Etimating Movie Rating From User Rating with Demographic Information

Jun Zhang, Galina Meyer, Yiming Sun

December 7, 2015

Abstract

In an expansive multi-billion international film industry, online reviews inform audiences which films may be the most appealing, determining in part the success a film or franchise. Online reviewers and even review aggregations are biased by social or financial factors; this bias clouds the value of movie reviews by prioritizing and propagating highly visible user's ratings of films, as compared to an ideal movie review independent of the individual's social connections. This paper seeks to create a movie rating prediction model that minimizing the social bias of online movie ratings, while providing metrics that can adjust movie satisfaction ratings according to a user's demographic and preference; these two approaches capture an effort to comprehensively capture the social impact on movie ratings. To compensate for the influence of hyper-visible user's views on a film, we leverage user-movie rating matrix completion to penalize the weight of similar and related users. Then, to capture information on the social networks of online movie critics, to predict a demographic's future reviews of a film, we use social network features such as influential factors, centrality, and more. Finally, we establish a statistical inference model that builds a rating for a film theoretically isolated from social bias.

1 Introduction

The film industry is large and powerful. More than an engine of incredible wealth, film communicates and reproduces cultural norms through narratives. Responses to various films have been motivated by social factors since the first feature length film: a pro-Ku Klux Klan film named A Birth of a Nation, which was lauded by the white intellectual elite for its advancements in art but is now wholly criticized for its content. Parallel reactions to films have continued today, with polarization surrounding films such as Stonewall, Da Vinci Code, Roman Polanski films, and so on; of course, the responses aren't split in two by in many fragments, but many.

Capturing information about movie response is currently one-dimensional. Characterizing relevant underlying social structures can be infinitely complex, but here we focus on two aspects: how a movie rating is impacted by influential similar users (community) and how different users in social clusters would likely influence others' judgment in different level (centrality). The methodology described here differs from Netflix-based recommendation engines because they are not inferred from similar users by movie taste, but by social connections and community.

2 Review of Prior Work

Rating prediction and analysis of users' habit have became a major tool for helping companies to realize the demand of consumers, then effectively improve the services they provided. Accordingly, researches in related area drew our attention for better understanding the method to solve these problems and the idea to predict the result more precisely.

2.1 Matrix Completion

It is well known that we could treat rating system as two dimension matrix with dimensions as user and movie, and the aim is to minimize a certain norm after filling the "missing data", in other words, we got the prediction of rating for all the movies from each person. In [1], Asendorf et al. discussed various ways to solve this optimization problem with low rank constraints, which provides detailed discussion on various setting of low rank approximation, especially some ways to do variable selection based on singular value threshold. These methods could avoid over-fitting and lead to robustness. However, their discussion is on a very mature topic, and all the theory in fact heavily based on the sparse structure which may not be satisfied by reality.

2.2 Linear Regression

In [2], Armstrong and Yoon employed very classic linear regression techniques, including kernel regression, step-wise forward search, ANOVA table to predict the movie rating. The focus is on feature selection by greedily minimize the p value and residuals.The advantage for this work is that it is very simple to implement and crystally clear for understanding. All the decisions made are based on rigorous classic statistic testing like F test. However, firstly it failed to consider the correlation of each user or movie which could be analyzed by social network technique. Secondly, those standard statistical tests may fail in big data setup. Especially the greedy search based p value may be problematic for large data set.

Based on previous two works, we think borrowing strength from the social network could improve the matrix completion prediction and simply linear regression rating estimation, since people with similar taste on movie may give each other a good predictor. Also, we should give weights for each reviewers, which may also be conducted based on their social network backgrounds, for example, assigning larger weights to those who are more reliable (actively rated more movies than others and shared properties with large portion of population). Therefore, we went through more works about social network technology.

2.3 Social Network Analysis

In [3] and [4], Yang et al. and Ma et al. both incorporate social science into the matrix completion. For example, Yang et al. added a penalty function to ensure people who are close to each other in their social network also make similar ratings. As such, they assigned weights to each user's rating based on their social influence. However, neither of them presented a convincing measure of objective "correctness", which means their methodology was limited to demonstrating a possible method for predicting movie ratings, rather than testing and measuring how precisely this model could predict an underlying "real" rating for movies.

To avoid the same pitfall, this paper predicts the rating based on sets of training data, which includes the rater's demographic information, while testing our model with another sets of data, which are authority ratings for the film industry.

3 Methodology

3.1 Establishing the Social Network and Detecting Community

In the first step, we incorporate demographic information and users' rating history into estimating how a user may be biased by their local social system. As a primary step, we establish the social networks for users based on their connections, where distance is indicated by shared demographic properties and common movies rated. Demographics here capture location (by city), occupation, age groups, and more; this is our best attempt at describing where a user may be located in the digital and social landscape. Naturally, users' rating history showed their interest in different types of movies.

If users share same or similar properties (a subset of the vector that represents the user), a weighted edge is assigned between them. Then the community detection algorithm, called Louvain community detection algorithm, would run over this set of weights to form several communities of users.

$$Q = \frac{1}{2m}\sum_{i,j}[A_{ij} - \frac{k_i k_j}{2m}]\delta(c_i c_j)$$

where A_{ij} represents the weight of the edge between i and j, $k_i = \sum_j A_i j$ is the sum of the weights of the edges attached to vertex i, c_i is the community to which vertex i is assigned, the δ-function $\delta(u, v)$ is 1 if $u = v$ and 0 otherwise and $m = \frac{1}{2}\sum_{ij} A_{ij}$.

Fig 1 is the visualization for the resultant community detectioned by user demographic network, and the community information would be utilized in the matrix completion algorithm.

3.2 Matrix Completion with Demographic Similarity Regularization

Since we believe that social network information could improve the missing rating filling procedure, we implement the following model to capture that information. The usual matrix completion is that we try to find a low rank approximation to the user rating matrix:

$$\min_{U,V} \frac{1}{2}\sum_{i\in I}\sum_{j\in J}(R(i,j) - U^T V(i,j))^2$$

where $R \in \mathcal{R}^{m\times n}, U \in \mathcal{R}^{\ell\times m}, V \in \mathcal{R}^{\ell\times n}$ are two users vectors, and set I, J represent the non-zero index: $R(i,j) \neq 0, i \in I, j \in J$. To overcome over-fitting, we also add two penalty function U, V, i.e.,

$$\min_{U,V} \frac{1}{2}\sum_{i\in I}\sum_{j\in J}(R(i,j) - U^T V(i,j))^2 + \frac{\lambda_1}{2}\|U\|_F + \frac{\lambda_2}{2}\|V\|_F$$

Since we have obtained community information from the previous section, we add similarity regularization derived from the prior model into our matrix-completion. We require column U_i

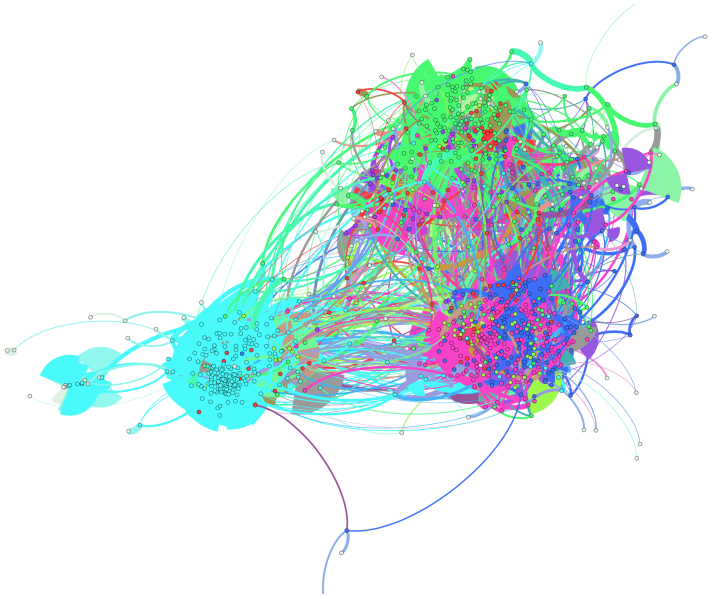

Figure 1: Community Detection for User Demographic Network (used as similarity penalty in matrix completion)

from U to be similar to U_j if i, j. Mathematically,

$$\min_{U,V} J(u,v) = \frac{1}{2}\sum_{i\in I}\sum_{j\in J}(R(i,j) - U^T V(i,j))^2 + \frac{\lambda_1}{2}\|U\|_F + \frac{\lambda_2}{2}\|V\|_F + \frac{\lambda}{2}\sum_{i\in C}\sum_{j\in C}\|U_i - U_j\|_2^2 \qquad (1)$$

where $(i,j) \in C$ means the ith, jth user belongs to the same community detected by the first section. At this point, we have a description of user communities.

3.3 How to Numerically Calculate Matrix Completion

A complicated optimization problem, we resort to gradient descent to solve the problem of calculating the matrix completion. For each column U_i,

$$\bigtriangledown_{U_i} J(u,v) = \sum_k (R_{i,k} - U_i^T V_k)V(k) - \lambda_1 U_i - \lambda \sum_{(i,j)\in C}(U_i - U_j)U_j$$

The first order method is efficient especially when the matrix is big. Similarly, we find the gradient for vector V_i. We run a simple gradient search by:

$$u^{(i+1)} = u^{(i)} - \eta \bigtriangledown_{U_i} J(u,v)$$

3.4 Building an Inference Model with Social Network Features

After fulfilling the user-rating matrix, we have obtained an estimated rating for each user for each movie. Now we want to combine all the user information to infer a "true rating." We utilize a

simple linear regression,

$$\min_{w} \sum_{j} \left(R_j - \sum_{i} w_i R_{i,j} \right)^2$$

where R_j is the "true rating" for movie j, and $R_{i,j}$ is the rating of ith user for movie j. w_i is the weight for each user based on their estimated status in the social network. The problem for this simple model is that there are too many weights to estimated relative to the size of the data. Therefore, we will express the weight vector as a linear combination of standardized pagerank, degree centrality, closeness centrality, betweenness centrality and some potential other social network features. Let w_i be each feature, then above model becomes:

$$\min_{\alpha} \sum_{j} \left(R_j - \sum_{i} \sum_{k} \alpha_k w_{ik} R_{i,j} \right)^2$$

We are faced with a simple least square problem to estimate the α. And following basics statistics principles,

$$\hat{\alpha} = (W^t R R^t W)^{-1} W^t R y.$$

Algorithm 1: Penalized Rating Estimation with Rating Matrix Completion Given Demographic Information

1. Build social network based on demographic information and the rating history;
2. Detect the communities for users, and calculate the centrality values for each user in this network;
3. Run gradient descent to solve matrix completion with similarity penalty based on shared communities;
4. Extract the penalized feature for each user, using pagerank, degree centrality, betweenness centrality and closeness centrality;
5. Fit linear regression for predicting the true rating corresponding to weighted average based on personalized feature;

4 Experiments on Real Data

4.1 Extract Data

We've used quotes around the term "true rating," since there is no way to authoritatively find this information; here we make our best estimate. To test our predicting model, we used the Internet Movie Database (IMDb) as our test data, due to its high volume of information and, more importantly, their weighted published rate, which allows our estimation to avoid flawed explanations compared to ratings only collected from users.

We extracted relevant data from "movielens.org" to build our training data set. This database not only included movie ratings data, but also users' demographic information and rating history. The data we collected includes over 6000 users and more than 4000 movies, of which we chose 2000 most-rated movies rated in the user pool; each user rated at least 20 movies in this list. This

rating frequency limitation serves to avoid using an extremely sparce matrix, which would propogate larger errors in the matrix completion method. Additionally, as a film was more frequently rated, the more meaningful the social networks between the ratings' creators. The result was a 6000*2000 user-rating matrix.

To constrast with these individual user's ratings, we extracted film's general ratings of these 2000 selected movies from IMDb as the test data.

4.2 Extract Features

In order to improve the sparse matrix completion method and better estimate ratings, we examined some social features that may impact a user's judgment of a film. We extracted select features to help us build a relevant social network capturing that information.

First, we weighed most heavily the number of movies users rated with a similar score. Second, age is an important social factor that informs a user's experience of a film or other product; for example, movie producers' first description of their target audience is an age range (e.g. 16-25 year olds). Finally, this data was augmented with occupation, location and gender information, but only if there already existed a connection between two users based on the previous two features. Of course, there are many other factors that may be relevant such as race, socioeconomic status, English fluency, and more, but none of these were represented in our database.

The network based on these criterion creates a model that can measure the influence of each individual in the network by calculate their pagerank, degree centrality, closeness centrality and betweenness centrality. Then the outcomes are respectively used as weighted factors w_i in our prediction model.

4.3 Training the model

Based on the algorithm presented in part 3, the first step in the algorithm is matrix completion, where a stochastic gradient search is run. The tuning parameter λ_i and the approximated rank are determined using K-fold cross validation.

The second step, estimating coefficients in the inference model given the weighted vector, is just a simple least square problem.

$$\hat{\alpha} = (W^t R R^t W)^{-1} W^t R y$$

Again, we use a cross-validation estimator to avoid overfitting.

5 Evaluation of the Model

5.1 Rating Matrix Completion with Social Similarity Penalty

Figure 2 is the error graph (sampled by the first 200 movies for clear visualization), displaying the residues between test data and the average rating calculated by the completed matrix based on similarity regularity (blue points), and the residues between the test data and the average rating calculated without the similarity regularity (yellow points). The latter residues are based on the method presented by [1] Asendorf et al. Errors with similarity regularity are more stable, which indicates the method we proposed is more robust. This is also evident in a histogram comparing

these two errors datasets. As before, the yellow histogram is for completion without similarity penalty, while the blue one is for the one with similarity penalty. Completion with similarity penalty makes the result more stable, but not universally better than the one without.

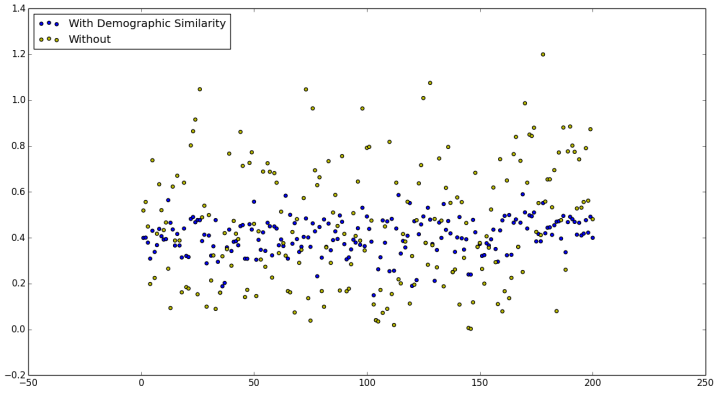

Figure 2: Error Graph (With Similarity vs Without Similarity)

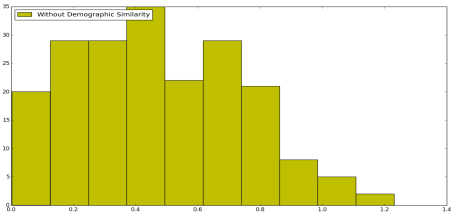

Figure 3: Without Similarity Penalty

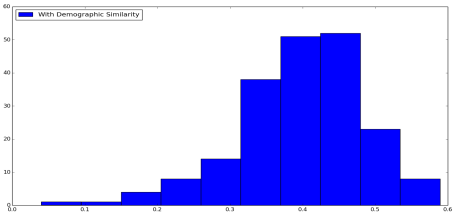

Figure 4: With Similarity Penalty

5.2 Regression Based on Personalized Weighting Rating

As previously outlined, a linear regression is run on demographic features as pagerank, degree centrality, betweenness centrality and closeness centrality. Here we presented the ANOVA table for this problem:

Coefficients:	Estimate	Std.Error	t	value Pr(>|t|)
(Intercept)	0.005806	0.001053	5.516	4.42e-08 ***
PageRank	2.146481	0.089829	23.895	< 2e-16 ***
Deg_Cen	0.115251	0.053078	2.171	0.03014 *
Clo_Cen	-0.013138	0.005447	-2.412	0.01604 *
Bet_Cen	0.151372	0.051827	2.921	0.00357 **

Through the above table, we could observe that these features are strongly correlated, especially with strong degree centrality–closeness centrality and betweenness centrality are not significant. Therefore, through the forward regression method, we think it is appropriate to only insert two features in our model: page rank and degree centrality.

The outcome is tested after a regression on two predictors and compared to the one without linear regression. Below is the figure comparing the errors, also sampled the first 200 movies.

In figure 5, improvement of the model based on personalized weighting is not clear, so it's then

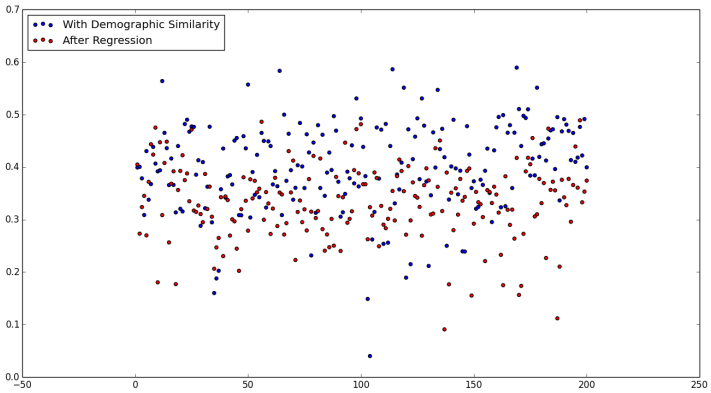

Figure 5: Error Graph (after regression vs before regression)

tested with the statistical technology of hypothesis testing.

First, the average error of the prediction, mean(X_1), before the linear regression is 0.399492 and the one after the linear regression, mean(X_2), is 0.3138296. Assume the Null Hypothesis is X_1=X_2 and run a two sample z-test. The calculated zeta equals 3.153572, which is greater than the value of the critical value zeta tabulated for α equal to 0.05. Therefore, we can reject the Null Hypothesis of $X_1 = X_2$ and say that the linear regression model improved our prediction statistically.

6 Conclusion and Future Work

In this paper, we presented a relatively new method to incorporate social analysis technology into some existing movie rating models, based on the potential connection between these two entities.

Although Demographic similarity penalty failed to meet our expectation of reducing the errors significantly, it did reduce the variance of estimation of matrix completion, which proved the robustness of our model. The major reason for this phenomena could be that the similarity regularization based on community factor converged the rating from people shared common community.

Also, the personalized rating based on the feature extracted from social network slightly improved our prediction statistically based on the two-sample test. This finding displayed the existence of influence from active users in rating system, which means people would follow the "authority" in their respect groups.

Future work may focus on more than weighing user features by also including information on connections between movies (actors, directors, producers), because these artists share similar audiences and visions. We believe that the feature extracted from the movies could further improve related predicting model.

References

[1] Nick Asendorf, Madison McGaffin, Matt Prelee, Ben Schwartz. *Algorithms for Completing a User Ratings Matrix.*
URL://web.eecs.umich.edu/cscott/eecs545f11/projects/AsendorfMcgaffinPressSchwartz.pdf

[2] Nick Armstrong, Kevin Yoon. *Movie Rating Prediction.*
URL://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.1964.rep=rep1.type=pdf

[3] Xiwang Yang, Yang Guo, Yong Liu, Harald Steck. *A survey of collaborative filtering based social recommender systems.*
URL://www.sciencedirect.com/science/article/pii/S0140366413001722

[4] Hao Ma, Dengyong Zhou, Chao Liu. *Recommender Systems with Social Regularization.*
URL://research.microsoft.com/en-us/um/people/denzho/papers/RSR.pdf

Individual Contribution:

Jun Zhang: Crawling the data, Coding up the algorithm, Running tests, Plotting graphs during data analysis, Writing the paragraph of extract data, feature, evaluation of model and conclusion, Revision, Make the poster.

Yiming Sun: Coming up with the algorithm, Running the linear regression tests, Writing the paragraph of methodology, experiment.

Galina Meyer: Coming up with the topic, Writing the paragraph of introduction and prior work review, Final revision