Collaborative Filtering on Skewed Datasets

Somnath Banerjee
Hewlett-Packard Labs
Bangalore, India
somnath.banerjee@hp.com

Krishnan Ramanathan
Hewlett-Packard Labs
Bangalore, India
krishnan.ramanathan@hp.com

ABSTRACT
Many real life datasets have skewed distributions of events when the probability of observing few events far exceeds the others. In this paper, we observed that in skewed datasets the state of the art collaborative filtering methods perform worse than a simple probabilistic model. Our test bench includes a real ad click stream dataset which is naturally skewed. The same conclusion is obtained even from the popular movie rating dataset when we pose a binary prediction problem of whether a user will give maximum rating to a movie or not.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: *Information Filtering*

General Terms
Algorithms, Experimentation

Keywords
Collaborative filtering, skewed dataset, pLSA.

1. INTRODUCTION
Collaborative filtering [1] is probably the most widely used technique in recommender systems. This technique finds interesting items for a user by utilizing the interest expressed by the users in the past on the items. No knowledge engineering issue (no rigorous description of user or item is required), novelty in recommendation etc. are the major advantages of this technique.

Most publications on collaborative filtering deal with rating datasets, e.g. movie rating dataset where the users have rated different movies in a certain scale. The evaluation is usually done in terms of error rate (e.g. Mean Absolute Error (MAE)). However, in many real problems the important question is to detect a particular class. In such scenario, often the dataset is skewed and the class of interest is the minority class. For example, in online advertisement the problem is to find the ads that a user is most likely to click. Here the class of interest is the 'click' which occurs much fewer times than the class 'not click'. Also the standard error rate based evaluation is not appropriate for such problems as it will not properly credit the algorithm that does lesser percentage of mistakes in detecting the class of interest.

Here we observed that in skewed datasets some state of the art, representative collaborative filtering methods are outperformed by a simple probabilistic model defined here. Our test bench includes a real ad click stream dataset that contains click information on a set of ads displayed to a set of users. The dataset is heavily skewed in a sense that very few times the displayed ads got clicked. We then tried to solve a binary prediction problem on the MovieLens movie rating dataset. The MovieLens (http://www.grouplens.org/node/73) dataset is widely used in the literature to evaluate collaborative filtering methods. The problem we posed here is to predict whether a user will give maximum rating to a movie or not. This problem is interesting as there is likely to be a correlation between high rating and the purchase behavior. Surprisingly we observed that our simple model still outperforms the advance collaborative filtering methods although the dataset is much less skewed now.

Earlier Hsu et al observed [4] that basic collaborative filtering or association rule mining do not perform well in skewed transaction datasets. A probabilistic graphical model was proposed as a remedy. As observed here that a state of the art probabilistic graphical model also not able to beat the simple model. In the following sections first we describe the methods we tried and then discuss the datasets and the results.

Copyright is held by the author/owner(s).
WWW 2008, April 21–25, 2008, Beijing, China.
ACM 978-1-60558-085-2/08/04.

2. METHODS
In this section, we briefly discuss the simple probabilistic model and the state of the art collaborative filtering methods we tried here. The methods described below are first trained on a training set containing a set of records where each record is a triplet of user (u), item (y), and rating (r). In our case the rating is binary, click vs. no-click or highest rating vs. not. We call these as positive and negative rating respectively. Once trained, each method can score a user-item pair indicating the likelihood that the user will give positive rating to the item. While testing a method this score is used to sort the user-item pairs in the test set. A ROC curve can be drawn using this sorted list and the ground truths that whether a user-item pair actually got a positive rating or not. ROC curve evaluation is more appropriate than that of accuracy, MAE etc as the cost of making "false positive" and "false negative" are different here.

2.1 Simple Probabilistic Model (SPM)
In this model, the score of a user-item pair is the sum of the probability of the item getting a positive rating and the probability the user gives positive rating to an item.

$$score(u, y) = P(r = 1 \mid y) + P(r = 1 \mid u) \qquad (1)$$

Here $P(r{=}1 \mid y)$ is the ratio of the number of times the item y got positive ratings to the number of records containing the item y. Similarly, $P(r{=}1 \mid u)$ is the portion of the time the user u has given positive ratings out of all his/her ratings.

2.2 Collaborative Filtering Methods
2.2.1 Latent Semantic Model
Successful application of the latent models in collaborative filtering can be found in many recent publications [1] and reported to be used in Google News [2]. Latent models assume the existence of latent class variables in the generation of user-item-rating triplets. For example, the pLSA (probabilistic latent semantic analysis) based latent semantic model [3] used here assumes the users belongs to different communities with certain probabilities. The rating of an item is determined by the communities and it is independent of the user given the community (graphical model is shown in the Figure 1). The

parameters of the model are estimated using EM algorithm by maximizing the likelihood of observing the data (details can be found in [3]). Once the parameters are estimated the score of a user-item pair is computed as given below. Here the z is the latent class variable (community).

$$score(u, y) = P(r = 1 \mid u, y) = \sum_z P(r = 1 \mid y, z) P(z \mid u) \quad (2)$$

One important thing to mention here that when z=1 the score becomes $P(r=1|y)$ as determined in the SPM, because for z=1 $P(z \mid u)=1$ and $P(r=1|y,z)=P(r=1 \mid y)$ in Eq. 2.

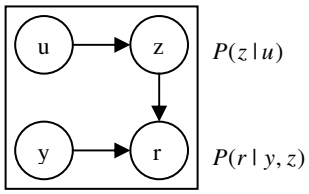

Figure 1 Graphical model of the pLSA model used here

We have tried also the other graphical models and the regularized version of the EM algorithm as proposed in [3] but did not observe much change in the results.

2.2.2 *Nearest Neighbor Algorithms*

Nearest neighbor or memory based algorithms predict the rating of a user-item pair based on the ratings given in the past by the similar users on that item [1]. There is another variation of this algorithm (know as item-to-item C.F.) where it first computes the similarity between the items. The rating of a user-item pair is computed based on the user ratings in the past on other items and the similarity of those items to the given item. We tried both these variations including some different methods of computing similarity between users or items. None of those methods were competitive at all.

3. DATASETS AND RESULTS

Ad click stream dataset was obtained from Komli, an online advertising agency in India. The dataset (available at http://www.komli.com/algogod/) contains 5,391,436 records. The rating here is binary clicked/not clicked. Although the dataset contains some color information about the ad (e.g. text color, background color etc) we did not use those here. There are 869,478 distinct users and 232 distinct ads in the dataset. The skewness of the data is prominent from the fact that the 99.23% entries are corresponding to 'not click'. 98 of the 232 ads have never been clicked by any user. Another interesting fact is that only 5 ads have received 80% of the clicks.

The MovieLens dataset contains 1 million records, 6400 users and 3900 movies. Ratings are made on a 5 star scale. Here we are interested in predicting whether a user-movie pair will have the rating 5 or not. Hence we change the rating to positive or negative based on whether it is equal to 5 or less respectively. There are more than 200 thousand records with positive ratings and 3232 movies that got at least one positive rating. Therefore, this dataset is much less skewed compared to the ad click stream dataset.

Each dataset is randomly split into 70% training and 30% test set. While splitting a dataset we ensure that for the users in the test set there is at least one record in the training set containing that user. From the test set, we then remove those records for which the same use-item pair is observed in a record in the training set. This is done as the ad click stream dataset contains multiple records for the same user-ad pair (same ad is shown multiple times to a user). Removing such records from the test set ensures that the evaluation is done only on the unobserved user-item pairs. The results reported here is the average of 10 random train-test split.

Figure 2 shows the ROC curves of the SPM, the latent semantic model (with z=1 and z=2) and item-to-item C.F. on the two datasets. In both the datasets, the SPM performed the best and the item-to-item C.F. was the worst.

For latent semantic model as we increase the value of z, opposite behaviors are observed on the two datasets. In the ad click stream dataset the area under the curve (AUC) decreases as z increases. For z=1 the latent model performs almost equally to the SPM but recall that for z=1 the latent model is basically $P(r=1|y)$ of SPM. On the other hand in the less skewed MovieLens dataset, the AUC increases as we increase the value of z. This means there is a value in assuming latent factors here. For very high value of z (~150) it almost matches the AUC of the SPM but then as z increases further the AUC starts decreasing. Also, a higher value of z has its own disadvantages in terms of computational complexity and number of parameters.

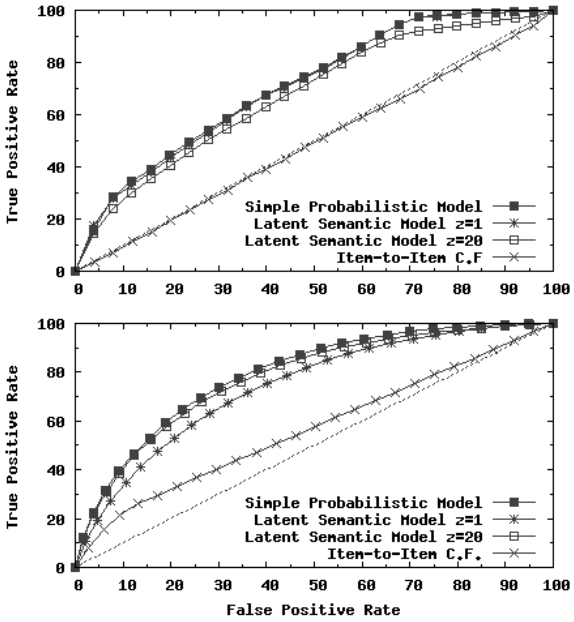

Figure 2 ROC curves for Ad click stream dataset (top) and MovieLens dataset (bottom)

4. CONCLUSION

Here we showed that when the dataset is skewed even the advanced collaborative filtering methods fail to beat a simple probabilistic model. We observed that this conclusion holds even in the popular test bench used in collaborative filtering literature when the problem is framed in a different but realistic way. We believe that the findings here will be useful for the researchers and practitioners in this area.

5. REFERENCES

[1] Adomavicius, G. and Tuzhilin, A. Toward the Next Generation of Recommender Systems: A Survey… IEEE Trans on KDE, 2005, Vol 17, 734-749.

[2] Das, A. Datar, M. Garg, A. and Rajaram, S. Google News Personalization: Online Collaborative Filtering. WWW, 2007.

[3] Hofmann, T. Latent Semantic Models for Collaborative Filtering. ACM TOIS, 2004, Vol 22, 89-115.

[4] Hsu, C.N. Chung, H.H. and Huang, H.S. Mining Skewed and Sparse Transaction Data for Personalized Shopping Recommendation. Machine Learning, Vol 57, 2004, 35-59.