Knowledge-Based Systems 000 (2016) 1–10

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Graph-based discriminative concept factorization for data representation

Huirong Li[a,b], Jiangshe Zhang[a,*], Junying Hu[a], Chunxia Zhang[a], Junmin Liu[a]

[a] The School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an,710049, China
[b] The School of Mathematics and Computer Application, Shangluo University, Shang luo, 726000, China

ARTICLE INFO

Article history:
Received 30 June 2016
Revised 8 November 2016
Accepted 13 November 2016
Available online xxx

Keywords:
Concept factorization
Semi-supervised learning
Data representation
Label information

ABSTRACT

Nonnegative Matrix Factorization (NMF) and Concept Factorization (CF) have been widely used for different purposes such as feature learning, dimensionality reduction and image clustering in data representation. However, CF is a variant of NMF, which is an unsupervised learning method without making use of the available label information to guide the clustering process. In this paper, we put forward a semi-supervised discriminative concept factorization (SDCF) method, which utilizes the limited label information of the data as a discriminative constraint. This constraint forces the representation of data points within the same class should be very close together or aligned on the same axis in the new representation. Furthermore, in order to utilize the local manifold regularization, we propose a novel semi-supervised graph-based discriminative concept factorization (GDCF) method, which incorporates the local manifold regularization and the label information of the data into the CF to improve the performance of CF. GDCF not only encodes the local geometrical structure of the data space by constructing K-nearest graph, but also takes into account the available label information. Thus, the discriminative abilities of data representations are enhanced in the clustering tasks. Experimental results on several databases expose the strength of our proposed SDCF and GDCF methods compared to the state-of-the-art methods.

1. Introduction

In recent years, data representation has attracted much more attention in various research fields like machine learning and computer vision[1–4]. Generally, we often need to confront with the high dimensional data, while the essential structures of images are laid in a relatively low dimensional space. Thus, many efforts have been devoted to seeking a suitable low dimensional representation for the high dimensional data. Matrix factorization is a popular technique for this data representation. In this paper, we focus on matrix factorization-based methods, which includes NMF [5], GNMF [6], CF [7], LCCF [8], etc. NMF decomposes a data matrix $\mathbf{X}$ into the product of two matrices $\mathbf{U}$ and $\mathbf{V}$ whose entries are constrained to be nonnegative [1,2,5]. The nonnegative constraints in NMF lead to a parts-based representation, which is widely used in various real word applications such as face recognition [9,10], document clustering [5,11], image representation[12]. However, NMF fails to be performed on negative data, and cannot perform effectively in the reproducing kernel Hilbert data space. To overcome these disadvantages, a variation of NMF, namely Concept Factorization (CF)[7], is proposed for data representation. In CF, each cluster is expressed by a linear combination of data points and each data point is repressed by a combination of cluster centers. Compared with NMF, CF can not only be kernelized, but also be performed [13]. To further improve the performance, GNMF [6] and LCCF [8] were proposed to incorporate the graph regularization terms into NMF and CF, respectively. GNMF constructs a nearest neighbor graph to encode the geometrical information of the data space, and aims to find a parts-based representation space in which two data points are close enough to each other when they are connected in the graph. Similar to GNMF, LCCF extracts the underlying concepts with respect to the intrinsic manifold structure by using a graph model and thus the data points with similar concepts can be well clustered.

All above aforementioned methods are completely unsupervised learning algorithms, and they cannot take advantage of the available label information. Many researchers have pointed out that when a small amount of labeled data is used in conjunction with unlabeled data, it can improve the performance of machine learning. There are also a few methods which can be viewed as semi-supervised learning methods [12,14–20]. The label information is added to NMF and CF in Refs. [12,18–22]. For examples, Discriminative Nonnegative Matrix Factorization (DNMF)[19] utilizes the la-

* Corresponding author.
E-mail address: jszhang@mail.xjtu.edu.cn (J. Zhang).

http://dx.doi.org/10.1016/j.knosys.2016.11.012
0950-7051/

bel information of a fraction of data as a discriminative constraint, which enforces the samples with the same label to be aligned on the axis in the new representation. Constrained Nonnegative Matrix Factorization (CNMF)[12] takes label information as hard constraints and the data points with the same class label must be strictly mapped to share the same representation in the new representation space. Discriminative concept factorization (DCF)[15] adopts a unified objective to combine the task of data reconstruction with the task of classification. Constrained Concept Factorization (CCF)[21] and Class-driven Concept Factorization (CDCF)[22] for image representation are such semi-supervised learning algorithms. Similar to CNMF, CCF provides a semi-supervised matrix decomposition for extracting the image concepts that are consistent with the know label information. The CCF can guarantee that the data points sharing the same label have the same concept in the low dimensional space. However, since CCF maps the images with the same label onto the same concept, it is infeasible when there is only one labeled data point to rely on. Therefore, CDCF associates the class labels of data points with their representations by introducing a class-driven constraint. This constraint forces the representations of data points to be more similar within the same class while different between classes.

CF is an unsupervised learning method, and pairwise constraints are incorporated into CF to guide the learning process in [8,13,16,21,22]. However, these methods cannot make use of the pairwise constraints as well as local manifold regularization to improve the performance of CF. Their performances could degrade to the level of the CF method when only limited label information is available. To address the above issue, we put forward a semi-supervised discriminative concept factorization (SDCF) method for data representation and clustering tasks. The SDCF method emphasizes that the data points belonging to same class should be very close together or aligned on the same axis in the new representation, but not to be merged into a single point. Besides, in order to consider the local manifold regularization, we propose a semi-supervised graph-based discriminative concept factorization (GDCF) method, which takes the limited label information and the local manifold regularization into account, simultaneously. Extensive experiments are conducted on publicly available databases, our experiments show that our proposed SDCF and GDCF perform better than other state-of-the-arts matrix factorization methods.

It is worth highlighting the contributions of this paper as follows:

- In our proposed SDCF, the label information constraint is introduced by coupling discriminative regularizer to the main objective function of CF.
- GDCF is also a semi-supervised learning method, which encodes the intrinsic geometrical and discriminative structures of the data space by constructing $k-$nearest neighbor graph, and takes the label information as additional constraints to decompose the data matrix. Specifically, we design a new concept factorization objective function incorporating the manifold regularization and the limited label information into it. So the new representation of GDCF can exhibit more discriminative power than others learning methods, such as DCF,DNMF, SDCF, and so on.
- We develop the corresponding multiplicative updating optimization schemes to derive the iterative updating rules of three matrices **W, V** and **A**. More importantly, the convergence proof of our proposed GDCF is provided.

The remainder of this paper is organized as follows. Section 2 briefly reviews some related work toward CF and several semi-supervised CF algorithms. The proposed SDCF and GDCF models as well as the optimization methods are described in Section 3 and the proof of convergence of the proposed GDCF in Section 4. Experimental results are reported in Section 5 with detailed analysis. Finally, we provide some concluding remarks in Section 6.

2. Related works

2.1. Concept factorization (CF)

CF is an efficient matrix factorization technique. It has been shown that CF is a very useful approach for applications like pattern recognition and image clustering [7,8]. Given a data matrix $\mathbf{X} = [x_1, x_2, \cdots, x_n] \in \mathbf{R}^{m\times n}$, x_i be the $m-$dimensional feature vector representing the data point i and u_j be the center of concept j, where $i = 1, 2, \cdots, n$, $j = 1, 2, \cdots, k$. Firstly, each underlying concept/cluster can be characterized by a linear combination of entire data points. This representation can be formulated as follows:

$$u_j = \sum_i w_{ij} x_i \tag{1}$$

where w_{ij} is a nonnegative association weight indicating the degree of data point i relating to concept j. Secondly, each data point can be approximated by a linear combination of all the concepts, we have:

$$x_i \approx \sum_j v_{ij} u_j \tag{2}$$

where v_{ij} is a nonnegative number indicating the projection value of x_i onto the base (or concept center) u_j. From the Eqs. (1) and (2), we have the following form:

$$\mathbf{X} \approx \mathbf{XWV^T} \tag{3}$$

where $\mathbf{W} \in \mathbf{R}^{n\times k}$ and $\mathbf{V} \in \mathbf{R}^{n\times k}$. The quality of the approximation can be quantified by using a cost function with Euclidean distance metric. Then it turns to minimize the following objective function:

$$\begin{aligned} &\min_{W,V} O = \|\mathbf{X} - \mathbf{XWV^T}\|_F^2 \\ &s.t \quad \mathbf{W} \geq 0, \mathbf{V} \geq 0 \end{aligned} \tag{4}$$

According to [7], the updating rules of CF in Eq. (4) can be introduced as follows:

$$w_{ij} \leftarrow w_{ij} \frac{(\mathbf{KV})_{ij}}{(\mathbf{KWV^TV})_{ij}}$$
$$v_{ij} \leftarrow v_{ij} \frac{(\mathbf{KW})_{ij}}{(\mathbf{VW^TKW})_{ij}}$$

where $\mathbf{K} = \mathbf{XX^T}$. We can construct a kernel matrix by using a kernel function. Thus, CF can easily be kernelized to enhance the performance in some case. Please refer to [7] for details.

2.2. Semi-supervised concept factorization

CF is an unsupervised learning algorithm, which cannot make use of the label information when such information is available. In order to incorporate the available label information, there are a few methods which can be viewed semi-supervised CF learning algorithms recently[13,14,21–23]. For example, Liu et al. [21] proposed a novel CCF that extracts the image concepts consistent with the known label information based on matrix factorization. CCF introduces a $p \times c$ label indicator matrix $\mathbf{C}$ with $\mathbf{C}_{i,j} = 1$ if x_i is labeled with class j and $\mathbf{C}_{i,j} = 0$ otherwise, where p is the number of labeled data points, c is the number of data classes. Based on the indicator matrix $\mathbf{C}$, a label constraint matrix $\mathbf{B}$ is defined as follows:

$$\mathbf{B} = \begin{pmatrix} \mathbf{C}_{p\times c} & 0 \\ 0 & \mathbf{I}_{n-p} \end{pmatrix}$$

Please cite this article as: H. Li et al., Graph-based discriminative concept factorization for data representation, Knowledge-Based Systems (2016), http://dx.doi.org/10.1016/j.knosys.2016.11.012