Soft-Constrained Nonnegative Matrix Factorization via Normalization

Long Lan, Student Member, IEEE Naiyang Guan, Member, IEEE Xiang Zhang, Student Member, IEEE Dacheng Tao, Senior Member, IEEE and Zhigang Luo Member, IEEE

Abstract-Semi-supervised clustering aims at boosting the clustering performance on unlabeled samples by using labels from a few labeled samples. Constrained NMF (CNMF) is one of the most significant semi-supervised clustering methods, and it factorizes the whole dataset by NMF and constrains those labeled samples from the same class to have identical encodings. In this paper, we propose a novel soft-constrained NMF (SCNMF) method by softening the hard constraint in CNMF. Particularly, SCNMF factorizes the whole dataset into two lower-dimensional factor matrices by using multiplicative update rule (MUR). To utilize the labels of labeled samples, SCNMF iteratively normalizes both factor matrices after updating them with MURs to make encodings of labeled samples close to their label vectors. It is therefore reasonable to believe that encodings of unlabeled samples are also close to their corresponding label vectors. Such strategy significantly boosts the clustering performance even when the labeled samples are rather limited, e.g., each class owns only a single labeled sample. Since the normalization procedure never increases the computational complexity of MUR, SCNMF is quite efficient and effective in practices. Experimental results on face image datasets illustrate both efficiency and effectiveness of SCNMF compared with both NMF and CNMF.

I. INTRODUCTION

C LUSTERING is an unsupervised learning task, which has been widely used to find natural groups and clusters of unlabeled samples based on similarity, proximity or continuation [1].However, in many applications, traditional clustering methods perform unsatisfactory on some datasets because they never utilize any prior knowledge.It is natural to use accessorial information such as class labels to improve the performance. In practice, it is extremely expensive to obtain labels of the whole dataset, but we usually can get labels for partial samples [2]. Recently, semi-supervised clustering [3][4][5] has attracted much attentions to solve this problem.

Semi-supervised clustering employs class labels to favorably boost clustering performance on unlabeled samples by transferring labels of partial samples to enhance their discriminative ability. Seeded K-Means [6] and constrained K-means [6] are two well-known semi-supervised clustering

Long Lan, Naiyang Guan, Xiang Zhang and Zhigang Luo are with the Science and Technology on Parallel and Distributed Processing Laboratory, School of Computer Science, National University of Defense Technology, Changsha, Hunan, P. R. China, 410073. Dacheng Tao is with the Centre for Quantum Computation & Intelligent Systems, and the Faculty of Engineering and Information Technology, University of Technology, Sydney, 235 Jones Street, Ultimo, NSW 2007, Australia.

This work was partially supported by Plan for Innovative Graduate Student at National University of Defense Technology and Australian Research Council Projects (FT-130101457 and DP-140102164) algorithms, both of which treat the labeled samples as seed sets. The semi-supervised algorithms iterate two steps: 1) the seed sets are used to initialize the centroids for each class, and the label of each sample is assigned according to its distance to centroids, and 2) the centroids are reestimated in virtue of current labels, until convergence. Seeded K-means re-estimates labels of all samples in each iteration round, while constrained K-Means keeps the labels of seeds unchanged throughout.Constrained normalized cuts (CNC) [7] incorporates the prior information into graph partition clustering algorithm, which is designed to deal with constraint pairs by specifying two must-link samples in clustering. CNC enforces the constrained samples to share identical cluster indicator, and thus the reminding samples take the advantage of the constrained samples.

Non-negative matrix factorization (NMF) [8], as a useful dimension reduction method, has attracted a lot of attentions in recent years. It has been practically proven that NMF is an effective clustering method [9][10][16]. NMF decomposed an original matrix into two lower-rank non-negative matrices, namely, basis matrix and coefficient matrix, where the basis matrix can be considered as centroids of clusters, and the coefficient matrix indicates membership of clusters for each sample. Recently, numerous works have addressed the semisupervised non-negative matrix factorization problem [11]-[15]. Some studies have shown that the classical clustering methods such as K-means and spectral clustering are theoretically equivalent to NMF [17]. Liu et al. [12] proposed a Constrained NMF (CNMF), which takes labels of partial samples as a hard constraint. In particular, CNMF decomposes the coefficient matrix into the product of two matrices including the label-constrained matrix and the indicator matrix. It forces samples with identical class label to have consistent coordinate in the reduced dimensional space, and thus samples show more discriminative. However, CNMF fails in case that each class has only one labeled sample, since the constraint matrix degenerates to an identity matrix and takes no effect. NMF- α [15] makes a good combination of NMF and SVM, which utilizes limited labeled samples to achieve the support vectors of large-margin classifiers. NMF- α assumes that the achieved support vectors show discriminative in the NMF subspace. However, limited labeled samples hardly form effective classifier for the whole dataset, and thus NMF- α fails in many real word applications.

In this paper, we propose Soft-Constrained NMF (SCNM-F) to solve the aforementioned problems by incorporating the labels of partial samples to boost the clustering performance on the unlabeled samples on the fly of multiplicative update rule (MUR) for the standard NMF. In particular, SCNMF normalizes the encoding of labeled samples to approximate their label vectors, and thus softens the label constraint. To keep the convergence of each step of MUR, SCNMF revises the centroids accordingly, and the revised centroids reshape the encoding of unlabeled samples in the next step.It is therefore reasonable to believe that the encoding of unlabeled samples also approximate their corresponding label vectors. The advantages of SCNMF lie in two-folds: 1) SCNMF works well when each class owns only a few labeled samples, e.g., there is single labeled sample in each class, because the approximate encodings distinctively differentiate samples in different classes, and 2) SCNMF is efficient because the normalization step does not increase the time complexity of MUR. The experimental results on classical face image datasets show that SCNMF outperforms the representative semi-supervised NMF clustering algorithms.

The remainder of this paper is organized as follows: Section 2 briefly reviews NMF and CNMF, and Section 3 presents SCNMF. Section 4 verifies the effectiveness of SCNMF and Section 5 concludes this paper.

II. RELATED WORKS

A. Non-negative Matrix Factorization

Non-negative matrix factorization (NMF) [8] approximates the data matrix, i.e., $V = [v_1, v_2, \cdots v_n] \in R^{m \times n}_+$ by the products of two lower-rank non-negative matrices, i.e., $W = [w_1, w_2, \cdots w_c] \in R^{m \times c}_+$ and $H = [h_1, h_2, \cdots h_n] \in R^{c \times n}_+$. where $c \ (c \ll \min(m, n))$ denotes the number of reduced dimensionalities. In mathematics, the objective function of NMF is

$$\min_{W,H} f(W,H) = ||V - WH||_F^2 \ s.t. \ W,H \ge 0.$$
(1)

NMF has been widely used in clustering tasks. If c equals to the number of clusters, W can be treated as cluster centroids and columns of H indicates the cluster membership of samples. Moreover, NMF has also been theoretically proven to be equivalent to classical clustering methods such as Kmeans and spectral clustering algorithms [17] and has been practically proven to be effective on document clustering [17] and image clustering [12]. The most popular NMF solver is multiplicative update rule (MUR):

$$W \leftarrow W \otimes \frac{VH^T}{WHH^T}, H \leftarrow H \otimes \frac{W^T V}{W^T WH}$$
(2)

where \otimes denotes the element-wise multiplication, which alternatively updates W and H until convergence. Other NMF solvers [18][19][21] have been proposed, but they all seek only local minimum because the objective function of NMF (1) is non-convex.

On the other hand, NMF is non-unique because $f(W, H) = f(WD^{-1}, DH)$ for any diagonal matrix D whose diagonal elements are positive. In this paper, we significantly boost the performance by determining D based on the labels of partial samples.

B. Constrained NMF

Constrained NMF (CNMF) [12] is a well-known semisupervised NMF algorithm, which decomposes the data matrix into the product of three parts and embeds the labels of partial samples in a constraint matrix. The objective function of CNMF is

$$\min_{W,H} ||V - WHA||_F^2 \ s.t. \ W, H \ge 0, \tag{3}$$

where A is a constraint matrix and constraints the coefficients of samples from the same class to be identical. However, CNMF fails if each class has only a single labeled sample because degenerates to an identity matrix in that case.

III. SOFT CONSTRAINED NMF

Soft-Constrained NMF (SCNMF) effectively utilizes labels of partial samples to constrain their indicators, and thus significantly improves the discriminative ability of unlabeled samples. Given an sample matrix $V = [v_1, v_2, \dots, v_n] \in \mathbb{R}^{m \times n}_+$ which consists of n samples, without loss of generality, we assume that the first l samples are labeled, and the rest u samples are unlabeled, where u = n - l. Supposing that there are c classes and each class has at least one labeled sample, we construct label matrix $Y_l \in \mathbb{R}^{c \times l}$ for labeled samples as follows:

$$(Y_l)_{ij} = \begin{cases} 1 & if \ v_j \ labeled \ with \ i \\ 0 & otherwise \end{cases}$$
(4)

where each column corresponds to a label vector. We decompose V into the product of W and H, i.e., $V \approx$ WH, like NMF. Different from NMF, we expect that the encodings of labeled samples to be as close as their label vectors. To this end, we introduce a diagonal matrix D with positive diagonal elements to normalize the decomposition, i.e., $WH = WD^{-1}DH$, and constrain that Y_l equals to the encodings of labeled samples H_l after normalization, i.e., $DH_l = Y_l$. The objective function of SCNMF is:

$$\min_{W \ge 0, H \ge 0, D} ||V - WD^{-1}DH||_F^2, \ s.t., \ D_{ii} > 0, DH_l = Y_l.$$
(5)

Note that $D^{-1}D = I$, the introduced matrix D does not change the objective value, and thus the traditional multiplicative update rule (MUR) can be adopted to solve SCNMF without any modifications, i.e.,

$$W = W \otimes \frac{VH^T}{WHH^T} \tag{6}$$

and

$$H = H \otimes \frac{W^T V}{W^T W H} \tag{7}$$

To meet the equality constraint in (5), we apply a normalization step after MUR and determine the matrix for normalization as follows:

$$\min_{D_{ii}>0} ||DH_l - Y_l||_F^2 \tag{8}$$

Let $\{d_1, \ldots, d_c\}$ denote the diagonal elements of D, by solving (8), we have

$$d_i = \left(\frac{Y_l H_l^T}{H_l H_l^T}\right)_{ii} \tag{9}$$

After obtaining the optimal D, we normalize both W and H by

$$W = W D^{-1} \tag{10}$$

and

$$H = DH \tag{11}$$

respectively.

On the fly of MUR, i.e., (6) and (7), followed by normalization, i.e., (9), (10), and (11), H_l gets closer and closer to Y_l . Since both H_u and H_l are normalized through the identical D, the encodings of unlabeled samples H_u approximates their encodings. Since the matrix D is calculated by the analytic formulation (9) and the normalization operators (10) and (11) never increase the computational overheads, SCNMF is as efficient as the standard NMF. However, SCNMF benefits much from the normalization step based on the labels of partial samples, and thus significantly boosts the clustering performance on the unlabeled samples.

IV. EXPERIMENTS

In this section, we conduct several experiments to verify SCNMF on four face image datasets including ORL [22], Yale [24], FERET [23] and UMIST [25]. The ORL dataset consists of 40 individuals and 10 images for each individual. The Yale dataset has 11 individuals and 15 images for each individual. The FERET dataset has 700 images and 7 images for each individual. And the UMIST dataset has 575 images taken from 20 individuals. We evaluate the effectiveness of SCNMF in image clustering at four aspects: 1) more representative cluster center, 2) better cluster performance, 3) more robust with variant label size, and 4) less running time. We use both accuracy (AC) and normalized mutual information (NMI) to compare different algorithms. The concrete definition of both metrics can be found in [10][26][27]. For the fairness of comparison, all algorithms are evaluated on the same labeled samples and unlabeled samples.

A. Illustration of Cluster Centers

We firstly investigate the cluster centroids, i.e., the basis matrix, of SCNMF compared with those obtained from both NMF and CNMF on ORL and FERET datasets. In this experiment, we randomly select two labeled samples for each class. Figure 1 and Figure 2 visualizes the basis images learned from ORL and FERET datasets, respectively.

From both Figure 1 and Figure 2, we can observe that the bases of SCNMF are more representative than those learned by NMF and CNMF. It shows that the reconstructed bases of SCNMF accurately describe the cluster centroids. However, the hard constraint of CNMF makes bases blurred. It is difficult to appoint each base to any specified cluster. The bases learned by SCNMF show a great superiority in



Fig. 1. The basis of SCNMF(3rd row), CNMF (4th row), and NMF (5th row) on the ORL dataset, and the top two rows display the two constrained labeled samples from each class



Fig. 2. The basis of SCNMF (3rd row), CNMF (4th row), and NMF (5th row) on the FERET dataset, and the top two rows display the two constrained labeled samples from each class

clustering. We will validate the effectiveness of these bases in the next section.

B. Clustering on Face Image Datasets

In this section, we evaluate the clustering performance of SCNMF on the ORL, FERET, YALE, and UMIST datasets by comparing with CNMF [12] and NMF- α [15]. In this experiment, according to [12], we vary the number of classes from 2 to 10 and randomly select two labeled samples for each class. Such trial was repeated ten times and averaged AC and averaged NMI were reported. Figures 3 to 6 show the clustering performance of SCNMF on the ORL, FERET, YALE, and UMIST datasets, respectively.



Fig. 3. The averaged AC (a) and averaged NMI (b) of SCNMF, NMF, NMF- α and CNMF versus the number of classes on the ORL dataset

Figures 4 to 6 show that SCNMF performs much better than NMF, CNMF and NMF- α . This observation confirms



Fig. 4. The averaged AC (a) and averaged NMI (b) of SCNMF, NMF, NMF- α and CNMF versus the number of classes on the FERET dataset



Fig. 5. The averaged AC (a) and averaged NMI (b) of SCNMF, NMF, NMF- α and CNMF versus the number of classes on the YALE dataset



Fig. 6. The averaged AC (a) and averaged NMI (b) of SCNMF, NMF, NMF- α and CNMF versus the number of classes on the UMIST dataset

that the label-constraint based normalization in SCNMF can effectively transduce the labels of few labeled samples to unlabeled samples, and thus significantly enhances the discriminative of unlabeled samples.

C. Performances versus Label Size

In this section, we evaluate the performance of SCNMF by varying the number of labeled samples in each class (denoted by label size). In this experiment, we selected a subset comprised of images taken 10 individuals of the ORL, FERET, YALE and UMIST datasets. Figures 7 to 10 depict the experimental results when the label size varies from 1 to 5. Various numbers of labeled samples were randomly selected from images of each individual and such trail was repeated ten times with the averaged AC and averaged NMI are reported.

From Figures 7 to 10, we can see that the curves of SCNMF rise rapidly as increasing of the label size. In contrast, both AC and NMI of CNMF and NMF- α increase



Fig. 7. AC and NMI of SCNMF, NMF- α and CNMF versus label size on ORL



Fig. 8. AC and NMI of SCNMF, NMF- α and CNMF versus label size on FERET



Fig. 9. AC and NMI of SCNMF, NMF- α and CNMF versus label size on YALE



Fig. 10. AC and NMI of SCNMF, NMF- α and CNMF versus label size on UMIST

rather gently as increasing the label size. It is reasonable to believe that SCNMF can more sufficiently take the effect of the provided labels than CNMF and NMF- α .

D. Complexity Analysis

SCNMF does not introduce auxiliary matrix, and thus never increase the time complexity of NMF, i.e., $\phi(mnc)$ for each iteration round. Note that both CNMF and NMF- α terribly increase the complexity of NMF due to the influence of the constraint matrix. In particular, the time complexity of CNMF is $\phi(m(n-l+c)(2n+2c))$, where *l* is the number of labeled samples. Since $l \ll n$ and $c \ll n$, SCNMF has much less time complexity than CNMF. The time complexity of NMF- α is $\phi(mn(c+n))$ which is much greater than SCNMF. We compare performance versus CPU time to validate the efficiency of SCNMF.



Fig. 11. AC, NMI and Time (second) versus iterations on ORL



Fig. 12. AC, NMI and Time (second) versus iterations on FERET



Fig. 13. AC, NMI and Time (second) versus iterations on YALE

Figures 11 to 14 show that SCNMF takes less CPU time than CNMF and NMF- α to achieve better clustering performance. By comparing the curves of SCNMF and NMF, we can see that the normalization step significantly boosts the clustering performance by incorporating the labels of partial samples without bringing in any extra computational overheads. Due to its simplicity and effectiveness, SCNMF provides a framework of boosting for semi-supervised NMF



Fig. 14. AC, NMI and Time (second) versus iterations on UMIST

algorithms. Hence, SCNMF can be extended to other NMF models such as manifold regularized NMF [26] and online NMF [20].

V. CONCLUSIONS

This paper introduces a soft-constrained non-negative matrix factorization (SCNMF) algorithm which boosts the clustering performance on unlabeled samples by incorporating the discriminative information of a few labeled samples. Since SCNMF smartly embeds the discriminative information of labeled samples in the normalization step of the multiplicative update rule in standard NMF without bringing in any extra overheads, it is quite efficient and effective, and provides a framework for boosting NMF algorithms. The experiments on real-world datasets confirm both efficiency and effectiveness of SCNMF.

REFERENCES

- J Shi. and J Malik., "Normalized Cuts and Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp.888-905, 2000.
- [2] D Cai., X He., and J Han., "Semi-supervised Discriminant Analysis," IEEE 11th International Conferenceon Computer Vision, pp. 1-7, 2007.
- [3] X Zhu., Z Ghahramani., and J Lafferty., "Semi-supervised Learning Using Gaussian Fields and Harmonic Functions," *International Conference on Machine Learning*, vol. 3, pp. 912-919, 2003.
 [4] D Zhang., Z Zhou ., and S Chen., "Semi-supervised Dimensionality
- [4] D Zhang., Z Zhou ., and S Chen., "Semi-supervised Dimensionality Reduction," *SIAM Conference on Data Mining*, vol. 3, pp. 912-919, 2003.
- [5] B Kulis., S Basu., I Dhillon., and Mooney R J. "Semi-supervised Graph Clustering: A Kernel Approach," *International Conference on Machine Learning*, vol. 74, no. 1, pp. 1-22, 2009.
- [6] S Basu., A Banerjee., and R J Mooney., "Semi-supervised Clustering bySeeding," *International Conference on Machine Learning*, vol. 2, pp. 27-34, 2002.
- [7] X Ji. and W Xu., "Document Clustering with Prior Knowledge," Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 405-412, 2006.
- [8] D D Lee. and H S Seung., "Learning the Parts of Objects by Nonnegative Matrix Factorization," *Nature*, vol. 401, no. 6755, pp. 788-791, 1999.
- [9] F Shahnaz., M W Berry., V P Pauca., and R J Plemmons. "Document Clustering using Nonnegative Matrix Factorization," *Information Processing and Management*, vol. 42, no. 2, pp. 373-386, 2006.
- [10] W Xu., X Liu., and Y Gong., "Document Clustering based on Non-negative Matrix Factorization," Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 267-273, 2003.
- [11] Y Chen., M Rege., M Dong., and J Hua., "Non-negative Matrix Factorization for Semi-supervised Data Clustering," *Knowledge and Information Systems*, vol. 17, no. 3, pp. 355-379,2008.
- [12] H Liu., Z Wu., X Li., D Cai.,and T S Huang "Constrained Nonnegative Matrix Factorizationfor Image Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 7, pp. 1299-1311, 2012.

- [13] N Guan., D Tao., Z Luo., and B Yuan., "Non-negative Patch Alignment Framework," *IEEE Transactions on Neural Networks*, vol. 22, no. 8, pp. 1218-1290, 2011.
- [14] X Zhang., N Guan., L Lan., D Tao., and Z Luo., "Box-constrained projective Nonnegative matrix factorization via Augmented Lagrangian Method,"*International Joint Conference on Neural Networks*, Preprint, 2014.
- [15] Y Cho. and L K Saul., "Nonnegative Matrix Factorization for Semisupervised Dimensionality Reduction," arXiv:1112.3714, 2011.
- [16] V P Pauca., F Shahnaz., M W Berry., R J Plemmons., "Text Mining usingNonnegative Matrix Factorizations," *Proceedings of the Fourth SIAMInternational Conference on Data Mining*, pp. 22-24,2004.
- [17] C Ding, X He, and H D Simon., "On the Equivalence of Nonnegative Matrix Factorization and Spectral Clustering," *SIAM Conference on Data Mining*, vol. 5, pp. 606-610, 2005.
- [18] D D Lee. and H S Seung., "Algorithms for Non-negative Matrix Factorization," Advances in Neural Information Processing Systems, vol. 13, pp. 556-562, 2001.
- [19] C J Lin. "Projected Gradient Methods for Nonnegative Matrix Factorization," *Neural Computation*, vol. 19, no. 10, pp. 2756-2779, 2007.
- [20] N Guan, D Tao, Z Luo., and B Yuan, "Online Nonnegative Matrix Factorization with Robust Stochastic Approximation," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 7, pp. 1087-1099, 2012.
- [21] N Guan., D Tao., Z Luo., and B Yuan., "NeNMF: An Optimal Gradient Methodfor Nonnegative Matrix Factorization," *IEEE Transactionson Signal Processing*, vol. 60, no. 6, pp. 2882-2898, 2012.
- [22] F S Samaria. and A C Harter., "Parameterisation of AStochastic Model forHuman Face Identification," *Applications of Computer Vision*, pp. 138-142, 1994.
- [23] P J Phillips., H Moon., S A Rizvi., and P Rauss., "The FER-ET EvaluationMethodology for Face Recognition Algorithms," *IEEE Transactions on Pattern Analysis andMachine Intelligence*, vol. 22, no. 10, pp. 1090-1104, 2000.
- [24] P N Belhumeur., J P Hespanha., and D J Kriegman., "Eigenfaces vs.Fisherfaces: Recognition using Class Specific Linear Projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp.711-720, 1997.
- [25] D B Graham. and N M Allinson., "Characterising Virtual Eigensignatures for General Purpose Face Recognition," *Face Recognition*, pp. 446-456, 1998.
- [26] D Cai., X He., X Wu., and J Han., "Non-negative Matrix Factorization onManifold," *Eighth IEEE International Conference on Data Mining*, pp. 63-72, 2008.
- [27] W Liu., N Zheng., and X Lu., "Non-negative Matrix Factorization for Visual Coding," *Acoustics, Speech, and Signal Processing*, vol. 3, no. 3, pp. 293-296, 2003.