An analysis of online ratings data

Chien-Lang Su, Sun-Hao Chang, Ruby Chiu-Hsing Weng*

Department of Statistics, National Chengchi University, Taipei, Taiwan

Abstract

Internet ratings data are everywhere and increasing rapidly. They are usually ordinal measurements from 1 to 5 or 1 to 10 rated by Internet users on the quality of all kinds of items; for example, movies, books, etc. The graphical displays of the ratings data have little change over the past years, and the traditional displays does not account for the inter-rater difference. To address this problem, Ho and Quinn 2008 proposed a model-based graphical display. However, in order to identify the model, certain parameters are constrained to be positive. In the present work we first show that such a restriction may have a great impact on the rankings of items. Then we addressed some issues concerning the estimation of model parameters. Two real data sets are used for illustration.

Key words: Ordinal item response; posterior distribution; ranking.

1 Introduction

Online ratings data are usually ordinal measurements (with an integer scale from 1 to 5 or 1 to 10) on the quality or reputation of all kinds of items. The ratings data have generated great interest in providing personalized recommendations. A vast literature are concerned about the issues of speed and prediction accuracy, and the techniques used are often from machine learning aspects. See, for example, Koren et al. 2009, ACM SIGCHI 2011, and references therein.

*Corresponding author; Email:chweng@nccu.edu.tw; Address: No. 64, Sec. 2, ZhiNan Rd., Wenshan District, Taipei City 116, Taiwan (R.O.C)

There is much less research on online ratings data from statistical perspectives. Among these studies, the issues concerned also varied. For example, Greaves et al. 2012 studied associations between Internet-based ratings and conventional surveys of patient experience; Zhou and Lange 2009 proposed a statistical model and an EM algorithm to identify the quirky raters; Ho and Quinn 2008 proposed model-based graphical displays of ratings data. For the past few decades, the graphical displays of ratings data typically represent the mean rating of a product by a number of stars, which lack a measure of uncertainty. Though some displays may also include the number of votes, or even add the frequency plot of ratings, they all ignore the *inter-rater difference*; for example, some raters may be more inclined to use high or low ratings while others may use all the rating categories but do not discriminate very well between products. Ho and Quinn 2008 proposed a model-based approach that incorporated statistical uncertainty in the ratings and can adjust for rater-specific factors.

To identify the model, they constrained some parameters to be positive. However, we found that this restriction may have a great impact on the ranking of the items. In the present paper we will address the following problems: Why are the rankings with and without the parameter constraint so different? Does the constraint really identify the model? Is it necessary to constrain these parameters?

2 Data

We consider two online ratings datasets: the ratings of news outlets from Mondo Times (`http://www.mondotimes.com/`) studied in Ho and Quinn 2008, and the ratings of movies from GroupLens Research Project at the University of Minnesota (`http://movielens.umn.edu`).

Mondo Times is an online company that disseminates information about media outlets such as newspapers, magazines, radio stations, and television stations in 211 countries. Raters submit five-point ratings of the content quality of news outlets from awful, poor, average, very good, to great. The dataset used in Ho and Quinn 2008, which consists of 4,511 ratings on 1,515 products (news outlets) from 946 raters, is available from their Ratings package. The average number of ratings for a product is 3.0(=4,511/1,515) and the average number rated by a rater is 4.8(=4,511/946). Obviously many y_{rp} are not observed, and the missingness rate is 0.997, obtained by $1 - 4,511/(1,515 \times 946)$.

As in Ratings of Ho and Quinn 2008, we remove raters who rate less than five products and remove products that are only rated by these raters. This ends up with 3,249 ratings from 232 raters on 1,344 products. The missingness rate of this reduced data is 0.990.

The movie ratings data was collected from September 19, 1997 through April 22, 1998. This dataset consists of 100,000 movie ratings collected from 943 users on 1,682 movies, also on a scale of 1 to 5. The average number of ratings for a product is 59.5(=100,000/1,682) and the average number rated by a rater is 106.0(=100,000/943). The missingness rate is 0.937. To produce a data matrix of size similar to the processed Mondo, we remove raters who rate less than 120 products and remove products that are only rated by these raters. This leaves 68,154 ratings from 305 raters on 1,657 movies, and the resulting missingness rate is 0.865.

3 Main results

We introduce the idea of connectivity of the data matrix, and conduct the hypothesis testing of correlation coefficient between each pair of raters. To further compare the models with $\beta \in \mathbb{R}^+$ and $\beta \in \mathbb{R}$, we assess their out-of-sample prediction ability. We found that the Mondo data is more sensitive to restricting $\beta \in \mathbb{R}^+$ or not, in terms of items ranking and prediction errors. This perhaps is due to the fact that a larger proportion of rater-pairs exhibited negative correlation.

In conclusion, to apply IRT model for online ratings data, we suggested checking connectivity and the correlation test first. If the data matrix is connected and not many rater-pairs are negatively correlated, we can do estimation with $\beta \in \mathbb{R}$ and identify the sign by constraining a particular parameter to be positive. If the matrix is not connected, we need to identify the connected submatrix, and ranking within this submatrix is still sensible. If the correlation test shows non-negligible negative correlation between rater-pairs, then we need more caution on the interpretation.

References

ACM SIGCHI. *RecSys: Proceedings of the 2011 ACM Conference on Recommender Systems.* Chicago, IL, 2011.

Felix Greaves, Utz J Pape1, Dominic King, Ara Darzi, Azeem Majeed, Robert M Wachter, and Christopher Millett. Associations between internet-based patient ratings and conventional surveys of patient experience in the english nhs: an observational study. *BMJ Quality and Safety*, forthcoming, 2012.

Daniel E. Ho and Kevin M. Quinn. Improving the presentation and interpretation of online ratings data with model-based figures. *The American Statistician*, 62(4):279–288, 2008.

Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.

Hua Zhou and Kenneth Lange. Rating movies and rating the raters who rate them. *The American Statistician*, 63(4):297–307, 2009.