Neurocomputing 87 (2012) 120–131

Contents lists available at SciVerse ScienceDirect

Neurocomputing

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

Clustering analysis using manifold kernel concept factorization

Ping Li*, Chun Chen, Jiajun Bu

Department of Computer Science, Zhejiang University, Hangzhou 310027, China

ARTICLE INFO

Article history:
Received 26 September 2011
Received in revised form
31 January 2012
Accepted 11 February 2012
Communicated by M. Wang
Available online 24 February 2012

Keywords:
Manifold kernel learning
Concept factorization
Graph Laplacian
Document clustering
Image clustering

ABSTRACT

Various exponential-growing documents and images have become omnipresent in past decades, and it is of vital importance to group them into clusters upon desired. Matrix factorization is exhibited to help yield encouraging clustering results in previous works, whereas the data manifold structure, which holds plentiful spatial model information, is not fully respected by most existing techniques. And kernel learning is advantageous for unfolding nonlinear structure. Therefore, in this paper we propose a novel clustering approach called *Manifold Kernel Concept Factorization* (MKCF) that incorporates the manifold kernel learning in concept factorization, which encodes the local geometrical structure in the kernel space. This method efficiently preserves the data semantic structure using graph Laplacian, and the nonlinear manifold learning in the warped RKHS potentially reflects the underlying local geometry of the data. Thus, the concepts consistent with the intrinsic manifold structure are well extracted, and this greatly benefits aggregating documents and images within the same concept into the same cluster. Extensive empirical studies demonstrate that MKCF owns the superiority of achieving the more satisfactory clustering performance as well as deriving the better-represented lower data space.

© 2012 Elsevier B.V. All rights reserved.

1. Introduction

Tremendous growths in the amount of text documents and images have been receiving more and more attentions during the last decade, especially in the data mining and information retrieval community. Clustering is regarded as a fundamental tool that has a broad range of applications in dealing with huge volumes of text documents and various images [1–6]. For a given data set, the task of clustering is to find good clusters, which enables easy organization and navigation of the data corpus. Roughly speaking, clustering methods can be categorized into two mainstreams: agglomerative clustering and partitioning approach. The former belongs to a bottom-up hierarchical type and the latter decomposes the given corpus into disjoint clusters. Both of them have been well studied and investigated in previous literatures [7–10].

For clustering, matrix factorization based techniques have attracted considerable interests from many researchers in this field. With regard to these methods, each text document or image in the corpus is often treated as a data point in the high dimensional linear space. Clustering analysis aims to look for similar data points and ensure them within the same cluster in maximum degree. Intuitively, similar samples are more likely to be grouped together than different ones, and this could be attributed to the fact that characteristics shared by similar ones in original data spaces are inherited by new representations in lower dimensional spaces, which makes the clustering more easily. There are particularly two popular matrix factorization methods widely applied to clustering analysis, i.e., Non-negative Matrix Factorization (NMF) [11,12] and Concept Factorization (CF) [3]. Generally, regardless of NMF or CF, they only consider using the global Euclidean geometry to find new basis vectors, according to how the new data representation is generated.

Previous studies show that the learning performance can be enhanced by taking advantage of the manifold geometry and the locally invariant idea [13–15], it is very natural to involve them in matrix factorization based techniques for clustering. It has been shown that CF can be kernerlized in [3] and kernel learning is helpful to discover nonlinear data structure, thus considering manifold kernel learning is an ideal choice. Fundamentally motivated by this, we present a novel clustering approach called *Manifold Kernel Concept Factorization* (MKCF). The goal of this algorithm is to extract the underlying concepts consistent with the manifold geometrical structure in the data space. The central idea is striving to incorporate the manifold kernel learning into concept factorization, which enables capturing the local latent semantic structure of the data [16,17]. This approach attempts to preserve the geometrical structure using graph Laplacian, and the nonlinear manifold learning in the warped RKHS essentially reflects the underlying local geometry in the data space. Thus, the concepts consistent with the intrinsic manifold structure are well extracted, and this significantly facilitates clustering. Our

* Corresponding author. Tel.: +86 13735843932.
E-mail address: patriclouis.lee@gmail.com (P. Li).

0925-2312/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2012.02.013

empirical evaluations on two text document corpora and two facial data sets suggest the proposed approach outperforms other clustering methods in terms of accuracy and normalized mutual information.

The rest of this paper is organized as follows. We provide a brief review of some related methods in Section 2. And Section 3 is primarily devoted to presenting our *Manifold Kernel Concept Factorization* approach as well as the detailed derivations. Experimental results are reported in Section 4 with considerable analysis and rigorous discussions. Lastly, Section 5 offers some concluding remarks and exploring directions in the future work.

2. Related works

In this section, we primarily review some related approaches to our research work, including some matrix factorization based techniques and manifold learning methods.

Non-negative Matrix Factorization (NMF) is a widely accepted matrix factorization method employed by many clustering and classification applications. It has been shown that NMF is able to obtain a parts-based representation since it only allows additive, not subtractive, combinations, which is in accordance with the psychological and physiological evidence for parts-based representation in human brain [11]. The non-negative constraints are enforced on the cluster centers results. As a result, NMF can only be performed in the original space but fails to work in the kernelized space, e.g., Reproducing Kernel Hilbert Space (RKHS). Xu et al. [12] propose one clustering method based on non-negative matrix factorization, where each axis captures the base topic of a particular document cluster and each document is represented as an additive combination of the base topics. They also put forward the other clustering method based on concept factorization in [3], which differs from NMF in that it can be applied to data containing negative values and be implemented in the kernel space. CF mainly strives to address the limitations and meanwhile inherits all the strengths of NMF, such as better semantic interpretation and easily derived clustering results. With this method, each concept or component is modeled as a linear combination of the data points while each data point consists of a linear combination of the concepts. It acquires the non-negative linear coefficients through minimizing the reconstruction error of the data points and derives the cluster label of each data point easily from these gained coefficients. Regarding NMF and CF, there exist many invariants and extensions [18,7,8]. In general, CF is more advantageous than NMF, because of its merits that it can be applied to any data points taking both positive and negative values. Recently, graph Laplacian which plays a role in regularizing the objective function is employed in [18]. Differently, we consider the manifold learning in the warped RKHS so that some nonlinear data structure could be well captured. Besides, there are two important parameters to be adjusted in [18] but only one for our method. In [8], the idea of locality preserving projection is merged into non-negative matrix factorization, which uses the KL-divergence to evaluate the similarity on the hidden topics. Ding et al. [19] give an orthogonal non-negative matrix factorization based clustering method since the orthogonal constraints leads to the more rigorous clustering interpretation.

Many researches have paid their attention on the case that the data is drawn from sampling a probability distribution, which supports on or near to a submanifold of the ambient Euclidean space [13,14,20]. Here, a submanifold with d dimensions is referred to a subset of a m-dimensional Euclidean space. Actually, it has somewhat difficulty filling the human generated text documents uniformly in high dimensional Euclidean spaces, which inspires us to consider the intrinsic manifold structure while obtaining the new lower data representations. In order to explore the underlying manifold geometry, there are many existing methods regarding manifold learning, such as Laplacian Eigenmap [13], Locally Preserving Projection [21] and Locally Linear Embedding [20]. One common feature shared by these algorithms is that the nearby points tend to have the similar labels, which is called *local invariance* [22]. Up to now, manifold learning methods have gained wide spread acceptance and achieved huge success in various kinds of applications [16,7,23,24,9]. In [16], active learning is performed in the manifold kernel space, and the most representative and informative data points are selected by minimizing the expected error. In [7], the graph structure is considered as a regularizer in non-negative matrix factorization, which makes the objective function and optimization process more complex. Different from the above methods, Yang et al. [24] employ local regression models to capture the manifold structure and impose a global regression regularized term to learn a model for out-of-sample data extrapolation. Additionally for image clustering, Yang et al. [4] propose a novel image clustering method, which learns a new Laplacian matrix by exploiting both manifold structure and local discriminant information.

3. Manifold Kernel Concept Factorization

This section is mainly devoted to our proposed method called *Manifold Kernel Concept Factorization* (MKCF). First, we simply describe the fundamental methods NMF and CF. Then the derivation of manifold adaptive kernel is given with the objective function. Third, we present the multiplicative update rules that consider both the positive and negative values for our approach. In addition, the complete MKCF algorithm and rigorous analysis on our method are provided in the end.

3.1. Review of NMF and CF

Given a non-negative matrix $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_n]\in\mathbb{R}^{m\times n}$, and each data vector is placed in one column. Suppose that a data corpus consists of p clusters with each of them corresponding to one coherent concepts, NMF aims to find two non-negative coefficient matrices $\mathbf{U}\in\mathbb{R}^{m\times p}$ and $\mathbf{V}\in\mathbb{R}^{n\times p}$. Its cost function is defined by minimizing the least squares error, shown as

$$J_{NMF}=\tfrac{1}{2}\|\mathbf{X}-\mathbf{U}\mathbf{V}^T\|^2, \tag{1}$$

where the matrix *Frobenius norm* $\|\cdot\|$ denotes the squared sum of all elements within the matrix. By this means, it easily derives an approximate factorization $\mathbf{X}\approx\mathbf{U}\mathbf{V}^T$, which is virtually a compressed approximation of the original matrix since p is usually much smaller than n and m, leading to a sparse encoding of the data. This cost function is non-convex in both variables $\mathbf{U}$ and $\mathbf{V}$ together but convex in an individual variable. Consequently, searching the global minima of this function is an impractical problem. The multiplicative update algorithm is expected to be an ideal solution [25], which states that J_{NMF} is non-increasing and convergent under these update rules. Each column vector of $\mathbf{X}$ is an approximately linear and additive combination of the corresponding columns in $\mathbf{U}$ with the components in $\mathbf{V}$ as the weights.

Concept Factorization (CF) is developed on the basis of NMF [3]. Each concept as a linear combination of the entire data points allows the formulated computation in the kernelized space. Let $\mathbf{r}_k$ be the center of the concept k, where $k=1,\ldots,p$, then we have $\mathbf{r}_k=\sum_{j=1}^{n}\mathbf{x}_j w_{jk}$. On the other hand, each data point can be approximated by a linear combination of all concepts, i.e., $\mathbf{x}_j=\sum_{k=1}^{p}\mathbf{r}_k v_{jk}$, where the non-negative weight w_{jk} represents the association degree of the data point $\mathbf{x}_j$ and the concept k. The

Download English Version:

https://daneshyari.com/en/article/407523

Download Persian Version:

https://daneshyari.com/article/407523

Daneshyari.com