Expert Systems with Applications 42 (2015) 5789–5799

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Matrix completion incorporating auxiliary information for recommender system design

Anupriya Gogna *, Angshul Majumdar

Indraprastha Institute of Information Technology – Delhi, Okhla Phase – III, New Delhi, Delhi 110020, India

ARTICLE INFO

Article history:
Available online 10 April 2015

Keywords:
Collaborative filtering
Matrix completion
Metadata
Latent factor model

ABSTRACT

Rating prediction accuracy of latent factor analysis based techniques in collaborative filtering is limited by the sparsity of available ratings. Usually more than 90% of the missing ratings need to be predicted from less than 10% of available ratings. The problem is highly under-determined. In this work, we propose to improve the prediction accuracy by exploiting the user's demographic information. We propose a new formulation to incorporate this information into the matrix completion framework of latent factor based collaborative filtering. The ensuing problem is efficiently solved using the split Bregman technique. Experimental evaluation indicates that the use of additional information indeed improves the accuracy of rating prediction. We also compared our proposed approach with an existing technique that incorporates auxiliary information using a graph-Laplacian framework and one utilizing neighborhood based approach; we find that our proposed method yields considerably superior results.

1. Introduction

With the ever growing density of online portals and e-commerce sites, customers can access anything online – from travel plans and conference alerts to movies and books. However, plaguing this ease of access is the information overload which a customer has to navigate through before finding the desired. This is where the role of a Recommender System (RS) (Hornick & Tamayo, 2012; Liu et al., 2014; Miller, Albert, Lam, Konstan, & Riedl, 2003) gains prominence for both customers as well as online portals. Most websites and service portals, be it movie rental services, online shopping sites or travel package providers and alike, offer some form of recommendations to the users. These recommendations provide the users more clarity, that too expeditiously and accurately in (shortlisting) limiting the items/information they need to search through, thereby improving the customer's experience. A relevant suggestion to the user improves user's satisfaction and hence popularity of online portals. Thus, design of an effective recommender system has sparked both academic and industrial interest, as it's linked directly to revenue generations for the e-commerce sites.

Most recommender system databases primarily consist of a partially filled rating matrix – containing ratings given by users to certain items. Ratings can either be explicit – like ratings given by users to items on a scale of 1–5 – or implicit – inferred from user's behavior such as browsing history or buying pattern. Explicit ratings are more dependable, but suffer from the sparsity problem – usually the task is to predict more than 90% of the ratings from less than 10% of available data. Implicit ratings are easier to come across, but less dependable. Moreover it is not possible to determine negative views from implicit ratings. In this work, we will concentrate on explicit ratings.

The problem of rating prediction is highly under-determined. In such a scenario, additional available information, apart from the rating database, can augment the basic model, improve performance and help achieve better Quality of Prediction (QoP). Most RS databases contain some secondary information such as user's demographic profile, item categories or genres and user's social network information. This information can be exploited to augment the rating information and improve the QoP.

Numerous approaches have been proposed for RS design (Adomavicius & Tuzhilin, 2005; Bobadilla, Ortega, Hernando, & Gutiérrez, 2013) – the most popular being collaborative filtering (CF) (Su & Khoshgoftaar, 2009) because of its superior performance over other methods. Conventional CF techniques use the ratings given by the user to predict his/her choice and make relevant suggestions. These methods can be further divided into memory based and (latent factor) model based approaches (Adomavicius & Tuzhilin, 2005).

* Corresponding author. Tel.: +91 9910622345.
E-mail addresses: anupriyag@iiitd.ac.in (A. Gogna), angshul@iiitd.ac.in (A. Majumdar).

http://dx.doi.org/10.1016/j.eswa.2015.04.012
0957-4174/

At the core, memory based methods (Bell & Koren, 2007) are linear interpolation techniques; they follow heuristic neighborhood based strategies relying on the assumption that if two users have rated certain items similarly, they will have similar choice on other items as well. A weighted average of ratings by similar users is used to make predictions for target user. Similar strategy can be extended to item-item similarity based approach (Sarwar, Karypis, Konstan, & Riedl, 2001) or a combination of user and item based approaches (Wang, De Vries, & Reinders, 2006). These techniques have the advantage of easier interpretability and analysis but are computationally slow and comparatively less accurate.

On the other hand, latent factor models (Hofmann, 2004) construct a lower dimensional model from the available dataset and use it for subsequent predictions. These methods are based on the premise that user's choice of an item is determined by a small number of factors/characteristics – the latent factors, thereby enabling a lower dimensional representation. It has been observed that the latent factor model provides better prediction than neighborhood based techniques. Also, they are able to provide much wider coverage and improved accuracy than their memory based counterparts (Adomavicius & Tuzhilin, 2005).

Most existing works using either of the two collaborative filtering methods utilize only the available (sparse) rating matrix, which poses a limitation on the prediction accuracy. In an attempt to improve coverage and accuracy, some works have been proposed that utilize secondary/auxiliary information, in addition to rating values, in either memory based (Bedi & Sharma, 2012; Lika, Kolomvatsos, & Hadjiefthymiades, 2014; Vozalis & Margaritis, 2007) or model based set ups (Gu, Zhou, & Ding, 2010; Koren, 2008; Ma, Zhou, Liu, Lyu, & King, 2011; Zhang, Chen, & Yin, 2013; Zhou, Shan, Banerjee, & Sapiro, 2012; Zhu, Xin, Wei, & Zhao, 2013). Although strategies which incorporate side information in memory based models, help improve accuracy and coverage to some extent, they still suffer from slow computational speeds.

The most commonly used formulation for latent factor model is the matrix factorization (MF) framework (Koren, Bell, & Volinsky, 2009) which aims to recover the rating matrix as a product of two matrices – item latent factor matrix and user latent factor matrix.

Matrix factorization, though fast, is bilinear and hence nonconvex; therefore there are no guarantees on global convergence. Recently, researchers have proposed an alternate formulation based on low rank matrix completion (Jaggi & Sulovsk, 2010; Lee, Recht, Srebro, Tropp, & Salakhutdinov, 2010; Shamir & Shalev-Shwartz, 2011), for recovering the full ratings prediction given the partial observations. It is formulated as a nuclear norm minimization problem (1)

$$\min_X \|Y - M \odot X\|_F^2 + \lambda \|X\|_* \tag{1}$$

where $X, Y \in \mathbb{R}^{m \times n}$ represent the completely filled (to be recovered) and the partially observed rating matrix respectively; λ is the regularization parameter and M a binary mask.

In case of latent factor models, as the number of independent variables (latent factors) is far less than the number of users or the number of items, the rating matrix has a low rank structure. Thus matrix completion approach (1) can be extended to RS design as well. The advantage of this approach is that it leads to a convex formulation unlike the matrix factorization framework.

In this work, our objective is to utilize auxiliary information about the users to improve rating prediction. Acquiring the auxiliary information incurs no extra cost. This is because for most ecommerce portals, before the user uses it, they need to sign-up; the supplementary information regarding the user is collected as a part of this process. Some works exist (Gupta & Gadge, 2015; Safoury & Salah, 2013) which use similar information but their focus is primarily to handle the cold start problem. We on the other hand solve a more general and challenging problem of improving overall rating prediction for all users.

So far there is no paper that incorporated auxiliary information into the matrix completion approach. Also, most existing works focus on exploiting user's social profile or network information to augment the base model; however in several cases such information is not available. In this work, we design a framework for using both (explicit) rating data and supplementary information, for improving accuracy, in a matrix completion framework. We focus on incorporating user demographic information – age, gender and occupation to augment the latent factor model; this information is more readily available to the portal than user's network structure.

Our design is based on the proposition that users belonging to same age group or sharing the same profession tend to have similar preferences. There exists few works that utilize similar arguments in either a neighborhood based set up (Vozalis & Margaritis, 2004) or using graphical modeling in a Non negative matrix factorization (NMF) framework (Zhu et al., 2013). We also design an efficient algorithm for our proposed framework. The novelty of our approach lies in presenting a new formulation (based on matrix completion) for incorporating user's demographic information. We focus on reducing the variability of (predicted) rating values amongst users grouped together by some demographic trait(s) by including suitable penalty terms in the regularized matrix completion formulation.

The main contribution of our work can be summarized as follows.

- Propose a new formulation by augmenting matrix completion framework to utilize user's demographic information in an attempt to improve prediction accuracy.
- Propose a generalized framework for the same which can be customized to include multiple information sources (like age, gender, etc.) as per available data.
- Propose an efficient algorithm using split Bregman technique (Goldstein & Osher, 2009), for solving our formulation, which is efficient both in terms of accuracy and processing speeds.
- Conduct extensive experimentation to study the impact of various kinds of demographic data on the QoP. We also propose a model including just the demographic information (without explicit ratings) which performs fairly well in terms of QoP and has a much lower complexity (run times) than existing works.

The remaining paper is organized as follows. In Section 2 we discuss MF formulation for latent factor and existing works incorporating auxiliary information in the same. We also review state of the art matrix completion algorithms. Section 3 describes our problem formulation and proposed algorithm. Section 4 contains the experimental setup, results and comparisons with existing CF techniques. Paper ends with conclusion and future direction in Section 5.

2. Related work

2.1. Latent factor model – matrix factorization framework

Latent factor model is the current de facto approach for recommender system design. As stated above, it is based on the belief that certain (handful of) features decide the user's preference for any particular item; these features are the latent factors. For example, in case of a book recommendation system, a user's liking for any book may be influenced by features such as author, genre, language etc. A book may also be described in terms of the degree to