Information Sciences 331 (2016) 86–98

Contents lists available at ScienceDirect

Information Sciences

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

Semi-supervised concept factorization for document clustering

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

[a] College of Computer Science and Technology, Soochow University, No.1 Shizi Street, Suzhou, Jiangsu 215006, China
[b] College of Computer Science and Technology, Jiangsu Normal University, No.101 Shanghai Rd, Tongshan New District, Xuzhou, Jiangsu 221116, China

ARTICLE INFO

Article history:
Received 29 August 2014
Revised 10 August 2015
Accepted 18 October 2015
Available online 30 October 2015

Keywords:
Concept factorization
Locally consistent concept factorization
Semi-supervised document clustering

ABSTRACT

Nonnegative Matrix Factorization (NMF) and Concept Factorization (CF) are two popular methods for finding the low-rank approximation of nonnegative matrix. Different from NMF, CF can be applied not only to the matrix containing negative values but also to the kernel space. Based on NMF and CF, many methods, such as Graph regularized Nonnegative Matrix Factorization (GNMF) and Locally Consistent Clustering Factorization (LCCF) can significantly improve the performance of clustering. Unfortunately, these are unsupervised learning methods. In order to enhance the clustering performance with the supervisory information, a Semi-Supervised Concept Factorization (SSCF) is proposed in this paper by incorporating the pairwise constraints into CF as the reward and penalty terms, which can guarantee that the data points belonging to a cluster in the original space are still in the same cluster in the transformed space. By comparing with the state-of-the-arts algorithms (KM, NMF, CF, LCCF, GNMF, PCCF), experimental results on document clustering show that the proposed algorithm has better performance in terms of accuracy and mutual information.

1. Introduction

Document clustering is a text mining technique used to group similar documents into a single cluster, whose approaches can be mainly divided into three types: document partitioning (flat clustering), agglomerative (bottom-up hierarchical) clustering and matrix factorization-based approaches [3,5–7,9,13,18,23,24,29,31,33]. In this paper, we focus on matrix factorization-based approaches, which includes NMF [24], CF [29], GNMF [6], LCCF [7] etc. NMF is widely used in learning the parts of objects like text documents and human faces. Given a nonnegative document data matrix **X**, NMF factorizes **X** into two nonnegative matrices **U** and **V** such that their product **UV** can well approximate the matrix **X**. The column vectors of matrix **U** and **V** are often considered as basis vectors coordinates, respectively. Then, cluster membership of each document can be determined by finding the base topic with which the document has the largest projection value. The nonnegative constraints in NMF lead to a parts-based representation. Previous studies show that there is physiological and psychological evidence for parts-based representation in human brain [21]. However, one issue with NMF is that it could not perform effectively in the reproducing kernel Hilbert (RKHS) data space. To address this issue, Xu and Gong [29] proposed the CF method for data clustering which makes full use of the strength of NMF, but also can applied to any data representations, either in the original space or RKHS. In CF method, each data point is represented by a linear combination of the cluster centers, and vice versa. The document clustering

[*] Corresponding author at: College of Computer Science and Technology, Soochow University, No.1 Shizi Street, Suzhou, Jiangsu 215006, China. Tel.: +86 18896581225.
E-mail address: louisazhaoxiaolu@sina.com.cn, lumei@jsnu.edu.cn (M. Lu).

http://dx.doi.org/10.1016/j.ins.2015.10.038

can be accomplished by computing the two sets of coefficients. Since the linear coefficients have explicit semantic meanings, the label of each document can be easily derived from these coefficients. To further improve the performance, GNMF [6] and LCCF [7] were proposed to incorporate the regularization terms into NMF and CF, respectively. Literatures have shown that human-generated text data are sampled from a submanifold of the ambient Euclidean space [4,22,26]. By combining the graph structure with NMF objective function, GNMF can 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 builds a graph model to extract the underlying concepts which are consistent with the local geometry structure, thus the documents associated with similar concepts can be well clustered.

All above methods are completely unsupervised, and they may yield inferior results due to their ignoring the a priori knowledge [10]. There are also a few methods which can be viewed as semi-supervised [2,8,11,13–15,17,19,20,25,27,28]. A priori knowledge is added to NMF algorithm in [8,17,19,28]. For example, Semi-Supervised Nonnegative Matrix Factorization (SSNMF) [17] is formulated as a joint factorization of the data matrix and the label matrix. Constrained Nonnegative Matrix Factorization (CNMF)[19] takes label information as hard constraints and the data points with the same class label need to be mapped strictly to share the same representation in the transformed space. Although CNMF obeys the strict mapping rule, it neglects the intra-class, which weakens its clustering ability. Nonnegative Matrix Factorization for Semi-Supervised data clustering (NMFSS) [8] constructs symmetric similarity matrix with pairwise constraints, which clusters data by using iterative solver with symmetric tri-factorization of the nonnegative similarity matrix in order to minimize the constraint violations. Although pairwise constraints have been used in semi-supervised learning, to our best knowledge, they have hardly been incorporated into the CF framework. Constrained Concept Factorization (CCF) [20] and Pairwise Constrained Concept Factorization for data representation(PCCF) [11] are such semi-supervised learning algorithms. Similar to CNMF, CCF provides a semi-supervised matrix decomposition method which takes the label information as additional constraints. The drawback of CCF is that it neglects the intra-class variance, which will weaken its representation ability. PCCF incorporates pairwise constraints into the concept factorization framework. PCCF imposes a constant penalty for all pairwise constraints. Each cluster has an intra-cluster variance, which means not all pairs of data points that belong to the same cluster should be forced with the same intensity to be near in the new representation space.

Motivated by the above analyses, we propose in this paper a Semi-Supervised Concept Factorization (SSCF) method, based on the penalized matrix factorization for document clustering. SSCF integrates reward and penalty items provided by pairwise constraints with CF framework. Supervisory information is provided in the form of two sets of pairwise constraints: *must-link*constraints C_{ML} and *cannot-link*constraints C_{CL}. For a pair of documents $\mathbf{x}_i$ and $\mathbf{x}_j$, $(\mathbf{x}_i, \mathbf{x}_j) \in C_{ML}$ and $(\mathbf{x}_i, \mathbf{x}_j) \in C_{CL}$ imply that they belong to the same and different clusters, respectively. Our SSCF provides a dynamic penalty mechanism which better accounts for the intra-cluster variance, i.e., allowing dissimilar data points with the same cluster label to be mapped farther than similar ones. In this way, we can learn more reasonable clustering structures in the representation space. To infer the document clusters, an iterative algorithm is proposed to perform the term-document matrix factorization, and the convergence of SSCF is proved in the main theorems. To validate our proposed method, extensive experiments are conducted on publicly available data sets, our experiments show the superior performance of SSCF for document clustering.

The rest of this paper is organized as follows. Section 2 reviews the related works. SSCF for document clustering is proposed in Section 3. Experimental results are illustrated in Section 4, followed by the conclusions and future work in Section 5.

2. Related works

2.1. Document representation

The entire document collection is typically represented in vector space [30] as a term-document matrix $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in \mathrm{R}^{m\times n}$, where columns and rows denote the documents and the terms appearing in the documents, respectively. Let $\mathbf{F} = \{f_1, f_2, \ldots, f_m\}$ be the complete vocabulary set of the document corpus. The term-frequency vector $\mathbf{x}_i$ of document d_i is defined as

$$\begin{aligned} \mathbf{x}_i &= \{x_{1i}, x_{2i}, \ldots, x_{mi}\}^T, \\ x_{ji} &= t_{ji} \cdot \log\left(\frac{n}{idf_j}\right), \end{aligned} \tag{1}$$

where t_{ji} denotes the frequency of term $f_j \in \mathbf{F}$ in document d_i, idf_j denotes the number of documents containing term f_j, n denotes the total number of documents in the corpus, and m is the cardinal number of $\mathbf{F}$. In addition, $\mathbf{x}_i$ is usually normalized to unit Euclidean length.

With the above notations, a text document can be represented as a point in a high-dimensional linear space, with each dimension corresponding to a term. The document clustering is performed in the low-dimensional manifold space which is constructed by some connected components of matrix factorization of the original data matrix.

2.2. CF

Recently, CF has received many attention and has been demonstrated to be very useful for applications like pattern recognition and text mining. As Xu and Gong [29] pointed out, data clustering problem can be modeled by using one of the following two