Subtractive Initialization of Nonnegative Matrix Factorizations for Document Clustering

Gabriella Casalino[1], Nicoletta Del Buono[2], and Corrado Mencar[1]

[1] Dipartimento di Informatica,
[2] Dipartimento di Matematica,
Università degli Studi di Bari Aldo Moro, Via E. Orabona 4, I-70125 Bari, Italy
gabriella.casalino@gmail.com, delbuono@dm.uniba.it, mencar@di.uniba.it

Abstract. Nonnegative matrix factorizations (NMF) have recently assumed an important role in several fields, such as pattern recognition, automated image exploitation, data clustering and so on. They represent a peculiar tool adopted to obtain a reduced representation of multivariate data by using additive components only, in order to learn parts-based representations of data. All algorithms for computing the NMF are iterative, therefore particular emphasis must be placed on a proper initialization of NMF because of its local convergence. The problem of selecting appropriate starting initialization matrices becomes more complex when data possess special meaning, and this is the case of document clustering. In this paper, we present a new initialization method which is based on the fuzzy subtractive scheme and used to generate initial matrices for NMF algorithms. A preliminary comparison of the proposed initialization with other commonly adopted initializations is presented by considering the application of NMF algorithms in the context of document clustering.

1 Introduction

Several applications store pertinent information in a huge matrix which is often non-negative. Examples are documents in document collections, which are stored as columns of term-by-document matrix, whose elements count the number of times (possibly weighted) a corresponding term appears in a selected document. Similarly, in image collections, each image is represented by a vector whose elements correspond to the intensity and/or the color of the image pixels. In recommender systems, the information for a purchase history of customers or ratings on a subset of items is stored in a non-negative sparse matrix.

Three common goals can be identified when mining information from non-negative matrices: to automatically cluster similar items into groups, to retrieve items most similar to a user query, to identify interpretable critical dimensions within the collection.

Taking into account the non-negativity constraint, benefits in terms of meaningful interpretations of the obtained model can be added to any data analysis process. Nevertheless, classical tools are not able to guarantee the conservation of the non-negativity.

A. Petrosino et al. (Eds.): WILF 2011, LNAI 6857, pp. 188–195, 2011.

Recently, non-negative matrix factorization (NMF) received an increasing attention from the data analysis community due to its capabilities of obtaining a reduced representation of data only using positive restrictions [9]. These constraints led to a part-based representation, because they allow nonnegative linear combination of a set of nonnegative "bases" that represent realistic "building blocks" for the original data. More formally, given an initial set of data expressed by a $n \times m$ matrix X, whereas each entries X_{ij} represents in a broad sense the score obtained by the entity j on the variable i, a NMF consists in approximating (generally, in terms of Frobenius norm) the matrix X with the product of two reduced rank nonnegative matrices, the $n \times r$ basis matrix W, and the $r \times m$ encoding matrix H (with rank factor $r < \min(m, n)$), so that $X \approx WH$.

In this way, the perception of the whole, being it an image or a document in a collection, becomes a combination of its parts represented by basis vectors. Particularly, in the standard vector space model context, when X represents a term-by-document matrix, the basis vectors identify a set of words denoting a particular concept or topic and each column of H contains an encoding of the linear combination of basis vectors approximating the corresponding column of X. Hence, each document is viewed as combination of basis vector and it can be categorized as belonging to a specific topic. So, nonnegative factors of NMF can be directly applied to perform partitional clustering that identifies semantic features in a document collection and groups the documents into clusters on the basis of shared semantic features [8,10,11]. Moreover, this factorization can be used to compute a low rank approximation of a large sparse matrix along with preservation of natural data non-negativity.

All algorithms for computing the NMF are iterative and require initialization of the basis and encoding matrices. Therefore, the efficiency of many NMF algorithms is affected by the selection of the starting matrices: poor initialization often results in slow convergence or lower error reduction. Furthermore, the problem of selecting appropriate initializations becomes more complicated when certain structures or constraints are imposed on the factorized matrices or when the data possess special meaning as in the context of document clustering. Different initialization mechanisms have been proposed in literature: some of them lead to rapid error reduction and faster convergence of the adopted NMF algorithm, others lead to better overall error accuracy at convergence. However, there does not exist a definitive suggestion about the best initialization strategy to be adopted for different NMF algorithms [6].

In this paper, we propose the use of the subtractive clustering [3], a fast method for estimating clusters in the data, as a basis scheme to generate the initial matrices $W^{(0)}$ and $H^{(0)}$ for any NMF algorithm. Each obtained cluster center can be directly translated into columns of the initial basis matrix $W^{(0)}$, while elements in the encoding matrix can be obtained as fuzzy membership degree of each data to each cluster. With respect to other cluster methods, such as k-means, widely used for NMF initialization, subtractive clustering could also be used to suggest the proper rank factor, when average distance between document data is estimated.

The rest of the paper is organized as follows. In the next section, we briefly review some of the NMF algorithms commonly used for document clustering together with some initialization strategies. Then, we illustrate the subtractive clustering and how it can be adopted to generate $W^{(0)}$ and $H^{(0)}$. In section 4, we report the results obtained from different NMF algorithms initialized by the proposed approach in clustering a subset of the Reuters data corpus. Comparisons with some usually adopted initialization techniques and evaluations of the obtained clusters are also reported. Finally, some conclusive remarks and guidelines for future work are sketched in section 5.

2 NMF Algorithms and Classical Initializations

A NMF of a given data matrix X can be obtained by finding a solution of a non-linear optimization problem over a specified error function. The most frequently adopted error function is the squared Euclidean distance which leads to the minimization of the functional $\|X - WH\|_F^2$ subject to the non-negativity constraints over the elements W_{ij} and H_{ij}.

The most popular approach to numerically solve the NMF optimization problem is the multiplicative update algorithm (NMFLS) proposed in [9]. It can be shown that, starting from some nonnegative initial matrices, the square Euclidean distance is non-increasing under the following iterative update rules:

$$H_{ij} \leftarrow H_{ij} \frac{(W^\top X)_{ij}}{(W^\top WH)_{ij} + \varepsilon} \qquad W_{ij} \leftarrow W_{ij} \frac{(XH^\top)_{ij}}{(WHH^\top)_{ij} + \varepsilon} \tag{1}$$

where ε is a small positive parameter used to avoid division by zero.

Algorithms following an alternating process, approximating (in the sense of mean squared error) firstly W, then H, and so on, can be also adopted to obtain a NMF of X. Particularly, starting from some nonnegative initialization of W, an elementary Alternate Least Square algorithm (ALS) [2] for minimizing the square Euclidean distance measure is:

$$\begin{array}{l} -\text{Solve matrix equation}: W^\top WH = W^\top X \text{ w.r.t } H \\ -\text{Set to 0 negative elements in } H \ (\textit{projection step}) \\ -\text{Solve matrix equation}: HH^\top W^\top = HX^\top \text{w.r.t } W \\ -\text{Set to 0 negative elements in } W \ (\textit{projection step}) \end{array} \tag{2}$$

Different modifications of the standard cost functions have been proposed to include further constraints on the factors W and/or H, such as sparsity or orthogonality. A nonnegative sparse encoding scheme (NMFSC), proposed in [7], has the peculiarity of controlling the statistical sparsity of the H matrix in order to discover parts-based representations that are qualitatively better than those given by standard NMF. Orthogonal nonnegative matrix algorithms (ONMF) attempt instead, to obtain the basis or the encoding matrix with columns as orthogonal as possible, to minimize the number of basis components required to represent the data and the redundancy between different bases [4].

2.1 Initialization Mechanisms

All algorithms for NMF are iterative and require the computation of initial matrices $W^{(0)}$ and/or $H^{(0)}$ by some numerical mechanism and then alternately update W and H until there is no further appreciable change in the objective function, yielding locally optimal solutions. The initial pair $(W^{(0)}, H^{(0)})$ plays a crucial role for the convergence speed of the iterative algorithm and to improve algorithm performance. Moreover, when NMF is applied to document clustering, initial matrices should also posses meaningful interpretations.

Initialization schemes can be classified in simple mechanisms, based on some kind of randomization, and complex schemes based on some alternative low rank factorization or clustering algorithms. The former class includes: (i) the random initialization which produces dense matrices $W^{(0)}$ and $H^{(0)}$ of dimension $n \times r$ and $r \times m$, respectively, with elements randomly generated in [0,1], (ii) several variants of random choices of columns in X used to build $W^{(0)}$ together with random or zeros initialization of $H^{(0)}$. Complex initialization strategies exploit clustering algorithms or some alternative low rank factorization scheme to construct the initial pair $(W^{(0)}, H^{(0)})$. Among this class, we can enumerate spherical k-means initialization (kmeans) [12], Fuzzy C-Means (FCM) initialization [13], and Nonnegative Double Singular Value Decomposition (NNDSVD) based on two SVD processes [1]. Generally speaking, complex initialization strategies, require a higher computational costs, but they produce a fast error reduction, a high convergence rate in NMF algorithms and reduce to the minimum or definitely do not require the use of any randomization step.

3 Initialization by Subtractive Clustering

In this section, we briefly describe the initialization scheme based on the subtractive clustering (SC) ([3]) and we illustrate how to generate the initial $(W^{(0)}, H^{(0)})$ for any NMF iterative algorithm. It should be pointed out that, all clustering methods adopted to initialize NMF algorithms need to fix the number of clusters corresponding to the rank factor r, defining the dimensionality of the subspace approximating the data. The SC, instead, is able to automatically discover the most appropriate value of r, when an estimation of distance among data is provided.

Consider the data matrix $X = [X_1, X_2, \ldots, X_n]$, where without loss of generality each column vector $X_j \in \mathbb{R}^m$ is assumed to be normalized to have unit l_2 norm.

The SC assumes each data point is a potential cluster center and calculates a measure of the likelihood that each data point would define the cluster center, based on the potential of surrounding data points as follows:

$$P_j = \sum_{k=1}^{n} \exp\left(-\frac{4}{r_a^2}\|X_j - X_k\|^2\right), \tag{3}$$

being r_a a positive constant representing a normalized radius defining a neighborhood. According to (3), high potential values correspond to a data point with many neighborhood data points. Hence, we compute the potential of each data point and then we select the point with the highest potential as the first cluster center. Then, in order to avoid that points near to the first cluster center could be selected as another center cluster, we subtract from each data point an amount of potential proportional to its distance from the first cluster center. After the potential reduction, we select the data point with the highest remaining potential as the second cluster and we further reduce the potential of each data point according to their distance to the second cluster center. Generally, after the k-th center cluster $\tilde{X}_k$ has been obtained with potential $\tilde{P}_k$, we reduce the potential of each data point by:

$$P_j \leftarrow P_j - \tilde{P}_k \exp\left(-\frac{4}{r_b^2}\|X_j - \tilde{X}_k\|^2\right), \quad j = \ldots, n, \tag{4}$$

where r_b is a positive constant (typically chosen as $r_b = 1.25r_a$). The process of finding new cluster center and reducing potential of all data iterates until the remaining potential of all data points is bounded by some fraction of the potential $\tilde{P}_1$ of the first center cluster. The stopping criterion usually adopted is $\tilde{P}_k < 0.15\tilde{P}_1$.

After the stopping criterion is satisfied, the SC applied to a term-by-document matrix provides: the number r of clusters, the cluster centroids and their potential value $\tilde{P}_k$, $k = 1, \ldots, r$.

The initial matrices $W^{(0)}$ and $H^{(0)}$ are constructed as follows. The basis matrix collects the cluster centroid vectors $\tilde{X}_k$ ordered by decreasing values of their potential $\tilde{P}_k$, i.e., $W^{(0)} = [\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_r]$. The encoding matrix $H^{(0)}$ provides the degree to which each document is assigned to each cluster. Particularly, the elements $H_{kj}^{(0)}$, $k = 1, \ldots, r$ and $j = 1, \ldots, m$, provide the fuzzy membership value for the j-th document in the k-th cluster and are computed by

$$H_{kj}^{(0)} = \frac{\exp\left(-\frac{1}{2}\frac{\|X_j - W_{*k}^0\|^2}{\sigma^2}\right)}{\sum_{i=1}^{r} \exp\left(-\frac{1}{2}\frac{\|X_j - W_{*i}^0\|^2}{\sigma^2}\right)} \tag{5}$$

being r the total number of clusters and $\sigma^2 = \frac{{r_a}^2}{8}$. The denominator inside the previous formula represents a normalization which is needed since the reconstruction of a column X_i can be regarded as a weighted average of the centroids $W_{*i}^{(0)}$ with respect to membership values in $H^{(0)}$ (which act as weights). The sum of membership values must be equal to 1 in order to obtain a convex average value.

4 Numerical Experiments

In this section, we illustrate the performance of some NMF algorithms applied on a document clustering problem and we aim to compare the SC initialization with other complex initialization schemes. The initialization strategies have

Table 1. Performance of the NMF algorithms initialized with different strategies applied to cluster data with $k = 10$

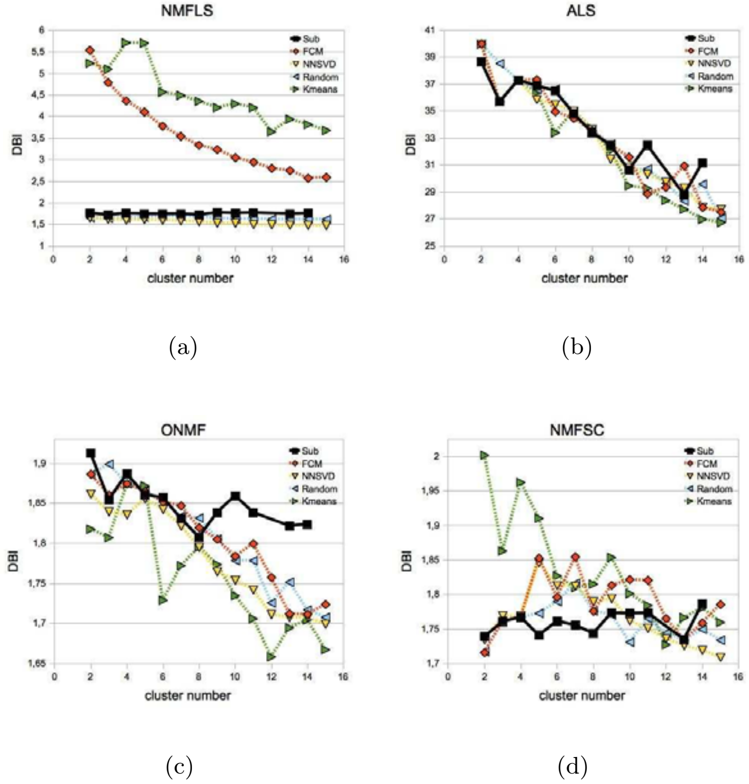
	Effectiveness of init.		NMFLS		ALS		ONMF		NMFSC	
Initial.	Time	Init. Err.	Err.	n iter	Err.	n iter	Err.	n iter	Err.	n iter
SC	0.6977	159.73	169.51	160	162.13	59	169.86	34	165.43	500
FCM	2.5053	190.74	163.92	406	164.10	270	165.50	309	167.22	500
NNSVD	0.7167	190.74	165.49	205	164.12	68	166.54	500	167.33	280
Rand	-	$39.51e^6$	164.40	500	164.48	122	165.58	500	167.73	500
K-Means	9.8546	168.73	167.25	81	164.34	170	169.55	66	169.32	156

been compared in terms of both the effectiveness of the starting pair $(W^{(0)}, H^{(0)})$ (evaluated by $\|X - W^{(0)}H^{(0)}\|_F$) and the run-time required to compute the initial factors (evaluated in seconds). The NMF algorithms with different initializations are compared in terms of error reduction and number of iterations. All the numerical results have been obtained by Matlab 7.7 codes implemented on an Intel Core Quad CPU Q6600 2.40 GHz.

The clustering problem is related to a subset of the Reuters data, consisting in 201 documents belonging to 10 categories. The dataset has been pre-processed to remove the stop words by means of a common words dictionary and by applying a stemming algorithm. The term-by-document matrix has been composed using the standards TF-IDF weights, and possesses the 97% of sparsity degree.

Table 1 reports the effectiveness of initialization strategies, together with their run-time values and the results obtained for each NMF algorithm (combined with different initializations) at the end of the learning phase over the Reuters dataset. For a fair comparison among all the algorithms, we adopted the same stopping criteria: a maximum number of iterations (`maxiter=500`) and a fixed tolerance ($toll = 10^{-6}$) for the difference between two subsequent values of the objective function. It should be pointed out that SC shows a good trade-off between run-time values and effectiveness of the initial pair $(W^{(0)}, H^{(0)})$, when compared with other complex initialization schemes. The run-time value for the random initialization has been omitted (being negligible), however this initialization produces full initial matrices with poor accuracy. Moreover, SC determines a higher convergence rate for some NMF algorithms.

The effectiveness of the clusters provided by NMF algorithms, with different initializations, has been evaluated by the Davies-Bouldin index (DBI) [5], given by $\frac{1}{r}\sum_{i=1}^{r} \max_{j:j\neq i} \frac{S_i+S_j}{M_{ij}}$, where S_i and S_j are the inter cluster similarity of the ith and jth cluster and M_{ij} is their intra-cluster separability. Figure 1 illustrates the behavior of the DBI, when the cluster number increases. Small values of DB correspond to clusters that are compact and whose centers are far away from each other. The graphs also suggest that, for almost all initializations, the most appropriate cluster number is 13. It should be also pointed out that SC is able to automatically discover the proper number of clusters when the r_a parameter is set to the mean distance among documents.



(a) (b)

(c) (d)

Fig. 1. DBI behavior related to the increasing number of clusters for (a) NMFLS , (b) ALS, (c) ONMF and (d) NMFSC, with different initializations

Table 2. Cluster accuracy and mutual information results for the NMF algorithms initialized with different strategies

	NMFLS		ALS		ONMF		NMFSC	
Initial.	Accuracy	MI	Accuracy	MI	Accuracy	MI	Accuracy	MI
SC	0.45	1.20	0.48	1.27	0.44	1.17	0.48	1.49
FCM	0.45	1.35	0.50	1.54	0.44	1.27	0.45	1.36
NNDSVD	0.48	1.48	0.46	1.38	0.48	1.48	0.50	1.60
Rand	0.41	1.19	0.42	1.25	0.43	1.32	0.47	1.42
kmeans	0.41	1.33	0.45	1.31	0.41	1.33	0.41	1.33

Table 2 reports a further evaluation of the obtained clusters with respect to the set of clusters defined by the original categorized documents in terms of cluster accuracy and mutual information measures. As it can be observed from the table, even if there are no appreciable differences among different initializations, SC and FCM provide the better results for almost all NMF algorithms.

5 Conclusive Remarks

In this paper, we proposed the fuzzy subtractive scheme as initialization for NMF algorithms in the context of document clustering. The proposed method is very fast in constructing initial pairs for NMF algorithms; in some cases it is able to increase the performance of NMF methods both in terms of number of iterations and accuracy of the final approximation. Moreover, differently from other clustering initialization strategies (such as K-Means), the SC method is able to predict the proper number of clusters, and consequently the rank factor for the low rank factorization, when mean distance among documents is provided. Future work can be addressed to assess the performance of NMF algorithms with SC initialization on different datasets as well as to further investigate its capability of predicting the most appropriate factor rank for data.

References

1. Boutsidis, C., Gallopoulos, E.: Svd based initialization: ahead start for nonnegative matrix factorization. Pattern Recognition 41(4), 1350–1362 (2008)
2. Cichocki, A., Zdunek, R., Phan, A.H., Amari, S.I.: Alternating Least Squares and Related Algorithms for NMF and SCA Problems. In: Nonnegative Matrix and Tensor Factorizations. John Wiley & Sons, UK (2009)
3. Chiu, S.L.: Fuzzy Model Estimation based on Cluster Estimation. J. Intelligent and Fuzzy Systems 2, 267–278 (1994)
4. Choi, S.: Algorithms for orthogonal nonnegative matrix factorization. Proc. Intern. Joint Conf. Neural Networks (2008)
5. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Machine Intell. 1(4), 224–227 (1979)
6. Del Buono, N., Lucarelli, M.: Comparative studies on initializations for nonnegative matrix factorization algorithms, Tech. Rep. 17/10, Univ. Bari, Italy (2010)
7. Hoyer, P.O.: Non-negative matrix factorization with sparseness constraints. J. Machine Learning Research 5, 1457–1469 (2004)
8. Lazar, C., Doncescu, A., Kabbaj, N.: Non Negative Matrix Factorization clustering capabilities; application on multivariate image segmentation. Int. J. of Business Intel. Data Mining 5(3), 285–296 (2010)
9. Lee, D.D., Seung, S.H.: Algorithms for non-negative matrix factorization. In: Proc. Adv. Neural Information Proc. Syst. Conf., vol. 13, pp. 556–562 (2000)
10. Shahnaz, F., Berry, M.W., Pauca, M.P., Plemmons, R.J.: Document clustering using nonnegative matrix factorization. Information Processing and Managements: Intern. J. 42(2), 373–386 (2006)
11. Xu, W., Liu, X., Gong, Y.: Document clustering based on nonnegative matrix factorization. In: Proc. SIGIR, pp. 267–273 (2003)
12. Xue, Y., Tong, C.S., Chen, Y., Chen, W.-S.: Clustering-based initialization for non-negative matrix factorization. Appl. Math. and Comp. 205, 525–536 (2008)
13. Zhenga, Z., Yang, J.: Initialization enhancer for non-negative matrix factorization. Eng. Appl. Art. Int. 20, 101–110 (2007)