Latent Factor Analysis for Low-dimensional Implicit Preference Prediction

Zili Zhou*†, Guandong Xu†, Xiao Zhu†, Shaowu Liu†
*School of Computer Engineering and Science, Shanghai University.
†Advanced Analytics Institute, University of Technology Sydney.
Email: zili.zhou@student.uts.edu.au, Guandong.Xu@uts.edu.au, chooxxxmail@gmail.com, shaowu.liu@uts.edu.au

***Abstract*—User preference prediction aims to predict a users future preferences on a large number of items according to his/her preference history. To achieve this goal, many models have been proposed, but mainly for explicit preference data, such as 5-star ratings. Nevertheless, real-world data are often in implicit format, such as purchase action, and the number of items is not always large. In this paper, we demonstrate the use of latent factor models for solving the task of predicting user preferences on implicit and low-dimensional dataset.**

I. Introduction

Preference prediction is a typical user modeling task to predict the preference of user based on their historical feedback on items. Existing works mainly focus on preference dataset of explicit ratings and with high dimensionality. However, these two constraints have limited broader adoption of the models as real-world datasets are often in implicit format and sometimes have low dimensionality.

Previous user preference prediction tasks normally use explicit user feedback, such as movie rating datasets from movielens. Users give the information of both they like or dislike the item. The most common example is that user rating with a score from 1 to 5, low score means "dislike", high score means "like". While implicit user feedback data can be automatically collected from many user behaviors including user purchase items and user click webpages. The implicit data amount is larger and the data collection efficency is higher.

Despite of preference format, another assumption made by existing models is that the dataset contains a large number of users and items. With user historical purchase records, a high-dimensional user-item matrix can be generated. User preference can be predicted based on latent item co-occurrence information contained in user-item matrix. While in some cases the number of items is limited, small item number makes the item dimension of matrix low, it is a challenge to discover user preference pattern from such low-dimensional data.

To deal with the challenge implicit low-dimensional user preference prediction, we test a few latent factor models with a few customizations. The performance of models will be compared and analyzed.

II. Related Work

There are some previous works tried to do the prediction by using some machine learning and data mining methods, such as collaborative filtering[1]. While the solutions for implicit low-dimensional data are rare. This paper focus on introducing a solution for user preference prediction on implicit low-dimensional data.

Some methods based on Latent Semantic Analysis (LSA) is used to solve the preference prediction problem [2]. In this paper we will use Latent Factor Analysis (LFA) models which are similar with Latent Semantic Analysis models, while the challenge is that our data is implicit and low-dimensional. We do a few customization on the existed Factorization models to fit the data. The models used in the experiment include Matrix Factorization (MF) [3], Sparse Coding (SC) [4], Dictionary Learning (DL) [5] and Restricted Boltzmann Machine (RBM) [6].

III. Our Method

We consider our problem as a collaborative filtering (CF) problem. Some LFA model based CF solutions are given in some previous works [7], [3]. Because of the implicitity, our training data definition is different from some other methods. We consider all the values of the known user-item pairs in user-item matrix as 1, and we assume all the values of unknown pairs as 0.

We use MF model and DL model on our datasets, we also try to using Kmeans clustering algorithm to predefine the dictionary of SC, and define it as KM+SC. The objective functions for MF, DL are following.

$$O_{MF} = \|X - UV\|_2^2 \tag{1}$$

$$O_{DL} = \|X - UV\|_2^2 + \alpha\|U\|_1 \tag{2}$$

X means the original data, U means the relation matrix between users and latent factors, V means the relation matrix between items and latent factors, α means the parameter which control the weight of sparse restriction. The objective functions of MF and SC are need to be minimized, then we can recover the data by dot operation between two matrices, UV. The energy function of RBM, which needs to be minimized, is given.

$$E_\theta(v,h) = -\sum_{i=1}^{n_v} a_i v_i - \sum_{j=1}^{n_h} b_j h_j - \sum_{i=1}^{n_v}\sum_{j=1}^{n_h} w_{i,j} v_i h_j \tag{3}$$

v and h vector represent the values of visible and hidden layer units, W matrix represents the links between two layers, a and b vector represent bias values. By minimizing the energy

TABLE I
AUC RESULTS FOR 5-ITEM DATA (6347 USERS AND 28 PRODUCTS)

	Latent Factor Number	AUC	Speed(sec)
MF	20	0.66	2.86
DL	20	0.76	4.72
KM+SC	20	0.81	2.33
RBM	**20**	**0.95**	**10.99**
MF	100	0.87	11.87
DL	100	0.76	7.65
KM+SC	100	0.79	6.19
RBM	**100**	**0.95**	**30.14**

TABLE II
AUC RESULTS FOR 2-ITEM DATA (350592 USERS AND 32 PRODUCTS)

	Latent Factor Number	AUC	Speed(sec)
MF	20	0.67	279.46
DL	20	0.63	124.54
KM+SC	20	0.66	137.41
RBM	**20**	**0.94**	**617.04**
MF	100	0.74	304.47
DL	100	0.60	148.57
KM+SC	100	0.63	393.58
RBM	**100**	**0.95**	**1662.97**

function, similar with MF and SC, we learn a matrix H for users and latent factors and a matrix W for items and latent factors. We do the recovery based on RBM model. By factorization and recovery, the value of potential user-item pairs should increase, and our prediction is based on the values in final recovered matrix.

Most of previous CF works used dimensionality reduction for high dimensional data. For low-dimensional user-item matrix, we use high latent factor number which is close to even larger than item number. We believe with large number of latent factors, even larger than item number, the abstract latent factors can still be learnt, the relation between user and item can still be inferred by factorized matrices.

IV. EXPERIMENTS

We extract 2 datasets from original data, 5-items dataset and 2-items dataset, which only remains the users who bought more than 5 or 2 types of products. The 5-items dataset is more dense, but with less users. The 2-items dataset remains data for most of users, while the data matrix is sparse. We will test the performance of selected models on this two different types of datasets.

Because of using implicit user feedback, we evaluate the result by AUC value. We compute the mean AUC values of each model for comparison.

The results of 5-items data are shown in table 1 and figure 1, the results of 2-items data are shown in table 2 and figure 2. We test 4 models, MF, DL, RBM and KM+SC, with 20, 50, 80 and 100 latent factors. Summarily, the accuracy performance of RBM is best, while the speed of other 3 models are faster. The choice of model should be based on detail requirement of data analytics.

V. CONCLUSION AND FUTURE WORK

In this paper, we try 4 models on implicit low-dimensional dataset for user preference prediction, the latent factor models

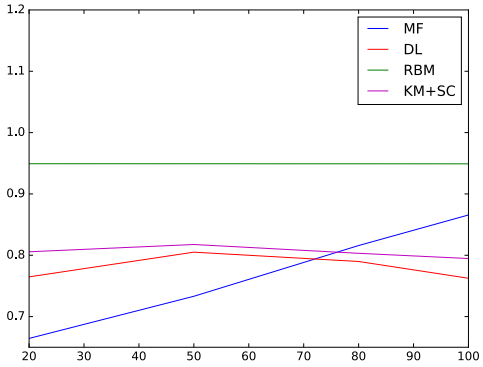

Fig. 1. Result of 4 models on 5-item dataset

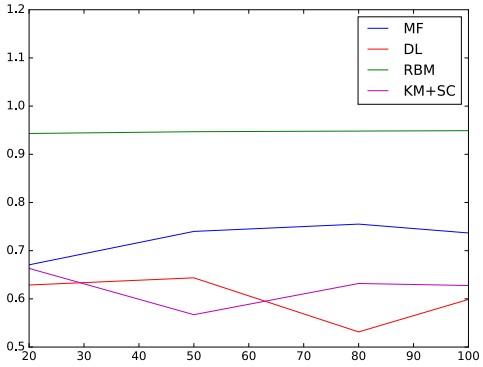

Fig. 2. Result of 4 models on 2-item dataset

are still effective. We will try to integrate more side information of users and items into the model in next step.

VI. ACKNOWLEDGEMENT

The authors thank the reviewers for their helpful comments. This work was partially supported by the Major Research Plan of National Science Fundation of China [No. 91630206].

REFERENCES

[1] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001, pp. 285–295.

[2] T. Hofmann, "Latent semantic models for collaborative filtering," *ACM Transactions on Information Systems (TOIS)*, vol. 22, no. 1, pp. 89–115, 2004.

[3] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, 2009.

[4] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," *Advances in neural information processing systems*, vol. 19, p. 801, 2007.

[5] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 689–696.

[6] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[7] R. Salakhutdinov, A. Mnih, and G. Hinton, "Restricted boltzmann machines for collaborative filtering," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 791–798.