Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

Transfer Learning in Collaborative Filtering with Uncertain Ratings

Weike Pan, Evan W. Xiang and Qiang Yang
Department of Computer Science and Engineering
Hong Kong University of Science and Technology, Hong Kong
{weikep, wxiang, qyang}@cse.ust.hk

Abstract

To solve the sparsity problem in collaborative filtering, researchers have introduced transfer learning as a viable approach to make use of auxiliary data. Most previous transfer learning works in collaborative filtering have focused on exploiting point-wise ratings such as numerical ratings, stars, or binary ratings of likes/dislikes. However, in many real-world recommender systems, many users may be unwilling or unlikely to rate items with precision. In contrast, practitioners can turn to various non-preference data to estimate a range or rating distribution of a user's preference on an item. Such a range or rating distribution is called an uncertain rating since it represents a rating spectrum of uncertainty instead of an accurate point-wise score. In this paper, we propose an efficient transfer learning solution for collaborative filtering, known as *transfer by integrative factorization* (TIF), to leverage such auxiliary uncertain ratings to improve the performance of recommendation. In particular, we integrate auxiliary data of uncertain ratings as additional constraints in the target matrix factorization problem, and learn an expected rating value for each uncertain rating automatically. The advantages of our proposed approach include the efficiency and the improved effectiveness of collaborative filtering, showing that incorporating the auxiliary data of uncertain ratings can really bring a benefit. Experimental results on two movie recommendation tasks show that our TIF algorithm performs significantly better over a state-of-the-art non-transfer learning method.

Introduction

Recently, researchers have developed new methods for collaborative filtering (Goldberg et al. 1992; Koren 2008; Rendle 2012). A new direction is to apply transfer learning to collaborative filtering (Li, Yang, and Xue 2009b; Pan et al. 2011b), so that one can make use of auxiliary data to help improve the rating prediction performance. However, in many industrial applications, precise point-wise user feedbacks may be rare, because many users are unwilling or unlikely to express their preferences accurately. Instead, we may obtain estimates of a user's tastes on an item based on the user's additional behavior or social connections. For example, suppose that a person Peter is watching a 10-minute video. Suppose that Peter stops watching the video after the first 3 minutes. In this case, we may estimate that Peter's preference on the movie is in the range of 1 to 2 stars with a uniform distribution. As another example in social media, suppose that Peter reads his followees' posts in a microblog about a certain movie[1]. Suppose that his followee John posts a comment on the movie with 3 stars. In addition, Peter's other followees Bob gives 4 stars, and Alice gives 5 stars. Then, with this social impression data, we should be able to obtain a potential rating distribution for Peter's preference on the movie. We call such a rating distribution as an uncertain rating, since it represents a rating spectrum involving uncertainty instead of an accurate point-wise score.

To leverage such uncertain ratings as described above, we plan to exploit techniques in transfer learning (Pan and Yang 2010). To do this, we have to answer two fundamental questions: "what to transfer" and "how to transfer" in transfer learning (Pan and Yang 2010). In particular, we have to decide (1) what knowledge to extract and transfer from the auxiliary uncertain ratings, and (2) how to model the knowledge transfer from the auxiliary uncertain rating data to the target numerical ratings in a principled way. As far as we know, there has not been existing research work on this problem.

Several existing works are relevant to ours. Transfer learning approaches are proposed to transfer knowledge in latent feature space (Singh and Gordon 2008; Yoo and Choi 2009; Pan et al. 2010; Cao, Liu, and Yang 2010; Pan et al. 2011b; Vasuki et al. 2011), exploiting feature covariance (Adams, Dahl, and Murray 2010) or compressed rating patterns (Li, Yang, and Xue 2009a; 2009b). In collaborative filtering, transfer learning methods can be adaptive (Li, Yang, and Xue 2009a; Pan et al. 2010) or collective (Singh and Gordon 2008; Li, Yang, and Xue 2009b; Yoo and Choi 2009; Cao, Liu, and Yang 2010; Pan et al. 2011b; Vasuki et al. 2011). Other works, such as that by Ma et al. (Ma, King, and Lyu 2011), tend to use auxiliary social relations and extend the rating generation function in a model-based collaborative filtering method (Salakhutdinov and Mnih 2008). Zhang

[1]For example, Tencent Video http://v.qq.com/ and Tencent Weibo (microblog) http://t.qq.com/ are connected.

et al. (Zhang et al. 2010) generate point-wise virtual ratings from sentimental polarities of users' reviews on items, which are then used in memory-based collaborative filtering methods for video recommendation. However, these works do not address the uncertain rating problem.

In this paper, we develop a novel approach known as TIF (*transfer by integrative factorization*) to transfer auxiliary data consisting of uncertain ratings as constraints to improve the predictive performance in a target collaborative filtering problem. We assume that the users and items can be mapped in a one-one manner. Our approach runs in several steps. First, we integrate ("how to transfer") the auxiliary uncertain ratings as constraints ("what to transfer") into the target matrix factorization problem. Second, we learn an expected rating for each uncertain rating automatically. Third, we relax the constraints and introduce a penalty term for those violating the constraints. Finally, we solve the optimization problem via stochastic gradient descent (SGD). We conduct empirical studies on two movie recommendation data sets of MovieLens10M and Netflix, and obtain significantly better results of TIF over other methods.

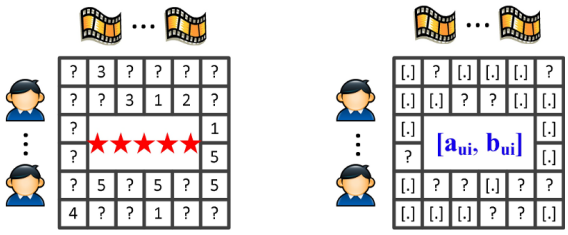


Figure 1: Illustration of transfer learning in collaborative filtering from auxiliary uncertain ratings (left: target 5-star numerical ratings; right: auxiliary uncertain ratings represented as ranges or rating distributions). Note that there is a one-one mapping between the users and items from two data.

Transfer by Integrative Factorization

Problem Definition

In our problem setting, we have a target *user-item* numerical rating matrix $\mathbf{R} = [r_{ui}]_{n\times m} \in \{1,2,3,4,5,?\}^{n\times m}$, where the question mark "?" denotes a missing value. We use an indicator matrix $\mathbf{Y} = [y_{ui}]_{n\times m} \in \{0,1\}^{n\times m}$ to denote whether the entry (u,i) is observed ($y_{ui} = 1$) or not ($y_{ui} = 0$), and $\sum_{u,i} y_{ui} = q$. Besides the target data, we have an auxiliary *user-item* uncertain rating matrix $\tilde{\mathbf{R}} = [\tilde{r}_{ui}]_{n\times m} \in \{\lfloor a_{ui}, b_{ui}\rceil, ?\}^{n\times m}$ with $\tilde{q}$ observations, where the entry $\lfloor a_{ui}, b_{ui}\rceil$ denotes the range of a certain distribution for the corresponding rating located at (u,i), where $a_{ui} \le b_{ui}$. The question mark "?" represents a missing value. Similar to the target data, we have a corresponding indicator matrix $\tilde{\mathbf{Y}} = [\tilde{y}_{ui}]_{n\times m} \in \{0,1\}^{n\times m}$ with $\sum_{u,i} \tilde{y}_{ui} = \tilde{q}$. We also assume that there is a one-one mapping between the users and items of $\mathbf{R}$ and $\tilde{\mathbf{R}}$. Our goal is to predict the missing values in $\mathbf{R}$ by exploiting uncertain ratings in $\tilde{\mathbf{R}}$.

The difference between the problem setting studied in this paper and those of previous works like (Pan et al. 2011b) is that the auxiliary data in this paper are uncertain and represented as ranges of rating distributions instead of accurate point-wise scores. We illustrate the new problem setting in Figure 1.

Model Formulation

Koren (Koren 2008) proposes to learn not only user-specific latent features $U_{u\cdot} \in \mathbb{R}^{1\times d}$ and item-specific latent features $V_{i\cdot} \in \mathbb{R}^{1\times d}$ as that in PMF (Salakhutdinov and Mnih 2008), but also user bias $b_u \in \mathbb{R}$, item bias $b_i \in \mathbb{R}$ and global average rating value $\mu \in \mathbb{R}$. The objective function of RSVD (Koren 2008) is as follows,

$$\min_{U_{u\cdot},V_{i\cdot},b_u,b_i,\mu} \sum_{u=1}^{n}\sum_{i=1}^{m} y_{ui}(\mathcal{E}_{ui} + \mathcal{R}_{ui}) \quad (1)$$

where $\mathcal{E}_{ui} = \frac{1}{2}(r_{ui} - \hat{r}_{ui})^2$ is the square loss function with $\hat{r}_{ui} = \mu + b_u + b_i + U_{u\cdot}V_{i\cdot}^T$ as the predicted rating, and $\mathcal{R}_{ui} = \frac{\alpha_u}{2}||U_{u\cdot}||^2 + \frac{\alpha_v}{2}||V_{i\cdot}||^2 + \frac{\beta_u}{2}b_u^2 + \frac{\beta_v}{2}b_i^2$ is the regularization term used to avoid overfitting. To learn the parameters $U_{u\cdot}, V_{i\cdot}, b_u, b_i, \mu$ efficiently, SGD algorithms are adopted, in which the parameters are updated for each randomly sampled rating r_{ui} with $y_{ui} = 1$.

In our problem setting, besides the target numerical ratings $\mathbf{R}$, we have some auxiliary uncertain ratings represented as ranges of rating distributions $\tilde{\mathbf{R}} \in \{\lfloor a_{ui}, b_{ui}\rceil, ?\}^{n\times m}$. The semantic meaning of $\lfloor a_{ui}, b_{ui}\rceil$ can be represented as a constraint for the predicted rating $\hat{r}_{ui} \in \mathcal{C}(a_{ui}, b_{ui})$, e.g., $\hat{r}_{ui} = (a_{ui} + b_{ui})/2$ or $a_{ui} \le \hat{r}_{ui} \le b_{ui}$. Based on this observation, we extend the optimization problem (Koren 2008) as shown in Eq.(1), and propose to solve the following optimization problem,

$$\begin{aligned} \min_{U_{u\cdot},V_{i\cdot},b_u,b_i,\mu} \quad & \sum_{u=1}^{n}\sum_{i=1}^{m} y_{ui}(\mathcal{E}_{ui} + \mathcal{R}_{ui}) \quad (2)\\ \text{s.t.} \quad & \hat{r}_{ui} \in \mathcal{C}(a_{ui}, b_{ui}),\\ & \forall\, \tilde{y}_{ui} = 1, u = 1,\ldots,n, i = 1,\ldots,m \end{aligned}$$

where the auxiliary domain knowledge involving uncertain ratings is transferred to the target domain, via integration of constraints into the target matrix factorization problem: $\hat{r}_{ui} \in \mathcal{C}(a_{ui}, b_{ui}), \tilde{y}_{ui} = 1$. For this reason, we call our approach *transfer by integrative factorization* (TIF). The knowledge, $\mathcal{C}(a_{ui}, b_{ui})$, from the auxiliary uncertain ratings can be considered as a rating spectrum with lower bound value of a_{ui} and upper bound value of b_{ui}, which can be equivalently represented as a rating distribution of $r \sim P_{ui}(r)$ over $\lfloor a_{ui}, b_{ui}\rceil$.

The optimization problem with a hard constraint $\hat{r}_{ui} \in \mathcal{C}(a_{ui}, b_{ui})$ as shown in Eq.(2) is difficult to solve. We relax this hard constraint, move it to the objective function, and derive the following new objective function with an additional penalty term,

$$\begin{aligned} \min_{U_{u\cdot},V_{i\cdot},b_u,b_i,\mu} \quad & \sum_{u=1}^{n}\sum_{i=1}^{m}[y_{ui}(\mathcal{E}_{ui} + \mathcal{R}_{ui})\\ & +\lambda\tilde{y}_{ui}(\tilde{\mathcal{E}}_{ui} + \tilde{\mathcal{R}}_{ui})] \quad (3) \end{aligned}$$

where $\tilde{\mathcal{E}}_{ui}$ includes the predicted rating $\hat{r}_{ui}$ and the observed uncertain rating $\lfloor a_{ui}, b_{ui} \rceil$. The tradeoff parameter λ is used to balance two loss functions for target data and auxiliary data. We use the same regularization terms $\tilde{\mathcal{R}}_{ui} = \mathcal{R}_{ui}$ for simplicity. We now show that the distribution $r \sim P_{ui}(r)$ in $\tilde{\mathcal{E}}_{ui}$ can be simplified as an expected rating value.

Theorem 1. *The penalty term $\tilde{\mathcal{E}}_{ui}$ over the rating spectrum $\lfloor a_{ui}, b_{ui} \rceil$ can be equivalently represented as $\frac{1}{2}(\bar{r}_{ui} - \hat{r}_{ui})^2$, where $\bar{r}_{ui} = \int_{a_{ui}}^{b_{ui}} P_{ui}(r) \cdot r \mathrm{d}r$ is the expected rating of user u on item i.*

Proof. Similar to the square loss used in RSVD (Koren 2008), the penalty over rating spectrum $\lfloor a_{ui}, b_{ui} \rceil$ can be written as $\tilde{\mathcal{E}}_{ui} = \frac{1}{2}\int_{a_{ui}}^{b_{ui}} \left[P_{ui}(r) \cdot (r - \hat{r}_{ui})^2\right] \mathrm{d}r$, where $P_{ui}(r)$ is the probability of rating value r by user u on item i . We thus have the gradient formula:

$$\begin{aligned}
\frac{\partial \tilde{\mathcal{E}}_{ui}}{\partial \hat{r}_{ui}} &= \frac{\partial \frac{1}{2}\int_{a_{ui}}^{b_{ui}} \left[P_{ui}(r) \cdot (r - \hat{r}_{ui})^2\right] \mathrm{d}r}{\partial \hat{r}_{ui}} \\
&= -(\int_{a_{ui}}^{b_{ui}} P_{ui}(r) \cdot r \mathrm{d}r - \hat{r}_{ui} \int_{a_{ui}}^{b_{ui}} P_{ui}(r) \mathrm{d}r) \\
&= \frac{\partial \frac{1}{2}(\int_{a_{ui}}^{b_{ui}} P_{ui}(r) \cdot r \mathrm{d}r - \hat{r}_{ui})^2}{\partial \hat{r}_{ui}} \\
&= \frac{\partial \frac{1}{2}(\bar{r}_{ui} - \hat{r}_{ui})^2}{\partial \hat{r}_{ui}},
\end{aligned}$$

which shows that we can use the expected rating $\bar{r}_{ui}$ to replace the rating distribution $r \sim P_{ui}(r)$ over $\lfloor a_{ui}, b_{ui} \rceil$ since it results in the exactly the same gradient. Hence, parameters learned using the same gradient in the widely used SGD algorithm framework in matrix factorization (Koren 2008) will be the same. □

However, we still find it difficult to obtain an accurate rating distribution $r \sim P_{ui}(r)$ or the expected rating $\bar{r}_{ui}$, because there is not sufficient information besides a rating range $\lfloor a_{ui}, b_{ui} \rceil$. One simple approach is to assign the same weight on a_{ui} and b_{ui}, that is $\bar{r}_{ui} = \frac{1}{2}(a_{ui} + b_{ui})$. But such a straightforward approach may not accurately reflect the true expected rating value. Furthermore, static expected value may not well reflect personalized taste. Instead, we learn the expected rating value automatically,

$$\bar{r}_{ui} = \left[s(a_{ui})a_{ui} + s(b_{ui})b_{ui}\right] / \left[s(a_{ui}) + s(b_{ui})\right], \qquad (4)$$

where $s(x) = \exp(-|\hat{r}_{ui} - x|^{1-\rho})$ is a similarity function, and $s(a_{ui}) / \left[s(a_{ui}) + s(b_{ui})\right]$ is the normalized weight or confidence on rating a_{ui}. The parameter ρ can be considered as an uncertainty factor, where a larger value means higher uncertainty. At the start of the learning procedure, we are uncertain of the expected rating, and thus we may set $\rho = 1$ and $\bar{r}_{ui} = (a_{ui} + b_{ui})/2$. In the middle of the learning procedure, we may gradually decrease the value of ρ as we are more sure of the expected rating. Note that the similarity function $s(x)$ in Eq.(4) can be other forms if we have additional domain knowledge. We illustrate the impact of ρ when we estimate the expected rating in Figure 2 ($a_{ui} = 4, b_{ui} = 5$).

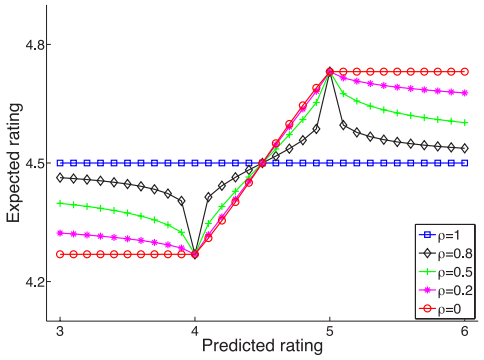


Figure 2: Illustration of the expected rating estimated using Eq.(4) with $a_{ui} = 4$ and $b_{ui} = 5$.

Learning the TIF

Denoting $f_{ui} = y_{ui}(\mathcal{E}_{ui} + \mathcal{R}_{ui}) + \lambda \tilde{y}_{ui}(\tilde{\mathcal{E}}_{ui} + \tilde{\mathcal{R}}_{ui})$ as part of the objective function in Eq.(3), we have the gradients $\nabla U_{u\cdot} = \frac{\partial f_{ui}}{\partial U_{u\cdot}}$, $\nabla V_{i\cdot} = \frac{\partial f_{ui}}{\partial V_{i\cdot}}$, $\nabla b_u = \frac{\partial f_{ui}}{\partial b_u}$, $\nabla b_i = \frac{\partial f_{ui}}{\partial b_i}$, $\nabla \mu = \frac{\partial f_{ui}}{\partial \mu}$ as follows,

$$\nabla U_{u\cdot} = \begin{cases} -e_{ui}V_{i\cdot} + \alpha_u U_{u\cdot}, & \text{if } y_{ui} = 1 \\ -\lambda \tilde{e}_{ui}V_{i\cdot} + \lambda\alpha_u U_{u\cdot}, & \text{if } \tilde{y}_{ui} = 1 \end{cases} \qquad (5)$$

$$\nabla V_{i\cdot} = \begin{cases} -e_{ui}U_{u\cdot} + \alpha_v V_{i\cdot}, & \text{if } y_{ui} = 1 \\ -\lambda \tilde{e}_{ui}U_{u\cdot} + \lambda\alpha_v V_{i\cdot}, & \text{if } \tilde{y}_{ui} = 1 \end{cases} \qquad (6)$$

$$\nabla b_u = \begin{cases} -e_{ui} + \beta_u b_u, & \text{if } y_{ui} = 1 \\ -\lambda \tilde{e}_{ui} + \lambda\beta_u b_u, & \text{if } \tilde{y}_{ui} = 1 \end{cases} \qquad (7)$$

$$\nabla b_i = \begin{cases} -e_{ui} + \beta_v b_i, & \text{if } y_{ui} = 1 \\ -\lambda \tilde{e}_{ui} + \lambda\beta_v b_i, & \text{if } \tilde{y}_{ui} = 1 \end{cases} \qquad (8)$$

$$\nabla \mu = \begin{cases} -e_{ui}, & \text{if } y_{ui} = 1 \\ -\lambda \tilde{e}_{ui}, & \text{if } \tilde{y}_{ui} = 1 \end{cases} \qquad (9)$$

where $e_{ui} = r_{ui} - \hat{r}_{ui}$, $\tilde{e}_{ui} = \bar{r}_{ui} - \hat{r}_{ui}$ are the errors according to the target numerical rating and the auxiliary expected rating, respectively, and $\bar{r}_{ui}$ is estimated via Eq.(4) using the parameters learned in the previous iteration. We thus have the update rules used in the SGD algorithm framework,

$$U_{u\cdot} = U_{u\cdot} - \gamma \nabla U_{u\cdot} \qquad (10)$$

$$V_{i\cdot} = V_{i\cdot} - \gamma \nabla V_{i\cdot} \qquad (11)$$

$$b_u = b_u - \gamma \nabla b_u \qquad (12)$$

$$b_i = b_i - \gamma \nabla b_i \qquad (13)$$

$$\mu = \mu - \gamma \nabla \mu. \qquad (14)$$

When there are no auxiliary uncertain ratings, our update rules in Eq.(10-14) reduce to that of RSVD (Koren 2008).

Finally, we obtain a complete algorithm as shown in Figure 3, where we update the parameters $U_{u\cdot}$, $V_{i\cdot}$, b_u, b_i and μ for each observed rating. Note that the stochastic gradient descent algorithm used in RSVD (Koren 2008) is different from ours, since we have auxiliary uncertain ratings, and learn and transfer the expected ratings $\bar{r}_{ui}$. TIF inherits

Input: The target *user-item* numerical rating matrix $\mathbf{R}$, the frontal-side auxiliary *user-item* uncertain rating matrix $\tilde{\mathbf{R}}$.

Output: The user-specific latent feature vector $U_{u\cdot}$ and bias $b_{u\cdot}$, the item-specific latent feature vector $V_{i\cdot}$ and bias $b_{i\cdot}$, the global average μ, where $u = 1, \ldots, n$, $i = 1, \ldots, m$.

For $t = 1, \ldots, T$
  For $iter = 1, \ldots, q + \tilde{q}$
  Step 1. Randomly pick up a rating from $\mathbf{R}$ or $\tilde{\mathbf{R}}$;
  Step 2. If $\tilde{y}_{ui} = 1$, estimate the expected rating $\bar{r}_{ui}$ as shown in Eq.(4);
  Step 3. Calculate the gradients as shown in Eq.(5-9);
  Step 4. Update the parameters as shown in Eq.(10-14).
  End
Decrease the learning rate γ and uncertainty factor ρ.
End

Figure 3: The algorithm of *transfer by integrative factorization* (TIF).

the advantages of efficiency in RSVD, and reduces to RSVD when there are only target 5-star numerical ratings. The time complexity of TIF is $O(T(q + \tilde{q})d)$, where T represents the number of scans over the whole data and is usually smaller than 100, $q + \tilde{q}$ denotes the number of observed ratings from both target and auxiliary data, and d is the number of latent dimensions. Similar to RSVD, TIF can also be implemented in a distributed platform like Map/Reduce.

Experimental Results

In this section, we plan to evaluate the effectiveness of the TIF algorithm and compare it with some well known benchmark approaches. We start by describing the experimental data.

Data Sets and Evaluation Metrics

MovieLens10M Data (ML) The MovieLens[2] rating data contains more than 10^7 ratings with values in $\{0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4, 4.5, 5\}$, which are given by more than 7.1×10^4 users on around 1.1×10^4 movies between 1995 and 2009. We preprocess the MovieLens data as follows: first, we randomly permutate the rating records since the original data is ordered with user ID; second, we use the official linux shell script[3] to generate 5 copies of training data and test data, where in each copy $4/5$ are used for training and $1/5$ for test; third, for each copy of training data, we take 50% ratings as auxiliary data, and the remaining 50% ratings as target data; fourth, for each copy of auxiliary data, we convert ratings of $0.5, 1, 1.5, 2, 3, 3.5$ to uncertain ratings $\lfloor 0.5, 3.5 \rceil$ with uniform distribution, and ratings of $4, 4.5, 5$ to $\lfloor 4, 5 \rceil$.

[2]http://www.grouplens.org/node/73/

[3]http://www.grouplens.org/system/files/ml-10m-README.html

Netflix Data (NF) The Netflix[4] rating data contains more than 10^8 ratings with values in $\{1, 2, 3, 4, 5\}$, which are given by more than 4.8×10^5 users on around 1.8×10^4 movies between 1998 and 2005. The Netflix competition data contains two sets, the training set and the probe set, and we randomly separate the training set into two parts, 50% ratings are taken as auxiliary data, and the remaining 50% ratings as target data. For the auxiliary data, to simulate the effect of rating uncertainty, we convert ratings of $1, 2, 3$ to $\lfloor 1, 3 \rceil$, and ratings of $4, 5$ to $\lfloor 4, 5 \rceil$. We randomly generate the auxiliary data and target data for three times, and thus get three copies of data.

We summarize the final data in Table 1.

Table 1: Description of MovieLens10M data ($n = 71,567, m = 10,681$) and Netflix data ($n = 480,189, m = 17,770$). Sparsity refers to the percentage of observed ratings in the *user-item* preference matrix, e.g. $\frac{q}{nm}$ and $\frac{\tilde{q}}{nm}$ are sparsities for target data and auxiliary data, respectively.

	Data set	Form	Sparsity
ML	target (training)	$\{0.5, ..., 5, ?\}$	0.52%
	target (test)	$\{0.5, ..., 5, ?\}$	0.26%
	auxiliary	$\{\lfloor 0.5, 3.5 \rceil, \lfloor 4, 5 \rceil, ?\}$	0.52%
NF	target (training)	$\{1, 2, 3, 4, 5, ?\}$	0.58%
	target (test)	$\{1, 2, 3, 4, 5, ?\}$	0.017%
	auxiliary	$\{\lfloor 1, 3 \rceil, \lfloor 4, 5 \rceil, ?\}$	0.58%

Evaluation Metrics We adopt two evaluation metrics: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE),

$$MAE = \sum_{(u,i,r_{ui}) \in T_E} |r_{ui} - \hat{r}_{ui}| / |T_E|$$

$$RMSE = \sqrt{\sum_{(u,i,r_{ui}) \in T_E} (r_{ui} - \hat{r}_{ui})^2 / |T_E|}$$

where r_{ui} and $\hat{r}_{ui}$ are the true and predicted ratings, respectively, and $|T_E|$ is the number of test ratings.

Baselines and Parameter Settings

We compare our TIF method with a state-of-the-art method in Netflix competition, RSVD (Koren 2008). For both TIF and RSVD, we use the same statistics of target training data only to initialize the global average rating value μ, user bias b_u, item bias b_i, user-specific latent feature vector $U_{u\cdot}$, and

[4]http://www.netflix.com/

item-specific latent feature vector $V_{i\cdot}$,

$$\begin{aligned}
\mu &= \sum_{u=1}^{n}\sum_{i=1}^{m} y_{ui} r_{ui} / \sum_{u=1}^{n}\sum_{i=1}^{m} y_{ui} \\
b_u &= \sum_{i=1}^{m} y_{ui}(r_{ui}-\mu) / \sum_{i=1}^{m} y_{ui} \\
b_i &= \sum_{u=1}^{n} y_{ui}(r_{ui}-\mu) / \sum_{u=1}^{n} y_{ui} \\
U_{uk} &= (r-0.5)\times 0.01, k=1,\ldots,d \\
V_{ik} &= (r-0.5)\times 0.01, k=1,\ldots,d
\end{aligned}$$

where r ($0 \le r < 1$) is a random value.

For both TIF and RSVD, the tradeoff parameters and learning rate are set similarly to that of RSVD (Koren 2008), $\alpha_u = \alpha_v = 0.01$, $\beta_u = \beta_v = 0.01$, $\gamma = 0.01$. Note that the value of learning rate γ decreases after each scan of the whole rating data (Koren 2008), $\gamma \leftarrow \gamma \times 0.9$. For MovieLens10M data, we set the number of latent dimensions as $d = 20$ (Zhou et al. 2009); and for Netflix data, we use $d = 100$ (Koren 2008). For TIF, we first fix $\lambda = 1$ when comparing to RSVD, and later study the effect of λ with different values of $\lambda \in \{0.1, 0.5, 1\}$.

To study the effectiveness of learning an expected rating for each uncertain rating, we also report the result of using static average rating $\bar{r}_{ui} = (a_{ui} + b_{ui})/2$ with $\tilde{y}_{ui} = 1$, which is denoted as TIF(avg.).

The uncertainty factor ρ in TIF is decreased in a similar way as that of the learning rate γ, which is updated after every 10 scans of the whole data, $\rho \leftarrow \rho \times 0.9$.

Summary of Experimental Results

The prediction performance of RSVD, TIF(avg.) and TIF are shown in Table 2 and 3. We can have the following observations,

1. TIF is significantly better than TIF(avg.) and RSVD in both data sets, which clearly shows the advantage of the proposed transfer learning approach in leveraging auxiliary uncertain ratings; and
2. for TIF, the parameter λ is important, since it determines how large impact will the auxiliary uncertain data make on the target data. TIF with $\lambda = 0.5$ or $\lambda = 1$ is much better than that of $\lambda = 0.1$, which shows that a medium value between 0.5 and 1 is likely to have the best result.

To gain a deep understanding of the performance of RSVD, TIF(avg.) and TIF, we show the prediction performance against different iteration numbers in Figure 4, from which we can have the following observations,

1. For RSVD, TIF(avg.) and TIF, the prediction performance becomes relatively stable after 50 iterations; and
2. TIF performs better than RSVD and TIF(avg.) after 20 iterations in both data sets, which again shows the advantages of the proposed transfer learning approach with the ability of leveraging auxiliary uncertain ratings.

Table 2: Prediction performance of RSVD, TIF(avg.) and TIF on MovieLens10M data (ML) and Netflix data (NF). The tradeoff parameter λ is fixed as 1, and the number of iterations is fixed as 50.

Data	Metric	RSVD	TIF(avg.)	TIF
ML	MAE	0.6438±0.0011	0.6415±0.0008	**0.6242**±0.0006
	RMSE	0.8364±0.0012	0.8188±0.0009	**0.8057**±0.0007
NF	MAE	0.7274±0.0005	0.7288±0.0002	**0.7225**±0.0004
	RMSE	0.9456±0.0003	0.9323±0.0002	**0.9271**±0.0002

Table 3: Prediction performance of TIF on MovieLens10M data (ML) and Netflix data (NF) with different value of λ. The number of iterations is fixed as 50.

Data	Metric	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1$
ML	MAE	0.6399±0.0003	0.6280±0.0007	**0.6242**±0.0006
	RMSE	0.8307±0.0008	0.8131±0.0007	**0.8057**±0.0007
NF	MAE	0.7233±0.0006	**0.7172**±0.0004	0.7225±0.0004
	RMSE	0.9377±0.0003	**0.9242**±0.0002	0.9271±0.0002

We further study the prediction performance on different user groups with respect to the users' activeness. For MovieLens10M data, we categorize the users in the test data into 10 groups, where users in different groups have different numbers of ratings. Users in training and test data have similar activeness, according to the data generation procedure. From the results as shown in Figure 5, we can see,

1. TIF performs best on all user groups; and
2. TIF(avg.) and TIF are more useful for users with fewer ratings, which shows the effect of sparsity reduction of transfer learning methods in collaborative filtering.

Note that the results of MAE and RMSE in Figure 5 is calculated over rating instances of users in the same group.

Related Works

Collaborative Filtering Collaborative filtering (Goldberg et al. 1992; Koren 2008; Rendle 2012) as an intelligent component in recommender systems (Linden, Smith, and York 2003; Zheng and Xie 2011) has gained extensive interest in both academia and industry, while most previous works can only make use of point-wise ratings. In this paper, we go one step beyond and study a new problem with uncertain ratings via transfer learning techniques, as shown in Figure 1.

Transfer Learning Transfer learning (Caruana 1993; Pan and Yang 2010) as a new learning paradigm extracts and transfers knowledge from auxiliary data to help a target learning task (Evgeniou and Pontil 2004; Dai et al. 2007; Pan et al. 2011a). From the perspective of model-based transfer, feature-based transfer and instance-based transfer (Pan and Yang 2010), TIF can be considered as a *rating instance*-based transfer. We make a link between traditional transfer learning methods in text classification and transfer learning methods in collaborative filtering from a unified view, which is shown in Table 4.

Table 4: Overview of TIF in a big picture of traditional transfer learning and transfer learning in collaborative filtering.

Transfer learning approaches	Text classification	Collaborative filtering
Model-based Transfer	MTL (Evgeniou and Pontil 2004)	CBT, RMGM: cluster-level rating patterns DPMF: covariance matrix (operator)
Feature-based Transfer	TCA (Pan et al. 2011a)	CST, CMF, WNMCTF: latent features
Instance-based Transfer	TrAdaBoost (Dai et al. 2007)	**TIF**: rating instances

Table 5: Summary of TIF and other transfer learning methods in collaborative filtering.

	Knowledge (what to transfer)	Algorithm style (how to transfer) Adaptive	Collective	Integrative
PMF family	Covariance		DPMF (Adams, Dahl, and Murray 2010)	
	Latent features	CST (Pan et al. 2010),	CMF (Singh and Gordon 2008)	
	Constraints			**TIF**
NMF family	Codebook	CBT (Li, Yang, and Xue 2009a)	RMGM (Li, Yang, and Xue 2009b)	
	Latent features		WNMCTF (Yoo and Choi 2009)	

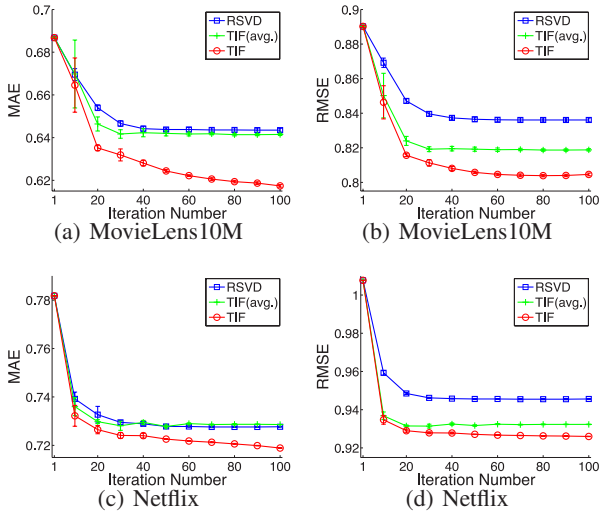
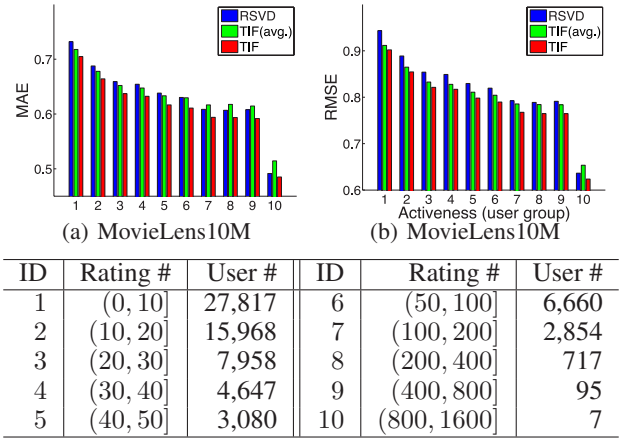


Figure 4: Prediction performance of RSVD, TIF(avg.) and TIF with different iteration numbers (the tradeoff parameter λ is fixed as 1).



ID	Rating #	User #	ID	Rating #	User #
1	$(0, 10]$	27,817	6	$(50, 100]$	6,660
2	$(10, 20]$	15,968	7	$(100, 200]$	2,854
3	$(20, 30]$	7,958	8	$(200, 400]$	717
4	$(30, 40]$	4,647	9	$(400, 800]$	95
5	$(40, 50]$	3,080	10	$(800, 1600]$	7

Figure 5: Prediction performance of RSVD, TIF(avg.) and TIF on different user groups (using the first fold of the MovieLens10M data). The tradeoff parameter λ is fixed as 1, and the number of iterations is fixed as 50.

Transfer Learning in Collaborative Filtering There are some related work of transfer learning in collaborative filtering, CMF (Singh and Gordon 2008), CBT (Li, Yang, and Xue 2009a), RMGM (Li, Yang, and Xue 2009b), WNMCTF (Yoo and Choi 2009), CST (Pan et al. 2010), DPMF (Adams, Dahl, and Murray 2010), etc. Please see (Pan et al. 2011b) for a detailed analysis and comparison from the perspective of "what to transfer" and "how to transfer" in transfer learning (Pan and Yang 2010).

Comparing to previous works on transfer learning in collaborative filtering, we can categorize TIF as an *integrative* style algorithm ("how to transfer") via transferring knowledge of constraints ("what to transfer"). We thus summarize the related work as discussed in (Pan et al. 2011b) and our TIF method in Table 5, where we can see that TIF extends previous works from two dimensions, "what to transfer" and "how to transfer" in transfer learning (Pan and Yang 2010).

Conclusions and Future Work

In this paper, we study a new problem of transfer learning in collaborative filtering when the auxiliary data are uncertain ratings. We propose a novel efficient transfer learning approach, *transfer by integrative factorization* (TIF), to leverage auxiliary data of uncertain ratings represented as rating distributions. In TIF, we take the auxiliary uncertain ratings as constraints and integrate them into the optimization problem for the target matrix factorization. We then reformulate the optimization problem by relaxing the constraints and in-

troducing a penalty term. The final optimization problem inherits the advantages of the efficient SGD algorithm in large-scale matrix factorization (Koren 2008). Experimental results show that our proposed transfer learning solution significantly outperforms the state-of-the-art matrix factorization approach without using the auxiliary data.

In the future, we plan to study the performance of TIF in industry recommender systems with uncertain ratings estimated from users' social impressions in the connected video streaming and microblogging systems.

Acknowledgments

We thank the support of Hong Kong RGC GRF Projects 621010 and 621211.

References

Adams, R. P.; Dahl, G. E.; and Murray, I. 2010. Incorporating side information into probabilistic matrix factorization using Gaussian processes. In *In Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence*, 1–9.

Cao, B.; Liu, N. N.; and Yang, Q. 2010. Transfer learning for collective link prediction in multiple heterogenous domains. In *Proceedings of the 27th International Conference on Machine Learning*, 159–166.

Caruana, R. 1993. Multitask learning: A knowledge-based source of inductive bias. In *Proceedings of the 10th International Conference on Machine Learning*, 41–48.

Dai, W.; Yang, Q.; Xue, G.-R.; and Yu, Y. 2007. Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*, ICML '07, 193–200. New York, NY, USA: ACM.

Evgeniou, T., and Pontil, M. 2004. Regularized multi–task learning. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, 109–117. New York, NY, USA: ACM.

Goldberg, D.; Nichols, D.; Oki, B. M.; and Terry, D. 1992. Using collaborative filtering to weave an information tapestry. *Communnication of ACM* 35:61–70.

Koren, Y. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, 426–434. New York, NY, USA: ACM.

Li, B.; Yang, Q.; and Xue, X. 2009a. Can movies and books collaborate? cross-domain collaborative filtering for sparsity reduction. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, 2052–2057.

Li, B.; Yang, Q.; and Xue, X. 2009b. Transfer learning for collaborative filtering via a rating-matrix generative model. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 617–624.

Linden, G.; Smith, B.; and York, J. 2003. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing* 7(1):76–80.

Ma, H.; King, I.; and Lyu, M. R. 2011. Learning to recommend with explicit and implicit social relations. *ACM Transactions on Intelligent Systems and Technology (ACM TIST)* 2:29:1–29:19.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering* 22(10):1345–1359.

Pan, W.; Xiang, E. W.; Liu, N. N.; and Yang, Q. 2010. Transfer learning in collaborative filtering for sparsity reduction. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 230–235.

Pan, S. J.; Tsang, I. W.; Kwok, J. T.; and Yang, Q. 2011a. Domain adaptation via transfer component analysis. *IEEE Transactions on Neural Networks* 22(2):199–210.

Pan, W.; Liu, N. N.; Xiang, E. W.; and Yang, Q. 2011b. Transfer learning to predict missing ratings via heterogeneous user feedbacks. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, 2318–2323.

Rendle, S. 2012. Factorization machines with libfm. *ACM Transactions on Intelligent Systems and Technology (ACM TIST)* 3:19:1–19:22.

Salakhutdinov, R., and Mnih, A. 2008. Probabilistic matrix factorization. In *Annual Conference on Neural Information Processing Systems 20*, 1257–1264. MIT Press.

Singh, A. P., and Gordon, G. J. 2008. Relational learning via collective matrix factorization. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, 650–658. New York, NY, USA: ACM.

Vasuki, V.; Natarajan, N.; Lu, Z.; Savas, B.; and Dhillon, I. 2011. Scalable affiliation recommendation using auxiliary networks. *ACM Transactions on Intelligent Systems and Technology (ACM TIST)* 3:3:1–3:20.

Yoo, J., and Choi, S. 2009. Weighted nonnegative matrix co-tri-factorization for collaborative prediction. In *Proceedings of the 1st Asian Conference on Machine Learning: Advances in Machine Learning*, ACML '09, 396–411. Berlin, Heidelberg: Springer-Verlag.

Zhang, W.; Ding, G.; Chen, L.; and Li, C. 2010. Augmenting chinese online video recommendations by using virtual ratings predicted by review sentiment classification. In *Proceedings of the 2010 IEEE International Conference on Data Mining Workshops*, ICDMW '10, 1143–1150. Washington, DC, USA: IEEE Computer Society.

Zheng, Y., and Xie, X. 2011. Learning travel recommendations from user-generated gps traces. *ACM Transactions on Intelligent Systems and Technology (ACM TIST)* 2(1):2:1–2:29.

Zhou, T. C.; Ma, H.; King, I.; and Lyu, M. R. 2009. Tagrec: Leveraging tagging wisdom for recommendation. In *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 04*, 194–199. Washington, DC, USA: IEEE Computer Society.