EUSIPCO 2013 1569741067

CLUSTERING BY NON-NEGATIVE MATRIX FACTORIZATION WITH INDEPENDENT PRINCIPAL COMPONENT INITIALIZATION

Liyun Gong[1], *Asoke K. Nandi*[2,3]

[1]Department of Electrical Engineering and Electronics, Liverpool University, Liverpool, L69 3BX, UK;
[2]Department of Electronic and Computer Engineering, Brunel University, Uxbridge, UB8 3PH, UK;
[3]Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland;
l.gong@liv.ac.uk; asoke.nandi@brunel.ac.uk;

ABSTRACT

Non negative matrix factorization (NMF) is a dimensionality reduction and clustering method, and has been applied to many areas such as bioinformatics, face images classification, and so on. Based on the traditional NMF, researchers recently have put forward several new algorithms on the initialization area to improve its performance. In this paper, we explore the clustering performance of the NMF algorithm, with emphasis on the initialization problem. We propose an initialization method based on independent principal component analysis (IPCA) for NMF. The experiments were carried out on the four real datasets and the results showed that the IPCA-based initialization of NMF gets better clustering of the datasets compared with both random and PCA-based initializations.

Index Terms— Non-negative matrix factorization; Principal component analysis; Independent component analysis; Independent principal component analysis

1. INTRODUCTION

Principal component analysis (PCA) and Independent component analysis (ICA) are two of the most popular dimensionality reduction methods used for visualizing high-throughput dataset in two or three dimensions. They keep the most information about dataset in the lower dimensional space so that the similarities within the dataset can be easily visualized. Recently, Yao et al. has proposed independent principal component analysis (IPCA) which combines the advantages of both PCA and ICA [1]. It uses ICA as a de-noising process of the basic matrix produced by PCA to highlight the important structure of the dataset [1].

Another dimensionality reduction method called non-negative matrix factorization (NMF) has been proposed by Lee and Seung [2, 3]. It is different from PCA and ICA, with the added non-negative constraints [4]. Recently it has been applied to many areas such as bioinformatics, face images classification, and so on. To bioinformatics, Pascual-Montano et al. proposed a versatile tool called bioNMF based on NMF to cluster and bicluster gene expression data [5]. In [6], NMF was used for recognizing protein sequence patterns. In [7, 8], clustering results of gene expression data obtained by NMF were compared with hierarchical clustering and self-organizing maps. To improve on the traditional NMF, some researchers have also proposed several different algorithms such as Least squares-NMF [9], Weighted-NMF [10] and Local-NMF [11], leading to enhanced convergence rates. Recently researchers have paid much attention to the NMF initialization problem. Wild proposed the initialization method based on spherical k-means clustering [12]. Langville et al. compared the six initialization methods, including random initialization, centroid initialization, SVD-centroid initialization, random acol initialization, random C initialization, and co-occurrence initialization [13]. Boutsidis et al. proposed the initialization method based on singular value decomposition [14]. In this paper, we apply an initialization method based on IPCA for NMF and results are compared with PCA-based initialization [15] and random initialization, using the RAND index [16]. The experiments were carried out on the four real datasets from UCI machine learning repository [17] and the results showed that the IPCA-based initialization of NMF gets better clustering of the datasets compared with the other two methods.

The rest of the paper is organized as follows. In Section 2 we review the basic knowledge of non-negative matrix factorization method (NMF) and the PCA-based initialization, and describe the IPCA-based initialization. The RAND index used for the comparison is described in Section 3. Experimental results based on the three initialization methods are evaluated and analysed in Section 4. Finally, conclusion is drawn in Section 5.

2. NMF, PCA-BASED NMF, AND IPCA-BASED NMF

2.1. NMF

Here we briefly review the basic idea of NMF as follows:

Given a non-negative matrix X with m rows and n columns, each column represents the data points which need to be clustered. The NMF algorithm seeks to find non-negative factors W and H such that

$$X \approx WH \tag{1}$$

where W is an $m \times k$ matrix and H is a $k \times n$ matrix. Each column of W is considered as the basic vectors while each column of H contains the encoding coefficient. All the elements in W and H represent non-negative values.

Many algorithms have been proposed to obtain W and H [15]. In this paper, we use the multiplication update rule to minimize an objective function which is Euclidean distance measure. The formulae are given as follows.

$$H \leftarrow H \frac{W^T X}{W^T W H}$$
$$W \leftarrow W \frac{X H^T}{W H H^T} \tag{2}$$

Here NMF is used for both dimensionality reduction and clustering analysis. An element of H, h_{ij}, describes the degree of the point *j* belonging to the cluster *i*. If the point *j* belongs to cluster *i*, then h_{ij} will have a larger value compared with the rest of the elements in *j*'th column of H.

NMF is a nonconvex programing in the iteration process, thus it may lead to different solutions with the different initial values of W and H. In this paper, we apply two initialization methods to improve the performance of NMF. The details are described below.

2.2. PCA-based initialization

PCA is the dimensionality reduction method in which the lower-dimensional representation of the dataset preserves as much of its variation as possible to highlight its similarities and differences. PCA-based initialization method to NMF has been proposed in [15] and here we briefly review the basic idea of this method. Given the $m \times n$ matrix X as that of in NMF, and its pseudo inverse is set as A. We first apply PCA on the matrix A to obtain the eigenvectors and eigenvalues. The initial values W and H of NMF then can be described below.

- The initial matrix W of NMF is constructed by keeping the first k eigenvectors (corresponding to the k largest eigenvalues) obtained from PCA as column vectors. k is the number of cluster classes for each dataset which is already known in this paper (shown in Table 1).
- The initial matrix H of NMF is the $k \times n$ matrix which can be denoted by $H = W^T X$.

Because these two initial matrices obtained above may contain negative elements, we use the absolute value [12] for all elements in W and H in order to satisfy the initial constraint of NMF. Finally, we apply the NMF algorithm with the initial values of W and H obtained above to create the clustering results of the datasets.

2.3. IPCA-based initialization

ICA is another dimensionality reduction method in which the goal is to find a linear representation of non-gaussian signal so that the components are statistically independent. So ICA can be treated as the method to remove most of the noise from the signal (when the noise has a Gaussian distribution). Yao et al. recently proposed an approach called IPCA which combines the advantage of both PCA and ICA [1]. ICA used in IPCA is a de-noising process of the basic matrix W produced by PCA [1]. Once the basic matrix W is denoised, we expect it to be non-gaussian with no noise included. In this section, we use IPCA method as the initialization of NMF instead of PCA and the details of IPCA-based initialization method is described as follows.

STEP1: Given the $m \times n$ matrix X as that of in NMF, and its pseudo inverse is set as A. Apply PCA on the matrix A to generate the basic matrix W (the same as in section 2.2).

STEP2: Whiten the basic matrix W obtained above by using the eigenvalue decomposition of the covariance matrix of W.

STEP3: Implement ICA algorithm on the whitened matrix W and obtain the independent basic matrix W^*.

STEP4: obtain the matrix H^* which is calculated by $H^* = W^{*T} X$.

STEP5: We take the absolute value for all elements in W^* and H^*.

STEP6: Apply NMF algorithm with the initial values W^* and H^* and obtain the final H value H^{final}.

STEP7: Obtain the cluster labels of the dataset from H^{final}.

3. CLUSTERING VALIDATION

As many clustering algorithms have been proposed for the analysis of datasets, it is necessary to find a way to assess these algorithms. Clustering validation is available to do this. In this paper, the RAND index [16] is adopted to evaluate and to compare the clustering performance of the three initialization methods in the four datasets.

RAND is defined as the probability of correction for the cluster results. It handles two partition matrices defined as T and Q of the same dataset. T encodes the k known cluster labels and Q records the cluster labels obtained

from clustering algorithms. So the RAND index $w \in [0,1]$ is then defined as

$$w = \frac{a+d}{a+b+c+d} \times 100\% \quad (3)$$

where a represents the number of pairs of data points belonging to the same cluster both in T and in Q, b represents the number of pairs of data points belonging to the same cluster in T but different clusters in Q, c represents the number of pairs of data points belonging to different clusters in T but the same cluster in Q, and d represents the number of pairs of data points belonging to different clusters both in T and in Q. Note that a RAND value closer to one suggests the better cluster result.

4. EXPERIMENT RESULTS AND ANALYSIS

4.1. Datasets

Four datasets used in this paper are all from UCI machine learning repository [17]. Some properties of these datasets are presented in Table 1 and described after that.

Table 1: Properties of the datasets

Name	Pattern	Attribute	Class
Balance	625	4	3
Cancer	683	9	2
Dermatology	358	34	6
Iris	150	4	3

Balance: This dataset is based on balance a scale weight and distance. It contains 625 patterns which are classified as having the balance scale tip to the right, tip to the left, or be balanced. It has 4 attributes and 3 classes.
Cancer: This dataset is based on the diagnosis of breast cancer at Wisconsin. There are 683 patterns and each pattern has one of 2 possible classes: benign or malignant.
Dermatology: This dataset contains 358 patterns, 34 attributes, and 6 classes. It is based on the different types of Erythemato-squamous disease.
Iris: This dataset contains 3 classes of 50 instances each, where each class refers to a type of iris plant (setosa, virginica or versicolor).

4.2. Results and analysis

The experiments were carried out by using the above four datasets. We applied three different initialization methods - random, PCA-based, and IPCA-based initialization to improve the clustering performance of NMF. In order to avoid the influence of the randomness, each initialization method was run 20 times and the total number of iterations for each run of NMF was set to 500 in this paper. The rank (dimensionality) k for each dataset is set to the number of cluster of the corresponding dataset which is shown in Table1.

Figure 1 shows the initial RAND values at the first iteration from the 20 runs for each initialization method. The IPCA-based initialization always gets the highest RAND values compared with the other two methods on the four datasets. The details of the initial average RAND values of three different initialization methods are shown in Table 2. The bold values in the table represent the largest RAND value for each dataset. It can be seen from Figure 1 and Table 2 that in these four datasets, the initial RAND values obtained by the three different initialization methods always satisfy the inequality IPCA > PCA > Random. The main reason is that the random initialization has nothing to do with the initial values of W and H while the PCA-based and IPCA-based initialization already works for clustering with predefined values of W and H, so the initial RAND value of random initialization is the smallest. IPCA-based initialization adds the de-noising process using ICA, so its initial RAND value is larger than PCA-based initialization. From Table 2, we summarize that the IPCA-based initialization has the best clustering performance at the beginning of the NMF iteration compared with the other two methods.

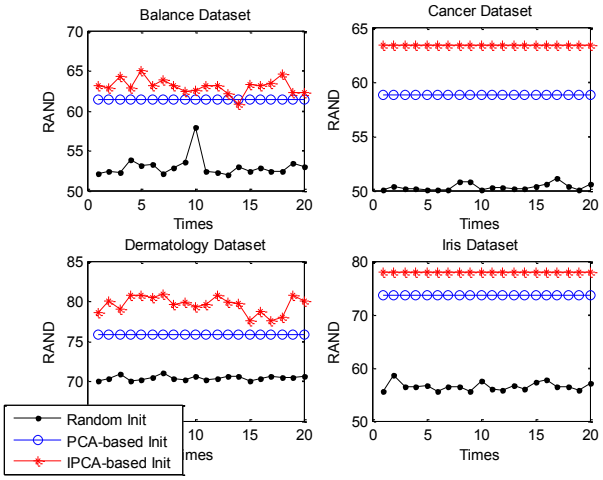

Figure 1: The initial RAND values of different initialization methods in 20 times (iteration = 1).

Table 2: The average of initial RAND values of different initialization methods (iteration = 1, time = 20).

Name	Random	PCA-based	IPCA-based
Balance	53.0	61.4	**63.1**
Cancer	50.3	58.8	**63.4**
Dermatology	70.4	75.8	**79.6**
Iris	56.5	73.5	**78.0**

Figure 2 shows the RAND values from the 20 runs for each initialization method at the 500th iteration. It can be seen that

the final RAND values of PCA-based initialization have no change during the 20 runs. This is because this initialization method computes the same initial values of W and H each time. On the contrary, the final RAND value of random initialization changes greatly during the 20 runs, as the random initialization method has no contributions to the initial values of NMF. IPCA-based initialization only varies much in the dermatology. It may due to the dimension of the dermatology dataset is much higher (34) compared to the other three datasets so that it may varies a lot for the different IPCA process. At the same time, we can see that the most of the final RAND values of the random initialization in cancer and dermatology are lower than the other two methods. This means that in these two datasets the IPCA-based and PCA-based initialization have the better performance in clustering analysis compared with the random initialization.

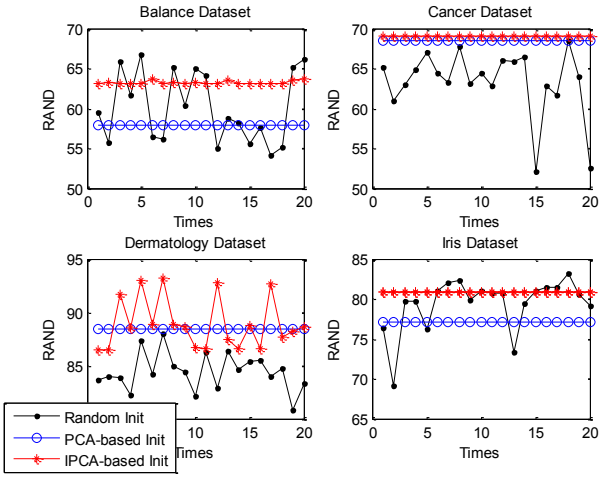

Figure 2: The final RAND values of different initialization methods in 20 times (iteration = 500).

Table 3 shows the average RAND values of three different initialization methods at the 500th iteration. In balance and iris datasets, the random initialization achieves a larger RAND value than the PCA-based initialization. This is because NMF algorithm with PCA-based initialization using Euclidean distance measure cannot pull the factorization out of local minima in these datasets [15]. However, IPCA-based initialization can perform well in these two datasets which has a higher RAND value than random initialization. In dermatology dataset, although the average of final RAND values of IPCA-based initialization is similar with that of PCA-based initialization, its final RAND values sometimes achieve above 90 during the 20 runs which is much higher than the other two methods (see Figure 2). In cancer dataset, there is no variation in PCA-based and IPCA-based initializations shown in Figure 2 and the IPCA-based initialization always gets slightly higher RAND value than PCA-based initialization during the 20 runs. By analyzing Figure 2 and Table 3, we can conclude that after 500 iterations the NMF based on the IPCA initialization can obtain higher average RAND values, which is, clustering results are better compared with the other two methods.

Table 3: The average of final RAND values of different initialization methods (iteration = 500, time = 20).

Name	Random	PCA-based	IPCA-based
Balance	60.1	57.9	**63.3**
Cancer	63.3	68.5	**69.0**
Dermatology	84.4	88.4	**88.9**
Iris	79.4	77.1	**80.9**

The RAND value from these three initialization methods increases fast before 100 iterations while it increases slowly after that, so we focus on analyzing the performance of these methods during the first 100 iterations. We draw the RAND values of different initialization methods with the increasing iteration number in Figure 3. In dermatology datasets, it shows that the PCA-based and IPCA-based initialization have a similar clustering performance and always get the higher RAND value than the random initialization as the NMF algorithm progresses. Compared with the PCA-based initialization, the IPCA-based initialization has a better start at the beginning (79.6% shown in Table 3). In balance and iris datasets, although the PCA-based initialization enhances the initial values of W and H, it still gets the lower RAND values than the random initialization after number of iterations. However, the IPCA-based initialization can solve this problem which has the higher clustering performance than the random one all the time. In cancer dataset, the IPCA-based initialization keeps the highest RAND values at the head start and maintains this advantage until about 20 iterations. In this case, IPCA-based initialization can be used in the short term with the less computational complexity.

By studying Figures 1 to 3 as well as Tables 2 and 3, it is clear that IPCA-based initialization achieves the highest RAND value in the short term and still remain the highest in the long term while PCA-based initialization gets the bad cluster results on some datasets in the long term even though it enhances the initial values of W and H. So we conclude that the NMF based on the IPCA initialization gets better clustering of the datasets compared with random and PCA-based initializations.

5. CONCLUSION AND DISCUSSIONS

Researchers often use random initializations when utilizing NMF. To improve the performance of NMF, we have proposed an initialization method based on IPCA for NMF in this paper. Altogether, we have explored the NMF algorithm with the three different initialization methods. The initialization methods are based on random, PCA, and IPCA. The experiments were carried out on four real datasets from

UCI machine learning repository [17] and we assessed the clustering performance using the RAND index [16]. From the experimental results, we see that the performance of IPCA-based NMF in balance and iris datasets is comparable to random and PCA-based NMF, while its performance in cancer and dermatology datasets is roughly comparable to PCA-based NMF in the long term. Most importantly, IPCA-based NMF can achieve faster convergence in all four datasets. So we conclude that the proposed IPCA-based initialization of NMF gets better clustering of the datasets compared with both random and PCA-based initialization. Here we only compared the three initialization methods (two standard and one new) together. As there are other good initialization methods in the literature, comparing these initialization methods would be considered in the future work.

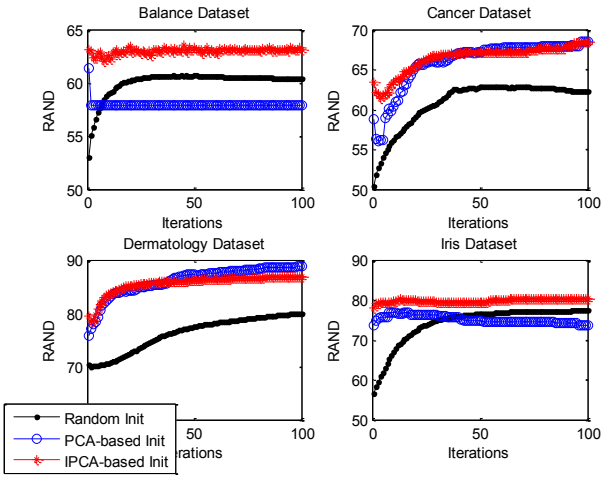

Figure 3: The RAND values of different initialization methods with the increasing iteration number.

6. ACKNOWLEDGEMENT

Asoke K. Nandi would like to thank TEKES for their award of the Finland Distinguished Professorship.

7. REFERENCES

[1] F Z Yao, J Coquery and K A Le Cao, "Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets", BMC Bioinformatics, 2012, **13**:24, doi:10.1186/1471-2105-13-24.

[2] D D Lee, H S Seung, "Learning the parts of objects by nonnegative matrix factorization", Nature, vol. 401, pp. 788-791, 1999.

[3] D D Lee, H S Seung, "Algorithm for non-negative matrix factorization", Advances in Neural Information Processing Systems, vol. 13, pp. 556-562, 2001.

[4] M S Barlett, H M Lades, T J Sejnowski, "Independent component representations for face recognition", Neural Computing, vol. 7, pp. 1129-1159, 1988.

[5] A Pascual-Montano, P Carmona-Saez, M Chagoyen, F Tirado, J M Carazo, R D Pascual-Marqui, "bioNMF: a versatile tool for nonnegative matrix factorization in biology", BMC Bioinformatics, vol. 7(1), pp. 366-374, 2006.

[6] A Heger, L Holm, "Sensitive pattern discovery with fuzzy alignments of distantly related proteins", Bioinformatics, vol. 19(90001), pp. 130-137, 2003.

[7] J P Brunet, P Tamayo, T R Golub, J P Mesirov, "Metagenes and molecular pattern discovery using matrix factorization", The National Academy of Sciences, vol. 101(12), pp. 4164-4173, 2004.

[8] Y Gao, G Church, "Improving molecular cancer class discovery through sparse non-negative matrix factorization", Bioinformatics, vol. 21, pp. 3970-3975, 2005.

[9] G Wang, A V Kossenkov, M F Ochs, "LS-NMF: a modified nonnegative matrix factorization algorithm utilizing uncertainty estimates", BMC Bioinformatics, 2006, **7**:175, doi:10.1186/1471-2105-7-175.

[10] D Guillamet, J Vitria, B Scheile, "Introducing a weighted nonnegative matrix factorization for image classification", Pattern Recognition Letters, vol. 24, pp. 2447-2454, 2003.

[11] S Z Li, X Hou, H Zhang, Q Cheng, "Learning spatially localized parts-based representation", Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 207–212, 2001.

[12] S Wild, "Seeding non-negative matrix factorizations with the spherical k-means clustering", Master of Science Thesis, University of Colorado, 2003.

[13] A N Langville, C D Meyer, R Albright, "Initializations for the nonnegative matrix factorization", Proceeding of the Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006.

[14] C Boutsidis, E Gallopoulos, "SVD based initialization: A head start for nonnegative matrix factorization", Pattern Recognition, published by Elsevier, vol. 41, pp. 1350-1362, 2007.

[15] Z Zheng, J Yang, Y Zhu, "Initialization enhancer for non-negative matrix factorization", Engineering Applications of Artificial Intelligence, published by Elsevier, vol. 20, pp. 101–110, 2007.

[16] W M Rand, "Objective Criteria for the Evaluation of Clustering Methods", Journal of the American Statistical Association, Theory and Methods Section, vol. 66, pp. 846-850, 1971.

[17] Datasets from UCI machine learning repository (http://archive.ics.uci.edu/ml/)