DISCRIMINATIVE FEATURE LEARNING FROM SAR IMAGES

Mohammadreza Babaee[2,3], *Stefanos Tsoukalas*[2], *Gerhard Rigoll*[2], *Mihai Datcu*[1,3]

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

ABSTRACT

Clustering of Earth Observation (EO) images has gained a high amount of attention in remote sensing and data mining. Here, each image is represented by a high-dimensional feature vector which could be computed as the results of coding algorithms of extracted local descriptors or raw pixel values. In this work, we propose to learn the features using discriminative Nonnegative Matrix factorization (DNMF) to represent each image. Here, we use the label of some images to produce new representation of images with more discriminative property. To validate our algorithm, we apply the proposed algorithm on a dataset of Synthetic Aperture Radar (SAR) and compare the results with the results of state-of-the-art techniques for image representation. The results confirm the capability of the proposed method in learning discriminative features leading to higher accuracy in clustering.

Index Terms— Feature Learning, Nonnegative Matrix Factorization, Clustering

1. INTRODUCTION

Data representation and features learning is considered as a challenging problem in data mining and pattern recognition. The problem is how to represent a data term that in one side it has discriminative property and on the other side it is a compact lower dimensional representation. Matrix factorization techniques are intensively employed in data analysis due to their capability in capturing the most useful representation of the data. Perhaps, the most well-known matrix factorization techniques are Principal Component Analysis (PCA), Singular Value Decomposition (SVD), and Nonnegative Matrix Factorization [1]. The goal of these methods is to provide a compact low dimensional representation of original data for further processes such as learning and visualization. NMF is an unsupervised learning algorithm, that decomposes a nonnegative data matrix into two (or three) nonnegative factors, one of which is considered as a new representation of original data [1]. This factorization leads to a parts-based representation of data, which is widely used in different applications such as face recognition [1] and clustering [2]. Since the invention of NMF, many variants of this algorithms have been proposed to get a customized representation of data. For example, Graph regularized NMF (GNMF) [3] preserves the locality property of data by utilizing the Laplacian of neighborhood graph in its regularization term. In Constrained NMF (CNMF) [4] those points which have the same label information are forced to have the same representation.

In this paper, we propose a label constrained NMF, namely Discriminative Nonnegative Matrix Factorization, which utilizes the label information of a fraction of data in a regularization term. The key idea of our approach is to use the raw data in order to create a discriminative representation of image for clustering. This regularizer increases the discriminative property of data points in the new representation space, controlled by a parameter.

The rest of paper is organized as follows: In Section 3 the details of our proposed approach is provided. Section 4 some preliminary experimental results applied on a dataset of SAR images are presented.

2. RELATED WORK

Feature learning or image representation has been intensively considered in data mining and machine learning [5]. A variety of techniques such as matrix factorization [6], dictionary learning [7], neural networks [8] are used to represent the data. Additionally, feature learning methods could be categorized in supervised or unsupervised methods. K-means clustering is considered as an unsupervised feature learning by producing K cluster centers and using them to create feature of the size K to represent the data [9]. It has gained a high amount of attention in object recognition and image classification. However, it is an unsupervised method and the leaned features dont have necessarily discriminative property. In order to learn discriminative features, some works have utilized the label information of a fraction of data points in the body of learning process. For example, in [10] the label information are used to form a regularized coupled to the main objective function of K-SVD. Nonnegative matrix factorization (NMF) has been successfully applied on data to gain a part based representation. However, the main NMF framework does not consider any semantic information in the new representation. In this work, we couple a regularizer to the main function of NMF in order to increase the discriminative property of features.

3. APPROACH

In an optimal NMF factorization we would expect each of the dataset classes to be placed in a clearly separated cluster in the resulting vector V. To enforce this property, based on the available label information, we introduce the matrix $Q \in \mathbb{R}^{S\times N}$ as follows:

$$Q_{i,j} = \begin{cases} 1 & \text{if sample } j \text{ is labeled and belongs to class } i \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

For example, consider the case of $N = 8$ samples, out of which $N_l = 5$ are labeled, with the sample categories $c_1 = 1$, $c_2 = 2$, $c_3 = 1$, $c_4 = 3$, $c_5 = 2$. In this case the matrix Q would take the form:

$$Q = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{2}$$

Based on the introduced matrix we add the following term to the Frobenius-NMF objective:

$$O_L = \alpha \left\|Q - AV_l^{\mathrm{T}}\right\|^2, \tag{3}$$

with $V_l = [\boldsymbol{v}_1, ..., \boldsymbol{v}_{N_l}, \boldsymbol{0}, ..., \boldsymbol{0}]^{\mathrm{T}} \in \mathbb{R}^{N\times K}$ and the matrix $A \in \mathbb{R}^{S\times K}$, which linearly transforms and scales the vectors in the new representation, in order to obtain the best fit for the matrix Q. The matrix A is allowed to take negative values and is computed as part of the NMF minimization. We arrive at the following minimization problem:

$$\begin{aligned} &\min O = \left\|X - UV^{\mathrm{T}}\right\|^2 + \alpha \left\|Q - AV_l^{\mathrm{T}}\right\|^2 \\ &\text{s.t.} \quad U = [u_{ik}] \geq 0 \\ &\qquad\;\; V = [v_{jk}] \geq 0. \end{aligned} \tag{4}$$

3.1. Update rules

For the derivation of the update rules we expand the objective to

$$\begin{aligned} O =& \mathrm{Tr}\left(XX^{\mathrm{T}}\right) - 2\mathrm{Tr}\left(XVU^{\mathrm{T}}\right) + \mathrm{Tr}\left(UV^TVU^T\right) \\ &+ \alpha\mathrm{Tr}\left(QQ^{\mathrm{T}}\right) - \alpha 2\mathrm{Tr}\left(QV_lA^{\mathrm{T}}\right) + \alpha\mathrm{Tr}\left(AV_l^TV_lA^T\right) \end{aligned} \tag{5}$$

and introduce Lagrange multipliers $\Phi = [\phi_{ik}]$, $\Psi = [\psi_{jk}]$ for the constraints $[u_{ik}] \geq 0$, $[v_{jk}] \geq 0$ respectively. Adding the Lagrange multipliers and ignoring the constant terms leads to the Lagrangian:

$$\begin{aligned} \mathcal{L} =& -2\mathrm{Tr}\left(XVU^{\mathrm{T}}\right) + \mathrm{Tr}\left(UV^TVU^T\right) + \mathrm{Tr}\left(\Phi U\right) + \mathrm{Tr}\left(\Psi V\right) \\ &- \alpha 2\mathrm{Tr}\left(QV_lA^{\mathrm{T}}\right) + \alpha\mathrm{Tr}\left(AV_l^TV_lA^T\right). \end{aligned}$$

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

$$\frac{\partial\mathcal{L}}{\partial U} = -2XV + 2UV^{\mathrm{T}}V + \Phi \tag{6}$$

$$\frac{\partial\mathcal{L}}{\partial V} = -2X^TU + 2VU^TU - 2\alpha Q^TA + 2\alpha V_lA^TA + \Psi \tag{7}$$

$$\frac{\partial\mathcal{L}}{\partial A} = -2QV_l + 2AV_l^{\mathrm{T}}V_l \tag{8}$$

For the derivation of the update rules for U and V we solve the equations $\frac{\partial\mathcal{L}}{\partial U} = 0$ and $\frac{\partial\mathcal{L}}{\partial V} = 0$ in terms of Φ and Ψ respectively and apply the KKT-conditions $\phi_{ik}u_{ik} = 0$, $\psi_{jk}v_{jk} = 0$ [11]. For A, since we have no Lagrange multiplier, the update rules can be derived directly by setting $\frac{\partial\mathcal{L}}{\partial A} = 0$ and solving for A. We arrive at the following equations:

$$u_{ik} \leftarrow u_{ik}\frac{[XV]_{ik}}{[UV^{\mathrm{T}}V]_{ik}} \tag{9}$$

$$v_{jk} \leftarrow v_{jk}\frac{[X^{\mathrm{T}}U + \alpha(V_lA^{\mathrm{T}}A)^- + \alpha(Q^{\mathrm{T}}A)^+]_{jk}}{[VU^{\mathrm{T}}U + \alpha(V_lA^{\mathrm{T}}A)^+ + \alpha(Q^{\mathrm{T}}A)^-]_{jk}} \tag{10}$$

$$A \leftarrow QV_l(V_l^{\mathrm{T}}V_l)^{-1} \tag{11}$$

where for a matrix M we define M^+, M^- as $M^+ = (|M| + M)/2$ and $M^- = (|M| - M)/2$. As expected, the update rule for U remains the same as in the original algorithm [12], since the newly introduced terms depend only on the variables V and A. We have the following theorem, that guarantees convergence to a local minimum:

4. EXPERIMENTS

4.1. Dataset

To validate our approach on EO images, we use a dataset of SAR images. The used SAR dataset contains 3434 TerraSAR-X satellite images of size 160×160 which are grouped in 15 classes. We use two different types of representation as original data. In the first representation, we use the pixel values (raw data) and in the second representation, we use the Bag-of-Word model of SIFT features.

4.2. Setup

In order to quantify the accuracy of the proposed algorithm, we label a fraction of the data points and use DNMF to get a discriminative representation of all data. Then we apply Kmeans clustering algorithms on original and new representation and compare the results by computing the Accuracy (AC) and Mutual Information (MI), as evaluation metrics, of clustering result. We run the experiment with different number of classes and each time we set the dimension of new representation equal to the number of classes. We change the number of classes from 2 to 10. Since Kmeans algorithm is an heuristic algorithm, we run it 10 times and compute the average of results. Additionally, we apply PCA, NMF, and GNMF to get other representations of the data and apply the Kmeans on these representations. The results of experiments are depicted in Fig. 1. In Fig. 1.a and Fig. 1.b sample images of dataset and the corresponding learned based using DNMF are presented. The results of clustering applied on new representation based on raw data is given in Fig.1a-b. Fig.1.e-f show the result of clustering on new representation based on SIFT features.

4.3. Evaluation Metrics

We use two metrics to evaluate the performance of the compared algorithms, namely accuracy (AC) and normalized mutual information (nMI) [13]. The accuracy computes the percentage of correctly predicted groups, compared to the true labels. The normalized mutual information is based on the mutual information, which is used in clustering applications, to measure the similarity of two clusters. Given two sets of clusters $C = \{c_1, ..., c_k\}$ and $\acute{C} = \{\acute{c}_1, ..., \acute{c}_{\acute{k}}\}$, the mutual information metric is computed by

$$MI(C, \acute{C}) = \sum_{c_i \in C, \acute{c}_j \in \acute{C}} p(c_i, \acute{c}_j) \log \frac{p(c_i, \acute{c}_j)}{p(c_i)p(\acute{c}_j)}, \qquad (12)$$

where $p(c_i), p(\acute{c}_j)$ represent the probability, that an arbitrarily selected data point belongs to the clusters C or $\acute{C}_j$, respectively and $p(c_i, \acute{c}_j)$ represents the joint probability, that the point belongs to both clusters simultaneously. As the similarity of the two clusters increases, the mutual information $MI(C, \acute{C})$ takes increasing values between 0 and $\max\left\{H(C), H(\acute{C})\right\}$. There, $H(C), H(\acute{C})$ represent the entropy of the clusters $C, \acute{C}$ respectively. Dividing the the mutual information by $\max\left\{H(C), H(\acute{C})\right\}$ leads to the normalized mutual information, which takes values between 0 and 1:

$$nMI(C, \acute{C}) = \frac{MI(C, \acute{C})}{\max\left\{H(C), H(\acute{C})\right\}}. \qquad (13)$$

4.4. Discussion

As it can be inferred from the results, the proposed method outperforms the others in terms of clustering accuracy. It is clear that the learned representations using DNMF in different dimensions have more discriminative property than the others. However, the accuracy decreases by increasing the dimension and also the number of clusters. Additionally, the learned representation based on raw data outperforms the learned representation based on extracted SIFT features. This is quite useful since we can skip the feature extraction process and use the raw data directly into the algorithm. All

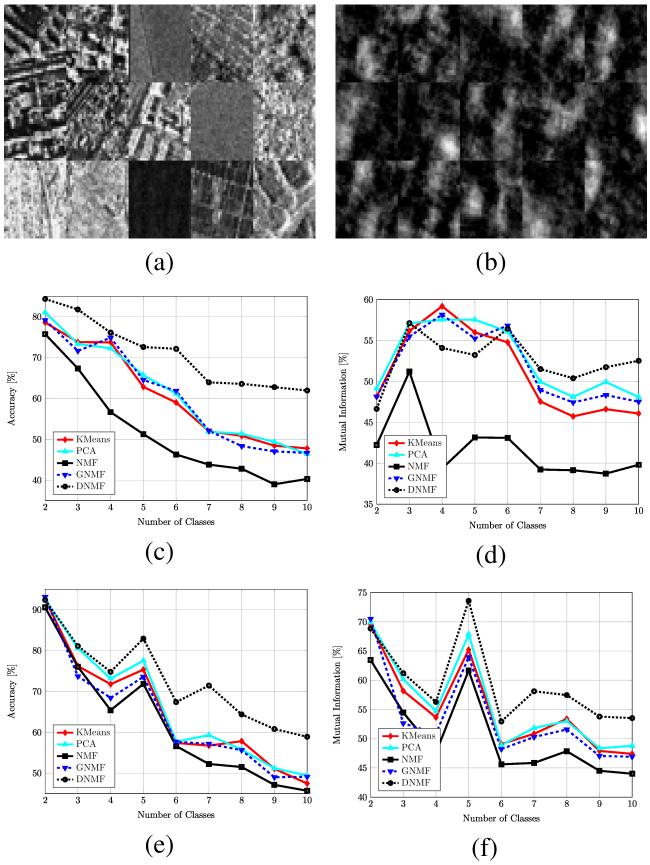

Fig. 1. (a) Sample images of dataset; (b) the corresponding learned bases using DNMF; (c)-(d) Accuracy and MI of clustering on DNMF using raw data as original representation. (e)-(f) Accuracy and MI of DNMF result based on SIFT data representation.

plots confirm that the NMF algorithm has the worst performance compare to other, which means that introducing a proper regularizer to the NMF can increase its performance significantly.

5. CONCLUSIONS

We presented a variant of NMF algorithm, the so-called discriminative NMF (DNMF), to learn discriminative features from SAR images in order to increase the performance of clustering. We provided the main objective function of DNMF along with its updating rules. Experimental results on both raw SAR data and extracted SIFT features confirm the performance of algorithm in comparison to others. Additionally, we found that running DNMF algorithm on raw data results in more discriminative features than on extracted SIFT features. This readily implied the power of algorithm even on raw data. As future work, it could be suggested to run the DNMF on other features such as Gabor or Weber.

6. REFERENCES

[1] 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, October 1999.

[2] Jialu Liu, Chi Wang, Jing Gao, and Jiawei Han, "Multi-view clustering via joint nonnegative matrix factorization," in *Proc. of SDM*. SIAM, 2013, vol. 13, pp. 252–260.

[3] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2011.

[4] Haifeng Liu, Zhaohui Wu, Xuelong Li, Deng Cai, and Thomas S Huang, "Constrained nonnegative matrix factorization for image representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 7, pp. 1299–1311, 2012.

[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1798–1828, 2013.

[6] Nathan Srebro, Jason Rennie, and Tommi S Jaakkola, "Maximum-margin matrix factorization," in *Advances in neural information processing systems*, 2004, pp. 1329–1336.

[7] Michal Aharon, Michael Elad, and Alfred Bruckstein, "-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, 2006.

[8] Alexey Dosovitskiy, Jost Tobias Springenberg, Martin Riedmiller, and Thomas Brox, "Discriminative unsupervised feature learning with convolutional neural networks," *arXiv preprint arXiv:1406.6909*, 2014.

[9] Adam Coates and Andrew Y Ng, "Learning feature representations with k-means," in *Neural Networks: Tricks of the Trade*, pp. 561–580. Springer, 2012.

[10] Zhuolin Jiang, Zhe Lin, and Larry S Davis, "Label consistent k-svd: learning a discriminative dictionary for recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 11, pp. 2651–2664, 2013.

[11] Stephen Boyd and Lieven Vandenberghe, *Convex Optimization*, Cambridge University Press, New York, NY, USA, 2004.

[12] Daniel D. Lee and H. Sebastian Seung, "Algorithms for non-negative matrix factorization," in *In NIPS*. 2001, pp. 556–562, MIT Press.

[13] 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.