FARNESS PRESERVING NON-NEGATIVE MATRIX FACTORIZATION

Mohammadreza Babaee[1,*], *Reza Bahmanyar*[2,*], *Gerhard Rigoll*[1], *Mihai Datcu*[2,*]

[1]Institute for Human-Machine Communication,
Technische Universität München, Munich, Germany
[2]Remote Sensing Technology Institute (IMF),
German Aerospace Center (DLR), Oberpfaffenhofen, Germany

ABSTRACT

Dramatic growth in the volume of data made a compact and informative representation of the data highly demanded in computer vision, information retrieval, and pattern recognition. Non-negative Matrix Factorization (NMF) is used widely to provide parts-based representations by factorizing the data matrix into non-negative matrix factors. Since non-negativity constraint is not sufficient to achieve robust results, variants of NMF have been introduced to exploit the geometry of the data space. While these variants considered the local invariance based on the manifold assumption, we propose *Farness preserving Non-negative Matrix Factorization (FNMF)* to exploits the geometry of the data space by considering non-local invariance which is applicable to any data structure. FNMF adds a new constraint to enforce the far points (i.e., non-neighbors) in original space to stay far in the new space. Experiments on different kinds of data (e.g., Multimedia, Earth Observation) demonstrate that FNMF outperforms the other variants of NMF.

Index Terms— Non-negative Matrix Factorization, Farness Preserving, Clustering.

1. INTRODUCTION

The exponential growth of the available data (e.g., multimedia, Earth Observation) increases the demand for efficient methods to provide a compact and informative representation of the data contents. Nowadays, the contents of the data is represented to data mining algorithms by a matrix composed of high-dimensional vectors of the most descriptive features of the data samples, so-called *feature vectors*. However, high-dimensional data leads to the storage problem, the curse of dimensionality which increase the compuation effort, and the limited degree of freedom [1].

Matrix factorization methods have shown impressive performance in addressing these problems by providing a compact representation of the original high-dimensional data. The main idea behind matrix factorization is to decompose a matrix into two or three lower-dimensional matrix factors such that their product is a good approximation of the original matrix. A variety of these methods have been proposed using different constraints on matrix factors such as PCA [2], SVD [3], and VQ [4].

Because the main goal in data mining algorithms is to provide human understandable results, developing methods which perform similar to human brain have attracted a great attention in recent years. *Non-negative Matrix Factorization (NMF)* is a widely used matrix factorization method following the parts-based perception behavior of the human brain, i.e., perception of an object by combining the perceptions of its parts [1, 5, 6]. The parts-based representation is achieved by enforcing non-negativity constraint to the matrix factors which only allows additive combinations of the original data. However, the non-negativity constraint is not enough to achieve robust data representations [1]. Therefore, wide range of applicability of NMF motivated the researchers to improve it by introducing additional constraints. For example, many of the recent works focused on preserving the intrinsic geometry of the data space by defining new objective functions. GNMF, proposed by Cai et al. [6], considers the local invariance by constructing a nearest neighbor graph and encoding the geometrical information of the data space. Liu and Wu [7] introduced CNMF to constrain NMF to use the prior annotation of the data. This enforces the points from same class to be encoded similarly. Gu and Zhou [8] used local linear embedding assumption to propose NPNMF. They introduced a new constraint to allow each data point to be presented by its neighbors. All the mentioned variants of NMF assume that the data points are sampled form a sub-manifold of the ambient space, i.e., data points form a flat high-dimensional euclidean space. Therefore, the proposed constraints were applied only to the neighboring points to exploit the geometry of the data space.

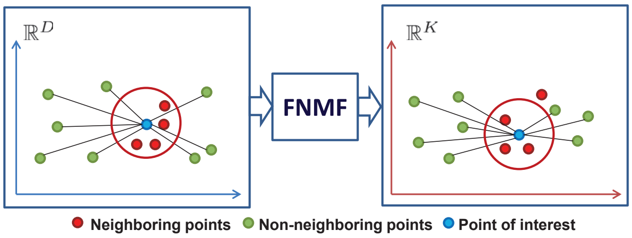


Fig. 1. Data from high-dimensional ($\mathbb{R}^D$) space are mapped to low-dimensional one ($\mathbb{R}^K$), where $K \ll D$, such that the farness property is respected. The red points located in the red circle are the neighboring points of the blue interest point where the green points are the non-neighboring points.

In this paper, we introduce a novel algorithm, called Farness preserving Non-negative Matrix Factorization (FNMF) to exploit the geometry of the data space by introducing a new constraint to enforce the far points (i.e., non-neighbors) in original space to be also far in the new space. While the previous variants of NMF consider the similarity of the neighboring points based on manifold structure assumption, FNMF considers the discrimination of the points

[*] The authors are also affiliated by Munich Aerospace Faculty, Munich, Germany.

which is applicable to any structure of the data space. Fig 1 shows the mapping of the data points from high-dimensional space to the lower-dimensional one using FNMF. The red points located in the red circle are the neighboring points of the blue interest point where the green points are the non-neighboring points. FNMF enforces the green points to stay far away from the interest point. Experiments on different real word datasets demonstrate that FNMF outperforms NMF and its manifold-based variant as well as other state-of-the-art dimensionality reduction methods.

The rest of the paper is organized as follows. Section 2 presents a review of NMF. In Section 3, we explain our algorithm followed by its optimizing rules. The experimental results are provided in Section 4. Finally, in Section 5 we draw our conclusion.

2. A REVIEW OF NMF

Non-negative Matrix Factorization (NMF) is a widely used matrix factorization method which provides a part-based representation of data by enforcing non-negative constraint to the matrix factors. Given a non-negative matrix $X = [x_1, \dots, x_N] \in \mathbb{R}^{D\times N}$, where each column is a feature vector representing a data sample, the goal of NMF is to factorize X into two non-negative matrices U and V such that:

$$X \approx UV^T, \quad \text{where} \quad U \in \mathbb{R}^{D\times K}, \quad V \in \mathbb{R}^{N\times K} \tag{1}$$

This factorization is a constrained non-convex optimization problem with the cost function equal to:

$$\begin{aligned} F &= \|X - UV^T\|_F^2 \\ \text{s.t.} \quad U &= [u_{ik}] \geq 0, \\ V &= [v_{jk}] \geq 0 \end{aligned} \tag{2}$$

The cost function is convex only in U or V, but not convex in both together. Therefore, there is no global solution for the algorithm, but Lee and Seung [5, 9] presented an iterative update algorithm to find a local minimum as follows:

$$u_{ik} \leftarrow \frac{(XV)_{ik}}{(UV^TV)_{ik}}, \quad v_{jk} \leftarrow v_{jk}\frac{(X^TU)_{jk}}{(VU^TU)_{jk}} \tag{3}$$

It is proved that the updating rules can converge to a local minimum of the cost function [9].

3. FARNESS PRESERVING NON-NEGATIVE MATRIX FACTORIZATION

NMF is an unsupervised learning algorithm which represents the data in parts. Additionally, specific constraints can also be imposed to its main objective function in order to hold some properties of the original data space during mapping to the new space. Possibility to add new constraints makes NMF a powerful framework to derive new data representation algorithms. In this section, we introduce our *Farness Preserving Non-negative Matrix Factorization (FNMF)* algorithm which is proposed to exploit the geometry of the data space by enforcing the far points in the original data space to stay far in the new one.

3.1. Objective Function

The input data is represented as a matrix $X = [x_1, ..., x_N] \in \mathbb{R}^{D\times N}$, $x_i \in \mathbb{R}^D$, where N denotes the number of samples and D represents the feature dimension.

In spectral graph theory [10] and manifold learning [11] K-nearest neighbor (Knn) graph represents the locality of data points. This graph is represented as an adjacency matrix $W \in \mathbb{R}^{N\times N}$, whose elements are 0 or 1. Evidently, the complement of W (i.e. $\overline{W}$) represents the far points, where $\overline{W}_{ij} = 1$ states that point j is far from point i (i.e., non-neighboring point). For simplicity in the equations, we use $G = \overline{W}$. Using matrix G, in FNMF, a new regularizer is added to the main objective function of NMF to consider the farness of the points. This regularizer increase the value of the objective function when the far points become close. The new objective function is represented by Equation (4) which should be minimized.

$$\begin{aligned} C &= \|X - UV^T\|_F^2 + \lambda R \\ &= \sum_i^N \sum_j^N (x_{ij} - \sum_{k=1}^K u_{ik}v_{jk})^2 + \lambda R \end{aligned} \tag{4}$$

In this equation R is the regularizer which is defined as an exponential function (see in Equation (5)). Parameter λ controls the contribution of the regularizer in the objective function.

$$\begin{aligned} R &= \exp(-\frac{\beta}{2}\sum_{i=1}^N\sum_{j=1}^N \|v_i - v_j\|^2 G_{ij}) \\ &= \exp(-\beta\sum_{i=1}^N v_i^T v_j D_{ii} + \beta \sum_{i=1}^N\sum_{j=1}^N v_i^T v_j G_{ij}) \\ &= \exp(-\beta Tr(V^TDV) + \beta Tr(V^TGV)) \\ &= \exp(-\beta Tr(V^TLV)) \end{aligned} \tag{5}$$

In Equation (5), $Tr(.)$ denotes the trace of a matrix, D is a diagonal matrix whose elements are sum of the rows of the matrix G, and $L = D - W$ is the laplacian of the matrix G. Therefore, the objective function of FNMF can be written as

$$C = \|X - UV^T\|^2 + \lambda \mathrm{e}^{-\beta Tr(V^TLV)} \tag{6}$$

3.2. Optimizing rules

To minimize the the cost function, Equation (6), we first expand it to

$$\begin{aligned} C &= Tr((X - UV^T)(X - UV^T)^T) + \lambda \mathrm{e}^{-\beta Tr(V^TLV)} \\ &= Tr(XX^T) - 2Tr(XVU^T) + Tr(UV^TVU^T) \\ &+ \lambda \mathrm{e}^{-\beta Tr(V^TLV)}. \end{aligned} \tag{7}$$

We define Lagrange multiplier α_{ik} and β_{jk} for the constraints $u_{ik} \geq 0$ and $v_{jk} \geq 0$, respectively. Therefore, by defining $A = [\alpha_{ik}]$ and $B = [\beta_{jk}]$, the Lagrangian $\mathcal{L}$ is

$$\begin{aligned} \mathcal{L} &= Tr(XX^T) - 2Tr(XVU^T) + Tr(UV^TVU^T) \\ &+ \lambda \mathrm{e}^{-\beta Tr(V^TLV)} + Tr(AU) + Tr(BV). \end{aligned} \tag{8}$$

The partial derivatives of $\mathcal{L}$ with respect to U, V are

$$\frac{\partial \mathcal{L}}{\partial U} = -2XV + 2UV^TV + A \tag{9}$$

$$\frac{\partial \mathcal{L}}{\partial V} = -\, 2X^TU + 2VU^TU \\ -\, 2\beta LV\mathrm{e}^{-\beta Tr(V^TLV)} + B \tag{10}$$

Using the Karush-Kuhn-Tucker (KKT) conditions [12], where $\alpha_{ij}u_{ij} = 0$ and $\beta_{ij}v_{ij} = 0$, the following equations are obtained:

$$-(XV)_{ik}u_{ik} + (UV^TV)_{ik}u_{ik} = 0 \tag{11}$$

$$[-X^TU + VU^TU - \beta LV\mathrm{e}^{-\beta Tr(V^TLV)}]_{jk}v_{jk} = 0 \tag{12}$$

Introducing $L = L^+ - L^-$, where $L_{ij} = (|L_{ij}| + L_{ij})/2$ and $L_{ij} = (|L_{ij}| - L_{ij})/2$, we come up with the following updating rules for U and V

$$u_{ik} \leftarrow u_{ik}\frac{(XV)_{ik}}{(UV^TV)_{ik}} \tag{13}$$

$$v_{jk} \leftarrow v_{jk}\frac{(X^TU + \beta L^+V\mathrm{e}^{-\beta Tr(V^TLV)})_{jk}}{(VU^TU + \beta L^-V\mathrm{e}^{-\beta Tr(V^TLV)})_{jk}} \tag{14}$$

The convergence of updating rules can be proved using an auxiliary function similar to the one used in [13].

4. EXPERIMENTS AND RESULTS

In this section, we evaluate the performance of the proposed FNMF for data clustering on four different real world datasets. The statistics of these datasets can be seen in Table 1. We compare our method with NMF and its related variants such as GNMF and NPNMF.

4.1. Datasets and feature descriptors

In our experiments we used three different datasets, two multimedia datasets, namely AT&T ORL[1] and Caltech-101[2], and an Earth Observation dataset, namely TerraSAR-X[3]. Figure 3 shows some representative samples of these datasets.
AT&T ORL contains 400 images of 40 individuals, where each person has 10 images of size 32 × 32 pixels. **C**altech-101 contains 101 non-equal size categories of images of size roughly 300×200 pixels. In our experiments, 10 largest categories are used containing 3379 images. **T**erraSAR-X contains 3434 TerraSAR-X satellite images of size 160 × 160 which are grouped in 15 classes.

To input the images to the algorithms, they are represented by vectors of their most representative features, so-called feature vectors. For each image in AT&T ORL, all the pixel values are considered as feature values. For Caltech-101 data, rgb-Histograms are extracted locally from the images, then Bag-of-Words (BoW) model of the images are used for data representation. For each image of TerraSAR-X, the pixel values of local windows of size 32 × 32 are used as local feature vectors. Then the images are represented using BoW model.

dataset	size (N)	dimensionality	# classes
AT&T ORL	400	1024	40
Caltech-101	3379	64	10
TerraSAR-X	3434	64	15

Table 1. The statistics of the datasets used in the experiments.

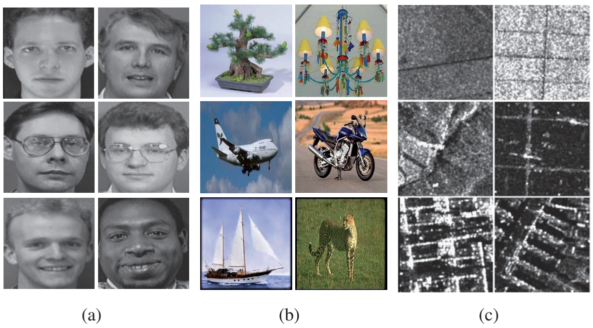


Fig. 2. Representative samples from (a) AT&T ORL, (b) Caltech-101, and (c) TerraSAR-X datasets.

4.2. Evaluation metrics

Using the provided annotations, the performance of the clusterings are evaluated by two widely used metrics, namely Accuracy (AC) and normalized Mutual Information (nMI) [14]. Additionally, in order to compare our results with others we need to use these metrics as well.

Assuming there are N data points, c_i is the provided label of data point P_i and l_i is the label assigned by the clustering algorithm (e.g., Kmeans) in the new space. To compare the prior labels to the newly assigned ones, the $mapfunction$ finds the best mapping using Kuhn-Munkers algorithm [15]. Then AC is computed according to Equation (15) where $\delta(x, y)$ is the delta function which is equals to one if $x = y$.

$$AC = \frac{\sum_{i=1}^{N}\delta(c_i, map(l_i))}{N} \tag{15}$$

Normalized mutual information is another widely used metric to evaluate clusterings by measuring the similarity between two sets of clusters. Given two sets of clusters $C = \{c_1, c_2, ..., c_k\}$ and $C' = \{c'_1, c'_2, ..., c'_{k'}\}$, the mutual information metric is computed by

$$MI(C, C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j).\log\frac{p(c_i, c'_j)}{p(c_i).p(c'_j)}, \tag{16}$$

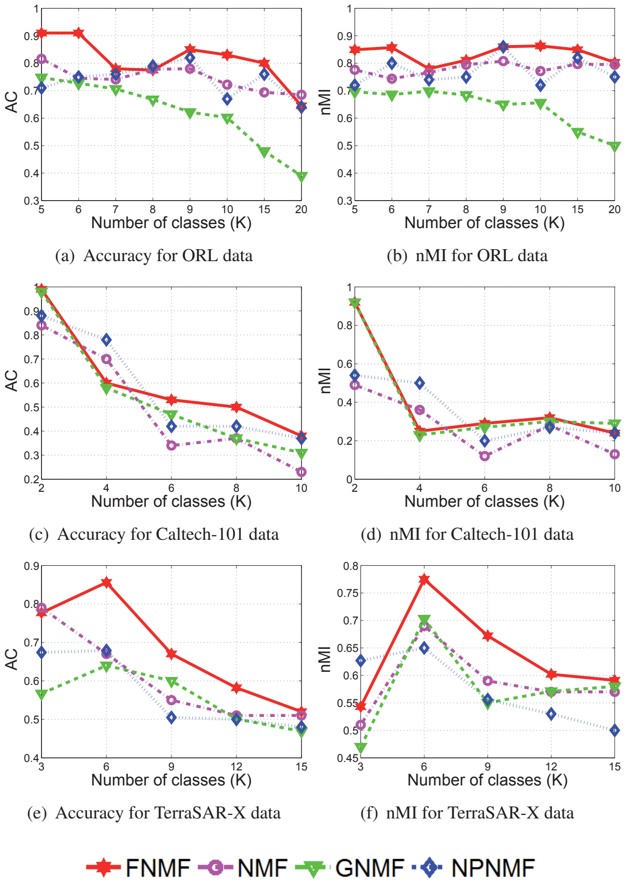
where $p(c_i), p(c'_j)$ represent the probability that an arbitrarily selected data point P_i belongs to the clusters c_i and c'_j, respectively. In this equation $p(c_i, c'_j)$ represents the joint probability that P_i belongs to the both clusters simultaneously. Due to the fact that $MI(C, C')$ take a value between zero and $\max\{H(C), H(C')\}$, it is normalized by dividing by $\max(H(C), H(C'))$ to take a value between 0 and 1. Here, $H(C), H(C')$ represent the entropy of C and C', respectively. Consequently, the normalized mutual information is given by

$$nMI(C, C') = \frac{MI(C, C')}{max\{H(C), H(C')\}}. \tag{17}$$

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

[2] http://www.vision.caltech.edu/Image_Datasets/Caltech101

[3] The images are collected from TerraSAR-X data by Shiyong Cui, Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Germany, shiyong.cui@dlr.de.



(a) Accuracy for ORL data

(b) nMI for ORL data

(c) Accuracy for Caltech-101 data

(d) nMI for Caltech-101 data

(e) Accuracy for TerraSAR-X data

(f) nMI for TerraSAR-X data

Fig. 3. Graph representation of the clustering performance of FNMF, NMF, GNMF, and NPNMF on three different datasets using AC and nMI metrics.

4.3. Results and discussion

To perform the clustering on each dataset, we randomly select K classes and mix their images. After mapping the data samples to the new space using FNMF, NMF, GNMF, and NPNMF, the samples are clustered into K number of clusters. Finally, the resulted clusters are compared to the prior annotation of the samples using AC and nMI metrics. In the experiments, β set to 200 while the λ parameter varies and the best one is taken.

Figure 3 shows the clustering results for three different datasets. As the curves show, the clustering performances using all the data representation methods decrease by increasing the number of classes. This happened because increasing the number of classes increase not only the number of data points but also the complexity of the clusters. For example, for AT&T ORL data, by increasing the number of classes from 5 to 20 the number of data points is increased from 50 to 200. The results also verify that in all the three datasets FNMF outperforms NMF and its related variants, e.g., GNMF and NPNF. FNMF outperforms NMF due to considering the information about the geometry of the data space. FNMF even outperforms the geometry constrained variants of NMF such as GNMF and NPNMF because:

- GNMF and NPNMF consider the locality by constraining the neighboring points which is usually 1 - 2 % of the whole data points, where the rest 98 % of the points are not cared. However, FNMF considers the geometry of the most part of the data space by constraining much more number of points.
- GNMF and NPNMF enforce the neighboring points to stay close to each other during mapping to be assigned the same label by clustering, whereas the non-neighboring points are not cared. This can bring some non-neighboring points closer to the point of interest which can cause mislabeling. However, FNMF enforce the far points to stay away from the point of interest to avoid cluster confusion.
- GNMF and NPNMF constrain a small number of points in comparison to the size of the clusters. This limits these methods to enforce the points from one class to stay in the same class after being mapped to the new space. However, FNMF constrains large number of points from different classes to preserve the structure of data apace during the mapping.

Evaluation metric	FNMF	NMF	GNMF	NPNMF
AT&T ORL	**81.25**	74.51	61.79	73.75
Caltech-101	**60.00**	49.60	54.20	57.40
TerraSAR-X	**68.10**	60.60	55.54	56.76

Table 2. Average clustering accuracy (%) on the three datasets.

Evaluation metric	FNMF	NMF	GNMF	NPNMF
AT&T ORL	**83.40**	78.13	64.01	77.03
Caltech-101	**40.40**	27.60	40.20	35.00
TerraSAR-X	**63.66**	58.60	57.48	57.26

Table 3. Average nMI (%) on the three datasets.

Furthermore, comparing the results for different datasets helps to understand the structure of the data. For example, in AT&T ORL dataset, NMF performs comparably to its geometry respecting variants. This shows that the metric distances of the points have small correlation to the semantic distances (i.e., relation between the points respecting their labels). However, in TerraSAR-X data, the three constrained methods highly outperform NMF, which shows that the metric distances and the semantic distances are highly correlated. To compare the average performance of FNMF to the other methods, the average AC and the average nMI is presented in Table 2 and 3.

5. CONCLUSION

In this paper, we introduce Farness preserving Non-negative Matrix Factorization (FNMF) algorithm to represent the structure of high-dimensional data in a more compact and informative form by mapping the data to a lower-dimensional space. FNMF adds a constraint to the objective function of NMF to enforce the far points (e.g., non-neighbors) to stay far during mapping. Experimental results on real world datasets shows that FNMF outperforms the other variants of NMF which only preserve the locality conditions (i.e., neighboring points) such as GNMF and NPNMF. This confirms that non-local invariance is an important issue in data structure representation.

6. ACKNOWLEDGEMENT

The authors would like to thank Munich Aerospace for the financial support of this research.

7. REFERENCES

[1] H. Liu, Z. Yang, Z. Wu, and X. Li, "A-optimal non-negative projection for image representation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 1592–1599.

[2] I. T. Jolliffe, *Principal Component Analysis*, Springer, second edition, Oct. 2002.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*, Wiley-Interscience, 2000.

[4] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, Norwell, MA, USA, 1991.

[5] 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, 1999.

[6] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 8, pp. 1548–1560, 2011.

[7] H. Liu and Z. Wu, "Non-negative matrix factorization with constraints," in *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[8] Q. Gu and J. Zhou, "Neighborhood preserving nonnegative matrix factorization.," in *BMVC*, 2009, pp. 1–10.

[9] D. Seung and L. Lee, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, pp. 556–562, 2001.

[10] F. R. Chung, *Spectral graph theory*, vol. 92, AMS Bookstore, 1997.

[11] J. J. A. Lee and M. Verleysen, *Nonlinear dimensionality reduction*, Springer, 2007.

[12] S. P. Boyd and L. Vandenberghe, *Convex optimization*, Cambridge university press, 2004.

[13] A. P Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 1–38, 1977.

[14] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 267–273.

[15] M. D. Plummer and L. Lovász, *Matching theory*, Access Online via Elsevier, 1986.