Int J Software Informatics, Volume 7, Issue 3 (2013), pp.407–418 Tel: +86-10-62661040
International Journal of Software and Informatics, ISSN 1673-7288 http://www.ijsi.org

Email: ijsi@iscas.ac.cn, ijsi2007@gmail.com

Multi-Manifold Concept Factorization for Data Clustering

Ping Li, Jiajun Bu, and Deng Cai

(College of Computer Science, Zhejiang University, Hangzhou, China)

Abstract Clustering plays an important role in many fields, such as pattern recognition and data mining. Its goal is to group the collected data points into their respective clusters. To this end, a number of matrix factorization based methods have been developed to obtain satisfying clustering results by extracting the latent concepts in the data, *e.g.*, concept factorization (CF) and locally consistent concept factorization (LCCF). LCCF takes into account the local manifold structure of the data, but it is nontrivial to estimate the intrinsic manifold, reflecting the true data structure. To address this issue, we in this paper present a novel method called *Multi-Manifold Concept Factorization* (MMCF) to derive more promising clustering performance. Specifically, we assume the intrinsic manifold lies in a convex hull of some predefined candidate manifolds. The basic idea is to learn a convex combination of a group of candidate manifolds, which is utilized to approximate the intrinsic manifold of the data. In this way, the low-dimensional data representation derived from MMCF is able to better preserve the locally geometrical structure of the data. To optimize the objective function, we develop an alternating algorithm and learn the manifold coefficients using the entropic mirror descent algorithm. The effectiveness of the proposed approach has been demonstrated through a set of evaluations on several real-world data sets.

Key words: concept factorization; multi-manifold learning; locally geometrical structure; data clustering

Li P, Bu JJ, Cai D. Multi-Manifold concept factorization for data clustering. *Int J Software Informatics*, **Vol.7, No.3 (2013): 407–418. http://www.ijsi.org/1673-7288/7/i167.htm**

1 Introduction

Clustering is a fundamental problem in pattern recognition and data mining[1]. It groups the samples with similar features into the same cluster and the dissimilar samples into different clusters[2,3,4,5]. Generally, the collected databases have high dimensions and are even contaminated by some noises, which not only makes the clustering computationally expensive but also deteriorates the performance. Typical methods to address this issue include dimensionality reduction, feature selection and matrix factorization, *e.g.*, principal component analysis (PCA)[6], locality preserving projection (LPP)[7], singular value decomposition (SVD)[8], nonnegative matrix factorization (NMF)[9] and concept factorization (CF)[10]. In this work, we focus on

This work is sponsored by National Basic Research Program of China (2011CB302206) and the Fundamental Research Funds for the Central Universities (2013FZA5012).
Corresponding author: Ping Li, Email: lpcs@zju.edu.cn
Received 2012-11-30; Revised 2013-06-18; Accepted 2013-08-31

matrix factorization based method for data clustering, specifically extracting the underlying concepts consistent with the local manifold structure of the data.

Matrix factorization has gained considerable popularity in data representation for a very long time. Basically, it decomposes a given matrix into two or three sub-matrices, whose product is regarded as an approximation of the original data matrix. Taking NMF for example, for a data matrix $\boldsymbol{X}$, NMF factorizes it into one basis matrix $\boldsymbol{U}$ and one coefficient matrix $\boldsymbol{V}$, thus approximating $\boldsymbol{X}$ using $\boldsymbol{UV}$. The clustering results can be easily derived from the new low-dimensional data representation $\boldsymbol{V}$. But the nonnegative constraints are imposed on both $\boldsymbol{U}$ and $\boldsymbol{V}$, which makes NMF be unable to handle the input matrix containing negative entries. To overcome this difficulty, CF models each concept as a linear combination of the data points, and each data point is treated as a linear combination of the concepts. Assume each data point $\boldsymbol{x}_i$ is a m-dimensional column vector and there are c latent concepts, we have $\boldsymbol{x}_j = \sum_{k=1}^{c} \boldsymbol{u}_k \boldsymbol{v}_{jk}$. CF has a major merit that it can be performed in either original space or the kernel space, *e.g.*, reproducing kernel Hilbert space (RKHS). Both of NMF and CF only consider the global data structure and do not fully respect the locally geometrical structure of the data. To take into account the intrinsic manifold structure, graph regularized nonnegative matrix factorization (GNMF)[11] and locally consistent concept factorization (LCCF)[12] were proposed by adding the manifold regularizer on the new data representation derived from NMF and CF, respectively. However, in many real-world applications, data points might be sampled from different data distributions, and it is nontrivial to estimate the intrinsic manifold in a systematical way.

To address this problem, inspired by the work in Ref. [13], we propose a novel method named *Multi-Manifold Concept Factorization* (MMCF) to equip the low-dimensional data representation with more locally consistent structure. Assuming that the intrinsic manifold lives in a convex hull of a set of pre-defined candidate manifolds, we aim to approximate the intrinsic data manifold through a linear combination of these candidate manifolds, *i.e.*, multi-manifold ensemble. In this way, the locally geometrical structure of the data can be better preserved since more diverse structural information is available from different manifolds. Therefore, the extracted concepts are characteristic of more local consistency. Besides, we develop an alternating algorithm to optimize the objective function and adopt the entropic mirror descent algorithm[14] to learn the coefficients for the multi-manifold ensemble. Experiments were conducted on several real-world databases to show the superiority of the proposed method.

The rest of the paper is organized as follows. Section 2 gives a brief review of the related works. In Section 3, we introduce the proposed multi-manifold concept factorization method. Then, we report the experimental results with analysis in Section 4. Finally, the concluding remarks are provided in Section 5.

2 Related Works

In this section, we briefly review some works closely related to our work. Matrix factorization has established itself as a very useful tool for data clustering in the past decades. Many researchers have devoted themselves to developing a series of matrix factorization based techniques for clustering analysis[3,11,15]. Among them, the most

widely used one is nonnegative matrix factorization (NMF)[16].

NMF aims to decompose a nonnegative matrix into one basis matrix and one coefficient matrix, on both of which are imposed the nonnegative constraints, leading to a parts-based representation because they allow only additive, not subtractive, combinations[9]. The optimal coefficient matrix can be obtained by minimizing the reconstruction error of the data points. The multiplicative update rules[16] and the projected descent rules[17] are often used to optimize the objective function. One advantage of NMF over SVD is that the factorization result has better semantic interpretation. But it has some limitations, *e.g.*, NMF requires the input data is nonnegative which is unnecessarily satisfied in many other fields rather than documents, and it cannot be kernelized directly due to the nonnegative constraints. To handle these problems, Xu *et al.*[11] proposed concept factorization (CF) to model each cluster as a linear combination of the data points, and each data point as a linear combination of the cluster centers. CF not only inherits the advantages of NMF but also has more merits, *e.g.*, it can deal with data points containing negative elements and be performed in the kernel space.

Both of NMF and CF only consider the global Euclidean structure of the data points but neglect the locally geometrical structure, which is more important in many scenarios. Recent works have shown the great success of manifold learning in various applications, *e.g.*, face recognition[18], data representation[19,20,21,22]. This learning paradigm assumes the data points are sampled from a submanifold from the ambient Euclidean space and nearby data points are more likely to be closer than those with large distances[23,24]. Inspired by this, Cai *et al.*[11,12] imposed the manifold regularizer on NMF and CF, *i.e.*, GNMF and LCCF, respectively. Thus, the local consistency of the data can be well guaranteed in the low-dimensional data space. Since CF can be kernerlized, Li *et al.*[15] employed the manifold kernel in concept factorization, which learns the new data representation in the warped RKHS. However, these methods fail to consider the case when data points reside on the overlapped manifolds. Actually, this makes the task of estimating the intrinsic manifold very challenging. Geng *et al.*[13] utilized a convex combination of some pre-defined candidate manifolds to approximate the intrinsic manifold of the data collection. Motivated by this, we propose to incorporate the multi-manifold ensemble learning into concept factorization to better preserve the local structure of the data.

Besides, there are many other extensions of NMF and CF. For example, to further ensure the parts-based representation of NMF, Hoyer *et al.*[25] explicitly imposed the sparseness constraints on the coefficient matrix. Zhang *et al.*[26] attempted to deal with the corrupted data matrix using a robust nonnegative matrix factorization method. To enhance the sparsity of the new data representation, Liu *et al.*[27] enforced a locality constraint onto CF by requiring the concepts to be as close to the original data points as possible. Thus, each data can be represented by a linear combination of only a few basis concepts. In addition, Hua *et al.*[28] took advantage of the available label information to equip the new data representation with more discriminating power in concept factorization.

3 Multi-Manifold Concept Factorization

This section is devoted to our proposed multi-manifold concept factorization approach. First, we give a concise description about concept factorization followed with the introduction of multi-manifold learning. Then, the objective function of MMCF is provided with its optimization framework. Finally, we present the complete algorithm of MMCF.

3.1 A review of CF

Concept factorization (CF)[10] is an extension of nonnegative matrix factorization. Given a data collection of n samples, each of which has m features, our goal is to group them into c clusters. Mathematically, we denote the input data matrix by $\boldsymbol{X} \in \mathcal{R}^{m\times n}$ and each data point by $\boldsymbol{x}_i \in \mathcal{R}^m$. The basic idea of CF is to represent every sample as the linear combination of the concepts (*i.e.*, cluster centers), each of which is the linear combination of samples. Essentially, CF seeks an approximation of three matrices, *i.e.*,

$$\boldsymbol{X} \approx \boldsymbol{X}\boldsymbol{W}\boldsymbol{V}^T, \tag{1}$$

where the basis matrix in NMF is reconstructed by the product of the original data matrix $\boldsymbol{X}$ and its association matrix $\boldsymbol{W} \in \mathcal{R}^{n\times k}$, and the new data representation is shown by the low-dimensional coefficient matrix $\boldsymbol{V}^{n\times k}$. This assembles the convex-NMF[29], which interprets the column of basis matrix in NMF as weighted sums of certain data points. The two nonnegative variables $\boldsymbol{W}$ and $\boldsymbol{V}$ can be readily obtained by optimizing the cost function measured by Euclidean distance, *i.e.*,

$$\min_{\boldsymbol{W},\boldsymbol{V}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}\boldsymbol{V}^T\|_F, \quad s.t.,\ \boldsymbol{W} \geq 0, \boldsymbol{V} \geq 0, \tag{2}$$

where $\|\cdot\|_F$ denotes the *Frobenius norm*.

Defining the kernel matrix as $\boldsymbol{K} = \boldsymbol{X}^T\boldsymbol{X}$, the update rules of CF can be shown as[10]

$$w_{jk} \leftarrow w_{jk}\frac{(\boldsymbol{K}\boldsymbol{V})_{jk}}{(\boldsymbol{K}\boldsymbol{W}\boldsymbol{V}^T\boldsymbol{V})_{jk}}, \qquad v_{jk} \leftarrow v_{jk}\frac{(\boldsymbol{K}\boldsymbol{W})_{jk}}{(\boldsymbol{V}\boldsymbol{W}^T\boldsymbol{K}\boldsymbol{W})_{jk}}. \tag{3}$$

As can be seen from the above rules, CF can easily substitute the data-dependent adaptive kernel to improve the performance in resolving different problems.

3.2 Multi-Manifold learning

Traditional manifold learning considers the case that data points reside on a submanifold of the ambient Euclidean space. However, in many real-world applications, some data points might be sampled from different distributions. Therefore, inspired by Geng's work[13], we propose to employ *multi-manifold learning* (MML) to approximate the intrinsic manifold using a subset of candidate manifolds, so as to better reflect the locally geometrical structure by graph Laplacian[19].

To consider different data sources, we assume the intrinsic manifold of the collected data points lives in a convex hull $\boldsymbol{C}$ of a group of candidate manifolds, each of which indicates one kind of data distribution. MML essentially learns an

approximated intrinsic manifold by integrating the diverse manifold information of candidates in a linear manner. We regard this linear combination of candidate manifolds as *manifold ensemble*. Let $\boldsymbol{L}$ be the intrinsic manifold, $\boldsymbol{L}_i$ be the i-th candidate manifold, and there are q candidates corresponding to different data distributions. Here, we use p-nearest neighbor graph to encode the data structure and p is kept small to ensure the local preserving property. In general, the typical schemes to construct the weight matrix $\boldsymbol{S}$ are binary weighting, gaussian weighting and cosine similarity weighting[11]. Each candidate manifold is a graph Laplacian obtained by $\boldsymbol{L}_i = \boldsymbol{D}_i - \boldsymbol{S}_i$, where $\boldsymbol{D}_i$ is a diagonal matrix with its entries by the column or row sum of the weight matrix $\boldsymbol{W}_i$.

Multi-manifold learning represents the manifold ensemble $\boldsymbol{L}$ by a linear combination of the pre-defined candidate manifolds. Each candidate $\boldsymbol{L}_i$ is associated with a coefficient μ_i, which is shown by

$$\boldsymbol{L} = \sum_{i=1}^{q} \mu_i \boldsymbol{L}_i, \quad s.t. \ \sum_{i=1}^{q} \mu_i = 1, \mu_i \geq 0. \tag{4}$$

Since $\boldsymbol{L}$ is in a convex hull of q candidate graph Laplacians, it is also a graph Laplacian. The coefficients are imposed by the simplex constraints.

3.3 The objective function

To preserve the locally geometrical structure of the data space, we impose the multi-manifold regularizer $\mathrm{Tr}(\boldsymbol{V}^T \sum_{i=1}^{q} \mu_i \boldsymbol{L}_i \boldsymbol{V})$ onto concept factorization. Here, $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix. Moreover, we introduce the l_2 *norm* of the variable $\boldsymbol{\mu}$ (*i.e.*, $\|\boldsymbol{\mu}\|^2$) to avoid overfitting on only one manifold as in Ref. [13]. Therefore, the objective function of MMCF is formulated as

$$\begin{aligned} &\min_{\boldsymbol{W},\boldsymbol{V},\boldsymbol{\mu}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{W}\boldsymbol{V}^T\|_F + \alpha \mathrm{Tr}(\boldsymbol{V}^T \sum_{i=1}^{q} \mu_i \boldsymbol{L}_i \boldsymbol{V}) + \beta \|\boldsymbol{\mu}\|^2, \\ &s.t., \boldsymbol{W} \geq 0, \boldsymbol{V} \geq 0, \sum_{i=1}^{q} \mu_i = 1, \mu_i \geq 0, \end{aligned} \tag{5}$$

where the parameter α is used to tradeoff the contribution of the multi-manifold regularizer and β controls the regularization term $\|\boldsymbol{\mu}\|^2$.

The objective function incorporates the multi-manifold learning into concept factorization, which enables better extracting the concept with the local consistency, thus yielding more satisfactory clustering results. In the following, we provide the optimization framework to solve this problem.

3.4 The optimization framework

The objective function in Eq. (5) is not convex jointly, but convex for each variable individually. So, we present an alternating algorithm to solve it, *i.e.*, optimize one variable while fixing others. Thus, the locally optimal solutions can be achieved.

Through some algebraic transformations, we can rewrite the objective function of MMCF as

$$\mathcal{O}_{MMCF} = \mathrm{Tr}(\boldsymbol{K}) - 2\mathrm{Tr}(\boldsymbol{V}\boldsymbol{W}^T\boldsymbol{K}) + \mathrm{Tr}(\boldsymbol{V}\boldsymbol{W}^T\boldsymbol{K}\boldsymbol{W}\boldsymbol{V}^T) + \alpha \mathrm{Tr}(\boldsymbol{V}^T\boldsymbol{L}\boldsymbol{V}) + \beta(\boldsymbol{\mu}^T\boldsymbol{\mu}), \tag{6}$$

where $\boldsymbol{K} = \boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{L} = \sum_{i=1}^{q} \mu_i \boldsymbol{L}_i$. Let $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$ be the Lagrange multiplier for the two nonnegative constraints for $\boldsymbol{W}$ and $\boldsymbol{V}$, respectively. We define $\boldsymbol{\Theta} = [\theta_{jk}]$ and $\boldsymbol{\Psi} = [\psi_{jk}]$. Then, the Lagrangian function of MMCF can be expressed by

$$\mathcal{L} = \mathcal{O}_{MMCF} + \mathrm{Tr}(\boldsymbol{\Theta}\boldsymbol{W}^T) + \mathrm{Tr}(\boldsymbol{\Psi}\boldsymbol{V}^T). \tag{7}$$

Taking the first-partial derivatives of $\mathcal{L}$ with respect to $\boldsymbol{W}$ and $\boldsymbol{V}$, respectively, we can easily arrive at

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{W}} = -2\boldsymbol{K}\boldsymbol{V} + 2\boldsymbol{K}\boldsymbol{W}\boldsymbol{V}^T\boldsymbol{V} + \boldsymbol{\Theta}, \tag{8}$$

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{V}} = -2\boldsymbol{K}\boldsymbol{W} + 2\boldsymbol{V}\boldsymbol{W}^T\boldsymbol{K}\boldsymbol{W} + 2\alpha\boldsymbol{L}\boldsymbol{V} + \boldsymbol{\Psi} \tag{9}$$

Using the Karush-Kuhn-Tucker (KKT) conditions[30], *i.e.*, $\theta_{jk}w_{jk} = 0$ and $\psi_{jk}v_{jk} = 0$. Then, we can get the update rules

$$w_{jk} \leftarrow w_{jk}\frac{(\boldsymbol{K}\boldsymbol{V})_{jk}}{(\boldsymbol{K}\boldsymbol{W}\boldsymbol{V}^T\boldsymbol{V})_{jk}}, \qquad v_{jk} \leftarrow v_{jk}\frac{(\boldsymbol{K}\boldsymbol{W} + \alpha\boldsymbol{S}\boldsymbol{V})_{jk}}{(\boldsymbol{V}\boldsymbol{W}^T\boldsymbol{K}\boldsymbol{W} + \beta\boldsymbol{D}\boldsymbol{V})_{jk}}, \tag{10}$$

where $\boldsymbol{D} = \sum_{i=1}^{q} \mu_i \boldsymbol{D}_i$ and $\boldsymbol{S} = \sum_{i=1}^{q} \mu_i \boldsymbol{S}_i$.

Now, we can update $\boldsymbol{W}$ or $\boldsymbol{V}$ while fixing the rest. How to optimize the multi-manifold coefficient seems to be a tough problem. It can be solved using quadratic programming (QP) or the entropic mirror descent algorithm (EMDA)[14] to optimize $\boldsymbol{\mu}$ when the other two variables are held. Here, we adopt the latter in the sense that EMDA is provably with a global efficiency estimate and is mildly dependent on the problem size. But QP might bear the computational burden when the problem size scales up. While fixing $\boldsymbol{W}$ and $\boldsymbol{V}$, the objective function reduces to

$$\min_{\boldsymbol{\mu}} \sum_{i=1}^{q} \mu_i h_i + \beta\|\boldsymbol{\mu}\|^2, \quad s.t., \ \sum_{i=1}^{q} = 1, \mu_i \geq 0, \tag{11}$$

where $h_i = \mathrm{Tr}(\boldsymbol{V}^T\boldsymbol{L}_i\boldsymbol{V})$. If β equals 0, then $\boldsymbol{\mu}$ will take trivial solutions 0 and 1. If β approaches infinity, the candidate manifolds will be treated equally. Hence, we should assign a proper value to β to guarantee the effectiveness of multi-manifold learning.

As a kind of nonlinear projected-subgradient methods, EMDA can be achieved using a general distance-like function rather than Euclidean squared distance. Since the constraints imposed on $\boldsymbol{\mu}$ is a unit simplex $\triangle = \{\boldsymbol{\mu} \in \mathbb{R}^q : \sum_{i=1}^{q} \mu_i = 1, \boldsymbol{\mu} \succeq 0\}$, it makes sense to choose EMDA thanks to its natural advantages over convex problems[14]. EMDA requires the objective function f to be a convex Lipschitz continuous function with Lipschitz constant L_f w.r.t. a fixed norm. For MMCF, this Lipschitz constant is computed by $\|\nabla f(\boldsymbol{\mu})\|_1 \leq 2\beta + \|\boldsymbol{h}\|_1 = L_f$, where $\boldsymbol{h} = \{h_1, \ldots, h_q\}$. The pseudocode of EMDA is shown in Algorithm 1.

3.5 The MMCF approach

In the optimization framework, we have shown an alternating algorithm which updates the three variables $\boldsymbol{W}$, $\boldsymbol{V}$ and $\boldsymbol{\mu}$ iteratively. To have an overview of our multi-manifold concept factorization approach, we provide the complete procedures in Algorithm 2.

Algorithm 1 Entropic Mirror Descent Algorithm

Input: Lipschitz constant L_f, β, $\boldsymbol{h}$.

Output: Multi-manifold ensemble coefficient $\boldsymbol{\mu}$.

Procedure:

Initialize μ_i with identical weights $1/q$.
for $i = 1$ to q **do**
  repeat
    $t_m = \sqrt{\frac{2 \ln q}{m L_f^2}}$, where m is the m-th iteration.
    $\mu_i^{m+1} \leftarrow \frac{\mu_i^m \exp[-t_m f'(\mu_i^m)]}{\sum_{i=1}^q \mu_i^m \exp[-t_m f'(\mu_i^m)]}$.
  until convergence
end for

Algorithm 2 Multi-Manifold Concept Factorization

Input: Data collection $\mathbf{X} \in \mathbb{R}^{m \times n}$, parameters α and β, the number of concepts c.

Output: New data representation $\boldsymbol{V}$.

Initialize $\boldsymbol{W}$ and $\boldsymbol{V}$ with random values between 0 and 1.
Construct the a set of candidate manifolds $\{\boldsymbol{L}\}_{i=1}^q$.
repeat
  Conduct multi-manifold learning using EMDA in Algorithm 1 to obtain $\boldsymbol{\mu}$.
  Update $\boldsymbol{W}$: $w_{jk} \leftarrow w_{jk} \frac{(\boldsymbol{KV})_{jk}}{(\boldsymbol{KWV}^T\boldsymbol{V})_{jk}}$.
  Update $\boldsymbol{V}$: $v_{jk} \leftarrow v_{jk} \frac{(\boldsymbol{KW}+\alpha\boldsymbol{SV})_{jk}}{(\boldsymbol{VW}^T\boldsymbol{KW}+\beta\boldsymbol{DV})_{jk}}$.
until convergence

The convergence of our method can be easily proved using an auxiliary function as in Refs. [12,16]. As mentioned earlier, MMCF inherits the advantage of CF and LCCF, it can also handle the input data matrix with negative entries by using the multiplicative updates for nonnegative quadratic programming[31]. Readers are referred to Ref. [12] for more details. If the MMCF algorithm converges in t times and EMDA converges in z times, then the time complexity of MMCF is $\boldsymbol{O}(tn^2k + n^2m + n^2pq + qz)$. Compared to LCCF (*i.e.*, $\boldsymbol{O}(tn^2k + n^2m + n^2p)$), it requires to construct q ($q \ll n$) candidate manifolds and learn $\boldsymbol{\mu}$ in linear time. In consequence, MMCF, LCCF and CF (*i.e.*, $\boldsymbol{O}(tn^2k + n^2m)$) have the same time complexity using the big $\boldsymbol{O}$ notation.

4 Experiments

To investigate the clustering performance of the proposed MMCF method, we have conducted extensive experiments on several databases. First, we give brief descriptions about the data collections and the evaluation metrics. Second, the performance comparison and corresponding results are shown. Third, some parameter selections are illustrated by curves.

4.1 Data sets

In total, three image databases and one document data were used in our test. ORL[1] is composed of 400 face images, which belong to 10 different persons. Every

[1] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

person has 40 distinct samples. All images were taken against a dark homogeneous background with the subjects in an upright, frontal position. After preprocessing, each image is represented by a 1,024-dimensional vector. UMIST[2][32] is a face image database that contains 575 multi-view face images of 20 people, referring to a range of poses from profile to frontal views. Each image is rescaled to 28×23 pixels. The COIL20[3][33] image library contains 1,440 images, which refer to 20 different objects viewed from varying angles. Each object has 72 gray scale images with the size of 32×32. WebKB[4][34] is a document database involving the home pages of 4 universities. These web pages were collected from computer science departments and were classified into seven topics: student, faculty, staff, department, course, project, and other. We used a subset of 814 samples in this test. The statistics of the data sets are given in Table 1.

Table 1 Statistics of the data sets

Database	Samples	Features	Classes
ORL	400	1,024	40
UMIST	575	644	20
COIL20	1,440	1,024	20
WebKB	814	4,029	7

4.2 Evaluation metrics

We employ *accuracy* (AC) and the *normalized mutual information* (NMI)[4,12] to evaluate the clustering performance of the compared algorithms.

AC is the percentage of correctly estimated labels in the total data collection. Given a data point $\boldsymbol{x}_i$, let a_i and g_i be the estimated and true label respectively, then we have

$$\mathrm{AC} = \frac{\sum_{i=1}^{n} \delta(g_i, map(a_i))}{n}, \tag{4.12}$$

where $\delta(\cdot,\cdot)$ is an indicator function. The mapping function $map(a_i)$ permutes the cluster labels and maps each label a_i to its equivalent one in the database.

NMI measures the ability of the clustering algorithm to reconstruct the underlying label distribution of the data. Denote the cluster sets from the ground truth and the clustering method by C and C' respectively, then we calculate NMF using

$$\mathrm{NMI}(C, C') = \frac{\sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \cdot \log_2 \frac{p(c_i, c'_j)}{p(c_i) \cdot p(c'_j)}}{\max(H(C), H(C'))}, \tag{4.13}$$

where $p(c_i)$, $p(c'_j)$ indicate the probabilities of a sample belonging to the clusters c_i and c'_j respectively. $p(c_i, c'_j)$ is the joint probability suggesting the specified sample belongs to c_i and c'_j at the same time. $H(\cdot)$ denotes the entropy. Generally, $\mathrm{NMI}(C, C') \in [0, 1]$, and the larger value means better clustering performance.

[2]http://images.ee.umist.ac.uk/danny/database.html
[3]http://www1.cs.columbia.edu/CAVE/software/softlib/coil-20.php
[4]http://www-2.cs.cmu.edu/ webkb/

4.3 Performance comparison

To demonstrate the advantages of MMCF, we compare it with the following methods:

- Traditional k-means clustering method (Kmeans).
- Principle component analysis (PCA)[6].
- Non-negative matrix factorization (NMF)[16].
- Concept factorization (CF)[10].
- Convex non-negative matrix factorization (ConNMF)[29].
- Graph regularized non-negative matrix factorization (GNMF)[11].
- Locally consistent concept factorization (LCCF)[12].

For all the algorithms, the number of concepts is set to the number of classes in the data collection. The number of nearest neighbor is set to 5 for the graph-based methods. For GNMF, LCCF and MMCF, we search the best parameters from the grid $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000\}$. The results derived from the optimal parameters are reported. For GNMF and LCCF, we use the gaussian weight to construct the weight matrix for the graph Laplacian, and the bandwidth is set to the inverse of the mean square of Euclidean distances between all samples, *i.e.*, $\frac{1}{\tau} = \frac{1}{n^2}\sum_{i,j=1}^{n} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2$. For MMCF, we adopted three weighting schemes to diversify the candidate manifolds for better approximating the intrinsic manifold. In concrete, for gaussian weighting, we vary the bandwidth in a broad range of area, *i.e.*, $\{\frac{\tau}{[45:-5:5]}, 1, \tau[5 : 5 : 30]\}$; for binary weighting, we vary the number of nearest neighbor from 1 to 5; for cosine similarity weighting, we directly compute the dot product of two samples. Thus, we totally have a set of 22 candidate manifolds for multi-manifold learning in MMCF.

For all the compared algorithms, k-means were run 20 times with different initializations, and we record the results corresponding to the minimum objective value. To randomize the experiments, the final results are averaged over 10 repeated test runs. All data points are normalized to unit Euclidean length and we perform clustering on the new low-dimensional representation.

4.4 Results

The clustering results are reported in Tables 2 and 3. The best results on each data are highlighted in boldface. From these tables, a number of interesting points can be observed as follows.

- MMCF enjoys more promising performances in comparison with other methods, which justifies the effectiveness of incorporating multi-manifold learning into concept factorization.
- MMCF significantly outperforms LCCF, which can be attributed to the fact that LCCF only uses one local manifold to represent the intrinsic manifold of the data. However, MMCF employs a set of diverse candidate manifold to

maximally approximate the true manifold, leading to better locally preserved structure in the data space.

- In most cases, GNMF performs better than NMF while LCCF performs better than CF and Convex NMF. This demonstrates that taking into account the locally geometrical structure indeed improves the clustering performance.
- On WebKB, the NMI values are very low for all the compared algorithms. The reason for this might be that the seven classes in this database are severely imbalanced since the number of some classes is much smaller than others.

Table 2 Clustering performance of the compared algorithms(AC: %)

Database	Kmeans	PCA	NMF	CF	ConNMF	GNMF	LCCF	MMCF
ORL	51.05	57.10	58.12	48.90	53.50	56.03	51.65	**60.05**
UMIST	39.77	44.33	41.17	38.77	43.48	54.92	49.13	**60.23**
COIL20	63.62	60.39	62.44	61.65	66.67	74.56	76.57	**78.84**
WebKB	35.54	35.36	33.82	35.47	40.26	46.78	46.01	**48.15**
Avg.	47.49	49.30	48.89	46.20	50.98	58.07	55.84	**61.82**

Table 3 Clustering performance of the compared algorithms (NMI: %)

Database	Kmeans	PCA	NMF	CF	ConNMF	GNMF	LCCF	MMCF
ORL	71.06	74.78	74.27	66.95	73.04	74.10	69.57	**77.84**
UMIST	58.94	62.69	58.80	57.20	62.16	72.57	68.04	**76.68**
COIL20	74.31	72.86	72.67	72.05	75.29	84.58	87.41	**89.97**
WebKB	16.18	**17.48**	14.61	15.24	14.11	13.43	14.23	15.70
Avg.	55.12	56.95	55.09	52.86	56.15	61.17	59.81	**65.05**

4.5 Parameter selection

There are two important parameters in our approach, *i.e.*, α and β. In order to explore the influences of different parameters on clustering performance, we select the parameters from the wide range of grids $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 100, 1000\}$. Here we illustrate the results in terms of AC in Fig. 1. The depicted curves of α are obtained when fixing β as its best, and vice versa.

As vividly drawn in these figures, MMCF enjoys better clustering performance on the image database when α is large and β is small. On the document data WebKB, MMCF achieves the best performance when $\alpha = 1$ and β takes 1 or 10. Besides, it can be readily find that our method performs more robustly on UMIST since it has satisfactory performances in a wide range of parameters. Parameter selection has the similar behaviors on other data sets.

5 Conclusion

This paper presents a novel method called Multi-Manifold Concept Factorization (MMCF) for clustering analysis. It incorporates the multi-manifold ensemble learning into concept factorization, thus better respecting the locally geometrical structure of the data. Specifically, it approximates the intrinsic manifold by using a convex combination of some given candidate manifolds. When data points reside on multiple manifolds, MMCF is considerably superior to others only considering one manifold.

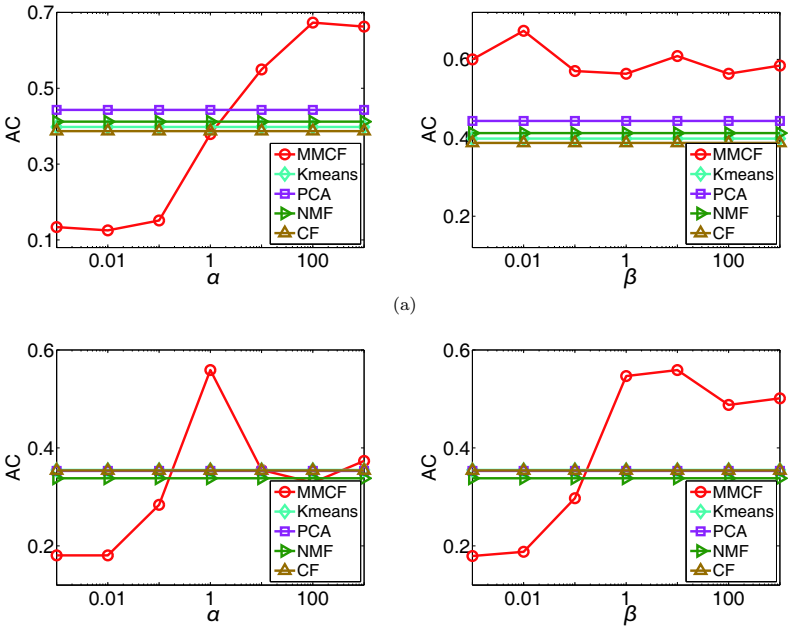

Figure 1. Clustering performance of MMCF with different α and β in a wide range. Here we use AC to evaluate the clustering performance. (a) UMIST; (b) WebKB.

To optimize the objective function, we present an alternative method to learn the variables, which are updated until it reaches the convergence. Experiments were conducted on several real-world databases. Results have shown that the proposed method enjoys more promising performance compared to some alternatives.

References

[1] Bishop C, et al. Pattern recognition and machine learning, Vol. 4. Springer New York. 2006.

[2] Slonim N, Tishby N. Document clustering using word clusters via the information bottleneck method. Proc. of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval. 2000. 208–215.

[3] Xu W, Liu X, Gong Y. Document clustering based on non-negative matrix factorization. Proc. of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval. 2003. 267–273.

[4] Cai D, He X, Han J. Document clustering using locality preserving indexing, IEEE Trans. on Knowledge and Data Engineering, 2005, 17(12): 1624–1637.

[5] Yang Y, Xu D, Nie F, Yan S, Zhuang Y. Image clustering using local discriminant models and global integration. IEEE Trans. on Image Processing, 2010, 19(10): 2761–2773.

[6] Abdi H, Williams L. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics, 2010, 2(4): 433–459.

[7] He X, Niyogi P. Locality preserving projections. Advances in Neural Information Processing Systems, 2003, 16: 153–160.

[8] Wall M, Rechtsteiner A, Rocha L. Singular value decomposition and principal component analysis. A Practical Approach to Microarray Data Analysis, 2003: 91–109.

[9] Lee D, Seung H, et al. Learning the parts of objects by non-negative matrix factorization. Nature, 1999, 401(6755): 788–791.

[10] Xu W, Gong Y. Document clustering by concept factorization. Proc. of the 27th ACM SIGIR Conference on Research and Development in Information Retrieval. ACM. 2004. 202–209.

[11] Cai D, He X, Han J, Huang T. Graph regularized nonnegative matrix factorization for data representation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2011, 33(8): 1548–1560.
[12] Cai D, He X, Han J. Locally consistent concept factorization for document clustering. IEEE Trans. on Knowledge and Data Engineering, 2011, 23(6): 902–913.
[13] Geng B, Tao D, Xu C, Yang L, Hua XS. Ensemble manifold regularization. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2012, 34(6): 1227–1233.
[14] Beck A, Teboulle M. Mirror descent and nonlinear projected subgradient methods for convex optimization. Operations Research Letters, 2003, 31(3): 167–175.
[15] Li P, Chen C, Bu J. Clustering analysis using manifold kernel concept factorization. Neurocomputing, 2012, 87: 120–131.
[16] Seung D, Lee L. Algorithms for non-negative matrix factorization. Advances in neural information processing systems, 2001, 13: 556–562.
[17] Lin C. Projected gradient methods for nonnegative matrix factorization. Neural computation, 2007, 19(10): 2756–2779.
[18] He X, Yan S, Hu Y, Niyogi P, Zhang H. Face recognition using laplacianfaces. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005, 27(3): 328–340.
[19] Belkin M, Niyogi P. Laplacian eigenmaps for dimensionality reduction and data representation. Neural computation, 2003, 15(6): 1373–1396.
[20] Gu Q, Zhou J. Co-clustering on manifolds. Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data mining. ACM. 2009. 359–368.
[21] Guan N, Tao D, Luo Z, Yuan B. Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent. IEEE Trans. on Image Processing, 2011, 20(7): 2030–2048.
[22] Li P, Bu J, Yang Y, Ji R, Chen C, Cai D. Discriminative orthogonal nonnegative matrix factorization with flexibility for data representation. Expert Systems With Applications, 2014, 41(4): 1283–1293.
[23] Belkin M, Niyogi P, Sindhwani V. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. Journal of Machine Learning Research, 2006, 7: 2399–2434.
[24] Li P, Bu J, Chen C, Wang C, Cai D. Subspace learning via locally constrained a-optimal nonnegative projection. Neurocomputing, 2013, 115: 49–62.
[25] Hoyer PO, Dayan P. Non-negative matrix factorization with sparseness constraints. Journal of Machine Learning Research, 2004, 5: 1457–1469.
[26] Zhang L, Chen Z, Zheng M, He X. Robust non-negative matrix factorization. Frontiers of Electrical and Electronic Engineering in China, 2011, 6(2): 192–200.
[27] Liu H, Yang Z, Wu Z. Locality-constrained concept factorization. Proc. of the 22nd International Joint Conference on Artificial Intelligence. 2011. 1378–1383.
[28] Hua W, He X. Discriminative concept factorization for data representation. Neurocomputing, 2011, 74(18): 3800–3807.
[29] Ding CH, Li T, Jordan MI. Convex and semi-nonnegative matrix factorizations. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2010, 32(1): 45–55.
[30] Boyd S, Vandenberghe L. Convex optimization. Cambridge University Press. 2004.
[31] Sha F, Lin Y, Saul L, Lee D. Multiplicative updates for nonnegative quadratic programming. Neural Computation, 2007, 19(8): 2004–2031.
[31] Graham D, Allinson N. Face recognition: From theory to applications. NATO ASI Series F, Computer and Systems Sciences, 1998, 163: 446–456.
[32] Nene S, Nayar S, Murase H, et al. Columbia object image library (coil-20), Rapport interne CUCS-005-96, Columbia University Computer Science.
[33] Craven M, DiPasquo D, Freitag D, McCallum A, Mitchell T, Nigam K, Slattery S. Learning to construct knowledge bases from the world wide web. Artificial Intelligence, 2000, 118(1): 69–113.