ADAPTIVE GRAPH VIA MULTIPLE KERNEL LEARNING FOR NONNEGATIVE MATRIX FACTORIZATION

Jing-Yan Wang and Mustafa AbdulJabbar
King Abdullah University of Science and Technology (KAUST)
Mathematical and Computer Sciences and Engineering Division
Thuwal 23955-6900, Saudi Arabia
email: {jingyan.wang,mustafa.abduljabbar}@kaust.edu.sa

ABSTRACT

Nonnegative Matrix Factorization (NMF) has been continuously evolving in several areas like pattern recognition and information retrieval methods. It factorizes a matrix into a product of 2 low-rank non-negative matrices that will define parts-based, and linear representation of nonnegative data. Recently, Graph regularized NMF (GrNMF) is proposed to find a compact representation,which uncovers the hidden semantics and simultaneously respects the intrinsic geometric structure. In GNMF, an affinity graph is constructed from the original data space to encode the geometrical information. In this paper, we propose a novel idea which engages a Multiple Kernel Learning approach into refining the graph structure that reflects the factorization of the matrix and the new data space. The GrNMF is improved by utilizing the graph refined by the kernel learning, and then a novel kernel learning method is introduced under the GrNMF framework. Our approach shows encouraging results of the proposed algorithm in comparison to the state-of-the-art clustering algorithms like NMF, GrNMF, SVD etc.

KEY WORDS

Data Representation, Nonnegtive Matrix Factorization, Graph Regularization, Multiple Kernel Learning.

1 Introduction

Nonnegative matrix factorization (NMF) [16] has been introduced as a matrix factorization technique that produces a useful decomposition in the analysis of data. NMF decomposes the data as a product of two matrices that are constrained by having nonnegative elements. This method results in a reduced representation of the original data that can be seen either as a feature extraction or a dimensionality reduction technique, and have become popular in recent years for data representation for bioinformatics, medical imaging, pattern recognition and information retrieval. Recently, Cai et al. improved the transitional NMF to Graph regularized Nonnegative Matrix Factorization (GrNMF) in [4]. The basic idea is that the data is drawn from sampling a probability distribution that has support on or near to a sub-manifold of the ambient space. One then hopes to find a compact representation, which uncovers the hidden se-

mantics and simultaneously respects the intrinsic geometric structure. In GrNMF, the geometrical information of the data space is encode by constructing a nearest neighbor graph, and then the matrix factorization is sought respecting the graph structure.

The key component of GrNMF is the graph. In the original GrNMF algorithm, the graph is constructed according to the original input feature space. The nearest neighbors of a data point are found by comparing the Euclidean distances [23] between pairs of data points, while the weights of edges are also estimated in the Euclidean space. However, as is well known that in some data clustering and classification problems, using the original linear feature space directly is not appropriate because for many applications, the distribution of original data space is nonlinear, which brings problem for the graph construction for GrNMF. We can solve this problem by mapping the input data into a nonlinear feature space, where the mapping is represented by introducing a kernel. The graph should be constructed according to the new nonlinear data space represented by a kernel. In practice, the types and the parameters of the kernels must be selected. Unfortunately, the most suitable kernel for a particular task is often unknown in advance. Moreover, exhaustive search on a user-defined pool of kernels will be quite time-consuming when the size of the pool becomes large. Recently, the so-called multiple kernel learning method [20, 8, 7] have shown the necessity to consider multiple kernels or the combination of kernels rather than a single fixed kernel for data representation. In fact, multiple kernel learning, graph construction and NMF have been extensively studied for data representation in the literatures respectively, but have never been investigated in a uniform framework, thus the inherent relationship among them has been neglected.

In this paper, we try to investigate the inherent relationship between multiple kernel learning and NMF with graph regularization. The multiple kernel learning will provide a new data space for the graph construction of GrNMF, and GrNMF will also provide the criterion for feature selection/ multiple kernel learning. We will unify the multiple kernel learning and GrNMF within a single object function and repeat their optimizations alternately, so that they will effect the learning of each other. In this paper, we propose a unified multiple kernel learning and

graph regularization framework for NMF, referred to as **Adap**tive **G**raph **r**egularized **NMF** with Multiple Kernel (AdpaGrNMF$_{MultiK}$), for data representation of clustering and classification tasks. The main contributions of this paper include:

1. We propose the unified frameworks for multiple kernel learning with new matrix factorization objective functions and incorporate the graph structure into it.

2. GrNMF is improved by utilizing the graph adaptive to the new data space refined by multiple kernel learning.

3. A novel kernel learning method is proposed under the framework of GrNMF.

The rest of the paper is organized as follows: We briefly review the GrNMF in Section 2. In Section 3, we present AdapGrNMF$_{MultiK}$ algorithm to tackle the multiple kernel learning problem for GrNMF. We experimentally compare the proposed methods with other NMF learning methods on the two data sets for clustering and classification tasks in Section 4. Finally, conclusive remarks and future works are presented in Section 5.

2 Overview of Graph Regularized NMF

2.1 Nonnegative Matrix Factorization

Given N data points $\mathcal{X} = \{x_1, \cdots, x_N\} \in \mathbb{R}^D$ represented as a data matrix $X = [x_1, \cdots, x_n] \in \mathbb{R}^{D \times N}$, We consider factorizations of the form:

$$X \approx HW \tag{1}$$

where $X \in \mathbb{R}^{D \times N}$, $H \in \mathbb{R}^{D \times R}$, and $W \in \mathbb{R}^{R \times N}$. Commonly, we have $R \ll D$ and $R \ll N$. NMF can be written in this form, where the data matrix X is assumed to be nonnegative, as are the factors H and W [6].

NMF aims to find two nonnegative matrices H and W whose product can well approximate the original matrix X as in (1). In reality, each data vector x_n is approximated by a linear combination of the columns of H, weighted by the components of W, as

$$x_n \approx \sum_{r=1}^{R} h_r w_{rn} \tag{2}$$

Therefore, H can be regarded as containing a set of basis vectors. Let $w_n = [w_{n1}, \cdots, w_{nR}]^\top$ denote the n-th columns of W. w_n can be regarded as the coding vector or a new representation of the n-th data point with respect to the basis H.

The most commonly used cost function is the squared Euclidean distance between two matrices (the square of the Frobenius norm of two matrices difference):

$$\begin{aligned} O^{NMF}(H, W) =& ||X - HW||^2 \\ =& Tr(XX^\top) - 2Tr(XW^\top H^\top) \\ & + Tr(HWW^\top H^\top) \end{aligned} \tag{3}$$

where $Tr(\cdot)$ denotes the trace of a matrix. The above objective function can be minimized by the iterative update algorithm proposed by Lee and Seung [16].

2.2 Graph regularized NMF

By performing this learning in the Euclidean space, NMF fails to discover the intrinsic geometrical and discriminating structure of the data space [4]. To avoid this limitation Cai et al. [4] introduced the Graph regularized NMF (GrNMF) algorithm, by incorporating a geometrically based regularizer.

In [4] the Local Invariance Assumption (LIA) that was imposed to NMF as: if two data points x_n and x_m are close in the intrinsic geometry of the data distribution, then h_n and h_m, the coding vectors of these two points with respect to the new basis, are also close to each other; vice versa. Cai et al. modeled the local geometric structure by a P nearest neighbor graph $\mathcal{G}$ on a scatter of data points.

For each data point $x_n \in \mathcal{X}$, its P nearest neighbors $\mathcal{N}_n$ in $\mathcal{X}$ can be determined via squared Euclidean distance metric [23] as

$$\begin{aligned} d(x_n, x_m) &= ||x_n - x_m||^2 \\ &= x_n^\top x_n + x_m^\top x_m - 2x_n^\top x_m \end{aligned} \tag{4}$$

A P nearest neighbor graph is constructed for $\mathcal{X}$ as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, A\}$. The node set $\mathcal{V}$ corresponds to N data points. $\mathcal{E}$ is the edge set, and $(x_n, x_m) \in \mathcal{E}$ if $x_m \in \mathcal{N}_n$. $A \in R^{N \times N}$ is the weight matrix on the graph with A_{nm} equal to the weight of edge (x_n, x_m). There are many choices to define the weight matrix A. Two of the most commonly used are as follows:

0-1 Weighting

$$A_{nm} = \begin{cases} 1, & if \ (x_n, x_m) \in \mathcal{E}, \\ 0, & else. \end{cases} \tag{5}$$

Dot-Product Weighting

$$A_{nm} = \begin{cases} x_n^\top x_m, & if \ (x_n, x_m) \in \mathcal{E}, \\ 0, & else. \end{cases} \tag{6}$$

With the defined weight matrix A above, we can use the following **G**raph **r**egularization term to measure the smoothness of the low-dimensional coding vector representations in W

$$\begin{aligned} O^{Gr}(W; A) =& \frac{1}{2} \sum_{n,m=1}^{N} ||w_n - w_m||^2 A_{nm} \\ =& Tr(WDW^\top) - Tr(WAW^\top) = Tr(WLW^\top) \end{aligned} \tag{7}$$

where D is a diagonal matrix whose entries are column sums of A, $D_{nn} = \sum_{n=1}^{N} A_{nm}$ and $L = D - A$ is the graph Laplacian.

By minimizing $O^{Gr}(W; A)$ with respect to W, we expect that if two data points x_n and x_m are close (i.e.,A_{nm}

is big), w_n and w_m are also close to each other. This objective function is similar to the one used in LPP [9], in which it is assumed that the low-dimensional coordinates share the same linear construction weights with the high-dimensional coordinates. Differently, we assume that the sharing relation exists between the coding coefficient space of NMF and the feature space.

Combining this geometrically-based regularizer $O^{Gr}(W; A)$ with the original NMF objective function $O^{NMF}(H, W)$ leads to the loss function of GrNMF [4]:

$$
\begin{aligned}
O^{GrNMF}(H, W; A) =& O^{NMF}(H, W) + \alpha O^{Gr}(W; A) \\
=& Tr(XX^\top) - 2Tr(XW\top H^\top) \\
& + Tr(HWW^\top H^\top) + \alpha Tr(WLW^\top)
\end{aligned}
\tag{8}
$$

in which α is the tradeoff parameter to balance the two terms. Thus the GrNMF problem turns to a minimization problem as

$$
\begin{aligned}
&\min_{H,W} O^{GrNMF}(H, W; A) \\
&subject\ to\ H \geq 0,\ W \geq 0.
\end{aligned}
\tag{9}
$$

where H and W can be solved in a iterative way by updating them alternately [4].

3 Adaptive Graph Regularized NMF with Multi-Kernel Learning

In this section, we attempt to obtain an appropriate data representation for GrNMF in the Hilbert space of kernel methods. Accordingly, multiple kernel learning is considered. Moreover, we re-construct the new graph adaptive to the new data representation and re-regularize the NMF, and then re-estimate the kernel coefficients, resulting the novel iterative data representation algorithms by **r**egularizing **NMF** by **Adap**tive **G**raph — **AdapGrNMF**$_{MultiK}$.

3.1 Multiple Kernel Learning for NMF

Consider a nonlinear mapping $x_n \to \varphi(x_n)$ or $X \to \varphi(X) = [\varphi(x_1), \cdots, \varphi(x_N)]$. Then the kernel matrix $K \in \mathbb{R}^{N \times N}$ is given by $K = \varphi(X)^\top \varphi(X)$. A direct application of NMF to the feature matrix $\varphi(X)$ yields

$$
\varphi(X) \approx HW
\tag{10}
$$

While, in NMF there are no constraints on the basis vectors $H = [h_1, \cdots, h_R]$. For reasons of interpretability it may be useful to impose the constraint that the vectors defining H lie within the column space of $\varphi(X)$: $h_r = \sum_{n=1}^{N} f_{nr}\varphi(x_n)$ or

$$
H = \varphi(X)F
\tag{11}
$$

where f_{nr} is the (n, r)-th element of the matrix $F \in \mathbb{R}^{N \times K}, F \geq 0$. Substituting (11) to (3), we have the objective function for the kernelized version of NMF

$$
\begin{aligned}
O^{NMF_K}(F, W) =& ||\varphi(X) - \varphi(X)FW||^2 \\
=& Tr[(\varphi(X) - \varphi(X)FW)(\varphi(X) - \varphi(X)FW)^\top] \\
=& Tr[\varphi(X)(I - FW)(I - FW)^\top \varphi(X)^\top] \\
=& Tr[\varphi(X)^\top \varphi(X)(I - FW)(I - FW)^\top] \\
=& Tr[K(I - FW)(I - FW)^\top]
\end{aligned}
\tag{12}
$$

Suppose there are altogether L different kernel functions $\{K_l\}_{l=1}^{L}$ available for the NMF task in hand. Accordingly, there are L different associated nonlinear feature spaces. In general, we do not know which kernel space should be used. An intuitive way is to use them all by concatenating all feature spaces into an augmented Hilbert space, and associate each feature space a relevance weight τ_l, $\tau_l \geq 0$, $\sum_{l=1}^{L} \tau_l = 1$. We denote the kernel weights as a vector $\tau = [\tau_1, \cdots, \tau_L]^\top$. Performing the NMF in such feature space is equivalent to employing a combined kernel function for the NMF:

$$
K^\tau = \sum_{l=1}^{L} \tau_l K_l
\tag{13}
$$

Substitute this relation into (12) to obtain the objective function for **Multi**ple **K**ernel based **NMF** (NMF$_{MultiK}$):

$$
O^{NMF_{MultiK}}(F, W, \tau) = Tr\left[\sum_{l=1}^{L} \tau_l K_l(I - FW)(I - FW)^\top\right]
\tag{14}
$$

3.2 Graph Adaptive to Multiple Kernel Learning

To update the graph $\mathcal{G}$ regarding the multiple kernel space, given a τ, the P nearest neighbors $\mathcal{N}_n^\tau$ for the GrNMF algorithm will be re-found by the τ-weighted squared Euclidean distance in multiple kernel space, i.e.,

$$
\begin{aligned}
d_\tau(x_n, x_m) =& ||\varphi(x_n) - \varphi(x_m)||_\tau^2 \\
=& \varphi(x_n)^\top \varphi(x_n) + \varphi(x_m)^\top \varphi(x_m) - 2\varphi(x_n)^\top \varphi(x_m) \\
=& K^\tau(x_n, x_n) + K^\tau(x_n, x_m) - 2K^\tau(x_n, x_m) \\
=& \sum_{l=1}^{L} \tau_l[K_l(x_n, x_n) + K_l(x_n, x_m) - 2K_l(x_n, x_m)]
\end{aligned}
\tag{15}
$$

The corresponding P nearest neighbor graph adaptive to τ is donated as $\mathcal{G}^\tau = \{\mathcal{V}, \mathcal{E}^\tau, A^\tau\}$. Here we discuss the updating of 0-1 weighting and dot-product weighting for the weight matrix A^τ of adaptive graph with multiple kernel learning. 0-1 weighting is simply updated as $A_{nm}^\tau = 1$, if $(n, m) \in \mathcal{E}^\tau$; 0, otherwise. For dot-product weighting, $A_{nm}^\tau = \varphi(x_n)^\top \varphi(x_m) = K^\tau(x_n, x_m) = \sum_{l=1}^{L} \tau_l K_l(x_n, x_m)$, if $(n, m) \in \mathcal{E}^\tau$.

With the graph $\mathcal{G}^\tau$ adaptive to the multiple kernel space, we then re-regularize the NMF$_{MultiK}$ in the multiple kernel space. Similar to the GrNMF and AdapNMF$_{fs}$,

we propose the **Adap**tive **G**raph **r**egularization term as

$$O^{AdapGr}(W; A^\tau) = \frac{1}{2} \sum_{n,m=1}^{N} ||w_n - w_m||^2 A^\tau{}_{nm} \quad (16)$$
$$= Tr(WL^\tau W^\top)$$

where $L^\tau = D^\tau - A^\tau$ is the corresponding graph Laplacian.

By minimizing $O^{AdapGr}(W; A^\tau)$, we expect that if two data points $\varphi(x)_n$ and $\varphi(x)_m$ are close respecting to the new kernel regarding τ, then the representations w_n and w_m of these two points with respect to the new feature selected basis $H = \varphi(X)F$ are also close to each other.

3.3 AdapGrNMF Algorithm with Multiple Kernel Learning

To perform the multiple kernel representation together with the adaptive graph regularized NMF, we first propose the unified AdapGrNMF and multiple kernel learning object function for data representation by combining the loss function of NMF with multiple kernel learning and adaptive graph regularization term, and then develop an alternating update algorithm to estimate the basis matrix $H = \varphi(X)F$, coefficient matrix W and the kernel weight vector τ as follows.

3.3.1 Object function

Combining the adaptive graph-based regularizer defined in (16) with the NMF objective function with multiple kernel defined in (14) leads to the optimization problem of our AdapGrNMF with **Multi**ple **K**ernel learning — AdapGrNMF$_{MultiK}$:

$$\min_{F,W,\tau} O^{AdapGrNMF_{MultiK}}(F, W, \tau)$$
$$= O^{NMF_{MultiK}}(F, W, \tau) + \alpha O^{AdapGr}(W; A^\tau)$$
$$= Tr\left[\sum_{l=1}^{L} \tau_l K_l (I - FW)(I - FW)^\top \right]$$
$$+ \alpha Tr(WL^\tau W^\top)$$
$$= Tr\left[K^\tau (I - FW)(I - FW)^\top \right]$$
$$+ \alpha Tr(WL^\tau W^\top)$$
$$s.t.\ F \geq 0,\ W \geq 0,\ \tau \geq 0,\ \sum_{l=1}^{L} \tau_l = 1.$$
$$(17)$$

3.3.2 Optimization

Since direct optimization to (17) is difficult, we instead adopt an iterative, two-step strategy to alternately optimize (H, W) and τ. At each iteration, one of (H, W) and τ is optimized while the other is fixed, and then the roles of (H, W) and τ are switched. Iterations are repeated until convergence or a maximum number of iterations is reached.

- **On optimizing** (F, W): By fixing τ and updating the adaptive graph $\mathcal{G}^\tau$ and kernel matrix K^τ, the optimization problem (17) is reduced to

$$\min_{F,W} Tr\left[K^\tau (I - FW)(I - FW)^\top \right]$$
$$+ \alpha Tr(WL^\tau W^\top) \quad (18)$$
$$s.t.\ F \geq 0,\ W \geq 0.$$

The Lagrange $\mathcal{L}$ of the above optimization problem is

$$\mathcal{L} = Tr(K^\tau I) - 2Tr(K^\tau W^\top F^\top)$$
$$+ Tr(K^\tau FWW^\top F^\top) + \alpha Tr(WL^\tau W^\top)$$
$$+ Tr(\Phi F^\top) + Tr(\Psi W^\top)$$
$$(19)$$

where $\Phi = [\phi_{nr}]$ and $\Psi = [\psi_{rn}]$ are the lagrange multiplier matrices for constraint $F \geq 0$ and $W \geq 0$ respectively. As we mentioned before, it is often difficult to find a closed form for $O^{AdapGrNMF_{MultiK}}(F, W, \tau)$. Such difficulties often arise when one wishes to maximize or minimize a function subject to fixed outside conditions or constraints. Introducing Φ and Ψ as our Lagrange multipliers is important for solving this class of problems without the need to explicitly solve the conditions and use them to eliminate extra variables.

By setting the partial derivatives of $\mathcal{L}$ with respect to F and W to zero, we have

$$\frac{\partial \mathcal{L}}{\partial F} = -2K^\tau W^\top + 2K^\tau FWW^\top + \Phi = 0$$
$$\frac{\partial \mathcal{L}}{\partial W} = -2F^\top K^\tau + 2F^\top K^\tau FW + 2\alpha WL^\tau + \Psi = 0$$
$$(20)$$

Using the KKT conditions $\phi_{dr} f_{dr} = 0$ and $\psi_{rn} w_{rn} = 0$, we get the following equations for h_{dr} and w_{rn}:

$$(K^\tau FWW^\top)_{nr} f_{nr} - (K^\tau W^\top)_{nr} f_{nr} = 0$$
$$(F^\top K^\tau FW + \alpha WD^\tau)_{rn} w_{rn} - (F^\top K^\tau + \alpha WA^\tau)_{rn} w_{rn}$$
$$= 0$$
$$(21)$$

The chosen KKT (Karush-Kuhn-Tucker)conditions are satisfied at the minimum (F, W) of the given constrained optimization problem, given any constraints, provided that the intersection of the set of feasible directions with the set of descent directions coincides with the intersection of the set of feasible directions for linearized constraints with the set of descent directions. This rather technical regularity assumption holds for all classification problems, since the constraints are always linear. For convex problems (if the regularity condition holds), the KKT conditions are necessary and sufficient to find a solution [3].

These equations lead to the following updating rules:

$$f_{nr} \leftarrow \frac{(K^\tau W^\top)_{nr}}{(K^\tau FWW^\top)_{nr}} f_{nr}$$

$$w_{rn} \leftarrow \frac{(F^\top K^\tau + \alpha WA^\tau)_{rn}}{(F^\top K^\tau FW + \alpha WD^\tau)_{rn}} w_{rn} \tag{22}$$

- **On optimizing** τ: By fixing (F, W) and removing the irrelevant terms, the optimization problem (17) becomes

$$\min_\tau Tr\left[\sum_{l=1}^{L} \tau_l K_l (I - FW)(I - FW)^\top\right]$$

$$= Tr\left[\sum_{l=1}^{L} \tau_l K_l ZZ^\top\right] = \sum_{l=1}^{L} \tau_l g_l \tag{23}$$

$$s.t. \ \tau \geq 0, \ \sum_{l=1}^{L} \tau_l = 1.$$

where $Z = I - FW$ and $g_l = Tr\left[K_l ZZ^\top\right]$. The optimization of (23) with respect to the feature weights τ is a standard Linear Programming (LP) problem .

3.3.3 Algorithms

The iterative AdapGrNMF algorithm with multiple kernel learning (named as AdapGrNMF$_{MultiK}$) is summarized in Algorithm 1.

Algorithm 1 AdapGrNMF$_{MultiK}$ Algorithm.

Require: L base kernel matrices K_l, $l = 1, \cdots, L$;
Require: Initial factorization matrices F^0 and W^0;
Require: Tolerance stopping criterion ξ;
  Initialize the kernel weight variables as $\tau_l^0 = \frac{1}{L}, l = 1, \cdots, L$;
  Initialize $t = 1$;
  repeat
    Update the graph $\mathcal{G}^{\tau^t}$ and its corresponding Laplacian matrix L^{τ^t} according to τ^{t-1} as introduce in section 3.2;
    Update the factorization matrices F^t and W^t as in (22);
    Update the kernel weights τ^t as in (23);
    $t = t + 1$;
  until $O^{AdapGrNMF_{MultiK}}(F^t, W^t, \tau^t) \leq \xi$
  Output $F = F^{t-1}$, $W = W^{t-1}$ and $\tau = \tau^{t-1}$.

---

4 Experiments

In this section, we investigate the use of our proposed AdapGrNMF$_{MultiK}$ algorithms for document clustering and face recognition.

4.1 Experiment I: Document Clustering

In this section, we will evaluate our proposed AdapGrNMF$_{MultiK}$ for data representation in document clustering task.

4.1.1 TDT2 Document Dataset and Setup

The first data set is the NIST Topic Detection and Tracking (TDT2) corpus [1]. The TDT2 corpus consists of data collected during the first half of 1998 and taken from six sources. It consists of 11,201 on-topic documents, which are classified into 96 semantic categories. In this experiment, following [4], those documents appearing in two or more categories were removed and only the largest 30 categories were kept, thus leaving us with $N = 9,394$ documents in total. Each document is represented as a $D = 36,771$ dimension nonnegative feature vector.

We set the dimensionality of the new space R to be the same as the number of clusters. Assume that the document corpus is comprised of R clusters each of which corresponds to a coherent topic. We project a documents feature vector into a R-dimensional semantic space in which each axis corresponds to a particular topic. In this semantic space, each document can be represented as a linear combination of the R topics using NMF, GrNMF or our AdapGrNMF$_{MultiK}$. We apply different matrix factorization algorithms to obtain new data representations W. Therefore, this step maps the data from the original space to a low dimensional (R-dimensional) space. Kmeans [10] is then applied to the new data representation W for document clustering. We finally compare the obtained clusters with the original image category to compute the accuracy.

For this dataset, we applied our $AdapGrNMF_{MultiK}$ algorithm with 10 pre-computed base kernels altogether, i.e.,

- seven **RBF kernels** $K(x_n, x_m) = exp(-\frac{||x_n - x_m||^2}{2\sigma^2})$ with $\sigma = const \times \varrho$, where ϱ is the maximum distance between samples and $const$ varies in the pre-specified range of $\{0.01, 0.05, 0.1, 1, 10, 50, 100\}$.

- two **polynomial kernels** $K(x_n, x_m) = (1 + x_n^\top x_m)^p$ with degree $p = \{2, 4\}$, and

- a **cosine kernel** $K(x_n, x_m) = \frac{x_n^\top x_m}{||x_n|| \cdot ||x_m||}$

4.1.2 Experiment Resutls

We compare our $AdapGrNMF_{MultiK}$ algorithm to other state-of-the-art related clustering algorithms applied to this dataset. The algorithms that we evaluated are listed below:

- Kmeans [10] clustering with original data space,

- Singular Value Decomposition (SVD),

- Normalized Cut (NCut), a typical spectral clustering algorithms;

- Original Nonnegative Matrix Factorization based clustering (NMF) [6];

- Graph regularized Nonnegative Matrix Factorization (GrNMF) [4];

- Our AdapGrNMF$_{MultiK}$ algorithm.

In order to randomize the experiments, the evaluations are conducted with different numbers of clusters varying from 5 to 30. For the fixed cluster number R, we randomly choose R categories from the data set, and mix the images of these R categories as the collection X for clustering. This procedure is repeated 20 times with different initial points and the best result in terms clustering accuracy of Kmeans is recorded. Vectors of documents were created using the term frequency. Then they were clustered using the new and the old algorithms. The documents were labeled using the majority approach, i.e., if most of them assigned to a label belong to the cluster C, then the label of the document is designated as C. After this process is done, the documents get labeled and we know their classes. Then comparing the actual classes with the found classes we can obtain the number of correctly clustered documents. The accuracy of clustering is given by:

$$Accuracy = \frac{Number\ of\ Correctly\ Clustered\ Documents}{Total\ Number\ of\ Documents} \quad (24)$$
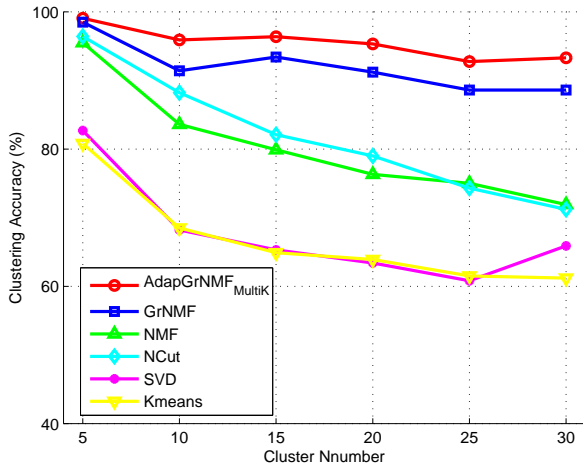

Figure 1. Comparison of clustering accuracy when the cluster number varies from 5 to 30 on TDT2 dataset.

The average results of all methods are presented in Fig. 1. From Fig. 1, we have the following observations:

1. NMF is much worse than GrNMF according to the clustering accuracies over all the 6 clustering number settings, which demonstrates that the W representation learned with original data space performs poorly on the clustering domain. The explanation is that the data distributions of TDT2 data set collected in different topics are quite different. It is interesting to observe that NMF outperforms SVD and Kmeans in terms of clustering accuracies, but NCut is better than NMF.

2. GrNMF is worse than AdapGrNMF$_{MultiK}$. The assumption in GrNMF is that the graph constructed from the original data space can reflect its manifold structure. When the data distributions of different features change considerably in cross-topics learning, the optimal combination coefficients W may not be effectively learned by using GrNMF methods based on the graph from original domains.

3. GrNMF and AdapGrNMF$_{MultiK}$ outperform NMF in terms of clustering accuracies from all the 6 groups of experiments, which demonstrates that the information from the graph can be effectively used in NMF to improve the clustering performance in the TDT2 dataset.

4. AdapGrNMF$_{MultiK}$ is better than GrNMF in terms of clustering accuracies over all the 6 groups of experiments. Moreover, AdapGrNMF$_{MultiK}$ and GrNMF outperform all other method. These results clearly demonstrate that the AdapGrNMF$_{MultiK}$ method can successfully minimize the data distribution mismatch between $\phi(X)$ and HW the structural risk functional through effective combination of multiple base kernels. AdapGrNMF$_{MultiK}$ is better than GrNMF because of the additional utilization of the Adaptive Graph. In addition, some concepts enjoy large performance gains. For instance, the accuracy for the task with 30 clusters significantly increases from 71.90% (NMF) to 93.28% (AdapGrNMF$_{MultiK}$), equivalent to a 21.38% relative improvement. Compared with the best results from the existing methods, AdapGrNMF$_{MultiK}$ (93.28%) enjoys a relative improvement of 4.68% over GrNMF (88.60%).

4.2 Experiment II: Face Recognition

We also evaluated the performance of AdapGrNMF$_{MultiK}$ as a feature representation method in the task of supervised face recognition.

4.2.1 Yale Face Dataset and Setup

The Yale database contains 165 gray scale images of 15 individuals [21]. There are 11 images per subject, one per different facial expression or configuration: centerlight, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink. Thus, each image is also represented by a 1024-dimensional vector in image space.

We divided facial images into training set X_{train} and test set X_{test}. The training set matrix are firstly decomposed using the AdapGrNMF$_{MultiK}$ as $\phi(X_{train}) \approx$

$\phi(X_{train})F_{train}W_{train}$. Then the feature matrix W_{test} in the case of test set X_{test} is computed by LS projection regarding to F_{train}. The kernels $K_{train} = \phi(X_{train})^\top \phi(X_{train})$ and $K_{test} = \phi(X_{test})^\top \phi(X_{train})$ are combination of the following 11 RBF kernel matrices with different bandwidths. We used 11 different values of $const = \{\frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, 16, 32\}$. The feature matrix will be used as features of Nearest Neighbor (NN) classifier and classification accuracies will be averaged over 20 independent runs for comparison.

4.2.2 Experiment Results

In this experiment, we compare our algorithm AdapGrNMF$_{MultiK}$ with the following related algorithms:

- Original NMF [6];

- Single kernel-based NMF (NMF$_K$) [17];

- Multiple Kernel-based NMF (NMF$_{MultiK}$) [2];

Fig. 2 shows the classification accuracy when the value of R varies from 20 to 200. Fig. 2 shows that the proposed AdapGrNMF$_{MultiK}$ consistently outperforms NMF$_{MultiK}$ for all numbers of features R. To further show the consistency of the AdapGrNMF$_{MultiK}$ in the accuracy improvement, Fig. 3 plots the curves for of classification accuracy with different training sample number. For each individual in the dataset, we assigned randomly-selected 2 (3, or 4) images as training samples into the training set and remaining images into the test set. It shows that AdapGrNMF$_{MultiK}$ consistently outperforms NMF$_{MultiK}$ for all different training sample numbers.

Although all the tested algorithms have the same number of basis vector R, Fig. 2 shows that AdapGrNMF$_{MultiK}$ outperforms all other algorithms for all R as the "larger" R has higher recognition accuracy, which is a strong proof of the power of multiple kernel and graph regularizer. However, if the dimension is overreduced, NMF$_{MultiK}$ may not outperform NMF$_K$, as shown in Fig. 2. Fig. 2 shows that the over-dimension reduction by NMF or NMF$_K$ results in sharp increase of the classification error while the discriminant methods NMF$_{MultiK}$ and AdapGrNMF$_{MultiK}$ are much less sensitive to the number of features. Note that although NMF$_K$ maps the data into a nonlinear data space, the discriminant analyses NMF$_K$ where the kernel selection is performed via cross-validation performs very badly if applying them directly on the original 1024-dimensional space, as shown in Fig. 2 and Fig. 3. However, after applying NMF$_{MultiK}$ or AdapGrNMF$_{MultiK}$ to remove the unreliable kernels, the discriminant analysis NMF$_{MultiK}$ or AdapGrNMF$_{MultiK}$ outperforms NMF or NMF$_K$.
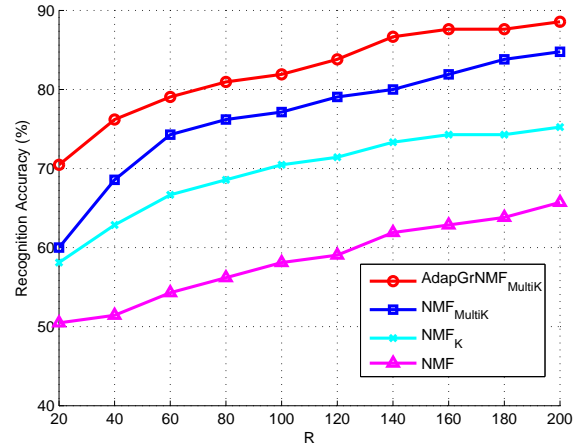

Figure 2. Comparison of recognition accuracy when the intrinsic dimension R varies from 20 to 200, in the case of Train-4 on Yale dataset.
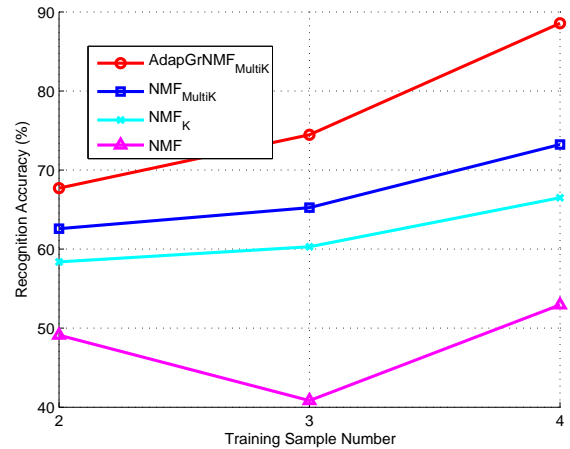

Figure 3. Comparison of recognition accuracy when the training sample number varies from 2 to 4 on Yale dataset.

5 Conclusion

There has been an increasing interest in the nonnegative matrix factorization method in the past few years due to its helpful capability in retrieving human intelligible features. Data analysis processing is a complex task, especially when high dimensional and noisy data is used, so that any method that helps in alleviating the interpretation of the data becomes very appealing. The method presented here is an attempt to improve the ability of the classical NMF algorithm. The proposed Adaptive graph regularized NMF with multiple Kernel learning has proven to be the most accurate among the algorithms that has been used to solve the multidimensional classification problem. It has surpassed the clustering accuracy of the best scoring GrNMF approach

by at least 10% even when the cluster number is increased, because of the additional utilization of adaptive graph. The main advantage of this algorithm is that it alternately optimizes (H, W) and τ. At each iteration, one of (H, W) and τ is optimized while the other is fixed, and then the roles of (H, W) and τ are switched. Iterations stop when reaching the needed accuracy which will allow us to refine the clustering techniques as our processing resources are increased or as our experience in the field becomes more mature.

In the future, more detailed investigation of the theoretical and biological basis is desired. The proposed AdapGrNMF$_{MultiK}$ algorithm can also be applied to other aspects, such as bioinformatics [30, 15, 14, 28, 35, 22], medical imaging [13, 12, 34, 27], biometrics [33, 31, 11, 37, 18, 36, 5, 26, 24, 25] and computer vision [32, 38, 29, 19].

Acknowledgments

References

[1] National Institute of Standards and Technology (NIST) . Topic detection and tracking —tdt phase 2, 1998.

[2] S. An, J.-M. Yun, and S. Choi. Multiple Kernel Nonnegative Matrix Factorization. In *2011 IEEE International Conference On Acoustics, Speech, And Signal Processing*, International Conference on Acoustics Speech and Signal Processing ICASSP, pages 1976–1979, 2011.

[3] A. Battermann and M. Heinkenschloss. Preconditioners for Karush-Kuhn-Tucker matrices arising in the optimal control of distributed systems. In *Control And Estimation Of Distributed Parameter Systems*, volume 126 of *International Series Of Numerical Mathematics*, pages 15–32, 1998.

[4] D. Cai, X. He, J. Han, and T. S. Huang. Graph Regularized Nonnegative Matrix Factorization for Data Representation. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 33(8):1548–1560, AUG 2011.

[5] Chao Wang and Xubo Song. Tracking facial feature points with prediction-assisted view-based active shape model. In *Proceedings 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG 2011)*, pages 259–64. IEEE, 2011 2011. 2011 IEEE International Conference on Automatic Face & Gesture Recognition (FG 2011), 21-24 March 2011, Santa Barbara, CA, USA.

[6] Y. Chen, L. Wang, and M. Dong. Non-Negative Matrix Factorization for Semisupervised Heterogeneous Data Coclustering. *IEEE Transactions on Knowledge And Data Engineering*, 22(10):1459–1474, OCT 2010.

[7] M. Gonen and E. Alpaydin. Multiple Kernel Learning Algorithms. *Journal Of Machine Learning Research*, 12:2211–2268, JUL 2011.

[8] Guohua Chen, Yong Tang, Yan Pan, and Qiang Deng. Question classification using multiple kernel learning and semantic information. *Journal of Computers*, pages 2325–34, 2011.

[9] X. He and P. Niyogi. Locality preserving projections. In *In Advances in Neural Information Processing Systems 16*. MIT Press, 2003.

[10] M. Hua, M. K. Lau, J. Pei, and K. Wu. Continuous K-Means Monitoring with Low Reporting Cost in Sensor Networks. *IEEE Transactions On Knowledge And Data Engineering*, 21(12):1679–1691, DEC 2009.

[11] Jingyan Wang, Yongping Li, and Chao Wang. How to handle missing data in robust multi-biometrics verification. *International Journal of Biometrics*, 3:265–83, 2011 2011.

[12] Jingyan Wang, Yongping Li, Ying Zhang, Honglan Xie, and Chao Wang. Bag-of-Features Based Classification of Breast Parenchymal Tissue in the Mammogram via Jointly Selecting and Weighting Visual Words. In *Proceedings of the Sixth International Conference on Image and Graphics (ICIG 2011)*, pages 622–7, 2011 2011.

[13] Jingyan Wang, Yongping Li, Ying Zhang, Honglan Xie, and Chao Wang. Boosted learning of visual word weighting factors for bag-of-features based medical image retrieval. In *Proceedings of the Sixth International Conference on Image and Graphics (ICIG 2011)*, pages 1035–40, 2011 2011.

[14] Jingyan Wang, Yongping Li, Ying Zhang, and Jianhua He. Semi-supervised Protein Function Prediction via Sequential Linear Neighborhood Propagation. In *Bio-Inspired Computing and Applications. 7th International Conference on Intelligent Computing, ICIC 2011.Revised Selected Papers*, pages 435–41, 2011 2011.

[15] Jingyan Wang, Yongping Li, Ying Zhang, Ning Tang, and Chao Wang. Class Conditional Distance Metric for 3D Protein Structure Classification. In *2011 5th International Conference on Bioinformatics and Biomedical Engineering*, page 4 pp., 2011 2011.

[16] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *NIPS*, pages 556–562, 2000.

[17] H. Lee, A. Cichocki, and S. Choi. Kernel nonnegative matrix factorization for spectral EEG feature extraction. *NEUROCOMPUTING*, 72(13-15):3182–3190, AUG 2009.

[18] J. Liu, Y. Hou, J. Wang, Y. Li, Q. Wang, J. Man, H. Xie, and J. He. Fusing Iris and Palmprint at Image Level for Multi-Biometrics Verification. In *FOURTH INTERNATIONAL CONFERENCE ON MACHINE VISION (ICMV 2011): COMPUTER VISION AND IMAGE ANALYSIS: PATTERN RECOGNITION AND BASIC TECHNOLOGIES*, volume 8350 of *Proceedings of SPIE*, 2012.

[19] Z. Liu, J. Wang, J. Man, Y. Li, X. You, and C. Wang. Self-adaptive Local Fisher Discriminant Analysis for semi-supervised image recognition. *International Journal of Biometrics*, 2012.

[20] Lixin Duan, I. Tsang, and Dong Xu. Domain Transfer Multiple Kernel Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 465–79, 2012.

[21] S. Mitra and M. Savvides. Gaussian mixture models based on the phase spectra for illumination invariant face identification on the Yale database. In *2007 First IEEE International Conference on Biometrics: Theory, Applications, and Systems - BTAS '07*, pages 225–30.

[22] H. Pei, J. Li, M. Lv, J. Wang, J. Gao, J. Lu, Y. Li, Q. Huang, J. Hu, and C. Fan. A graphene-based sensor array for high-precision and adaptive target identification with ensemble aptamers. *Journal of the American Chemical Society*, 2012.

[23] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn. Shape Analysis of Elastic Curves in Euclidean Spaces. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, 33(7):1415–1428, JUL 2011.

[24] C. Wang, W. Jiang, and X. Dong. Characterization of clustered microcalcifications in mammograms based on support vector machines with genetic algorithms. In *Proceedings of International Symposium on Biophotonics, Nanophotonics and Metamaterials*, pages 106–109. Zhejiang Univ; Royal Inst Technol; Chinese Univ Hong Kong; SUNY Buffalo; IEEE LEOS, 2006. International Symposium on Biophotonics, Nanophotonics and Metamaterials, Hangzhou, PEOPLES R CHINA, OCT 16-18, 2006.

[25] C. Wang and Y. Li. Combine image quality fusion and illumination compensation for video-based face recognition. *NEUROCOMPUTING*, 73(7-9, SI):1478–1490, MAR 2010.

[26] C. Wang, Y. Li, and C. Wang. An Efficient Illumination Compensation based on Plane-fit for Face Recognition. In *2008 10TH INTERNATIONAL CONFERENCE ON CONTROL AUTOMATION ROBOTICS & VISION: ICARV 2008, VOLS 1-4*, pages 939–943. IEEE, 2008. 10th International Conference on Control, Automation, Robotics and Vision, Hanoi, VIETNAM, DEC 17-20, 2008.

[27] J. Wang and I. Almasri. Modeling multiple visual words assignment for bag-of-features based medical image retrieval. In *Proceedings of the IASTED International Conference on Computer Graphics and Imaging, CGIM 2012*, pages 217 – 224, Crete, Greece, 2012.

[28] J. Wang, X. Gao, Q. Wang, and Y. Li. ProDis-ContSHC: learning protein dissimilarity measures and hierarchical context coherently for protein-protein comparison in protein database retrieval. *BMC BIOINFORMATICS*, 13(7), 2012.

[29] J. Wang and M. A. Jabbar. Multiple kernel learning for adaptive graph regularized nonnegative matrix factorization. In *Proceedings of the IASTED International Conference on Signal Processing, Pattern Recognition and Applications, SPPRA 2012*, pages 115 – 122, Crete, Greece, 2012.

[30] J. Wang and Y. Li. SEQUENTIAL LINEAR NEIGHBORHOOD PROPAGATION FOR SEMI-SUPERVISED PROTEIN FUNCTION PREDICTION. *JOURNAL OF BIOINFORMATICS AND COMPUTATIONAL BIOLOGY*, 9(6):663–679, DEC 2011.

[31] J. Wang, Y. Li, X. Ao, C. Wang, and J. Zhou. MULTIMODAL BIOMETRIC AUTHENTICATION FUSING IRIS AND PALMPRINT BASED ON GMM. In *2009 IEEE/SP 15TH WORKSHOP ON STATISTICAL SIGNAL PROCESSING, VOLS 1 AND 2*, pages 349–352, 2009.

[32] J. Wang, Y. Li, X. Bai, Y. Zhang, C. Wang, and N. Tang. Learning context-sensitive similarity by shortest path propagation. *PATTERN RECOGNITION*, 44(10-11, SI):2367–2374, OCT-NOV 2011.

[33] J. Wang, Y. Li, P. Liang, G. Zhang, and X. Ao. An Effective Multi-Biometrics Solution for Embedded Device. In *2009 IEEE INTERNATIONAL CONFERENCE ON SYSTEMS, MAN AND CYBERNETICS (SMC 2009), VOLS 1-9*, IEEE International Conference on Systems Man and Cybernetics Conference Proceedings, pages 917–922, 2009.

[34] J. Wang, Y. Li, E. Marchiori, and C. Wang. Iterated large-margin discriminant analysis for feature dimensionality reduction in medical image retrieval. *International Journal of Biomedical Engineering and Technology*, 7(2):116 – 134, 2011.

[35] J. Wang, Y. Li, Q. Wang, X. You, J. Man, C. Wang, and X. Gao. ProClusEnsem: Predicting membrane protein types by fusing different modes of pseudo amino acid composition. *COMPUTERS IN BIOLOGY AND MEDICINE*, 42(5):564–574, MAY 2012.

[36] J. Wang, Y. Li, Y. Zhang, and Y. Huang. Implementing multimodal biometric solutions in embedded systems. *Biometrics - Unique and Diverse Applications in Nature*, 2011.

[37] X. Zhai, Y. Zhao, J. Wang, Y. Li, Q. Wang, and H. Xie. Adaptive SVM Fusion for Robust Multi-Biometrics Verification with Missing Data. In *2011 INTERNATIONAL CONFERENCE ON ENERGY AND ENVIRONMENTAL SCIENCE-ICEES 2011*, volume 11 of *Energy Procedia*, pages 1006–1012, 2011.

[38] Zhonghua Liu, Jingyan Wang, Yongping Li, Ying Zhang, and Chao Wang. Quantized image patches co-occurrence matrix: a new statistical approach for texture classification using image patch exemplars. In *Proceedings of the SPIE - The International Society for Optical Engineering*, volume 8009, page 80092P (5 pp.), 2011 2011.