Neurocomputing 266 (2017) 540–549

Contents lists available at ScienceDirect

Neurocomputing

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

Deep learning based matrix completion

Jicong Fan*, Tommy Chow

Department of Electronic Engineering, City University of Hong Kong, Hong Kong SAR, China

ARTICLE INFO

Article history:
Received 9 January 2017
Revised 10 April 2017
Accepted 25 May 2017
Available online 2 June 2017

Communicated by Dr. Xin Luo

Keywords:
Matrix completion
AutoEncoder
deep learning
out-of-sample extension
image inpainting
collaborative filtering

ABSTRACT

Previous matrix completion methods are generally based on linear and shallow models where the given incomplete matrices are of low-rank and the data are assumed to be generated by linear latent variable models. In this paper, we first propose a novel method called AutoEncoder based matrix completion (AEMC). The main idea of AEMC is to utilize the partially observed data to learn and construct a nonlinear latent variable model in the form of AutoEncoder. The hidden layer of the AutoEncoder has much fewer units than the visible layers do. Meanwhile, the unknown entries of the data are recovered to fit the nonlinear latent variable model. Based on AEMC, we further propose a deep learning based matrix completion (DLMC) method. In DLMC, AEMC is used as a pre-training step for both the missing entries and network parameters; the hidden layer of AEMC is then used to learn stacked AutoEncoders (SAEs) with greedy layer-wise training; finally, fine-tuning is carried out on the deep network formed by AEMC and SAEs to obtain the missing entries of the data and the parameters of the network. In addition, we also provide out-of-sample extensions for AEMC and DLMC to recover online incomplete data. AEMC and DLMC are compared with state-of-the-art methods in the tasks of synthetic matrix completion, image inpainting, and collaborative filtering. The experimental results verify the effectiveness and superiority of the proposed methods.

1. Introduction

Matrix completion [1–5] is to recover or predict the missing or unknown entries of partially observed matrices. It has been widely applied to practical problems such as image inpainting [6,7], collaborative filtering [8–10], and classification [11]. Conventional matrix completion is also called low-rank matrix completion, in which the given incomplete matrix is assumed to be of low-rank. Low-rank matrix can be completed by matrix factorization [12–14], where the incomplete matrix is approximated with the multiplication of a thin matrix and a short matrix. In the matrix factorization based methods [15–18], the matrix rank should be given in advance. In [19], a method called low-rank matrix fitting algorithm (LMaFit) was proposed for factorization based matrix completion to dynamically and adaptively adjust the matrix rank. Matrix factorization based methods have low computational complexities because the major computation is the multiplication of a thin matrix and a short matrix in each iteration. However, they are nonconvex and their performances are sensitive to the given or estimated rank.

As the missing entries of a low-rank matrix can be optimized to make the matrix have lowest rank, nuclear-norm minimization based methods [1,2,20] were proposed for matrix completion. Nuclear-norm is defined as the sum of the singular values of a matrix and is a convex relaxation for matrix rank. Nuclear-norm minimization based matrix completion can be solved through different approaches such as the singular value thresholding (SVT) algorithm [21], inexact augmented Lagrange multiplier (IALM) method [22], and alternating direction method (ADM) [13,20]. Recently, a few extensions or improvements for nuclear-norm were applied to matrix completion [6,23–25]. For example, in [6], truncated nuclear-norm (TNN) minimization was proposed for matrix completion. Truncated nuclear-norm is defined as the sum of the smallest few singular values and is a better approximation than nuclear-norm for matrix rank. The reason is that the largest few singular values usually contain important information and should be preserved; on the other hand, the small singular values should be minimized. Compared with matrix factorization based methods, nuclear-norm minimization related methods have quite higher computational complexities because of the singular value decomposition (SVD) in each iteration even if economy or truncated SVD are utilized. However, nuclear-norm minimization related methods are convex, accurate, and do not require accurately estimated rank.

The matrix factorization and nuclear-norm minimization based methods are linear methods because the low-rankness is based on

* Corresponding author.
E-mail address: fanj.c.rick@gmail.com (J. Fan).

http://dx.doi.org/10.1016/j.neucom.2017.05.074
0925-2312/

linear latent variable model [26]. Therefore, they are not effective in recovering incomplete matrix in which the data are from nonlinear latent variable model [27–29]. To handle the non-linear problem, in [30], restricted Boltzmann machines (RBMs) was proposed for collaborative filtering. More recently, AutoEncoder based collaborative filtering [31–33] were proposed and outperformed RBMs based collaborative filtering. In [31], for a recommendation system (MovieLens datasets), it was proposed to train a different AutoEncoder for each user; all AutoEncoders have the same number of hidden units but may have different number of input units set as the number of ratings given by the user. The method avoided the influence of the missing ratings but one has to train a large number of AutoEncoders. In [32] and [33], the missing entries were replaced by pre-defined constants and all the data were used to train a single AutoEncoder in which only the reconstruction errors of the observed entries were considered. The method was called AutoEncoder based collaborative filtering (AECF). However, in AECF, the influence of the biases introduced by the pre-defined constants cannot be ignored. Moreover, the performances are quite sensitive to the pre-defined constants. At last, in these works of collaborative filtering, RBM and AutoEncoder were never compared with nuclear-norm related methods and the performance differences are unknown.

In this paper, we propose a novel method called AutoEncoder based matrix completion (AEMC). In AEMC, the parameters of the AutoEncoder and the missing entries of the data matrix are simultaneously optimized to minimize the reconstruction errors of the observed entries. In the AutoEncoder to be learned for AEMC, the hidden units are much fewer than the visible units. Such a structure indicates that the given variables are redundant and can be compressed into (called encoding) and represented by (called decoding) a fewer number of hidden or latent variables. It also indicates that a subset of the input variables is able to reconstruct the hidden or latent variables, which are further able to generate the remaining input variables. That is why AEMC is able to recover the missing entries of a matrix in which the data are assumed to be given by a nonlinear latent variable model. As deep-structure neural networks could be more effective than shallow-structure neural networks [34–37], AEMC is integrated with stacked AutoEncoders (SAEs) to form a deep learning based matrix completion (DLMC). In DLMC, first, an AEMC is implemented; then the hidden variables of AEMC is used to train SAEs; finally, fine-tuning is carried out on the deep-structure AEMC. To make the offline-learned models applicable to online missing entry recovery, out-of-sample extensions for AEMC and DLMC are provided in this paper. We compare AEMC and DLMC with state-of-the-art methods of matrix factorization, nuclear-norm minimization, truncated nuclear-norm minimization, and AECF in the tasks of synthetic matrix completion, image inpainting, and collaborative filtering. The experimental results show that AEMC and DLMC are more effective than other methods.

The contributions of this paper are summarized as follows. First, AEMC, as a novel method of matrix completion is proposed to recover the missing entries of incomplete matrix in which the data are given by nonlinear latent variable model. Second, AEMC is extended to DLMC, which is a deep learning method for matrix completion and is able to outperform shallow methods of matrix completion. Finally, out-of-sample extensions for AEMC and DLMC are proposed to recover online incomplete data.

The remaining content of this paper are organized as follows. In Section 2, the previous works of matrix completion are introduced. Section 3 details our proposed methods AEMC and DLMC. Section 4 compares the proposed methods with other state-of-the-art methods in the tasks of synthetic matrix completion, image inpainting, and collaborative filtering. Section 5 draws a conclusion for this paper.

2. Previous work of matrix completion

In this section, several representative methods of matrix completion will be introduced. Given that a low-rank matrix $X \in \mathbb{R}^{m\times n}$ is partially observed, the observed entries and their positions are noted as M and Ω, respectively. Then $X_{i,j} = M_{i,j}$ for each pair $(i, j) \in \Omega$. X can be completed by solving the following matrix factorization problem [9,12,16–18]:

$$\min_{X,L,R} \frac{1}{2}(\|L\|_F^2 + \|R\|_F^2), \quad s.t.\ X = LR,\ X_{i,j} = M_{i,j},\ (i,j) \in \Omega \tag{1}$$

where $L \in \mathbb{R}^{m\times r}$ is a thin matrix and $R \in \mathbb{R}^{r\times n}$ is a short matrix. The parameter r should be determined beforehand and the optimal value is the rank of X. In practice, because X is incomplete, it is difficult to know or estimate r. To cope with the problem, [19] proposed a low-rank matrix fitting (LMaFit) algorithm for matrix completion. In LMaFit, with the model of (1), the parameter r is dynamically and adaptively adjusted. These matrix factorization based methods are nonconvex and sensitive to the parameter r though they have low computational complexities.

Matrix completion can also be solved by rank minimization. However, direct rank minimization is NP-hard. As a convex relaxation of matrix rank, nuclear-norm can be applied to matrix completion by solving the following problem

$$\min_X \|X\|_*,\ s.t.\ X_{i,j} = M_{i,j},\ (i,j) \in \Omega \tag{2}$$

where $\|\cdot\|_*$ denotes the nuclear norm, the sum of the singular values (σ) of a given matrix, i.e., $\|X\|_* = \sum_{i=1}^{min(m,n)} \sigma_i(X)$. Problem (2) is convex and can be solved by techniques such as the singular value thresholding (SVT) algorithm [21], inexact augmented Lagrange multiplier (IALM) method [22], and alternating direction method (ADM) [20].

As known, the largest few singular values of a matrix often contain important information and should be preserved if possible. Therefore, in [6], the truncated nuclear norm (TNN), a better rank approximation than nuclear-norm, was applied to matrix completion. TNN is defined as the nuclear norm subtracted by the sum of the largest few singular values, i.e., $\|X\|_r = \|X\|_* - \sum_{i=1}^{r} \sigma_i(X) = \sum_{i=r+1}^{min(m,n)} \sigma_i(X)$. For matrix completion, [6] proposed to solve the following problem

$$\min_X \|X\|_r,\ s.t.\ X_{i,j} = M_{i,j},\ (i,j) \in \Omega \tag{3}$$

which can be reformulated as

$$\min_X \|X\|_* - Tr(U_l X V_l^T), \quad s.t.\ X_{i,j} = M_{i,j},\ (i,j) \in \Omega, \tag{4}$$

where U_l (V_l) are the first r columns of U (V) given by the singular value decomposition of X in the lth iteration, i.e., $X_l = USV^T$ [6]. The optimization of (4) can be solved by alternating direction method of multipliers (ADMM) [38].

Recently, neural networks were applied to collaborative filtering [30–33]. For example, in [32], AutoEncoder based collaborative filtering (AECF) was proposed. In AECF, first, the missing entries of the data were replaced with pre-defined constants; then the data were utilized to learn an AutoEncoder, for which the reconstruction errors of the observed entries were minimized; finally, the missing entries were set as the corresponding outputs of the network.

3. Deep learning based matrix completion (DLMC)

3.1. Matrix completion by AutoEncoder and deep learning

Assume that a set of variables or measurements are given by the following nonlinear latent variable model

$$x = f(z) + \epsilon, \tag{5}$$