A preliminary method to construct a movie ranking list

Yinqiu Xu

Abstract: It's pretty common to see a ranking list on the website. For instance, the hot searches on Zhihu, the famous music ranking list--the billboard and top250 movies on Douban are ranking lists we are familiar with. This project aims to use the observed rating data of users to and some rational idea of ranking algorithm to construct a ranking list. Generally, we use the dataset from the website, Movielens, and by doing interval estimation to the applause proportion of each movie. Then we can rank those movies. Considering our data set, we will do some modifications to users ratings to get the applause proportion of each movie. Compared with simply ranking the average rating score of movies, interval estimation is more rational because it amends some errors of the point estimation of applause proportion. This ranking method still has many deficiency, so I will provide further modifications of it.

Keywords: ranking algorithm, ranking list, interval estimation, confidence interval, Movielens

1 Introduction

It's pretty common for us to see a lot of ranking lists on many applications or website. For example, the hot searches on Zhihu, the famous music ranking list--the billboard and top250 movies on Douban are ranking lists we are familiar with. Ranking lists can show the hot topics and searches to us visually and immediately. So many online communities rank hot topics to help users catch current affairs and the fundamental function of recommending websites is to make ranking lists to help users choose ones they like. The motivation is that I want to know how they construct these ranking lists and what are the ranking algorithms behind them. Then I check the articles about it, and I find a method called Wilson score interval but it's beyond my ability. After consulting the professor, I decide to use Wilson's simplest idea, the interval estimation, to construct a ranking list. During the process, I have learned some other ranking algorithms which I will show in the third part. Dataset is collecting from the website called Movielens. After cleaning the data, I got 4674 movies to construct the ranking list.

This report shows the ranking algorithm and some further modifications following codes and data in the appendix.

2data

The data is collected from Grouplens.org. Grouplens Research has collected and made available rating data sets from the Movielens website. The data sets were collected over various periods of time, depending on the size of the set.

The data I choose is Movielens Latest Datasets. This dataset consists of 100,000 ratings and 1,300 tag applications applied to 9,000 movies by 700 users. It's last updated 10/2016.

In the file, I use two excels—ratings and movies. Ratings.csv consists of userId, movieId, rating and timestamps. Ratings are made on a 5-star scale, with half-star increments (0.5 stars - 5.0 stars). Movies.csv consists of movieId, title and genres. MovieId is consistence between "ratings.csv" and "movies.csv".

	A	B	C	D	E	F	G
1	movieId	title	genres				
2	1	Toy Story	Adventure|Animation|Children|Comedy|Fantasy				
3	2	Jumanji (	Adventure|Children|Fantasy				
4	3	Grumpier (	Comedy|Romance				
5	4	Waiting t	Comedy|Drama|Romance				
6	5	Father of	Comedy				
7	6	Heat (199	Action|Crime|Thriller				
8	7	Sabrina (	Comedy|Romance				
9	8	Tom and H	Adventure|Children				
10	9	Sudden De	Action				
11	10	GoldenEye	Action|Adventure|Thriller				
12	11	American	Comedy|Drama|Romance				
13	12	Dracula:	Comedy|Horror				
14	13	Balto (19	Adventure|Animation|Children				

Movies.xlsx

	A	B	C	D	E	F	G
1	userId	movieId	rating	timestamp			
2	7	1	3	851866703			
3	9	1	4	938629179			
4	13	1	5	1.331E+09			
5	15	1	2	997938310			
6	19	1	3	855190091			
7	20	1	3.5	1.239E+09			
8	23	1	3	1.149E+09			
9	26	1	5	1.36E+09			
10	30	1	4	944943070			
11	37	1	4	981308121			
12	43	1	4	974768260			
13	44	1	4	858707138			
14	47	1	5	832228931			

Ratings.xlsx

3Some natural ideas of ranking algorithm

When it comes to ranking movies, here is some natural ideas. We may rank them by the number of thumbs up per unit time of each movie. The movie gets more thumbs up ranking higher. Other idea considers that every article has its own temperature and it changes with time and the number of thumbs up and down. By Newton's law of cooling, we can obtain a score for each movie, then ranking them to get a list. The third idea is simply ranking by their average rating score like IMDb's

top250 movies list. They share the same key idea but IMDb does it more complicated considering those minor movies and the quality of users.

4Ranking algorithm

The ranking algorithm I use is to use interval estimation of the applause proportion of each movie and rank them by the lower bound of the interval estimate. Finally, we can construct a top50 ranking list.

4.1key idea

The key idea to construct the ranking list is ranking the lower bound of the confidence interval of p. P denotes the proportion of thumbs up evaluated by users or we call it applause proportion of each movie.

We use interval estimation to estimate the applause proportion.

4.2assumption

In my model, I do following assumptions.

First, I assume that users give their ratings independently. This assumption guarantees that random variables $\mathrm{Y_i}\ \mathrm{i} = 1,2 \ldots . \mathrm{n}$ are mutually independent and identical.

Second, I assume that users can only choose thumbs up or down for each movie. Because the ratings in Movielens dataset ranges from 1.0 to 5.0 with step is 0.5.

To fit with my model, I divide the ratings into two parts which is higher than or equal to 3 and lower than 3. If the rating is higher than or equal 3, that means thumbs up. Similarly, if the rating is lower than 3, that means thumbs down. After this simplification, I change $\mathrm{X} = \sum Y_i$ from a multinomial random variable to a binomial variable. The analysis following will base on binomial assumption.

Third, if the total number of users is n and k of them give thumbs up, then the applause proportion p is k/n. P is the value that we want to estimate.

4.3notations

$\mathrm{Y_i}\ i = 1,2, \ldots . n$ denotes the rating of each user. After my simplification, it

follows Bernoulli(p)

X denotes the number of thumbs up of each movie. After my simplification, it follows Binomial(n,p)

P denotes the applause proportion. If the total number of ratings is n and the number of thumbs up is k, the applause proportion is p.

4.4process

The number of thumbs up follows binomial distribution. Denoted as $\mathrm{X}\sim\mathrm{Bin}(\mathrm{n},\mathrm{p})$.

Recalling the process of constructing confidence interval of p.

By central limit theorem,

$$\frac{X-\mathrm{np}}{\sqrt{(\mathrm{np}(1-\mathrm{p})}}\rightarrow^{\mathrm{d}}\mathrm{N}(0,1)$$

Replacing p with$\frac{\mathrm{X}}{\mathrm{n}}$, for a sufficiently large n,

$$\frac{\frac{\mathrm{X}}{\mathrm{n}}-\mathrm{p}}{\sqrt{\frac{\left(\frac{\mathrm{X}}{\mathrm{n}}\right)\left(1-\frac{\mathrm{X}}{\mathrm{n}}\right)}{\mathrm{n}}}}\rightarrow^{\mathrm{d}}\mathrm{N}(0,1)$$

Confidence interval of p, for n sufficiently large,

$$\mathrm{P}\left\{-\mathrm{z}_{\frac{\alpha}{2}}\leq\frac{\frac{\mathrm{X}}{\mathrm{n}}-\mathrm{p}}{\sqrt{\frac{\left(\frac{\mathrm{X}}{\mathrm{n}}\right)\left(1-\frac{\mathrm{X}}{\mathrm{n}}\right)}{\mathrm{n}}}}\leq\mathrm{z}_{\frac{\alpha}{2}}\right\}\approx 1-\alpha$$

Rearranging the terms, we get the confidence interval of p.

$$[\frac{\mathrm{X}}{\mathrm{n}}-\mathrm{z}_{\frac{\alpha}{2}}\sqrt{\frac{\left(\frac{\mathrm{X}}{\mathrm{n}}\right)\left(1-\frac{\mathrm{X}}{\mathrm{n}}\right)}{\mathrm{n}}},\frac{\mathrm{X}}{\mathrm{n}}+\mathrm{z}_{\frac{\alpha}{2}}\sqrt{\frac{\left(\frac{\mathrm{X}}{\mathrm{n}}\right)\left(1-\frac{\mathrm{X}}{\mathrm{n}}\right)}{\mathrm{n}}}\ ]$$

Focusing on my project, first, we use data from the dataset to calculate the observed applause proportion of each movie $\frac{\mathrm{X}}{\mathrm{n}}$. Then since it's a rule of thumb, we usually calculate the 95% confidence interval of each movie's applause proportion p.

We can get the lower bound of the 95% confidence interval $\mathrm{lowerlim} = \frac{\mathrm{X}}{\mathrm{n}} - \mathrm{Z}_{\frac{\alpha}{2}}\sqrt{\left(\frac{X}{n}\right)\left(1-\frac{X}{n}\right)/n}$. Then we rank the movies by their lower bound and select top50 to be our ranking list.

4.5advantages

Our ranking algorithm based on interval estimation has some advantages over point estimation. On one hand, the length of confidence interval is associated with sample size negatively. If sample size is small, then the length of confidence interval increases. Even though the observed values of two sample are the same, the negative relation is correct. Therefore, if two movies get the same applause proportion, the one with small sample size will have a relatively lower value of lowerlim. In other word, using confidence interval is a kind of amending of credibility. If we draw 95% confidence interval of p repeatedly for a larger number of times. 95% of intervals will cover the true p. For two movie's p, if the whole interval of a movie's p is higher than another one's, then we can believe that it ranks higher.

4.6preliminary ranking list

Following our process, we get preliminary ranking list.

	movieId	likes	ratingnum	p	q	lowerlim	upperlim	title	genres
1	26	5	5	1	0	1	1	Othello (1995)	Drama
2	40	6	6	1	0	1	1	Cry, the Beloved Country (1995)	Drama
3	49	2	2	1	0	1	1	When Night Is Falling (1995)	Drama|Romance
4	53	1	1	1	0	1	1	Lamerica (1994)	Adventure|Drama
5	54	3	3	1	0	1	1	Big Green, The (1995)	Children|Comedy
6	59	2	2	1	0	1	1	Confessional, The (Confessionnal, Le) (1995)	Drama|Mystery
7	61	7	7	1	0	1	1	Eye for an Eye (1996)	Drama|Thriller
8	77	2	2	1	0	1	1	Nico Icon (1995)	Documentary
9	80	4	4	1	0	1	1	White Balloon, The (Badkonake sefid) (1995)	Children|Drama
10	83	3	3	1	0	1	1	Once Upon a Time... When We Were Colored (1995)	Drama|Romance
11	84	1	1	1	0	1	1	Last Summer in the Hamptons (1995)	Comedy|Drama
12	85	16	16	1	0	1	1	Angels and Insects (1995)	Drama|Romance
13	97	8	8	1	0	1	1	Hate (Haine, La) (1995)	Crime|Drama
14	98	1	1	1	0	1	1	Shopping (1994)	Action|Thriller
15	103	4	4	1	0	1	1	Unforgettable (1996)	Mystery|Sci-Fi|Thriller
16	108	1	1	1	0	1	1	Catwalk (1996)	Documentary
17	114	1	1	1	0	1	1	Margarets Museum (1995) Drama 116 Anne Frank Reme...	Crime|Drama
18	119	1	1	1	0	1	1	Steal Big, Steal Little (1995)	Comedy
19	121	5	5	1	0	1	1	Boys of St. Vincent, The (1992)	Drama
20	123	7	7	1	0	1	1	Chungking Express (Chung Hing sam lam) (1994)	Drama|Mystery|Romance
21	124	1	1	1	0	1	1	Star Maker, The (Uomo delle stelle, L) (1995) Drama 125 Fl...	Drama
22	211	3	3	1	0	1	1	Browning Version, The (1994)	Drama
23	219	4	4	1	0	1	1	Cure, The (1995)	Drama
24	229	12	12	1	0	1	1	Death and the Maiden (1994)	Drama|Thriller
25	232	24	24	1	0	1	1	Eat Drink Man Woman (Yin shi nan nu) (1994)	Comedy|Drama|Romance
26	233	13	13	1	0	1	1	Exotica (1994)	Drama
27	242	4	4	1	0	1	1	Farinelli: il castrato (1994)	Drama|Musical
28	243	1	1	1	0	1	1	Gordy (1995)	Children|Comedy|Fantasy
29	244	1	1	1	0	1	1	Gumby: The Movie (1995)	Animation|Children
30	245	1	1	1	0	1	1	The Glass Shield (1994)	Crime|Drama

Seeing the columns named likes and ratingnum, the sample size of these movies is too small which makes the ranking list less credible. Some only have 1 or 2 ratings. Restricted by the dataset, we have to delete those movies. Then we can get a more believable ranking list.

4.7final ranking list

	movieId	likes	ratingnum	p	q	lowerlim	upperlim	title	genres
100	1203	74	74	1.0000000	0.00000000	1.0000000	1.0000000	12 Angry Men (1957)	Drama
102	1252	75	75	1.0000000	0.00000000	1.0000000	1.0000000	Chinatown (1974)	c("Crime", "Film-Noir", "Mystery", "Thriller")
2034	50	197	200	0.9850000	0.01500000	0.9681540	1.0018460	Usual Suspects, The (1995)	c("Crime", "Mystery", "Thriller")
2035	68157	58	59	0.9830508	0.01694915	0.9501139	1.0159878	Inglourious Basterds (2009)	c("Action", "Drama", "War")
2036	457	206	212	0.9716981	0.02830189	0.9493751	0.9940211	Fugitive, The (1993)	Thriller
2037	111	113	116	0.9741379	0.02586207	0.9452537	1.0030221	Taxi Driver (1976)	c("Crime", "Drama", "Thriller")
2038	1270	218	226	0.9646018	0.03539823	0.9405106	0.9886929	Back to the Future (1985)	c("Adventure", "Comedy", "Sci-Fi")
2039	2599	77	79	0.9746835	0.02531646	0.9400444	1.0093227	Election (1999)	Comedy
2040	527	232	241	0.9626556	0.03734440	0.9387177	0.9865935	Schindlers List (1993) Drama|War 528 Scout, The (1994) C...	Drama
2041	1617	121	125	0.9680000	0.03200000	0.9371465	0.9988535	L.A. Confidential (1997)	c("Crime", "Film-Noir", "Mystery", "Thriller")
2042	1196	223	232	0.9612069	0.03879310	0.9363590	0.9860548	Star Wars: Episode V - The Empire Strikes Back (1980)	c("Action", "Adventure", "Sci-Fi")
2043	1193	139	144	0.9652778	0.03472222	0.9353761	0.9951795	One Flew Over the Cuckoos Nest (1975) Drama 1194 Che...	Comedy
2044	4973	116	120	0.9666667	0.03333333	0.9345497	0.9987836	Amelie (Fabuleux destin dAm茅lie Poulain, Le) (2001) Co...	c("Adventure", "Comedy", "Fantasy", "Sci-Fi")
2045	1197	157	163	0.9631902	0.03680982	0.9342840	0.9920964	Princess Bride, The (1987)	c("Action", "Adventure", "Comedy", "Fantasy", "Romance")
2046	58559	115	119	0.9663866	0.03361345	0.9340044	0.9987687	Dark Knight, The (2008)	c("Action", "Crime", "Drama", "IMAX")
2048	293	126	131	0.9618321	0.03816794	0.9290217	0.9946424	L茅on: The Professional (a.k.a. The Professional) (L茅on) (...	c("Action", "Crime", "Drama", "Thriller")
2049	1250	63	65	0.9692308	0.03076923	0.9272488	1.0112127	Bridge on the River Kwai, The (1957)	c("Adventure", "Drama", "War")
2052	48516	81	84	0.9642857	0.03571429	0.9246003	1.0039712	Departed, The (2006)	c("Crime", "Drama", "Thriller")
2053	36	100	104	0.9615385	0.03846154	0.9245789	0.9984981	Dead Man Walking (1995)	c("Crime", "Drama")
2055	1258	96	100	0.9600000	0.04000000	0.9215928	0.9984072	Shining, The (1980)	Horror
2056	1198	208	219	0.9497717	0.05022831	0.9208443	0.9786991	Raiders of the Lost Ark (Indiana Jones and the Raiders of ...	c("Action", "Adventure")
2058	349	110	115	0.9565217	0.04347826	0.9192498	0.9937937	Clear and Present Danger (1994)	c("Action", "Crime", "Drama", "Thriller")
2059	4226	126	132	0.9545455	0.04545455	0.9190112	0.9900797	Memento (2000)	c("Mystery", "Thriller")
2062	32	186	196	0.9489796	0.05102041	0.9181747	0.9797845	Twelve Monkeys (a.k.a. 12 Monkeys) (1995)	c("Mystery", "Sci-Fi", "Thriller")
2063	1291	140	147	0.9523810	0.04761905	0.9179551	0.9868068	Indiana Jones and the Last Crusade (1989)	c("Action", "Adventure")
2064	5669	55	57	0.9649123	0.03508772	0.9171449	1.0126796	Bowling for Columbine (2002)	Documentary
2066	1208	105	110	0.9545455	0.04545455	0.9156196	0.9934713	Apocalypse Now (1979)	c("Action", "Drama", "War")
2067	8961	118	124	0.9516129	0.04838710	0.9138443	0.9893815	Incredibles, The (2004)	c("Action", "Adventure", "Animation", "Children", "Comed...
2068	356	320	341	0.9384164	0.06158358	0.9129011	0.9639317	Forrest Gump (1994)	c("Comedy", "Drama", "Romance", "War")
2071	6	99	104	0.9519231	0.04807692	0.9108081	0.9930380	Heat (1995)	c("Action", "Crime", "Thriller")

After we delete the incredible outcomes, we finally get a ranking list like this. The values of the “likes” and “ratingnum” columns is larger making the outcome more credible.

5 modifications

This ranking algorithm still has a lot of improvements to do.

First, for those movies with extremely small samples, we could collect more data to increase their credibility enabling to add them back to the list.

Secondly, the cutoff 3 is simply the median of rating. To do more rational research, we can replace it with the average rating of all movies or other more rational number. For example, if 4 is the rating a good movie should get at least for a single person, then he can set the cutoff to be 4.

Thirdly, we can use the same idea of the ranking algorithm but delete the assumption of binomial distribution. Then we can calculate the 95% confidence interval of multinomial distribution parameters. This procedure makes the algorithm more rational.

Fourthly, focusing on those minor cinema, we can try bootstrap which a resampling method with replacement to estimate a parameter. Since the number of viewers of minor cinema is naturally small, we can resample it repeatedly for a large number of times. By extracting each sample value independently, we can get many samples. Then using these samples to estimate the applause proportion will be relatively more credible.

6references

[1]Wilson score interval:

https://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval

[2]data:

https://grouplens.org/datasets/movielens/

[3]基于用户投票的排名算法（五）威尔逊区间:

https://blog.csdn.net/real_myth/article/details/48974925

[4]基于用户投票的排名算法（六）贝叶斯平均:

https://blog.csdn.net/zhuhengv/article/details/50476645

[5]基于用户投票的排名算法（四）牛顿冷却定律:

https://blog.csdn.net/zhuhengv/article/details/50476306

[6]MIAO Mu.Applications for Bootstrap Method to the Estimation of the Standard Error and Confidence Intervals[J].Journal of Kunming Teachers College,2007,29(4):26-28