Nonnegative Matrix Factorization on Orthogonal Subspace with Smoothed L0 Norm Constrained

Jun Ye[1,2,*] and Zhong Jin[1]

[1] School of Computer Science & Technology, Nanjing University of Science and Technology, Nanjing, 210094, China
[2] School of Natural Sciences, Nanjing University of Posts & Telecommunications, Nanjing, 210003, China

Abstract. It is known that the sparseness of the factor matrices by Nonnegative Matrix Factorization can influence the clustering performance. In order to improve the ability of the sparse representations of the NMF, we proposed the new algorithm for Nonnegatie Matrix Factorization, coined nonnegative matrix factorization on orthogonal subspace with smoothed L0 norm constrained, in which the generation of orthogonal factor matrices with smoothed L0 norm constrained are the parts of objective function minimization. Also we develop simple multiplicative updates for our proposed method. Experiment on three real-world databases (Iris, UCI, ORL) show that our proposed method can achieve the best or close to the best in clustering and in the way of the sparse representation than other methods.

Keywords: NMF, Orthogonality, Clustering, Sparse representation, L0 norm.

1 Introduction

Nonnegative Matrix Factorization (NMF) is a recent method for finding a nonnegative decomposition of the original data matrix. Given an input data matrix $\boldsymbol{V}$, each column of which represents a sample, NMF produces two factor matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ using low-rank approximation such that $\boldsymbol{V} \approx \boldsymbol{WH}$. Each column of $\boldsymbol{W}$ represents a base vector, and each column of $\boldsymbol{H}$ describes how these base vectors are combined fractionally to form the corresponding sample in $\boldsymbol{V}$. All entries in matrices are required to be nonnegative. Nonnegativity enables a non-subtractive combination of parts to form a whole, and make the encoding of data easier to interpret [1]. So, NMF is useful for learning parts-based representation and can be able to generate sparse representations of data. This caused the NMF method have been widely used in many applications, such as data mining, pattern recognition.

However, NMF cannot always guarantee an intuitive sparse representations of data. The parts-based representation of some facial images datasets reported in literature [2] was global rather than local. Later a multitude of variants have been proposed to improve NMF. A notable stream of efforts attaches various regularization

J. Yang, F. Fang, and C. Sun (Eds.): IScIDE 2012, LNCS 7751, pp. 1–7, 2013.

terms to the original NMF objective to enforce higher sparseness [3]. Recently, it has been shown that the orthogonality constraint on factor matrices can enhance the sparseness. Taking data clustering for example, it is conducted on the clustering of the columns of the input data matrix, and indicated by the matrix $\boldsymbol{H}$, orthogonality on each row of $\boldsymbol{H}$ can improve clustering accuracy. It was proved that orthogonal NMF is equivalent to k-means clustering [4]. So how to enhance the ability of the sparseness representation of the data can be an important issue? Ding et al. [5] proposed Orthogonal Nonnegative Matrix Factorization (ONMF) firstly which orthogonality is achieved by solving an optimization problem with orthogonality constraints. However, their method requires an intensive computation, which is very expensive for clustering task. Zhao Li et al. [6] considered the deficiency of the computational complexity, they proposed a new method called NMF on Orthogonal Subspace (NMFOS), in which orthogonality constraint on one of the factor matrices is embedded as part of the objective function optimization. Thus, orthogonality is achieved through the process of factorization instead of using additional constraints.

To obtain the sparsest of the factor matrices, its essence is searching for a solution with minimal L0 norm for matrices, i.e., minimum number of nonzero components of $\boldsymbol{W}$ and $\boldsymbol{H}$. It is stated that searching the minimum L0 norm is an intractable problem as the dimension increases (because it requires a combinatorial search), and it is too sensitive to noise (because any small amount of noise completely changes the L0 norm of a vector) [7]. Consequently, researchers consider other approaches. Zuyuan Yang et al.[8] introduced smoothed L0 norm constraints to the original NMF, denoted NMF-SL0, to enhance the ability of the sparseness.

In this paper, as the smoothed L0 norm of the factor matrices can reflect the sparseness intuitively and it is easy to be optimized, we consider NMF on orthogonal subspace with smoothed L0 norm constraints, called smoothed L0 norm constrained nonnegative matrix factorization on orthogonal subspace (NMFOS-SL0), and its application to the task of clustering, where an orthogonality constraint and the smoothed L0 norm constraints are imposed on the nonnegative decomposition of an inputing data matrix. We develop new multiplicative updates for NMFOS-SL0. Experiments on three different datasets show our method perform better in clustering task, and sparseness of the factor matrices, compared to other methods.

The rest of this paper is organized as follows. In Section 2, NMF and NMFOS are presented. Section 3 describes the NMFOS-SL0, and give new multiplicative updates for it. Simulations using three real databases are presented in Section 4. Finally, conclusions are summarized in Section 5.

2 Related Work

2.1 Standard NMF

Consider a data matrix $\boldsymbol{V} = [v_1, v_2, \cdots, v_n] \in \boldsymbol{R}^{m \times n}$, each column of which consists of m features, and represents a sample such as a text focument, or a face image. NMF

aims to find two nonnegative matrices $\boldsymbol{W} \in \boldsymbol{R}^{m\times r}$, $\boldsymbol{H} \in \boldsymbol{R}^{r\times n}$, $r \ll \min\{m,n\}$, such that $\boldsymbol{V} \approx \boldsymbol{WH}$. So the objective optimization problem can be concluded:

$$\min_{\boldsymbol{W},\boldsymbol{H}} \left\|\boldsymbol{V}-\boldsymbol{WH}\right\|_2^2 \quad s.t.\ \boldsymbol{W},\boldsymbol{H} \geq 0 \tag{1}$$

The multiplicative update rules were investigated by Lee et al. [1], as follows:

$$\text{a)} \quad \boldsymbol{W} \leftarrow \boldsymbol{W} \otimes \frac{\boldsymbol{VH}^T}{\boldsymbol{WHH}^T}; \quad \text{b)} \quad \boldsymbol{H} \leftarrow \boldsymbol{H} \otimes \frac{\boldsymbol{W}^T\boldsymbol{V}}{\boldsymbol{W}^T\boldsymbol{WH}} \tag{2}$$

where $\otimes$ denote elementwise multiplication.

2.2 Nonnegative Matrix Factorization on Orthogonal Subspace (NMFOS)

The important application of NMF is the parts-based representation. But NMF cannot always guarantee the intuitive parts-based representation which is conducted on the clustering of the rows of the input data matrix, and is described by the factor matrix $\boldsymbol{W}$. Local representation requires the base vectors of the factor matrix $\boldsymbol{H}$, which represent the parts of data, to be distinct from each other. Local representation is related to orthogonality, the more orthogonal between the base vectors, the more distinct between the parts[6]. Ding et al [5] in order to enhance the orthogonality of the factor matrices, they introduced the orthogonality constraints to the original NMF. So the ONMF can be described as follows:

$$\min_{\boldsymbol{W},\boldsymbol{H}\geq 0} \left\|\boldsymbol{V}-\boldsymbol{WH}\right\|_2^2 \ s.t.\ \boldsymbol{W}^T\boldsymbol{W}=\boldsymbol{I} \quad , \qquad \min_{\boldsymbol{W},\boldsymbol{H}\geq 0} \left\|\boldsymbol{V}-\boldsymbol{WH}\right\|_2^2 \ s.t.\ \boldsymbol{H}^T\boldsymbol{H}=\boldsymbol{I} \tag{3}$$

Considering the deficiency of the computational complexity, Zhao Li et al [6] proposed the method of NMFOS, in that orthogonality constraint on one of the factor matrices is embedded as part of the objective function optimization. And it was described as follows:

$$\min_{\boldsymbol{W},\boldsymbol{H}\geq 0} \left\|\boldsymbol{V}-\boldsymbol{WH}\right\|_2^2 + \lambda\left\|\boldsymbol{W}^T\boldsymbol{W}-\boldsymbol{I}\right\|_2^2 \quad , \qquad \min_{\boldsymbol{W},\boldsymbol{H}\geq 0} \left\|\boldsymbol{V}-\boldsymbol{WH}\right\|_2^2 + \lambda\left\|\boldsymbol{HH}^T-\boldsymbol{I}\right\|_2^2 \tag{4}$$

Also Zhao Li et al [6] developed the multiplicative update rules for problem (4) and could be find from the literature [6].

3 Related Work Smoothed L0 Norm Constrained Nonnegative Matrix Factorization on Orthogonal Subspace (NMFOS-SL0)

3.1 Smoothed L0 Norm

Let the function $f_\sigma(s) = \exp\left(-\frac{|s|^2}{2\sigma^2}\right)$, then $\lim_{\sigma\to 0} f_\sigma(s) = \begin{cases} 1, if\ s=0 \\ 0, if\ s\neq 0 \end{cases}$, where σ is a positive constant and s is a variable. Let the function $J_{\boldsymbol{W}} = m\times r - \sum_{s=1}^{m}\sum_{t=1}^{r} f_\sigma(w_{st})$ as

measurement for the matrix $\boldsymbol{W}$. It is obviously that when $\sigma \to 0$, $J_{\boldsymbol{W}} \to \|\boldsymbol{W}\|_0$. Therefore, $J_{\boldsymbol{W}}$ is called the smoothed L0 norm [9]. In a similar way, we can define the smoothed L0 norm for the matrix $\boldsymbol{H}$ as: $J_{\boldsymbol{H}} = r \times n - \sum_{t=1}^{r} \sum_{u=1}^{n} f_\sigma(h_{tu})$.

3.2 NMFOS-SL0

As the smoothed L0 norm of the factorization matrices can reflect the sparseness intuitively and it is easy to be optimized, we consider NMF on orthogonal subspace with smoothed L0 norm constraints, called NMFOS-SL0. We introduce the measure functions $J_{\boldsymbol{W}}$ and $J_{\boldsymbol{H}}$ i.e. the smoothed L0 norm constraint with the factorization matrices to the NMFOS's objective function. And get the new problems as follows:

$$\min_{\boldsymbol{W},\boldsymbol{H} \geq \boldsymbol{0}} : F_{\boldsymbol{W}} = \|\boldsymbol{V} - \boldsymbol{W}\boldsymbol{H}\|_2^2 + \lambda \left\|\boldsymbol{W}^{\boldsymbol{T}}\boldsymbol{W} - \boldsymbol{I}\right\|_2^2 + \alpha_{\boldsymbol{W}} J_{\boldsymbol{W}} \tag{5}$$

$$\min_{\boldsymbol{W},\boldsymbol{H} \geq \boldsymbol{0}} : F_{\boldsymbol{H}} = \|\boldsymbol{V} - \boldsymbol{W}\boldsymbol{H}\|_2^2 + \lambda \left\|\boldsymbol{H}\boldsymbol{H}^{\boldsymbol{T}} - \boldsymbol{I}\right\|_2^2 + \alpha_{\boldsymbol{H}} J_{\boldsymbol{H}} \tag{6}$$

The partial derivatives of $F_{\boldsymbol{W}}$ in (**5**) with respect to $\boldsymbol{W}$ and $\boldsymbol{H}$ are as follows:

$$\begin{cases} \dfrac{\partial F_{\boldsymbol{W}}}{\partial \boldsymbol{W}} = -\boldsymbol{V}\boldsymbol{H}^T - 2\lambda \boldsymbol{W} + \boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T + 2\lambda \boldsymbol{W}\boldsymbol{W}^T\boldsymbol{W} - \dfrac{\alpha_{\boldsymbol{W}}}{\sigma^2}\boldsymbol{W} \otimes exp\left(-\dfrac{\boldsymbol{W} \otimes \boldsymbol{W}}{2\sigma^2}\right) \\ \dfrac{\partial F_{\boldsymbol{W}}}{\partial \boldsymbol{H}} = -\boldsymbol{W}^T\boldsymbol{V} + \boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H} \end{cases} \tag{7}$$

where the parameter $\alpha_{\boldsymbol{W}}$ is selected according to the following exponential rule: $\alpha_{\boldsymbol{W}} = \beta_{\boldsymbol{W}} \exp(-\tau_{\boldsymbol{W}} k)$, k is the iteration number, $\beta_{\boldsymbol{W}}$ and $\tau_{\boldsymbol{W}}$ are constants [10].

In order to constrain $\boldsymbol{W}$ and $\boldsymbol{H}$ to be nonnegative, let $\xi_{\boldsymbol{H}} = \dfrac{\boldsymbol{H}}{\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H}}$, $\xi_{\boldsymbol{W}} = \dfrac{\boldsymbol{W}}{\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T + 2\lambda \boldsymbol{W}\boldsymbol{W}^T\boldsymbol{W} - \dfrac{\alpha_{\boldsymbol{W}}}{\sigma^2}\boldsymbol{W} \otimes exp\left(-\dfrac{\boldsymbol{W} \otimes \boldsymbol{W}}{2\sigma^2}\right)}$, Then, based on the widely used alternate-least-squares multiplication updating rules, Substitute $\xi_{\boldsymbol{W}}$ and $\xi_{\boldsymbol{H}}$ to $\boldsymbol{W} = \boldsymbol{W} - \xi_{\boldsymbol{W}} \dfrac{\partial F_{\boldsymbol{W}}}{\partial \boldsymbol{W}}$ and $\boldsymbol{H} = \boldsymbol{H} - \xi_{\boldsymbol{H}} \dfrac{\partial F_{\boldsymbol{H}}}{\partial \boldsymbol{H}}$ respectively. We can give the multiplicative update rules of $\boldsymbol{W}$ and $\boldsymbol{H}$ for problem (**5**) as follows:

$$\text{a) } \boldsymbol{W} \leftarrow \boldsymbol{W} \otimes \frac{\boldsymbol{V}\boldsymbol{H}^T + 2\lambda \boldsymbol{W}}{\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T + 2\lambda \boldsymbol{W}\boldsymbol{W}^T\boldsymbol{W} - \frac{\alpha_W}{\sigma^2}\boldsymbol{W} \otimes exp\left(-\frac{\boldsymbol{W} \otimes \boldsymbol{W}}{2\sigma^2}\right)}; \text{ b) } \boldsymbol{H} \leftarrow \boldsymbol{H} \otimes \frac{\boldsymbol{W}^T\boldsymbol{V}}{\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H}} \quad (8)$$

In similar way , we get the multiplicative update rules for problem (6) as follows:

$$\text{c) } \boldsymbol{W} \leftarrow \boldsymbol{W} \otimes \frac{\boldsymbol{V}\boldsymbol{H}^T}{\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T}; \quad \text{d) } \boldsymbol{H} \leftarrow \boldsymbol{H} \otimes \frac{\boldsymbol{W}^T\boldsymbol{V} + 2\lambda \boldsymbol{H}}{\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H} + 2\lambda \boldsymbol{H}\boldsymbol{H}^T\boldsymbol{H} - \frac{\alpha_H}{\sigma^2}\boldsymbol{H} \otimes exp\left(-\frac{\boldsymbol{H} \otimes \boldsymbol{H}}{2\sigma^2}\right)} \quad (9)$$

Based on the analysis above, the NMFOS-SL0 algorithm can be concluded as :

Step1: Initialization: input the nonnegative matrix $\boldsymbol{V}$, and give the initial nonnegative matrices of $\boldsymbol{W}$, $\boldsymbol{H}$ randomly . And set the parameters $\lambda, \sigma, \beta_i, \tau_i, i = W, H$, respectively;

Step2: Updating: update $\boldsymbol{W}$, $\boldsymbol{H}$ using (a) and (b) (or (c) and (d)) respectively;

Step3: Stopping: if a stopping criterion is satisfied, the algorithm stops; otherwise, go to step 2.

4 Experiments

We tested our proposed method on three databases[11]. And we do the clustering and part-based representation experiment on these database. For comparisons, three other algorithms have been chosen: NMF [1], ONMF [5], and NMFOS [6].

Iris, a data set that contains 150 instances of four positive-valued attributes. The samples belong to three iris classes, each including 50 instances. This dataset is selected mainly for comparison with the following larger scale datasets.

• Digit, a subset containing “0,” “2,” “4,” and “6” selected from UCI optical handwritten digit database. There are 2237 samples of 62 nonnegative integer attributes. This dataset is used to demonstrate the method behavior when samples are much more than attributes.

• ORL Database, a set of face images at different times, varying the lighting, facial expressions and facial details. There are 400 grayscale images from 40 distinct subjects and of size 92*112. This data set is used to study the case where the dimensionality is much higher than the number of samples.

A. Clustering

Clustering is an important application of NMF and its variants. We have adopted two measurements, purity and entropy, which are widely used in nonnegative learning literature, for comparing clustering results. Suppose there is ground truth data that labels the samples by one of classes. Purity is given by $purity = \frac{1}{n}\sum_{k=1}^{r} \max_{1 \le l \le q} n_k^l$, where n_k^l is the number of samples in the cluster k that belong to original class l .

A larger purity value indicates better clustering performance. Entropy measures how classes are distributed on various clusters. The entropy of the entire clustering solution is given by $entropy = -\frac{1}{n \log_2 q} \sum_{k=1}^{r} \sum_{l=1}^{q} n_k^l \log_2 \frac{n_k^l}{n_k}$, where $n_k = \sum_l n_k^l$. Generally, a smaller entropy value corresponds to a better clustering quality. We set $r = q$ and repeated each algorithm on each data set 100 times with different random seeds for initialization. In our method, we set parameter $\lambda = 5$ and $\beta_i, \tau_i, i = W, H$ equal to 100 and 0.01, respectively. The parameter $\sigma \in [1,0.5,0.2,0.1,0.05,0.02,0.01]$ that we chosed following the literature [9] and then the best result was reported.The mean and standard deviation of the purities and entropies of each algorithm data set pair are shown in Table 1 and 2 , respectively. From the Table 1, 2, we can see that the NMFOS-SL0 algorithm improves clustering accuracy for Iris and ORL datasets, and on the Digit database our proposed method performs very closely to the best.

Table 1. Clustering performance of purity comparison on three database (mean± deviation)

Database	NMF[11]	ONMF[11]	NMFOS	NMFOS-SL0
Iris	0.78±0.05	0.85±0.03	0.86±0.02	**0.88±0.02**
Digit	**0.98±0.00**	**0.98±0.00**	0.95±0.01	0.96±0.01
ORL	0.47±0.03	0.72±0.02	0.74±0.03	**0.78±0.02**

Table 2. Clustering performance of entropy comparison on three database (mean± deviation)

Database	NMF[11]	ONMF[11]	NMFOS	NMFOS-SL0
Iris	0.42±0.08	0.30±0.05	0.26±0.04	**0.24±0.03**
Digit	**0.08±0.00**	**0.08±0.00**	0.16±0.01	0.14±0.01
ORL	0.34±0.02	0.17±0.01	0.16±0.01	**0.15±0.02**

B. Part-Based Representation

In order to study the sparseness ability of parts-based representation of proposed NMFOS-SL0 method. We introduce the sparseness according to the Hoyer[3], comparing the base sparseness ability with those learned by NMF, ONMF and the NMFOS on ORL database. And the sparseness was defined as:

$$sparseness(X) = \left(\sqrt{n} - \left(\sum_i |x_i| \right) \Big/ \sqrt{\sum x_i^2} \right) \Big/ \sqrt{n} - 1 \tag{10}$$

where n is the dimensionality of vector $\boldsymbol{X}$. Table 3 shows the average sparseness of the columns in the learned basis by NMF, ONMF, NMFOS and NMFOS-SL0. It can be seen that the sparseness of the factorization matrices that used the method of NMFOS and NMFOS-SL0 are sparser than NMF and ONMF's.

Table 3. The sparseness of the factories matrices comparison

Hoyer's method	NMF	ONMF	NMFOS	NMFOS-SL0
W	0.38	0.58	0.69	**0.75**
H	0.34	0.49	0.60	**0.62**

5 Conclusion

In this work, a new NMF method based on the orthogonal subspace with smoothed L0 norm constrained is proposed, which can enhance the sparseness of the factor matrices. And we develop the multiplicative updates for the new method. This method introduce the additional parameters that balance the sparseness and reconstruction. However, the selection of the parameter value usually relies on exhaustive methods, which hinders their application. In future work, more efficient learning algorithms will be exploited.

References

1. Lee, D.D., et al.: Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing (Proc. NIPS), vol. 13, pp. 556–562 (2000)
2. Li, S., Hou, X., Zhang, H., Cheng, Q.: Learning spatially localized, parts-based representation. In: IEEE Comput. Soc. Conf. Comput. Vision Pattern Recognition, vol. 1, pp. 207–212 (2001)
3. Hoyer, P.O.: Nonnegative Matrix Factorization with Sparseness Constraints. J. Machine Learning Research 5, 1457–1469 (2004)
4. Yang, Z., Laaksonen, J.: Multiplicative updates for non-negative projections. Neurocomputing 71(1-3), 363–373 (2007)
5. Ding, C., Li, T., Peng, W., Park, H.: Orthogonal nonnegative matrix trifactorizations for clustering. In: KDD 2006: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 126–135. ACM, New York (2006)
6. Li, Z., Wu, X., Peng, H.: Nonnegative Matrix Factorization on Orthogonal Subspace. Pattern Recognition Letters 31, 905–911 (2010)
7. Donoho, D.L., Elad, M., Temlyakov, V.: Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Trans. Info. Theory 52(1), 6–18 (2006)
8. Yang, Z., Chen, X., Zhou, G., Xie, S.: Spectral unmixing using nonnegative matrix factorization with smoothed L0 norm constraint. In: Proceedings of SPIE, vol. 7494 (2009)
9. Hosen Mohimani, G., Babaie-Zadeh, M., Jutten, C.: A fast approach for overcomplete sparse decomposition based on smoothed l0 norm. IEEE Transactions on Signal Processing 57(1), 289–301 (2009)
10. Zdunek, R., Cichocki, A.: Nonnegative matrix factorization with constrained second order optimization. Signal Processing 87, 1904–1916 (2007)
11. Yang, Z., Oja, E.: Linear and Nonlinear Projective Nonnegative Matrix Factorization. IEEE Trans. Neural. Networks 21(5), 734–747 (2010)