Collaborative Filtering on Amazon Product Rating Prediction

100060001　郁安安

Problem Description

- **Sentiment Analysis of User Comments for One-Class Collaborative Filtering over TED talks. SIGIR 2013**

predict if a user would “like” or “unlike” a TED talk by collaborative filtering, perform SANN analysis on text comments for missing ground-truth labels.

- **Differences**

1. Ratings in scores (1.0~5.0)
2. No ambiguity of unliking or neutral
3. Weird model with bad performance

Model

$$\hat{r}_{ui} = b_{ui} + \frac{\sum_{j \in D^k(u;i)} d_{ij}(r_{uj} - b_{uj})}{\sum_{j \in D^k(u;i)} d_{ij}}$$

$$b_{ui} = \mu + b_u + b_i$$

$$d_{ij} = s_{ij} \frac{n_{ij}}{n_{ij} + \lambda};$$

Baseline estimates $\hat{r}_{ui} = b_{ui}$

Paper:

$$\mu = \frac{1}{|I|} \sum_{i \in I} \frac{r_i}{r_{max}};\ b_u = \frac{r_u}{|I|};\ b_i = \frac{r_i}{r_{max}}$$

Propose:

$$\min_{b_*} \sum_{(u,i) \in \mathcal{K}} (r_{ui} - \mu - b_u - b_i)^2 + \lambda_1 (\sum_u b_u^2 + \sum_i b_i^2)$$

Solve Regression with std process

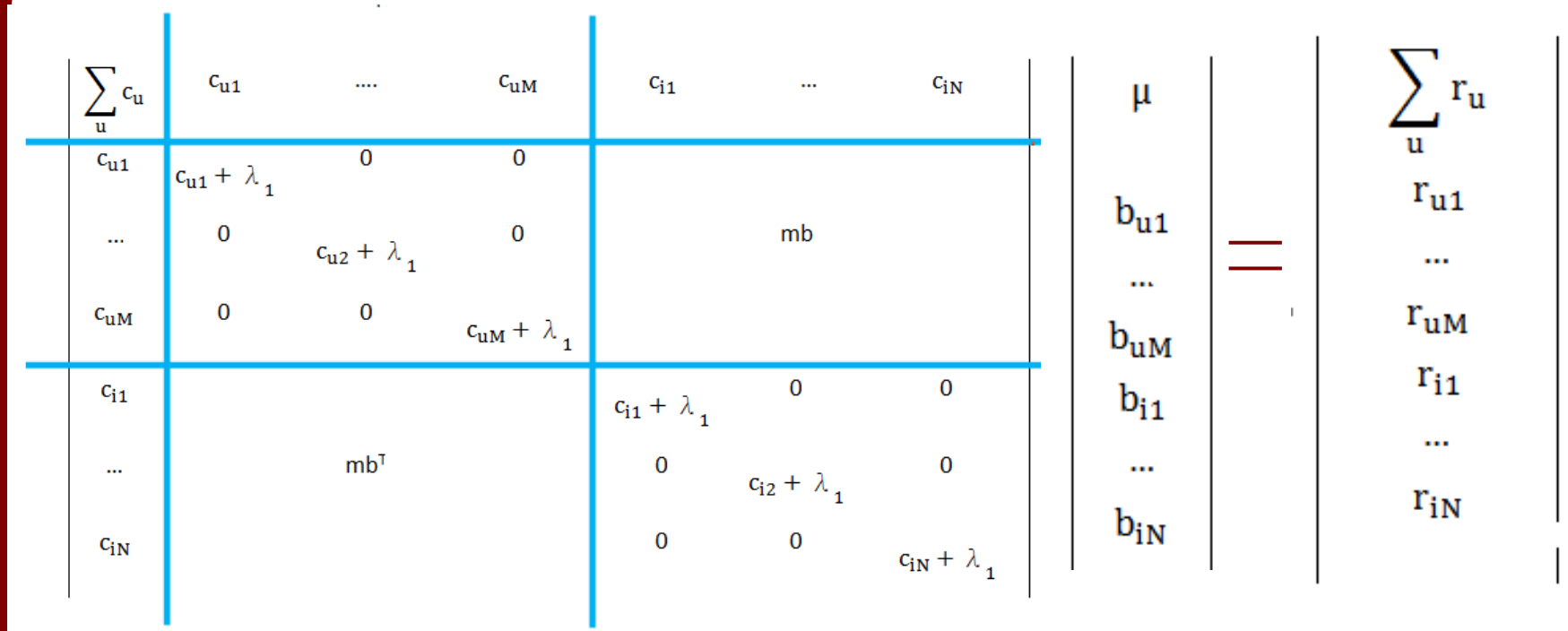
$$f = \sum_{u} (r_{ui} - \mu - b_u - b_i)^2 + \lambda_1 \left(\sum_{u} {b_n}^2 + \sum_{i} {b_i}^2 \right)$$

$$\frac{\partial f}{\partial \mu} = 0$$

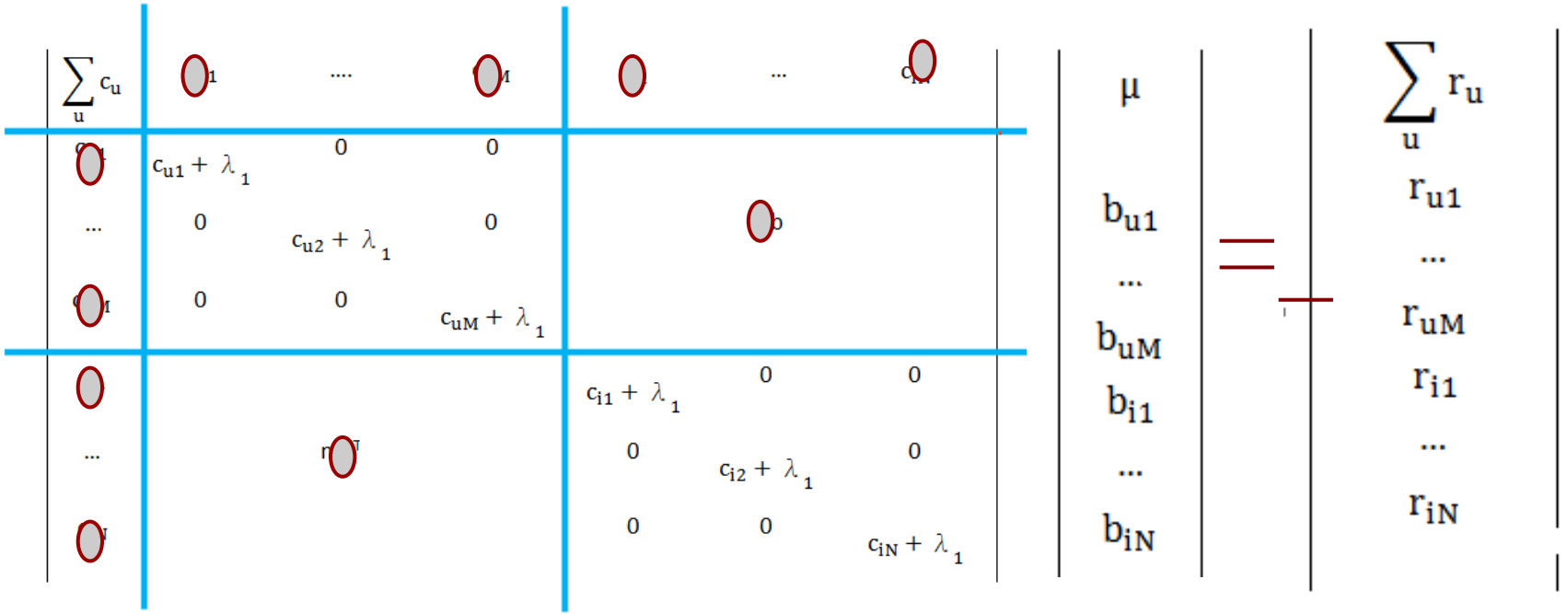
$$\frac{\partial f}{\partial b_{u1}} = 0 \ \ldots\ \frac{\partial f}{\partial b_{uM}} = 0$$

$$\frac{\partial f}{\partial b_{i1}} = 0 \ \ldots\ \frac{\partial f}{\partial b_{iN}} = 0$$

Obtain Matrix



G_G


$\sum_u c_u$
$c_{u1} + \lambda_1$
$c_{u2} + \lambda_1$
$c_{uM} + \lambda_1$
$c_{i1} + \lambda_1$
$c_{i2} + \lambda_1$
$c_{iN} + \lambda_1$
μ
b_{u1}
b_{uM}
b_{i1}
b_{iN}
$\sum_u r_u$
r_{u1}
r_{uM}
r_{i1}
r_{iN}

Results

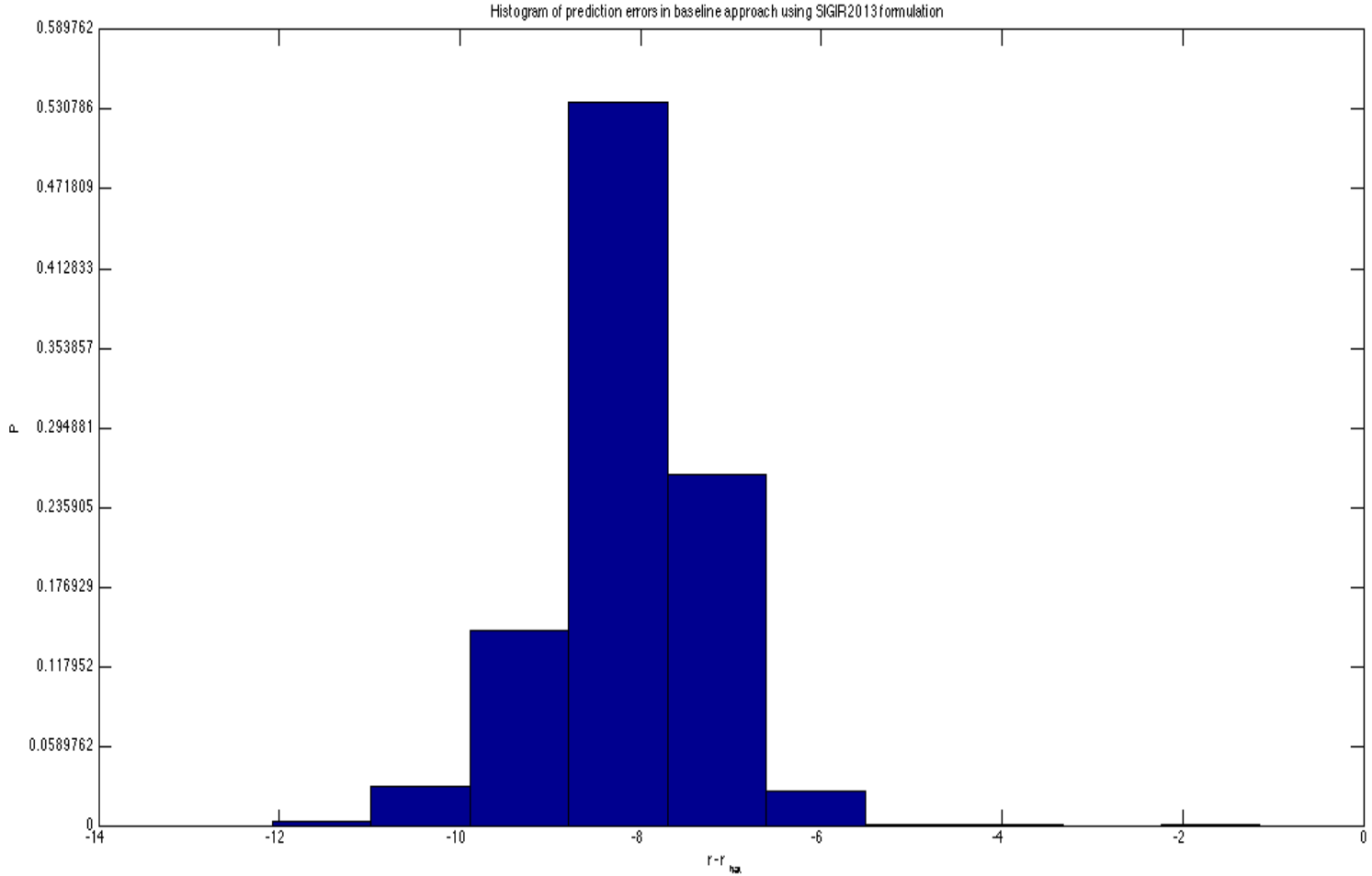
My bui:

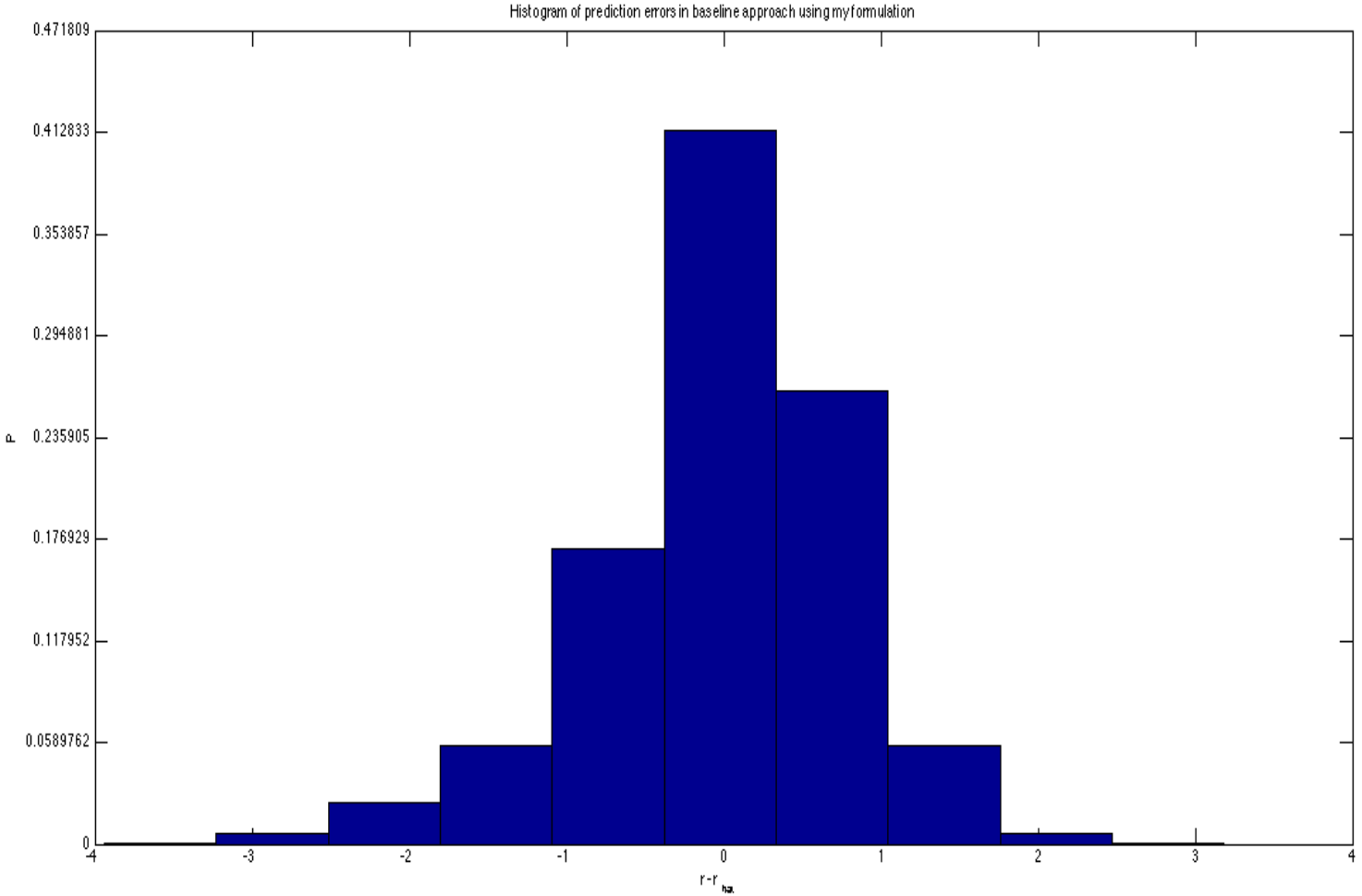
MSE = 0.616018723464

paper's bui:

MSE = 66.7697037634

Evaluation:
randomly partition a dataset of 42183 entries into 5 sets.
choose 4 sets as training data, 1 as testing set


Histogram of prediction errors in baseline approach using SIGIR2013 formulation
0.589762
0.530786
0.471809
0.412833
0.353857
0.294881
0.235905
0.176929
0.117952
0.0589762
0
P
-14
-12
-10
-8
-6
-4
-2
0
r - r hat


Histogram of prediction errors in baseline approach using my formulation
0.471809
0.412833
0.353857
0.294881
0.235905
0.176929
0.117952
0.0589762
0
p
-4
-3
-2
-1
0
1
2
3
4
r - r hat

Neighborhood Model

$$\hat{r}_{ui} = b_{ui} + \frac{\sum_{j \in D^k(u;i)} d_{ij}(r_{uj} - b_{uj})}{\sum_{j \in D^k(u;i)} d_{ij}}$$

$$b_{ui} = \mu + b_u + b_i$$

$$d_{ij} = s_{ij} \frac{n_{ij}}{n_{ij} + \lambda};$$

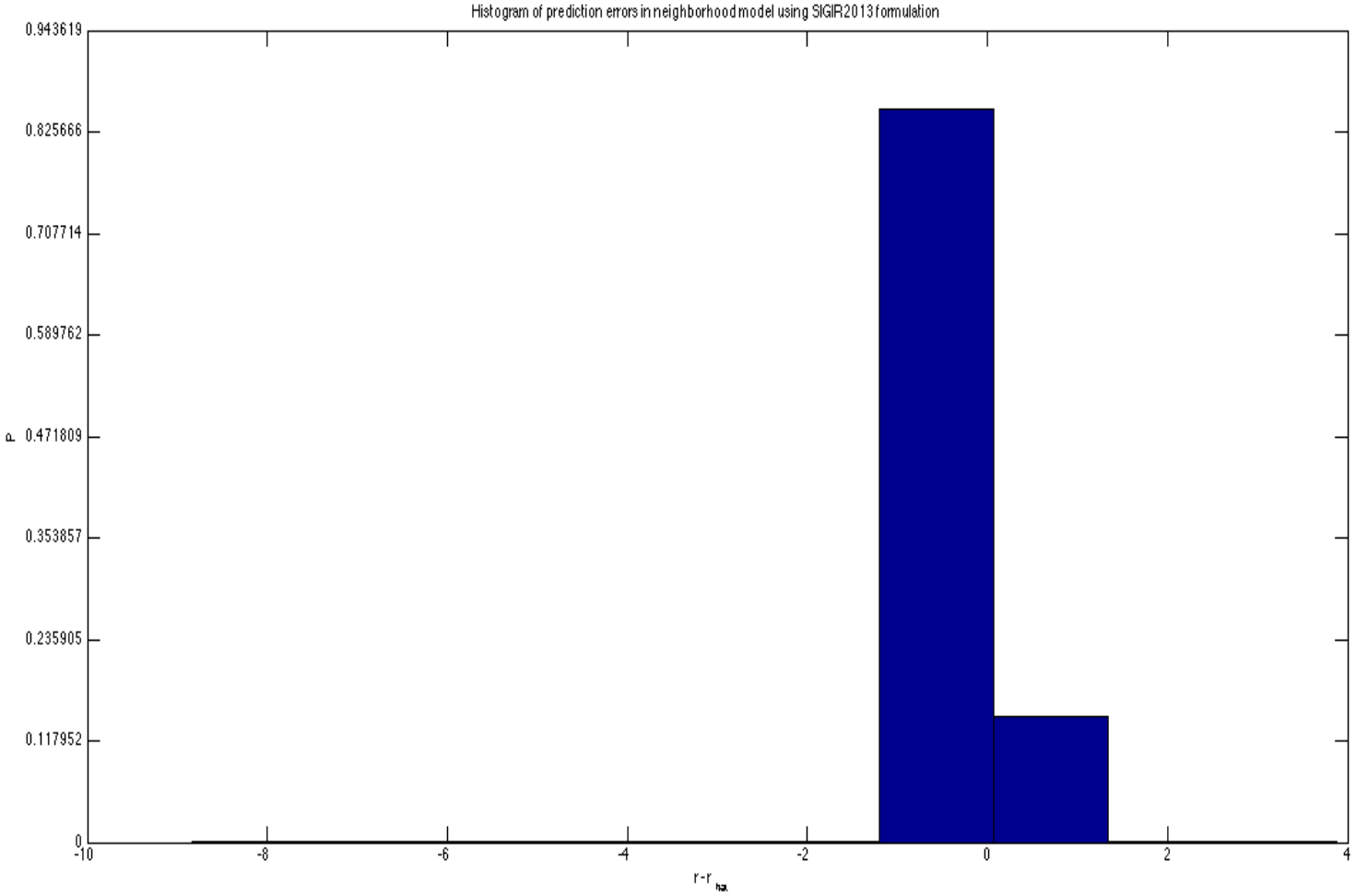
Results

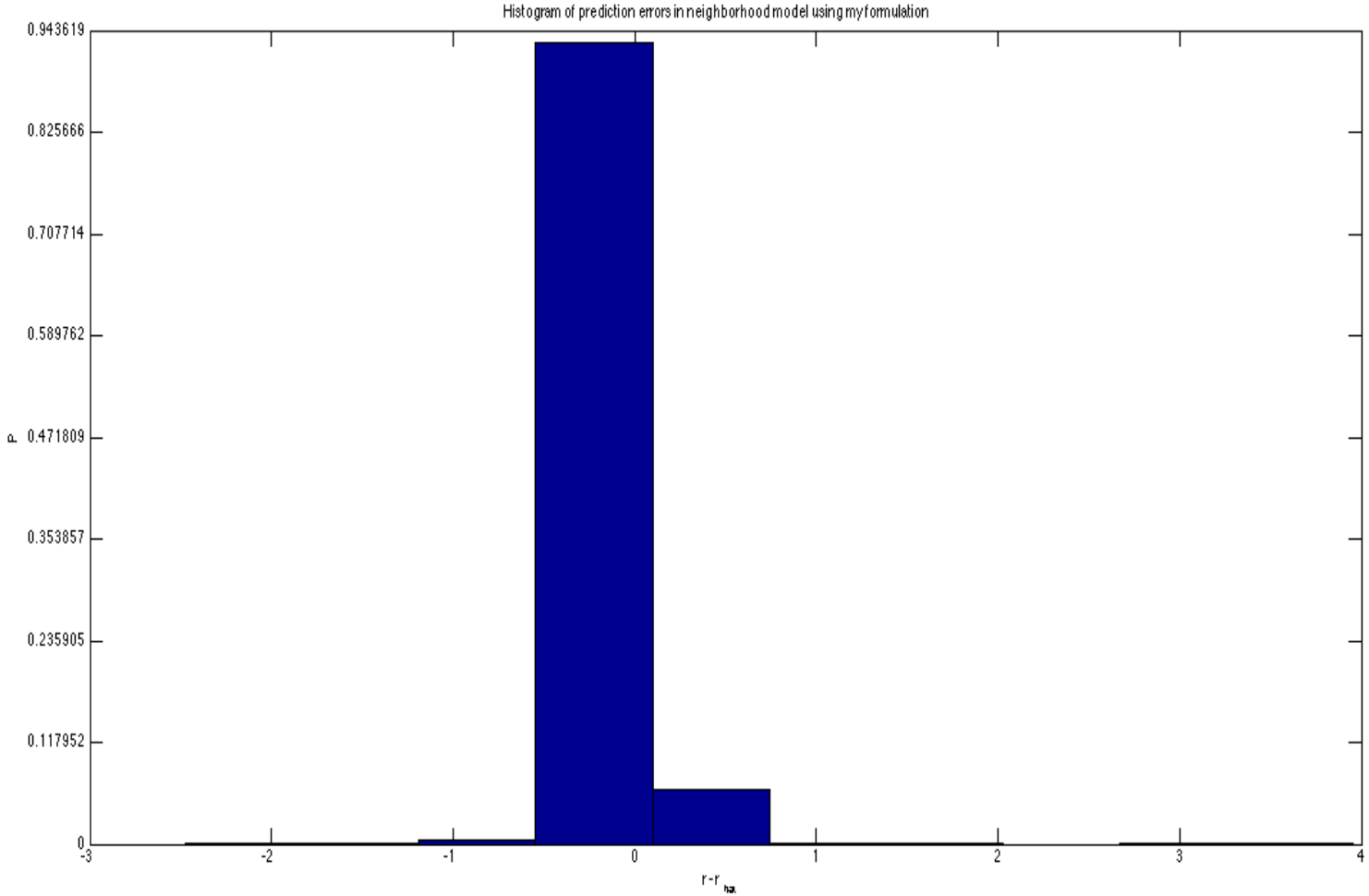
My bui:

MSE = 0.0187463518511

paper's bui:

MSE = 0.0648160577155


Histogram of prediction errors in neighborhood model using SIGIR2013 formulation
0.943619
0.825666
0.707714
0.589762
0.471809
0.353857
0.235905
0.117952
0
p
-10
-8
-6
-4
-2
0
2
4
r - r hat


Histogram of prediction errors in neighborhood model using my formulation
0.943619
0.825666
0.707714
0.589762
0.471809
0.353857
0.235905
0.117952
0
P
-3
-2
-1
0
1
2
3
4
r - r max