FEATURE EXTRACTION VIA MULTI-VIEW NON-NEGATIVE MATRIX FACTORIZATION WITH LOCAL GRAPH REGULARIZATION

Zhenfan Wang[1], *Xiangwei Kong*[1*], *Haiyan Fu*[1], *Ming Li*[1] *and Yujia Zhang*[2]

[1]School of Information and Communication Engineering
Dalian University of Technology, Dalian, Liaoning, 116024, China
[2]Fordham University, New York, 10023, USA
Email: zfwang@mail.dlut.edu.cn, {kongxw,fuhy,mli}@dlut.edu.cn, yzhang275@fordham.edu

ABSTRACT

Feature extraction is a crucial and difficult issue in pattern recognition tasks with the high-dimensional and multiple features. To extract the latent structure of multiple features without label information, multi-view learning algorithms have been developed. In this paper, motivated by manifold learning and multi-view Non-negative Matrix Factorization (NMF), we introduce a novel feature extraction method via multi-view NMF with local graph regularization, where the inner-view relatedness between data is taken into consideration. We propose the matrix factorization objective function by constructing a nearest neighbor graph to integrate local geometrical information of each view and apply two iterative updating rules to effectively solve the optimization problem. In the experiment, we use the extracted feature to cluster several realistic datasets. The experimental results demonstrate the effectiveness of our proposed feature extraction approach.

Index Terms— Feature extraction, multi-view learning, non-negative matrix factorization, graph regularization, clustering.

1. INTRODUCTION

High-dimensional input data is always a problem of computer vision and pattern recognition. An approximate solution is to project high-dimensional data into a low-dimensional subspace. Widely used dimensionality reduction techniques to appropriately represent original data include Principal Component Analysis (PCA), Singular Value Decomposition (SVD), Low-Rank Representation (LRR) [1], etc. In [1], low rank representation was used to recover the space of original space. It has been shown in [2], [3] that Non-negative Matrix Factorization (NMF) method provided an efficient representation of the data. As consequence, NMF has been widely used as a dimensionality reduction method in pattern recognition [4]-[6] and information retrieval [2], [7] and shown to be superior to the most state-of-the-art subspace learning methods.

In spite of the success of effective single view dimensionality reduction technologies, in real world the data can be represented by different views. For example, the image can be described as a variety of features (e.g., local features and global features), news can be reported by different sources or languages. Different features represent data in different views with complementary information [6]. Recently, integrating multi-views has shown to achieve better performance than single view [7]-[9]. In [8], A. Kumar *et al*. exploited global and local features in face recognition framework. In [9], S. Kong *et al*. extended sparse coding framework to study multiple views jointly. In [7], Akata *et al*. used co-regularization for each view, where coefficient matrix of each view was the same non-negative matrix. It was equivalent to first connect different views' features and then apply NMF. However, [6] thought the assumption was too strong and proposed a common consensus matrix as soft constraint and learned factors from different views with soft regularization. The computed coefficient matrices were equivalent to joint each view's low dimensional features, which were approximated to original data. Most multi-views methods have considered the relatedness between different views, but relationship between the inner-view can also benefit the performance of extracted features.

To discover the relationship of data, geometrically motivated approaches obtained great interest recently. In order to estimate geometrical properties of the sub-manifold, many manifold learning methods have been proposed, such as ISOMAP [15], Locality Preserving Projection (LPP) [16], Spectral Regression [5]. [14], [17] have shown that using the local invariance and geometrical structure would greatly improve the learning performance.

In this paper, we address the issue of the feature extraction of high-dimensional data with multi-view features, we propose a novel graph regularized multi-view NMF algorithm by taking latent local structure of each inner-view into consideration. Our goal is to generate feature representation in which

*The corresponding author.

similar data is connected nearly in the graph of each view. In our framework, we take geometrical structure of each view into a new matrix factorization objective function by constructing a nearest neighbor graph and propose two iterative updating rules to solve the optimization problem.

In the next section, we give a brief review of Multi-view NMF. In section 3, we introduce our graph regularized Multi-view NMF algorithm. Then in section 4, we present experiments on our proposed algorithm in data clustering problems. Finally, conclusion will summarize the contribution of this paper.

2. A BRIEF REVIEW OF MULTI-VIEW NON-NEGATIVE MATRIX FACTORIZATION

In this section, we briefly review multi-view matrix factorization algorithm. Some definitions are first provided. Given a matrix A, we define $A_{*.k}$ and $A_{k.*}$ as the k-th column and k-th row of matrix A respectively. $A \geq 0$ means that every entries of matrix A are non-negative. Given data collections with multiple views X, let $X^f = [X^f_{*.1}, ..., X^f_{*.N}] \in \mathbb{R}^{M_f \times N} (f \in [1, F])$ represent the feature matrix in the f-th view, with a total of F views. M_f is the feature dimension of the f-th view, N is the number of data. The idea of Non-negative Matrix Factorization is that if the data are non-negative, we can find two non-negative matrices to learn a part-based representation [5]. In Multi-view NMF [6], each X^f is also decomposed into two non-negative matrices : basis matrix $U^f \in \mathbb{R}^{M_f \times K}$ and coefficient matrix $V^f \in \mathbb{R}^{N \times K}$, whose product provides a good approximation to X^f, i.e. $X^f \cong U^f (V^f)^T$, where K is the number of latent concepts. In order to make the fusion of different views meaningful, [6] used ℓ_1 normalization with respect to basis vectors. After that, coefficient matrices of different views are comparable. Incorporating this idea, the Multi-view NMF framework is below:

$$\begin{aligned} \min_{\substack{U^f, V^f, V^* \\ f=1,...,F}} & \sum_{f=1}^{F} \|X^f - U^f (V^f)^T\|_F^2 + \lambda_f \|V^f - V^*\|_F^2, \\ \text{s. t.} \quad & U^f \geq 0, V^f \geq 0,\ V^* \geq 0, \\ & \|U^f_{*.k}\|_1 = 1,\ k = 1, ..., K,\ f = 1, ..., F. \end{aligned} \tag{1}$$

[6] suggested to normalize factorization U, V with diagonal matrix Q, $UV^T = UQ^{-1}QV^T$, where Q^f is defined as:

$$Q^f \triangleq Diag(\sum_{m=1}^{M} U^f_{m,1}, \sum_{m=1}^{M} U^f_{m,2} ... \sum_{m=1}^{M} U^f_{m,K}) \tag{2}$$

where $Diag(\cdot)$ denotes a diagonal matrix. After that, the absolute sum of each basis vector is 1, i.e., $\|U_{*.i}\|_1 = 1$. According to (2), the problem (1) is equivalent to the following optimization problem:

$$\begin{aligned} \min_{\substack{U^f, V^f, V^* \\ f=1,...,F}} & \sum_{f=1}^{F} \|X^f - U^f (V^f)^T\|_F^2 \\ & + \sum_{f=1}^{F} \lambda_f \|V^f Q^f - V^*\|_F^2, \\ \text{s. t.} \quad & U^f \geq 0,\ V^f \geq 0,\ f = 1, ..., F,\ V^* \geq 0. \end{aligned} \tag{3}$$

3. PROPOSED METHOD

Multi-view NMF learns a joint view representation. However, it fails to discover the geometrical structure of inner-view space. In real world, geometrical information of each view can improve learning performance. In this section, we introduce graph regularization into Multi-view NMF to avoid this limitation.

3.1. Graph Weight Matrix

In NMF based methods, we aim to find a new data representation to approximate the original data. It is nature to assume that if two data are close in latent distribution, the representations are also close to each other. With the studies of manifold learning theory [17], a nearest neighbor graph may be the approximate solution.

Each x_i is considered as a vertice, we find its k nearest neighbors and set edges weight between x_i and its neighbors. While there are many ways to compute the weight matrix W, we choose 0-1 weight to compute. Particularly, $W_{ij} = 1$ if and only if x_i is connected with x_j. Euclidean distance is used as the distance metric of original feature of data.

3.2. Graph Regularized Multi-view NMF

With the defined weight matrix W, we can use it to smooth the coefficient vectors. If two data x_i and x_j are close in the latent data distribution, then the low dimensional representation $V_{i.*}$ and $V_{j.*}$ are also close to each other. We use the Euclidean distance to measure the distance between basis vector $V_{i.*}$ and $V_{j.*}$, $\mathcal{D}(V_{i.*}, V_{j.*}) = \|V_{i.*} - V_{j.*}\|^2$. Then we define the smoothness penalty $\mathcal{R}^f$ of each view as below:

$$\begin{aligned} \mathcal{R}^f &= \sum_{j,l=1}^{N} \|V^f_{j.*} - V^f_{l.*}\|^2 W^f_{jl} \\ &= \text{Tr}((V^f)^T D^f V^f) - \text{Tr}((V^f)^T W^f V^f) \\ &= \text{Tr}((V^f)^T L^f V^f) \end{aligned} \tag{4}$$

where D is a diagonal matrix, $D_{(j,j)} \triangleq \sum_l W_{j,l}$, $L \triangleq D - W$. According to (4), we fuse local geometric structure of each view into MultiNMF, the objective function of the proposed

approach can be described as:

$$\min_{\substack{U^f,V^f,V^* \\ f=1,...,F}} \sum_{f=1}^{F} \|X^f - U^f(V^f)^T\|_F^2 + \lambda_f \|V^f Q^f - V^*\|_F^2 + \mu \sum_{f=1}^{F} \lambda_f \text{Tr}((V^f)^T L^f V^f), \quad \text{s. t.} \quad U^f \geq 0,\ V^f \geq 0,\ f = 1, ..., F,\ V^* \geq 0. \tag{5}$$

The problem (5) is difficult to be solved, because the loss function is not convex in both U and V together. As [5] suggested, let $\Psi_{i,k}$ and $\Phi_{j,k}$ be the Lagrange multiplier for constraint $U_{i,k}$ and $V_{j,k} \geq 0$. Then, the optimization problem 5 is equivalent to minimize the loss function (6) over U, V, V^* as:

$$Loss = \sum_{f=1}^{F} \|X^f - U^f V^{fT}\|_F^2 + \lambda_f \|V^f Q^f - V^*\|_F^2 + \mu \sum_{f=1}^{F} \lambda_f \text{Tr}((V^f)^T L^f V^f) + \sum_{f=1}^{F} (\text{Tr}(\Psi (U^f)^T) + \text{Tr}(\Phi (V^f)^T)). \tag{6}$$

3.3. Optimization of Minimizing Objective Function

To minimize the function (6), we adopt iterative updating procedure. Firstly, we keep V^* fixed and update U and V. Then, we minimize the loss function (6) over V^* while keeping U and V fixed.

Fixing V^*, update U and V: When V^* is given, each views are independent, for simplicity, U, V and Q represent U^f, V^f, Q^f. The objective function for each view is as follow:

$$L = \|X - UV^T\|_F^2 + \lambda_f \|VQ - V^*\|_F^2 + \mu * \lambda_f \text{Tr}(V^T L V) + \text{Tr}(\Psi U^T) + \text{Tr}(\Phi V^T). \tag{7}$$

The partial derivatives of *L* with respect to U and V are:

$$\frac{\partial L}{\partial U} = UV^T V + \lambda_f P - XV + \Psi$$
$$\frac{\partial L}{\partial V} = VU^T U - X^T U + \lambda_f (V - V^* + \mu LV) + \Phi \tag{8}$$

where $P \triangleq (\sum_{m=1}^{M} U_{m,k} \sum_{n=1}^{N} V_{n,k}^2 - \sum_{n=1}^{N} V_{n,k} V_{n,k}^*)$. Using the Karush-Kuhn-Tucker (KKT) conditions $\Psi_{i,k} U_{i,k} = 0$ and $\Phi_{j,k} U_{i,k} = 0$, we obtain the following updating rule:

$$U_{j,k} = U_{j,k} \times \frac{(XV)_{j,k} + \lambda_f \sum_{n=1}^{N} V_{n,k} V_{n,k}^*}{(UV^T V)_{j,k} + \lambda_f \sum_{n=1}^{N} V_{n,k}^2},$$
$$V_{j,k} = V_{j,k} \times \frac{(X^T U + \lambda_f V^* + \lambda_f \times \mu WV)_{j,k}}{(VU^T U + \lambda_f V + \lambda_f \times \mu DV)_{j,k}}. \tag{9}$$

When computing U^f and V^f, we first compute U^f and then normalize column vectors of U^f and V^f using Q^f as (2) defined:

$$U^f \leftarrow U^f (Q^f)^{-1}, V^f \leftarrow V^f Q^f. \tag{10}$$

Fixing U and V, update V^*: When U and V are computed over each view, we take the derivative of loss function (7) over V^* and get close-form solution to V^*:

$$V^* = \frac{\sum_{f=1}^{F} \lambda_f V^f Q^f}{\sum_{f=1}^{F} \lambda_f}. \tag{11}$$

After several iterations, the loss value can converge. The proposed method is summarized in Algorithm 1.

Algorithm 1 Our proposed framework

Input: Data feature of each view $\{X^1, X^2, ..., X^F\}$, parameters $\{\lambda_1, \lambda_2, ..., \lambda_F, \mu, k\}$.
Output: Basic matrices, coefficient matrices and center coefficient matrix$\{ U^1, V^1, U^2, V^2, ..., U^F, V^F, V^*\}$.
Initialize: For each view, $\|X^f\|_1 = 1$, use NMF to compute initial $\{U, V \text{ and } V^*\}$.
repeat
  for f=1 **to** F **do**
    repeat
      Fix V^* and update V^f, U^f by (9).
      Normalize V^f, U^f by (10).
    until (7) is converged.
  end for
  Fix V^f and U^f and update V^* by (11).
until (5) is converged.

4. EXPERIMENT

In this section, we apply the proposed framework as data representation method on clustering problem to demonstrate the effective performance of our feature extraction algorithm.

4.1. Experiment settings

For comparison, our metrics are the same as [6]. Three public datasets are used in the experiments. The first is text data 3-Sources, the last two are image datasets: CMU PIE face data and UCI handwritten digit data . The information of them are simply introduced as bellow.

- **3-Sources Text Dataset:** Collected from three well-known online news sources: BBC, Reuters and Guardian from the period February to April 2009 with totally 948 articles of 416 news. 169 news, which all three sources reported are testing samples. One of the six topical (business, entertainment, health, politics, sport and technology) is the label of the 169 testing samples [18].

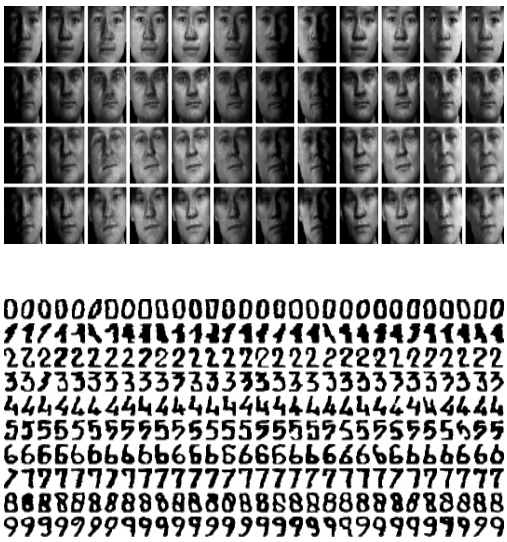

Fig. 1. Example of PIE and UCI Digit Dataset.

- **CMU PIE Dataset:** It originally contains 41,368 32×32 grayscale images of 68 people under 13 different poses, 43 different illumination condition [19]. We use one pose and randomly choose 42 images per person with totally 2856 images. In the terms of features, we use original pixel and HOG feature as multi-view data. The sample pictures of CMU PIE data are showed in Fig. 1.
- **UCI Handwritten Digit Dataset:** As [6] did, we use low Fourier factors and original pixel features as the different views. The sample images of UCI Handwritten Digit dataset are showed in Fig. 1. Because the original images were lost, we sample the pixel features in 15×16 pixels.

There are three kind parameters (λ_f, μ, k) in our framework. We set each $\lambda_f = 0.01$, $\mu = 10$ by cross-validation and the number of nearest neighbors $k = 5$. For comparison, the clustering results are evaluated by comparing the predicted labeled with the ground truth. We use two metrics: the accuracy (AC) and the normalized mutual information (NMI). Please refer to [2] for detail definitions.

We compare many algorithms, including Single View (BSV and WSV) [2], ContactNMF [2], ColNMF [7], Co-reguSC [8], MultiNMF [6] and SC-ML [21]. While single View algorithm use [2] to converge U, V and V^* and V^* is used to compute the best and worst performance refer to BSV and WSV respectively.

4.2. Results

In our experiments, we run 20 times and obtain the average and standard deviation performance. The clustering results of different algorithms on three datasets are showed in Table 1 and Table 2. From the tables, we can see that our proposed algorithm performs better in each dataset in terms of AC and NMI. Although other methods consider multiple feature integration, Co-reguSC and SC-ML use latent data relationship, the results demonstrate that our proposed Multi-view NMF with local graph regularization feature extraction framework can learn a better feature representation.

Table 1. The AC of different methods.

Algorithm	Accuracy(%)		
	3-Sources	PIE	Digit
BSV	60.8±.01	55.2±.02	68.5±.05
WSV	49.1±.03	47.6±.01	63.4±.04
ConcatNMF	58.6±.03	51.5±.00	67.8±.06
ColNMF	61.3±.02	56.3±.00	66.0±.05
Co-reguSC	47.8±.01	59.5±.02	86.6±.00
MultiNMF	68.4±.06	64.8±.02	88.1±.01
SC-ML	54.0±.00	72.3±.00	88.1±.00
Our Method	**72.6±.02**	**72.5±.02**	**95.1±.10**

Table 2. The NMI of different methods.

Algorithm	Normalized Mutual Information(%)		
	3-Sources	PIE	Digit
BSV	53.0±.01	74.1±.00	63.4±.03
WSV	44.1±.02	69.1±.02	60.3±.03
ConcatNMF	51.7±.03	70.5±.00	60.3±.03
ColNMF	55.2±.02	68.3±.00	62.1±.03
Co-reguSC	41.4±.01	80.5±.01	77.0±.00
MultiNMF	60.2±.06	82.2±.02	80.4±.01
SC-ML	45.5±.00	85.1±.00	87.6±.00
Our Method	**67.1±.02**	**90.2±.01**	**90.1±.04**

5. CONCLUSION

This paper proposed a new NMF-based algorithm by merging local geometrical structure information of each view in a multi-view feature extraction framework. Our model considered the inner-view relatedness between data, which can be approximated by a nearest neighbor graph. Experimental results demonstrated the effectiveness of our proposed NMF-based multiview feature extraction algorithms. Different local geometric structure models may result in different results. In the future work, we will exploit a new local geometric structure framework.

6. ACKNOWLEDGEMENTS

The work is supported by the Foundation for Innovative Research Groups of the NSFC (Grant No. 71421001), NSFC (Grant No. 61172109) and the Fundamental Research Funds for the Central Universities (DUT14QY03 and DUT14RC(3)103).

7. REFERENCES

[1] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *IEEE Int. Conf. Computer Vision (ICCV)*, Barcelona, Spain, Nov. 2011, pp. 1615–1622.

[2] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proc. ACM Conf. Research and Development in Information Retrieval (SIGIR)*, Toronto, Canada, July 2003, pp. 267–273.

[3] J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," in *Proc. Nat. Academy of Sciences*, Mar. 2004, vol. 101, pp. 4164–4169.

[4] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, Aug. 1999.

[5] D. Cai, X. He, J. Han, and T. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1548–1560, Aug. 2011.

[6] J. Liu, C. Wang, J. Gao, and J. Han, "Multi-view clustering via joint nonnegative matrix factorization," in *Proc. SDM*, Austin, Texas, May 2013, vol. 13, pp. 252–260.

[7] Z. Akata, C. Thurau, and C. Bauckhage, "Non-negative matrix factorization in multimodality data for segmentation and label prediction," in *16th Computer Vision Winter Workshop*, Mitterberg, Austria, Feb. 2011.

[8] A. Kumar, P. Rai, and H. Daume, "Co-regularized multi-view spectral clustering," in *NIPS*, Granada, Spain, Dec. 2011, pp. 1413–1421.

[9] S. Kong, X. Wang, D. Wang, and F. Wu, "Multiple feature fusion for face recognition," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, Shanghai, China, Apr. 2013, pp. 1–7.

[10] X. Jiang, L. Ma, and Y. Yang, "Cluster constraint based sparse nmf for hyperspectral imagery unmixing," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Paris, France, Oct. 2014, pp. 5107–5111.

[11] J. H. Zhou, H. Y. Fu, and X. W. Kong, "A balanced semi-supervised hashing method for CBIR," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Brussels, Belgium, Sept. 2011, pp. 2481–2484.

[12] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, Dec. 2000.

[13] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering.," in *NIPS*, Vancouver, Canada, Dec. 2001, vol. 14, pp. 585–591.

[14] D. Cai, X. He, and J. Han, "Spectral regression for efficient regularized subspace learning," in *IEEE Int. Conf. Computer Vision (ICCV)*, Rio de Janeiro, Brazil, Oct. 2007, pp. 1–8.

[15] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.

[16] X. F. He, S. C. Yan, Y. X. Hu, P. Niyogi, and H. J. Zhang, "Face recognition using laplacianfaces," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 328–340, Mar. 2005.

[17] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. ,Machine Learning Research*, vol. 7, pp. 2399–2434, Dec. 2006.

[18] D. Greene and P. Cunningham, "A matrix factorization approach for integrating multiple data views," in *Proc. Eur. Conf. Mach. Learn. Knowl. Discovery in Databases (ECML PKDD)*, pp. 423–438. Bled, Slovenia, Sept. 2009.

[19] T. Sim, S. Baker, and M. Bsat, "The cmu pose, illumination, and expression (PIE) database," in *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognition*, Pittsburgh, PA, Jan. 2001.

[20] M. Yin, S. Cai, and J. Gao, "Robust face recognition via double low-rank matrix recovery for feature extraction," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Melbourne, Australia, Sept. 2013, pp. 3770–3774.

[21] X. Dong, P. Frossard, P. Vandergheynst and N. Nefedov, "Clustering on multi-layer graphsvia subspace analysis on Grassmann manifolds," in *IEEE Trans. Signal Process.*, vol. 62, pp. 905–918, Feb. 2014.