The Comparison of Recommendation Approaches in MovieLens

Lichao Han
Ho Wai Lok
Outline

- Introduction
- Similarity Measures
- Implementation
- Results
- Reference
Introduction

- Recommender System (RS)
- Provides personalized suggestions for products or items to individual users
- Content-based VS Collaborating filtering (CF)
- Neighborhood based CF
Similarity Measures

- Jaccard Cosine • traditional
- PIP NHSM • co-rated
- Bhattacharyya Coefficient in CF (BCF) • not co-rated
Jaccard

\[
S(U, V)^{\text{Jacc}} = \frac{\left| I_U \cap I_V \right|}{\left| I_U \cup I_V \right|}
\]

- \( I_U \) is the set of items rated by user \( U \)
Cosine

\[
s(U, V)^{Cos} = \frac{(r_{UI})(r_{VI})}{\sqrt{\sum_{I \in I'} r^2_{UI} \sum_{I' \in I'} r^2_{VI}}}
\]

- \( r_{UI} \) is the rating made by user \( U \) on item \( I \) and \( I' \) is the set of co-rated items
### Jaccard & Cosine

<table>
<thead>
<tr>
<th>Similarity Measures</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaccard</td>
<td>Easy to compute</td>
<td>Suffers from few co-rated item problem; Does not take absolute value (rating) into account; Uses local information of the ratings only</td>
</tr>
<tr>
<td>Cosine</td>
<td>Easy to compute</td>
<td>Suffers from few co-rated item problem; Outputs high similarity in spite of significant difference in ratings; Uses local information of the ratings only</td>
</tr>
</tbody>
</table>
\[ s(U, V)^{PIP} = \sum_{p \in I} PIP(r_{U,P}, r_{V,P}) \]

where \( PIP(r_{U,P}, r_{V,P}) = \text{Proximity}(r_{U,P}, r_{V,P}) \times \text{Impact}(r_{U,P}, r_{V,P}) \times \text{Popularity}(r_{U,P}, r_{V,P}) \)
$Agreement(r_1, r_2) \begin{cases} 
\text{false} & \text{if} \ (r_1 > R_{\text{med}} \text{ and } r_2 < R_{\text{med}}) \text{ or} \\
& (r_1 < R_{\text{med}} \text{ and } r_2 > R_{\text{med}}) \\
\text{true} & \text{otherwise} 
\end{cases}$

Where $R_{\text{med}} = \frac{R_{\text{max}} + R_{\text{min}}}{2}$
$$Proximity(r_1,r_2) = \left\{ \begin{array}{ll} 2 \times (R_{\text{max}} - R_{\text{min}}) + 1 \right\} D(r_1,r_2)^2 \\
\text{where } D(r_1,r_2) = |r_1 - r_2| \text{ if } Agreement(r_1,r_2) \text{ is true} \\
= 2|r_1 - r_2| \text{ if } Agreement(r_1,r_2) \text{ is false} \end{array} \right.$$
if $Agreement(r_1, r_2)$ is true

$$Impact(r_1, r_2) = \left( |r_1 - R_{med}| + 1 \right) \left( |r_2 - R_{med}| + 1 \right)$$

if $Agreement(r_1, r_2)$ is false

$$Impact(r_1, r_2) = \frac{1}{\left( |r_1 - R_{med}| + 1 \right) \left( |r_2 - R_{med}| + 1 \right)}$$
if ($r_1 > k$ and $r_2 > k$) or ($r_1 < k$ and $r_2 < k$)

$Popularity(r_1, r_2) = 1 + \left( \frac{r_1 + r_2}{2} \right)^2$

otherwise

$Popularity(r_1, r_2) = 1$
\[ s(U, V)^{NHSM} = s(U, V)^{JPSS} \times s(U, V)^{URP} \], where

\[ s(U, V)^{JPSS} = s(U, V)^{PSS} \times s(U, V)^{Jacc'} \]

\[ s(U, V)^{URP} = 1 \frac{1}{1 + \exp\left( \left| \frac{U}{v_1} \right| \frac{1}{1 + \exp\left( \left| \frac{U}{v_2} \right| \frac{1}{1 + \exp\left( \left| \frac{U}{v_3} \right| \frac{1}{1 + \exp\left( \left| \frac{U}{v_4} \right| \right)} \right)} \right)} \right) \]
\[ s(U,V)^{PSS} = PSS(r_{U,p}, r_{V,p}) \]

where

\[ PSS(r_{U,p}, r_{V,p}) = \text{Proximity}(r_{U,p}, r_{V,p}) \times \text{Significance}(r_{U,p}, r_{V,p}) \times \text{Singularity}(r_{U,p}, r_{V,p}) \]

\[ \text{Proximity}(r_{U,p}, r_{V,p}) = 1 - \frac{1}{1 + \exp(-r_{U,p} - r_{V,p})} \]

\[ \text{Significance}(r_{U,p}, r_{V,p}) = \frac{1}{1 + \exp(-r_{U,p} - r_{med} \times r_{V,p} - r_{med})} \]

\[ \text{Singularity}(r_{U,p}, r_{V,p}) = 1 - \frac{1}{1 + \exp\left(-\frac{r_{U,p} + r_{V,p}}{2}\right)} \]

- \( p \) is the average rating of item \( p \)
- \( r_{U,p} \) is the rating of item \( p \) by user \( u \)
\[ s(U, V)^{JPSS} = s(U, V)^{PSS} \times s(U, V)^{Jacc'} \]

where
\[ s(U, V)^{Jacc'} = \frac{I_U}{I_U} \frac{I_V}{I_V} \]
\[ s(U, V)^{PSS} = \prod_{p} PSS(r_{U,p}, r_{V,p}) \]
$s(U, V)_{URP} = 1 \frac{1}{1 + \exp\left( \frac{u \times s(U) - v \times s(V)}{s(U) \times s(V)} \right)}$

- $u$ and $v$ is the mean rating of user $U$ and $P$ respectively.
- $u$ and $v$ represents the standard variance of user $U$ and $P$. 

NHSM
## PIP & NHSM

<table>
<thead>
<tr>
<th>Similarity Measures</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>PIP</td>
<td>Captures global information of the concerned item</td>
<td>Suffers from few co-rated item problem; Unnecessarily penalize more than once while computing <em>proximity</em> and <em>impact</em> factors</td>
</tr>
<tr>
<td>NHSM</td>
<td>Captures global information of the concerned item</td>
<td>Suffers from few co-rated item problem</td>
</tr>
</tbody>
</table>
Bhattacharyya Coefficient in CF (BCF)

\[ s(U, V)^{BC} = Jacc(U, V) + \sum_{i \in I_U, j \in I_V} BC(i, j) \cdot loc(r_{ui}, r_{vj}) \], where

\[ BC(i, j) = BC(\overline{p_i}, \overline{p_j}) = \prod_{h=1}^{m} \sqrt{p_{ih} \cdot p_{jh}} \]

\[ loc(r_{ui}, r_{vj}) = \frac{(r_{ui} - \overline{r_U})(r_{vj} - \overline{r_V})}{U \cdot V} \]
Bhattacharyya Coefficient in CF (BCF)

\[ BC(i, j) = BC(\widehat{p_i}, \widehat{p_j}) = \sqrt{\left(\frac{\widehat{p_{ih}}}{\#i}\right)\left(\frac{\widehat{p_{jh}}}{\#h}\right)} \]

where \( \widehat{p_{ih}} = \frac{\#h}{\#i} \)

- \( m \) is the number of bins
- \( \#i \) is the number of users rated the item \( i \)
- \( \#h \) is the number of users item \( i \) with rating value ‘\( h \)’
- \( \frac{\sum_{h=1}^{m} \widehat{p_{ih}}}{\sum_{h=1}^{m} \widehat{p_{jh}}} = 1 \)
**Bhattacharyya Coefficient in CF (BCF)**

\[
BC(i, j) = BC(p_i, p_j) = \prod_{h=1}^{m} \sqrt{(p_{ih})(p_{jh})}, \text{ where } \ p_{ih} = \frac{\#h}{\#i}
\]

- **Example:**

\[
I = (1, 0, 2, 0, 1, 0, 2, 0, 3, 0)^T, \quad J = (0, 1, 0, 2, 0, 1, 0, 2, 0, 3)^T
\]

\[
BC(i, j) = \prod_{h=1}^{3} \sqrt{I_{ih} J_{jh}} = \sqrt{\left(\frac{2}{5}\right) \left(\frac{2}{5}\right)} + \sqrt{\left(\frac{2}{5}\right) \left(\frac{2}{5}\right)} + \sqrt{\left(\frac{1}{5}\right) \left(\frac{1}{5}\right)} = 1
\]
### Bhattacharyya Coefficient in CF (BCF)

<table>
<thead>
<tr>
<th>Similarity Measures</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bhattacharyya Coefficient in CF (BCF)</td>
<td>Does not depend on co-rated items; Use both local and global information; Comprehensively use numerical values of the ratings</td>
<td>Relatively difficult to compute</td>
</tr>
</tbody>
</table>
Implementation

• Dataset
  • MovieLens 100K
  • 943 users
  • 1682 movie items
  • 100,000 ratings (scale: 1,2,3,4,5)
  • 80% training data; 20% testing data

• Programming Language
  • Python
Implementation

• Similarity Measures
  • Jaccard
  • Cosine
  • PIP
  • NHSM
  • Bhattacharyya
Rating Prediction

$$\hat{r}_{ui} = \bar{r}_u + \frac{1}{K} \sum_{k=1}^{K} s(U_u, U_k) \cdot \left( r_{ki} \cdot \bar{r}_{ku} \right)$$

- $\bar{r}_u$ is the average of the ratings made by the user $U_u$.
- $s(U_u, U_k)$ denotes the similarity value between user $U_u$ and its kth neighbor.
- $\bar{r}_{ku}$ is the average of the ratings made by kth neighbor of the user $U_u$.
- $r_{ki}$ is the rating made by kth neighbor on ith item.
Mean Absolute Error

\[ MAE = \frac{\sum_{i=1}^{MAX} |r_i - \hat{r}_i|}{MAX} \]

- \( r_i \): actual rating
- \( \hat{r}_i \): predicted rating by a CF algorithm
- \( MAX \): number of times the predictions are performed by a CF algorithm
Results

- Mean Absolute Error
Results

• Root Mean Squared Error

\[ RMSE = \sqrt{\frac{1}{MAX} \sum_{i=1}^{MAX} (r_i - \overline{r_i})^2} \]

\( r_i \): actual rating
\( \overline{r_i} \): predicted rating by a CF algorithm
MAX: number of times the predictions are performed by a CF algorithm
Experiment and Results

- Root Mean Squared Error
References


