Knowledge-Based Systems 102 (2016) 127–139

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Constrained neighborhood preserving concept factorization for data representation

Mei Lu[a,b,1,2,*], Li Zhang[a,1], Xiang-Jun Zhao[b,2], Fan-Zhang Li[a,1]

[a] College of Computer Science and Technology, Soochow University, Suzhou 215006, China
[b] College of Computer Science and Technology, Jiangsu Normal University, Xuzhou 221116, China

ARTICLE INFO

Article history:
Received 27 September 2015
Revised 24 March 2016
Accepted 3 April 2016
Available online 6 April 2016

Keywords:
Concept factorization
Locally consistent concept factorization
Semi-supervised document clustering
Neighborhood preserving
Data representation

ABSTRACT

Matrix factorization based techniques, such as nonnegative matrix factorization (NMF) and concept factorization (CF), have attracted a great deal of attentions in recent years, mainly due to their ability of dimension reduction and sparse data representation. Both techniques are of unsupervised nature and thus do not make use of a priori knowledge to guide the clustering process. This could lead to inferior performance in some scenarios. As a remedy to this, a semi-supervised learning method called Pairwise Constrained Concept Factorization (PCCF) was introduced to incorporate some pairwise constraints into the CF framework. Despite its improved performance, PCCF uses only a priori knowledge and neglects the proximity information of the whole data distribution; this could lead to rather poor performance (although slightly improved comparing to CF) when only limited a priori information is available. To address this issue, we propose in this paper a novel method called Constrained Neighborhood Preserving Concept Factorization (CNPCF). CNPCF utilizes both a priori knowledge and local geometric structure of the dataset to guide its clustering. Experimental studies on three real-world clustering tasks demonstrate that our method yields a better data representation and achieves much improved clustering performance in terms of accuracy and mutual information comparing to the state-of-the-arts techniques.

1. Introduction

Data representation is a key topic in machine learning and pattern recognition. Recent studies have shown that compact representation of data sets can greatly facilitate many learning tasks such as clustering and classification. For example, as demonstrated in [2,23], data similarity can be measured much more accurately in lower dimensional spaces. As one of the popular methods for dimension reduction, matrix factorization has received a great deal of attentions in recent years, and many techniques have been developed such as PCA [14], NMF [12,17,24,25,31,34], CF[30], etc. NMF focuses on the analysis of data matrices whose elements are nonnegative, and can be used to obtain a new part-based representation in some lower dimensional space. Resulting factorization from NMF often enables better semantic interpretation, and thus can be used to derive more accurate clustering. It has been shown that NMF provides better performance than PCA in face recognition [19] and document clustering [31]. A major limitation of NMF is that it could not effectively perform in some transformed spaces such as the reproducing kernel Hilbert space (RKHS) [3]. To address this issue, Xu and Gong [30] proposed CF, which works in any data representation space.

Another often used method for dimension reduction is manifold learning. Since the year of 2000, many manifold learning algorithms have been proposed, such as Locally Linear Embedding (LLE) [26], Laplacian Eigenmap [1], and ISOMAP [27]. All these methods are based on the idea of local invariant, which means that nearby points are likely to have similar embeddings [9]. A commonly used way for capturing the local invariant property of datasets is to construct a p-nearest neighbor graph and its corresponding Laplacian graph [3,4]. Combining these graphs with NMF and CF has resulted in two improved techniques called Graph Regularized Nonnegative Matrix Factorization(GNMF) [4] and Locally Consistent Concept Factorization (LCCF) [3]. A major benefit of such techniques is that their resulting data representations can better capture the geometric structures of the data space.

Many of the aforementioned learning methods are purely unsupervised. To further improve the quality of dimension reduction, an often used strategy is to incorporate some a priori knowledge into the learning process [6,8,10,20,22,28,29,32,33,35]. Such information is usually represented by pairwise constraints which means

* Corresponding author. Tel.: +8613852035062.
E-mail address: louisazhaoxiaolu@sina.com.cn (M. Lu).
[1] No.1 Shizi Street, Suzhou, Jiangsu, China
[2] No.101 Shanghai Rd, Tongshan New District, Xuzhou, Jiangsu, China

http://dx.doi.org/10.1016/j.knosys.2016.04.003
0950-7051/

that the constrained data pairs belong to either the same cluster or different clusters. Correspondingly, these constraints are called must-links constraints and cannot-links constraints, respectively.

Among these semi-supervised learning methods, Semi-Supervised Clustering via Matrix Factorization (SSMF) [28] and Constrained Non-Negative Matrix Factorization(CNMF) algorithm [20] both incorporate a priori knowledge into the NMF framework. SSMF incorporates some labeled data into traditional NMF by adapting the objective function to include penalties for violated constraints. CNMF incorporates the label information as additional hard constraints into NMF. As SSMF and CNMF are direct extension of NMF, the limitation of NMF that it could not effectively perform in RKHS still exits. Moreover, CNMF neglects the intra-class variance, which will weaken the representation ability of the method. Constrained Concept Factorization (CCF) [21] and PCCF [11] both incorporate a priori knowledge into the CF framework. CCF provides a semi-supervised matrix decomposition method which takes the label information as additional constraints. Similar to CNMF, the drawback of CCF is that it neglects the intra-class variance, which could weaken the representation ability of the method. PCCF augments the objective function of CF to ensure that data points with pairwise must-link constraints have the same class label and cannot-link constraints have different class labels. Both CCF and PCCF have difficulty to effectively capture the low dimensional manifold structure, due to their ignoring the proximity information of the dataset. Their performances could degrade to the level of the original CF method when only limited a priori information is available.

To fix the above issues, we propose in this paper a framework called Constrained Neighborhood Preserving Concept Factorization (CNPCF) on top of PCCF. CNPCF uses pairwise constraints and information related to local invariant for better learning performance, where local invariant is based on a generalized meaning of closeness which includes not only spatially close points in the geometric space but also points directly connected by must-links. To encode such information, we use a p-nearest neighbor graph [1,7] (which is mainly for capturing the local geometric structure of the dataset) and a membership graph (which preserves the similarity of those must-link constrained pairs). To take into consideration of all such information, a carefully designed objective function is used for the factorization process. Particularly, a penalty term is added to the objective function for each violation of the pairwise constraints. To preserve local invariant, a term corresponding to each of the p-nearest neighbor graph and the membership graph is added to the objective function. To optimize the objective function, we develop an iterative scheme, and show its convergence. Our generated data representation can well reflect the structure of the whole dataset.

The rest of the paper is organized as follows. In Section 2, a brief description of the related work is reviewed. Our proposed CNPCF is introduced, and an efficient iterative approach to solve the optimization problem of CNPCF is developed in Section 3. In Section 4 some experimental results are presented. Finally, some concluding remarks and suggestions for future work are provided in Section 5.

2. A brief overview of Concept Factorization (CF)

CF [3,30] is an efficient matrix decomposition technique. It has been shown that CF is a very suitable approach for data representation. Let $X = \{\mathbf{x}_i\}$ be a dataset of n data points, where each $\mathbf{x}_i \in \mathrm{R}^m$ is an m dimensional point represented by a column vector. CF aims to find nonnegative matrices $\mathbf{W} \in \mathrm{R}^{n \times k}$ and $\mathbf{V} \in \mathrm{R}^{n \times k}$ such that the product of **X, W** and **V** provides a good approximation to the matrix **X**, *i.e.,*

$$\mathbf{X} \approx \mathbf{XWV}^T \tag{1}$$

Each column of $\mathbf{V}^T$ is the k- dimensional representation of the original inputs in the transformed space. Since k is usually much smaller than m, CF can be regarded as a compressed approximation of the original matrix which leading to a sparse encoding of the data.

The objective function of CF uses the square of Frobenius norm to qualify the approximation. The Frobenius norm of matrix A is defined as

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}\left|a_{ij}\right|^2} = \sqrt{\mathrm{trace}(A^TA)} \quad ,$$

where A^Tdenotes the conjugate transpose of A, and the trace function is used. The objective function of CF has the following form,

$$O = \left\|\mathbf{X} - \mathbf{XWV}^T\right\|_F^2. \tag{2}$$

With the above formulation, the data representation problem is turned into the computation of the two matrices **W** and **V** that minimize this objective function. To minimize the function, [30] proposed an algorithm to iteratively update **W** and **V** as follows.

$$w_{jk}^{t+1} \leftarrow w_{jk}^{t}\frac{(\mathbf{KV})_{jk}}{(\mathbf{KWV}^T\mathbf{V})_{jk}}, \quad v_{jk}^{t+1} \leftarrow v_{jk}^{t}\frac{(\mathbf{KW})_{jk}}{(\mathbf{VW}^T\mathbf{KW})_{jk}}. \tag{3}$$

It has been proved that this objective function is convergent under the above update rules [30]. Note that since the kernel matrix $\mathbf{K} = \mathbf{X}^T\mathbf{X}$ computes the inner product in the original data space, CF can be effectively performed in the transformed data space by choosing a suitable kernel function to construct the kernel matrix. Please refer to [30] for details.

3. Constrained Neighborhood Preserving Concept Factorization (CNPCF)

As mentioned earlier, CF inherit all the strength of NMF and can be effectively performed in some transformed data spaces such RKHS. It could also suffer from a few limitations. For example, it learns effectively only in Euclidean space, which could prevent it from discovering the intrinsic geometrical structure of the input data. Moreover, since CF is an unsupervised method, it does not use any a priori information to guide the learning process. To address these issues, we propose in this section a new approach called CNPCF to incorporate information like some local geometric structure as well as pairwise constraints into the CF framework.

A priori information in this paper is provided as the form of must-link and cannot-link pairwise constraints. Let C_{ML} and C_{CL} be the sets of must-link and cannot-link constraints, and M_N and C_N be the number of the must-link and cannot-link constraints, respectively. If $\mathbf{x}_i$ and $\mathbf{x}_j$ are in the same cluster, then $(\mathbf{x}_i, \mathbf{x}_j) \in C_{ML}$, and if $\mathbf{x}_i$ and $\mathbf{x}_j$ are in different clusters, then $(\mathbf{x}_i, \mathbf{x}_j) \in C_{CL}$.

3.1. The object function

We formulate CNPCF matrix factorization for data representation as follows: Let $X = \{\mathbf{x}_i\}$ be a dataset of n data points, where each $\mathbf{x}_i \in \mathrm{R}^m$ is an m dimensional point represented by a column vector. CNPCF tries to find the new representations $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k] \in \mathrm{R}^{n\times k}$ of the original data which can best preserve the local structure as well as pairwise constraints in the lower-dimensional space.

To discuss this question, firstly we consider encoding the local geometric invariant information, *i.e.,* a p-nearest neighbor graph (see Section 1 for details). As mentioned earlier, this graph contains n vertices with each corresponding to a data point. We define its