}

支持度：0.982，置信度：0.983，lift评价：1.0，kulc评价：0.991

(12){'comments\_disabled=False'}—>{'ratings\_disabled=False', 'video\_error\_or\_removed=False'}

支持度：0.982，置信度：0.997，lift评价：1.002，kulc评价：0.992

(13){'ratings\_disabled=False'}—>{'video\_error\_or\_removed=False', 'comments\_disabled=False'}

支持度：0.982，置信度：0.986，lift评价：1.002，kulc评价：0.992

(14){'ratings\_disabled=False', 'comment\_count=[600,5700]'}—> {'comments\_disabled=False', 'likes=[5000,55000]'}

支持度：0.405，置信度：0.808，lift评价：1.594，kulc评价：0.804

(15){'comments\_disabled=False', 'likes=[5000,55000]'}—> {'ratings\_disabled=False', 'comment\_count=[600,5700]'}

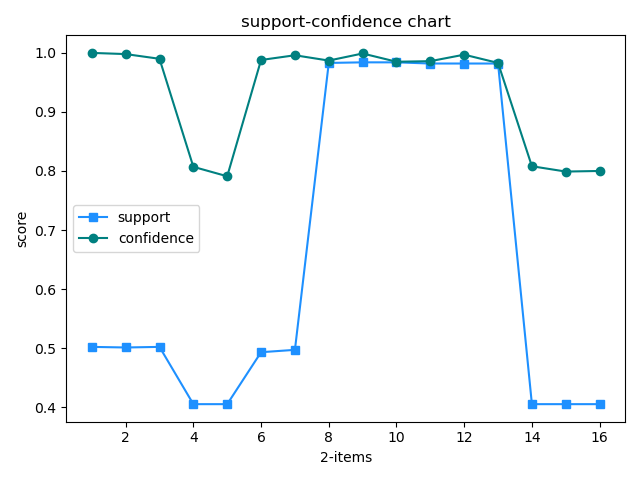
支持度：0.405，置信度：0.799，lift评价：1.595，kulc评价：0.804

(16){'video\_error\_or\_removed=False', 'comments\_disabled=False', 'likes=[5000,55000]'}—>{'ratings\_disabled=False', 'comment\_count=[600,5700]'}

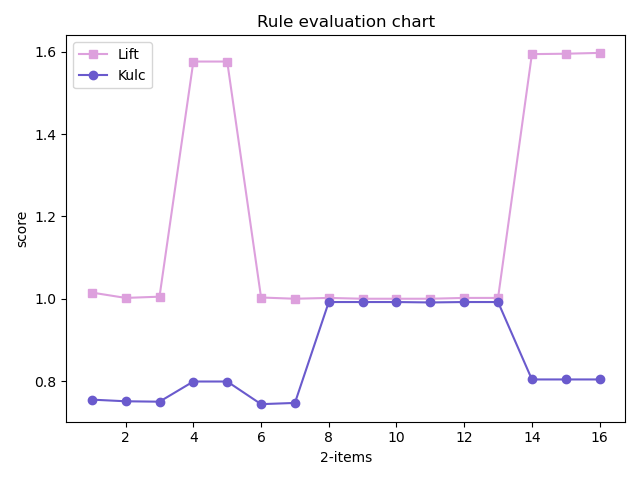
支持度：0.405，置信度：0.8，lift评价：1.597，kulc评价：0.804

（3）可视化

将其16条关联规则的支持度与置信度用折线图进行表示：



将其16条关联规则的使用lift评论和kulc评价的结果用折线图进行表示：



**四、**局部数据集挖掘结果分析

为了挖掘出更有趣且有价值的规则，除了对最小支持度和置信度阈值进行调整，还可以移除一些某一特定取值占比重偏大的属性，挑选几个感兴趣且具有代表性的属性组作为子数据集，这里对使用Apriori算法进行关联规则挖掘的结果进行着重分析，结果如下：

（1）属性组：{‘likes,’comment\_count’} min\_sup:0.4 min\_conf:0.8

频繁项集：{‘likes,’comment\_count’}，支持度：0.405

关联规则：{‘comment\_count=[600,5700]’}—>{‘likes=[5000,55000]’}，置信度：0.807

意义：在80.7%的评论数在[600,5700]区间的视频中，喜欢的数量在区间[5000,55000]。

（2）属性组：{‘likes’,’dislikes’} min\_sup:0.3 min\_conf:0.7

频繁项集：{‘dislikes=[200,2000]’,’likes=[5000,55000]’}，支持度：0.361

关联规则：{‘dislikes=[200,2000]’}—>{‘likes=[5000,55000]’}，置信度：0.712

意义：在71.2%的不喜欢数量在[200,2000]区间的视频中，喜欢的数量在区间[5000,55000]。

（3）属性组：{‘views’,’dislikes’} min\_sup:0.3 min\_conf:0.7

频繁项集：{‘dislikes=[200,2000]’,’views=[20000,1800000]’}，支持度：0.384关联规则：{‘dislikes=[200,2000]’}—>{‘views=[20000,1800000]’}，置信度：0.757

意义：在75.7%的不喜欢数量在[200,2000]区间的视频中，播放量在区间[20000,1800000]。

（4）属性组：{‘views’,’likes’,’dislikes’,’comment\_count’} min\_sup:0.3 min\_conf:0.7

频繁项集：

{‘views=[20000,1800000]’,’likes=[5000,55000]’}，支持度：0.387

{‘views=[20000,1800000]’,’dislikes=[200,2000]’}，支持度：0.384

{‘views=[20000,1800000]’,’comment\_count=[600,5700]’}，支持度：0.369

{‘dislikes=[200,2000]’,’likes=[5000,55000]’}，支持度：0.361

{‘comment\_count=[600,5700]’,’likes=[5000,55000]’}，支持度：0.405

{‘comment\_count=[600,5700]’,’dislikes=[200,2000]’}，支持度：0.379

{‘comment\_count=[600,5700]’,’dislikes=[200,2000]’,’likes=[5000,55000]’}，支持度：0.31

{‘views=[20000,1800000]’,’comment\_count=[600,5700]’,’dislikes=[200,2000]’}，支持度：0.307

{‘views=[20000,1800000]’,’comment\_count=[600,5700]’,’likes=[5000,55000]’}，支持度：0.319

{‘views=[20000,1800000]’,’dislikes=[200,2000]’,’likes=[5000,55000]’}，支持度：0.308

关联规则：

{‘dislikes=[200,2000]’}—>{‘comment\_count=[600,5700]’}，置信度：0.748

{‘comment\_count=[600,5700]’}—>{‘dislikes=[200,2000]’}，置信度：0.755

{‘likes=[5000,55000]’}—>{‘comment\_count=[600,5700]’}，置信度：0.791

{‘comment\_count=[600,5700]’}—>{‘likes=[5000,55000]’}，置信度：0.807

{‘likes=[5000,55000]’}—>{‘dislikes=[200,2000]’}，置信度：0.705

{‘dislikes=[200,2000]’}—>{‘likes=[5000,55000]’}，置信度：0.712

{‘comment\_count=[600,5700]’}—>{‘views=[20000,1800000]’}，置信度：0.735

{‘views=[20000,1800000]’}—>{‘comment\_count=[600,5700]’}，置信度：0.739

{‘dislikes=[200,2000]’}—>{‘views=[20000,1800000]’}，置信度：0.757

{‘views=[20000,1800000]’}—>{‘dislikes=[200,2000]’}，置信度：0.77

{‘views=[20000,1800000]’}—>{‘likes=[5000,55000]’}，置信度：0.776意义：由此可见，观看数、喜欢、不喜欢和评论数是息息相关的，一个视频观看数在区间[20000,1800000]，有77.6%的视频喜欢人数在区间[5000,55000]。一个视频喜欢人数少于5000，有77%不喜欢的人在区间[200,2000]，73.9%的视频评论数在区间[600,5700]。