Engineering Applications of Artificial Intelligence 68 (2018) 32–39

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Graph Regularized Nonnegative Matrix Factorization with Sample Diversity for Image Representation

Changpeng Wang [a,*], Xueli Song [a], Jiangshe Zhang [b]

[a] School of Mathematics and Information Science, Chang'an University, Xi'an 710064, China
[b] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an 710049, China

ARTICLE INFO

Keywords:
Nonnegative matrix factorization
Semi-supervised learning
Sample diversity
Clustering

ABSTRACT

Nonnegative Matrix Factorization (NMF) is an effective algorithm for dimensionality reduction and feature extraction in data mining and computer vision. It incorporates the nonnegativity constraints into the factorization, and thus obtains a parts-based representation. However, the existing NMF variants cannot fully utilize the limited label information and neglect the unlabeled sample diversity. Therefore, we propose a novel NMF method, called Graph Regularized Nonnegative Matrix Factorization with Sample Diversity (GNMFSD), which make use of the label information and sample diversity to facilitate the representation learning. Specifically, it firstly incorporates a graph regularization term that encode the intrinsic geometrical information. Moreover, two reconstruction regularization terms based on labeled samples and virtual samples are also presented, which potentially improve the new representations to be more discriminative and effective. The iterative updating optimization scheme is developed to solve the objective function of GNMFSD and the convergence of our scheme is also proven. The experiment results on standard image databases verify the effectiveness of our proposed method in image clustering.

1. Introduction

Matrix factorization is a useful tool for data representation as well as dimensionality reduction, and various matrix factorization based algorithms have been well studied. Perhaps, most well known matrix factorization based algorithms are Principal Component Analysis (PCA) (Jolliffe, 1989), Independent Component Analysis (Hyvarinen and Oja, 2000), Linear Discriminant Analysis (LDA) (Belhumeur et al., 1997), Local Linear Embedding (LLE) (Chen and Liu, 2011), and Nonnegative Matrix Factorization (NMF) (Lee and Seung, 1999). These methods aim to learn a compact low-dimensional representation of original data for further applications.

NMF is an unsupervised and effective analysis algorithm due to its theoretical interpretation and desired performance (Wang and Zhang, 2013). It aims to find a linear approximation to the original matrix by basis matrix and coefficient feature matrix, and simultaneously enforces the elements in both basis vectors and representation coefficients to be nonnegative. This constraint allows NMF with additive combination to approximation to the original data which accords with the cognitive process of human brain (Wachsmuth et al., 1994; Logothetis and Sheinberg, 1996). Thus, NMF and its variants have been widely used in different real-world applications, such as audiovisual document structuring (Essid and Fevotte, 2013), hyperspectral image analysis (Gillis and Plemmons, 2013), graph matching problem (Jiang et al., 2014), maintenance activities identification (Feng et al., 2016), face recognition (Zhi et al., 2011; Chen et al., 2016), data clustering (Li et al., 2014; Lu and Miao, 2016), etc. A comprehensive review about the theoretical research of NMF can be found in Wang and Zhang (2013).

Despite NMF has solid mathematical theory foundations and encouraging performances, it still have some limitations including neglecting intrinsic geometric structure of the data and lacking discriminative information for clustering. Recently, several variants of NMF have been proposed to improve the performance. Graph Regularized NMF (GNMF) (Cai et al., 2011) utilizes the Laplacian graph as a regularization term to exploit the locality property of the data. Dual GNMF (Shang et al., 2012) and Multiple GNMF (Wang et al., 2013) have been proposed by adding more constraints to the original objective function. To use the supervised information, Constrained NMF (CNMF) (Liu et al., 2011) encodes label information into NMF to make the data points share the same label in the new representation process. However, this constraint ignores the local geometrical structure of data set.

* Corresponding author.
E-mail address: cpwang@chd.edu.cn (C. Wang).

https://doi.org/10.1016/j.engappai.2017.10.018
Received 10 April 2017; Received in revised form 28 August 2017; Accepted 22 October 2017
0952-1976/

Furthermore, Graph Regularized and Sparse NMF with hard Constraints (GSNMFC) (Sun et al., 2016) integrates the geometrical structure and label information as well as sparse constraint in a joint framework. However, in GSNMFC, the unlabeled data are not fully utilized due to no constraints on them, and the sparsity constraint leads the intrasample structure information to be ignored during the learning process (Lu and Miao, 2016). Therefore, it is necessary to establish a semi-supervised framework for NMF which can fully utilize the label information and the diversity of unlabeled data simultaneously.

In this paper, we propose a novel NMF method, called Graph Regularized Nonnegative Matrix Factorization with Sample Diversity (GNMFSD), which adds three constraint conditions to ensure the effectiveness of the obtained representations. Our algorithm is motivated by recent study on dictionary learning, and particularly, sample diversity learning proposed by Xu et al. (2017). In our method, we incorporate label information into the graph to encode the intrinsic geometrical structures of the data space, and linear regression term based on the labeled samples to promote the discriminant power of the learned basis vectors. Furthermore, sample diversity term is used to try to obtain more effective representations of unlabeled data. By combining these three constraints with the graph-based regularizer, we expect to fully utilize the unlabeled data and the limited label information to improve the clustering performance. In addition, we discuss how to solve the corresponding optimization problem, and theoretically prove that our objective function is nonincreasing under the corresponding update rules. Finally, we conduct extensive experiments to validate the effectiveness of our GNMFSD method. The main contributions of this paper are summarized as follows:

1. The proposed GNMFSD method do not only take the diversity of unlabeled samples into account, but also characterize both the underlying local geometrical information and the global discriminative information of the samples with additional regularization terms. Hence, the learned effective representations have more discriminating power for the data representation, which could improve the performance on clustering accuracy and normalized mutual information.
2. The updating rules to solve the objective function and the convergence proof are provided. Experiments on the real databases are conducted to demonstrate the algorithm effectiveness quantitatively.

The rest of this paper is organized as follows. Section 2 reviews the related works including NMF, GNMF, CNMF. In Section 3, we introduce our proposed algorithm, as well as the optimization scheme and convergence study. Experimental results are presented for illustration in Section 4 and Section 5 concludes this paper.

2. Related work

In this section, we briefly discuss the important algorithms which are relevant to our work, including NMF (Lee and Seung, 1999), Graph Regularized NMF (GNMF) (Cai et al., 2011) and Constrained NMF (CNMF) (Liu et al., 2011).

2.1. NMF

Nonnegative matrix factorization (NMF) is linear model that learns a part-based representation of the data. it aims to find two nonnegative matrices $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$ to approximate the original data matrix $X = [x_1, \dots, x_n] \in \mathbb{R}^{m\times n}$:

$$X \approx UV^T. \tag{1}$$

In the above representation, U can be considered as a set of basis vectors and V as the representation of each sample with respect to these basis vectors. In order to measure the quality of the decomposition, the Euclidean distance based objective function is expressed as:

$$\min_{U,V} \|X - UV^T\|_F^2 \quad \text{s.t.} \quad U \geq 0, V \geq 0 \tag{2}$$

where $\|\cdot\|_F$ denotes the matrix Frobenius norm.

Because the objective function is not convex in both U and V, it is difficult to obtain the global minimum of the objective function. Lee and Seung (1999) provided update rules to obtain a local minimum and proved its convergence. The update rules for the objective function are given as:

$$u_{ik} \leftarrow u_{ik}\frac{(XV)_{ik}}{(UV^TV)_{ik}}, v_{jk} \leftarrow v_{jk}\frac{(X^TU)_{jk}}{(VU^TU)_{jk}}. \tag{3}$$

The iterative update procedure is executed repeatedly to decrease the approximation error, and final U and V are obtained when the given terminal condition is met. By NMF, each data point x_i is approximated by a linear combination of the columns of U, weighted by the ith column of V. The NMF has achieved good results in many practical applications due to its effectiveness and simple to implement.

2.2. GNMF

As aforementioned, NMF learns a part-based representation in Euclidean space, but it neglects the intrinsic geometric structure of the original data. In order to preserve this inherent structure information during the matrix decomposition, Cai et al. (2011) proposed a Graph Regularized Nonnegative Matrix Factorization (GNMF) algorithm. In GNMF, a nearest-neighbor graph is constructed to preserve the geometrical structure of data space, and the objective function of GNMF can be formulated as below:

$$\min_{U,V} \|X - UV^T\|_F^2 + \lambda Tr(V^TLV) \quad \text{s.t.} \quad U \geq 0, V \geq 0 \tag{4}$$

where $L = D - W$ is Laplacian matrix, W is the weight matrix to measure the similarity between the nearby data points, D is a diagonal matrix whose entries are column (or row) sums of W, $D_{jj} = \sum_l W_{jl}$. The λ is a regularization parameter which balance the reconstruction error and manifold term. The update rules to solve (4) are given below:

$$u_{ik} \leftarrow u_{ik}\frac{(XV)_{ik}}{(UV^TV)_{ik}}, v_{jk} \leftarrow v_{jk}\frac{(X^TU + \lambda WV)_{jk}}{(VU^TU + \lambda DV)_{jk}}. \tag{5}$$

With the local manifold learning in GNMF, the nearby data points are encouraged to be as close as possible in the new data space.

2.3. CNMF

In order to take advantage of the partial labeled data, CNMF (Liu et al., 2011) introduces a label constraint matrix A and takes the label information of data into account. Suppose there are c classes, the first l data points $x_1, \dots, x_l$ are labeled, and the rest of the $n - l$ data points $x_{l+1}, \dots, x_n$ are unlabeled. The $l \times c$ indicator matrix C is defined as below:

$$c_{ij} = \begin{cases} 1, & \text{if } x_i \text{ is labeled with the } j\text{th class} \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

With the indicator matrix C, the label constraint matrix A can be defined as below:

$$A = \begin{bmatrix} C_{l\times c} & 0 \\ 0 & I_{n-l} \end{bmatrix} \tag{7}$$

where I_{n-l} is an $(n-l)\times(n-l)$ identity matrix. With the introduced matrices, the original data X is approximated as $X \approx UV^T = U(AZ)^T$. The indicator matrix C means that if samples i and j have the same label, then their weighted coefficient vector are also same. The CNMF