A New Similarity Measure Based on Mean Measure of Divergence for Collaborative Filtering in Sparse Environment

Suryakant* and Tripti Mahara

IIT Roorkee, Roorkee 247 667, India

Abstract

Memory based algorithms, often referred to as similarity based Collaborative Filtering (CF) is one of the most popular and successful approaches to provide service recommendations. It provides automated and personalized suggestions to consumers to select variety of products. Typically, the core of similarity based CF which greatly affect the performance of recommendation system is to finding similar users to a target user. Conventional similarity measures like Cosine, Pearson correlation coefficient, Jaccard similarity suffer from accuracy problem under sparse environment. Hence in this paper, we propose a new similarity approach based on Mean Measure of Divergence that takes rating habits of a user into account. The quality of recommendation of proposed approach is analyzed on benchmark datasets: ML 100K, ML-1M and Each Movie for various sparsity levels. The results depict that the proposed similarity measure outperforms existing measures in terms of prediction accuracy.

Keywords: Collaborative Filtering; Information Filtering; Recommendation System; Recommendation Accuracy; Similarity Measure.

1. Introduction

A Recommendation system (RS) is a necessity and a popular technology to handle information explosion. It serves as an information filtering tool and is commonly used to assist the target user to filter through a large pool of products and present only those products that are of user’s interest. It gathers a large amount of data on the activities, inclinations, interest or taste of its client for a set of things i.e. movies, hardware item, garments and so forth and makes utilization of this gathered data to give suggestions to different clients.

In day to day life, we regularly depend on the opinion of like-minded individuals, individuals with comparable taste and preference or other trusted sources about the nature and quality of assets to get the suggestions for a product or item. A RS automates this word-of-mouth phenomenon and is generally utilized by the online shopping sites like Amazon to prescribe products of interest and by sites like Netflix to prescribe films to the users by giving personalization suggestions. The significance of RS increases inconceivably with the presence of long tail phenomenon (Anderson 2006). A physical retail store is characterized by shortage of resources. For instance, a fabric shop has constrained rack space and can show restricted assortment of items to a client. In the physical world it is impractical to
customize a physical retail store for every individual client, as it is either represented by deals figures (most prevalent products) or it relies on upon master judgment. Despite what might be expected, an online retail store can make everything accessible to the client that exists. This peculiarity to makes things accessible between the offline and on-line world has been termed as the long tail phenomenon. As per the long tail, offline retail stores just give the most popular things though on-line stores offer least popular items in addition most famous things. Henceforth, Park and Tuzhilin proposed the need of a RS to prescribe things to individual clients online as it is unrealistic to present every accessible thing to the client or things that match the essence of clients.

Till date, Collaborative filtering (CF) is the most successful and widely employed approach in RS. Bobadilla et al. is based on similar taste of users. It works on the fact that if a user had similar taste in past for a set of items, then they will share common taste in future. The information preferences for some resources are far more complex and hard to determine. It is due to the fact that sometimes preferences can’t be just defined by using a set of keywords or by quality and taste. It can be obtained by observing the preference, behaviour or taste of other users. Tapestry the first RS recommended documents, collected from newsgroup to a set of users by building database of contents and comments.

CF is one of the most widely used approaches to design a RS. However, sparsity is one of the major weaknesses of this prosperous approach. This problem inherently occurs in the system and is attributed to ever increasing number of users and items. Because of this numerous users may have evaluated or bought just couple of items from the total accessible items. Indeed, even extremely well known items may have been bought or evaluated by very few users. Hence it is difficult to compute similarity between users that leads to high sparsity in user-item ratings matrix. This affects the performance of a RS. In case, the system manages to evaluate similarity, their might exist a possibility that this similarity may be not reliable because of insufficient information processed. According to the density of matrix is lower than 1%.

In literature diverse similarity measures have been proposed but their performance is not very satisfactory for sparse matrix. Hence, the main aim of this paper is to build a new approach based on mean measure of divergence to find similarity between users to address the sparsity issue.

2. Related Work

One of the most critical factors that greatly affect the performance of CF is the computation of similarities between users. Cosine (COS), Pearson correlation coefficient (PCC), adjusted cosine measure (ACOS) and Spearman’s rank correlation (SRC) are the generic traditional measures for similarity computation. These similarity measures defined in Table 1 Patra et al.
Table 1. Traditional Similarity Measures.

<table>
<thead>
<tr>
<th>Similarity Measures</th>
<th>Computational formulae’s</th>
<th>Major Drawbacks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson Correlation (PCC)</td>
<td>$\text{sim}(u, u') = \frac{\sum_{i \in I} (r_{u,i} - \overline{r}<em>u)(r</em>{u',i} - \overline{r}<em>{u'})}{\sqrt{\sum</em>{i \in I} (r_{u,i} - \overline{r}<em>u)^2 \sum</em>{i \in I} (r_{u',i} - \overline{r}_{u'})^2}}$</td>
<td>It suffers from few co-rated item problem. It outputs high (low) similarity even if there exists significant difference in ratings.</td>
</tr>
<tr>
<td>Cosine (COS)</td>
<td>$\text{sim}(u, u') = \frac{\sum_{i = 1} {r_{u,i} \cdot r_{u',i}}}{\sqrt{\sum_{i = 1} (r_{u,i})^2 \sum_{i = 1} (r_{u',i})^2}}$</td>
<td>It suffers from few co-rated item problem. It gives high similarity. It outputs high similarity even if there exists significant difference in ratings.</td>
</tr>
<tr>
<td>Adjusted Cosine (ACOS)</td>
<td>$\text{sim}(u, u') = \frac{\sum_{u, u' \in U} (r_{u,i} - \overline{r}<em>u)(r</em>{u',i} - \overline{r}<em>{u'})}{\sqrt{(r</em>{u,i} - \overline{r}<em>u)^2 (r</em>{u',i} - \overline{r}_{u'})^2}}$</td>
<td>Unable to compute similarity if user U have rated only few items. It outputs high (low) similarity even if there exists significant difference in ratings.</td>
</tr>
<tr>
<td>Spearman’s Rank Correlation similarity</td>
<td>$\text{sim}(u, u') = 6 \sum_{i = 1} \frac{\text{rank}(r_{u,i})^2 - \text{rank}(r_{u',i})^2}{</td>
<td>I</td>
</tr>
</tbody>
</table>

Some researchers have also utilized the metaheuristic algorithms to find the similarity. For example, Bobadilla et al.⁴ have adapted Genetic algorithm to compute similarity between a pair of users. It combines values and weights to obtain this similarity. The similarity between each pair of users is obtained by first computing the values and then genetic algorithm is adapted for weights adjustment. The weights in the framework are computed just once.

3. Proposed Similarity Measure: CjacMD

The conventional similarity measures can’t be used in sparse environment, as they don’t utilize ratings data efficiently. Therefore, we propose our similarity measure: CjacMD (Cosine-Jaccard-Mean Measure of Divergence) suitable in sparse dataset. This measure combines Cosine, Jaccard and Mean Measure of Divergence to compute overall similarity.

The cosine similarity measure is a popular similarity measure. It treats each user as a vector in the space of items and uses the cosine between these vectors as a measure of similarity. However, it doesn’t compute the individual’s personal habits to express its preference. Hence, we propose to take into account the personal habits of a user. To achieve this Mean measure of divergence (MMD) is used. MMD¹² is the most common and most popular distance measures used for the computation of bio-distances based on no-metric traits.

Each individual has personal habits to express his preferences. Some user tends to give high or low ratings as compared to others. This biasness in ratings influences the relationship between the users. The traditional similarity neglects this factor. Mean measure of divergence is used to calculate likenesses in users ratings based upon rating habits. In this way, the $\theta_u$ represents a vector based on user $u$ ratings, $|I_u|$ represents a total number of ratings made.
by $u$ and $r$ represents number of co-rated item between $u$ and $u'$. For example, let ratings vector of $u$ and $u'$ are $I_u = (1, 5, 2, 0, 1, 5, 2, 0, 3, 0, 1, 4, 2)$ and $I_{u'} = (2, 4, 2, 1, 0, 0, 5, 2, 2, 0, 5, 5)$ respectively, where $m = 1$ and $M = 5$ (that is to say that ratings lie in $\{1, \ldots, 5\}$). Then $\theta_u$ can be represents as $\theta_u = (\theta_1^m, \ldots, \theta_M^m)$ where each component $\theta_m^m$ represents the number of times the rating value $m$ occurs in rating vector ($I_u$) of user $u$. Therefore $\theta_u$ and $\theta_{u'}$ can be computed as $\theta_u = (3, 2, 1, 2, 2)$ and $\theta_{u'} = (1, 4, 0, 1, 3)$. Therefore, $MMD$ between a pair of users can be computed as Eqs. 1:

$$sim(u, u')^{MMD} = \frac{1}{1 + \left( \frac{1}{r} \sum_{i=1}^{r} \left( \frac{\theta_u - \theta_{u'}}{\sqrt{|I_u| - |I_{u'}|}} \right)^2 \right)}$$

(1)

To give more significance to the co-rated items, we have added Jaccard similarity in our model. The Jaccard similarity between a two users is defined as Eqs. 2:

$$sim(u, u')^{Jaccard} = \frac{|I_u \cap |I_{u'}|}{|I_u \cup |I_{u'}|}$$

(2)

where $|I_u|$ is the cardinality of items rated by user $u$.

Taking cosine, jaccard and Mean measure of divergence based similarity into account we combine these similarities to calculate the final similarity between users. The hybrid user similarity between two users is given by Eqs. 3:

$$sim(u, u')^{CJacMD} = sim(u, u')^{COS} + sim(u, u')^{Jaccard} + sim(u, u')^{MMD}$$

(3)

After the similarities between users have been calculated the next step is to find the top K users with the highest similarity to active user. These users are called neighbour of an active user. In order to predict the ratings of the active user for the unrated items, the following function (Eqs. 4) is used:

$$Pred_{u,i} = \bar{r_u} + \frac{\sum_{u' \in N(u)} sim(u, u') \times (r_{u',i} - \bar{r_u})}{\sum_{u' \in N(u)} sim(u, u')}$$

(4)

where, $sim(u, u')$ indicates the similarity between two users, $\bar{r_u}$ is the mean rating of user $u$, $r_{u,i}$ is the rating given by user $u$ to item $i$ and $N(u)$ is the neighbors of user $u$.

The Algorithm 1 sketches the general scheme of proposed approach that is used make prediction for unknown ratings in user-item rating matrix.

---

1. Split dataset $\mathcal{R}_{train}$ into disjoint training ($\mathcal{Tr}_{train}$) and testing ($\mathcal{Tr}_{test}$) sets
2. for all users $u \in \mathcal{Tr}_{train}$ do
3. \hspace{1em} $sim(u, u')^{Jaccard} \leftarrow$ Compute the Jaccard similarities between user $u$ and $u'$ i.e target user and rest of the user respectively.
4. \hspace{1em} $sim(u, u')^{COS} \leftarrow$ Compute the COS similarities between $u$ and $u'$
5. \hspace{1em} $sim(u, u')^{MMD} \leftarrow$ Compute the MMD similarities between $u$ and $u'$
6. \hspace{1em} $sim(u, u')^{CJacMD} \leftarrow sim(u, u')^{COS} + sim(u, u')^{Jaccard} + sim(u, u')^{MMD}$
7. \hspace{1em} $K \leftarrow$ find the nearest neighbors of the active user $u$
8. \hspace{1em} $Pred_{u,i} \leftarrow$ Compute UserBased Prediction
9. end for
10. Evaluate RMSE and MAE of predicted ratings for $\mathcal{Tr}_{test}$

Algorithm 1. $CJacMD$ Algorithm
4. Experiments

4.1 Data sets

To evaluate the RS approach, a large number of benchmark datasets are publicly available. We evaluate the proposed model with three well known real-life datasets (i) MovieLens 100 K (ML-100 K). The ML-100K dataset, consists of 100,000 ratings on 1682 items (movies) made by 943 users. In this dataset, the sparsity rating is 93.69%, which means that 4.25% of the movie has been rated by users. (ii) MovieLens 1 M (ML-1 M), consists of approximately one million ratings from 6,040 users who reviewed 3,952 movies. The sparsity rating of ML-1 M is about 95.75%. In both the MovieLens dataset, each movie is rated on a scale from 1 to 5 and (iii) the each movie dataset where we extracted a subset of 1,004 users who reviewed 1,091 movies with rating scale from 1 to 6.

4.2 Evaluation measures

The effectiveness and accuracy of rating predictions of our proposed similarity measure and other traditional similarity measures was evaluated using most popular evaluation metrics namely Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

The Mean Absolute Error (MAE) is defined as Eqs. 5:

$$MAE = \frac{1}{|S|} \sum_{i=1}^{S} |Pred_i - r_i|$$

where $Pred_i$ is the predicted rating for a movie, $r_i$ the actual rating, and $|S|$ is the cardinality of the test ratings.

The RMSE is defined as Eqs. 6:

$$RMSE = \sqrt{\frac{1}{|S|} \sum_{i=1}^{S} (Pred_i - r_i)^2}$$

4.3 Experimental result and analysis

Assume $K$ denotes the number of nearest neighbours. In CF, the performance of recommendation can be greatly affected by the number of nearest neighbours ($K$). In order to examine the sensitivity of the neighbourhood size, we perform our experiment with different number of nearest neighbours. It is observed that the different number of neighbours ($K$) will give different predictions and accuracy. To prove the effectiveness of the proposed CJacMMD method, it is compared with four different similarity measures i.e. COS, PCC, ACOS and SRC.

![Fig. 1. Performance Comparison of Different Approaches on ML-100 K Dataset: (a) MAE; and (b) RMSE.](image-url)
Figure 1 illustrates the performance of different similarity method with varying the number of nearest neighbours on the ML 100k dataset. The evaluated MAE and RMSE of all the similarity measures decrease as the number of nearest neighbors increases.

Figure 2 elaborate the same conclusion hold using ML-1M dataset except change in the performance of different similarity measure.

Figure 3 shows the performance comparison of existing similarity measures on EachMovie dataset. CJacMMD outperforms the other approaches consistently, regardless of neighbourhood size. It is observed that the range of MAE and RMSE values in all the three dataset decreases as the value of K increases. That is because some similarity measures are sensitive to false neighbours, which results in the high similarity between two users, but in fact, their preference is not similar this often lead by the data sparsity.

5. Conclusions

CF has become an attractive subject for researchers because its capability to handle information overload efficiently. It provides personalized and automated suggestions to users based on their taste and those of similar interest. In order to overcome the problem of data sparsity, we proposed a new similarity measure based on Mean measure of divergence that considers the behaviour habit effects of rated items to measure the similarity between users. The detailed analysis
on three benchmark datasets indicates that the results of proposed approach are comparable with the traditional approach. The proposed technique provides better prediction accuracy. We plan to extend the proposed algorithm in a real-world recommender system.

References


