Neurocomputing 149 (2015) 1021–1037

Contents lists available at ScienceDirect

Neurocomputing

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

Label propagation based semi-supervised non-negative matrix factorization for feature extraction

Yugen Yi [a,b], Yanjiao Shi [a], Huijie Zhang [a], Jianzhong Wang [a,c,*], Jun Kong [b,**]

[a] College of Computer Science and Information Technology, Key Laboratory of Intelligent Information Processing of Jilin Universities, Northeast Normal University, Changchun, China
[b] School of Mathematics and Statistics, Northeast Normal University, Changchun, China
[c] National Engineering Laboratory for Druggable Gene and Protein Screening, Northeast Normal University, Changchun, China

ARTICLE INFO

Article history:
Received 3 September 2013
Received in revised form
16 March 2014
Accepted 22 July 2014
Communicated by S. Choi
Available online 2 August 2014

Keywords:
NMF
Feature extraction
Label propagation
LpSNMF
Classification
Clustering

ABSTRACT

As a feature extraction method, Non-negative Matrix Factorization (NMF) has attracted much attention due to its effective application to data classification and clustering tasks. In this paper, a novel algorithm named Label propagation based Semi-supervised Non-negative Matrix Factorization (LpSNMF) is proposed. For the sake of making full use of label information, our LpSNMF algorithm takes the distribution relationships between the labeled and unlabeled data samples into consideration and integrates the procedures of class label propagation and matrix factorization into a joint framework. Moreover, an iterative updating optimization scheme is developed to solve the objective function of the proposed LpSNMF and the convergence of our scheme is also proven. Extensive experimental results on several UCI benchmark data sets and four image data sets (such as Yale, CMU PIE, UMIST, and COIL20) demonstrate that by propagating the label information and factorizing the matrix alternately, our algorithm can obtain better performance than some other algorithms.

1. Introduction

For many real world applications such as face image recognition, computer vision, information retrieval, object recognition and text categorization, the observed data vectors often lie in a high-dimensional space, which leads to the increase of the storage space and computational cost. Thus, extracting the most useful low-dimensional representation from such high-dimensional data not only helps to avoid the "curse of dimensionality" problem [1,2], but also contributes to accomplish the data analysis tasks at a low computational cost [3,4]. Recently, feature extraction techniques have attracted more and more attentions as tools for data representation as well as dimensionality reduction [5]. Among these techniques, matrix decomposition based algorithms are well-studied. Nowadays, the most representative matrix decomposition based feature extraction algorithms are Principal Component Analysis (PCA) [6] and Non-negative Matrix Factorization (NMF) [5]. Unlike PCA which represents the original high-dimensional data as a linear combination of basis matrix and allows negative elements in the basis matrix and representation coefficients, the non-negative constraints imposed in NMF enforces the elements in both basis vectors and representation coefficients to be non-negative. Thus, it possesses better psychological and physiological interpretation for naturally occurring data whose representation may be parts-based in the human brain and has become an imperative tool in multivariate data analysis [5].

Recently, although the NMF algorithm has been successfully applied to many real-world applications [7–14], it still suffers from the following limitations. Firstly, the original NMF only concentrated on learning a non-negative parts-based representation of the high-dimensional data in Euclidean space. Thus, the intrinsic geometric structure of the data was neglected in it. Since researchers have shown that the observed high-dimensional data (such as face image, text document and so on) always lies on a nonlinear low-dimensional manifold [15–17], some extensions of NMF were proposed to preserve the intrinsic manifold structure of the high-dimensional data. In [18], a Locality Preserving Nonnegative Matrix Factorization (LPNMF) algorithm was presented by Cai et al. LPNMF characterized the local geometric structure of the input data by constructing a k nearest-neighbor (kNN) graph, which plays an essential role in graph-based manifold learning algorithms, such as Laplacian Eigenmap (LE) [17] and Constrained Large Margin Local Projection (CMLP) [19]. Similarly, Gu et al. proposed a Neighborhood Preserving NMF (NPNMF) algorithm [20] in which the intrinsic geometry of the input data was

* Corresponding author at: College of Computer Science and Information Technology, Key Laboratory of Intelligent Information Processing of Jilin Universities, Northeast Normal University, Changchun, China. Tel.: +86 43184536326.
** Corresponding author.
E-mail addresses: wangjz019@nenu.edu.cn (J. Wang), kongjun@nenu.edu.cn (J. Kong).

http://dx.doi.org/10.1016/j.neucom.2014.07.031
0925-2312/

captured by the assumption that each point can be reconstructed by the data points in its neighborhood. More recently, a Graph Regularized Nonnegative Matrix Factorization (GNMF) was proposed in [21]. In this algorithm, the geometrical information of the high-dimensional data was also explicitly considered by incorporating an additional graph regularization term into the objective function of NMF.

The second limitation of NMF is that it is an unsupervised matrix decomposition algorithm, so it cannot perform well for classification or recognition tasks due to lacking discriminative information. In order to improve the classification performance, many supervised extensions of NMF were proposed. In Ref. [22], a supervised extension of NMF termed Fisher NMF (FNMF) was presented. In FNMF, a regularization term which indicates the difference between inter-class and intra-class scatters was incorporated into the original NMF. Zafeiriou et al. proposed a Discriminant NMF (DNMF) in [23]. Similar to FNMF, DNMF also incorporated the inter-class and intra-class scatters into the NMF as a discriminant constraint. Though the label information was considered in FNMF and DNMF, these two algorithms were both inspired by Fisher Linear Discriminant Analysis and did not take the geometric structure of the input data into account. Thus, a new supervised NMF algorithm named NMF-KNN was proposed in [24]. In NMF-KNN, two nearest neighbor graphs (inter-class and intra-class) were constructed to exploit discriminative information as well as geometric structure of the high-dimensional data. In Ref. [25], Yang et al. presented a unified framework, called Non-negative Graph Embedding (NGE), for non-negative matrix factorization. By introducing the intrinsic and penalty graphs into their framework, the NGE algorithm can preserve the similarities measured by the graphs and minimize the matrix reconstruction error simultaneously, which makes it can be applied for both unsupervised and supervised learning tasks. In Ref. [26], Nikitidis et al. presented a Subclass Discriminant Nonnegative Matrix Factorization (SDNMF). Different from other supervised NMF algorithms which assume the underlying data distribution in each class is unimodal, SDNMF assumed that the data of each class comes from a multimodal distribution and incorporated a clustering based discriminant criterion into the NMF. Thus, it can remedy the limitation of other algorithm and obtain better classification performance.

Though the experimental results have shown that the supervised extensions of NMF outperformed the unsupervised ones in the classification and recognition tasks, these supervised NMF algorithms demand that the input training data is completely labeled, which may prevent their application in some problems. Since the acquisition of labeled data for some real world tasks often requires a skilled human agent or physical experiment, the cost associated with the labeling process may render a fully labeled training set time-consuming and infeasible. Conversely, acquisition of unlabeled data is relatively inexpensive. For instance, a large number of unlabeled images can be easily obtained from the Internet, or from a digital camera for surveillance or web chatting. Thus, in order to utilize the labeled data and unlabeled data simultaneously, numerous semi-supervised NMF algorithms have been proposed [27–29]. In Ref. [27], Chen et al. proposed a semi-supervised non-negative matrix factorization (SSNMF) approach for data clustering. In this algorithm, the users were able to provide some pairwise constraints on a few labeled and unlabeled data samples to indicate the similarity or dissimilarity between them. Liu et al. [28] also proposed a Constrained Non-negative Matrix Factorization (CNMF) algorithm, which incorporated the label information as additional hard constraints into NMF. Furthermore, another semi-supervised version of NMF (SNMF) has been developed in [29]. In SNMF, the partial class label matrix and data matrix are jointly incorporated into the original NMF to share the common factor matrix. Through the experiments in [27–29], it can be seen that when there are no sufficient labeled data samples available, the learning accuracy of the NMF algorithm can be considerably improved by employing the information of labeled and unlabeled data conjunctively. However, a major limitation in these semi-supervised NMF is that the label information is not fully used in them. In SSNMF, the pairwise constraints were given manually, thus only the label information of the data points with constraints was considered. In CNMF and SNMF, although the labeled and unlabeled data was both utilized during the matrix decomposition, the distribution relationships between the labeled and unlabeled data samples were not exploited.

In this paper, a novel feature extraction algorithm termed Label propagation based Semi-supervised Non-negative Matrix Factorization (LpSNMF) is presented to overcome the limitations of the existing semi-supervised NMF algorithms. Label propagation is a newly proposed semi-supervised learning framework [30], which is based on the “cluster assumption”. In other words, the nearby data points from the same global cluster should share similar labels in label propagation. Thus, by integrating the label propagation with the procedure of non-negative matrix factorization, the distribution relationship between the labeled and unlabeled data is explicitly considered in our LpSNMF, which makes our algorithm distinct from other semi-supervised NMF methods. Furthermore, compared with the Semi-supervised Orthogonal Discriminant Analysis (SODA) [31] which also combines label propagation with feature extraction, there are still two differences between it and our LpSNMF. Firstly, the label propagation and feature extraction procedures are implemented separately in SODA. Thus, the interaction between the label propagation and feature extraction is neglected. However, the procedures of label propagation and feature extraction are jointly executed in the proposed LpSNMF, which makes our algorithm more efficient. Secondly, since the low-dimensional features were extracted by Orthogonal Linear Discriminant Analysis [32] in SODA, it cannot obtain the parts-based representation of high-dimensional data. Nevertheless, the low-dimensional features in our algorithm are obtained by non-negative matrix factorization. Therefore, as can be seen from Section 4, the interpretability of the feature extraction results obtained by our LpSNMF is better than SODA.

The rest of the paper is organized as follows. In Section 2, NMF, GNMF and SNMF are briefly reviewed. The proposed LpSNMF algorithm is presented in Section 3. Section 4 shows the experimental results to evaluate the proposed LpSNMF algorithm. Finally, the conclusions are given in Section 5.

2. Related works

In this section, three existing algorithms including NMF [5], Graph Regularized NMF (GNMF) [21] and Semi-supervised NMF (SNMF) [29] are reviewed.

2.1. NMF

Non-negative matrix factorization (NMF) is an unsupervised learning method for low-rank approximation of nonnegative data [5]. Given a non-negative matrix $X=[x_1,x_2,...,x_N]\in R_+^{D\times N}$ which denotes N non-negative data points in a D dimensional space. NMF aims to find two low-rank non-negative matrices inducing basis matrix $A=[a_{ij}]\in R_+^{D\times R}$ and factor matrix $S=[s_{ij}]\in R_+^{N\times R}$ satisfying $X=AS^T$, where $R\ll \min(D,N)$. Therefore, the objective function of