Mining Patterns of Lung Infections in Chest Radiographs

Spyros Tsevas[1,*], Dimitris K. Iakovidis[1], and George Papamichalis[2]

1 Technological Educational Institute of Lamia, Dept. of Informatics and Computer Technology, GR-35100 Lamia, Greece

2 Chest Hospital of Athens "Sotiria"

{s.tsevas, dimitris.iakovidis}@ieee.org

Abstract Chest radiography is a reference standard and the initial diagnostic test performed in patients who present with signs and symptoms suggesting a pulmonary infection. The most common radiographic manifestation of bacterial pulmonary infections is foci of consolidation. These are visible as bright shadows interfering with the interior lung intensities. The discovery and the assessment of bacterial infections in chest radiographs is a challenging computational task. It has been limitedly addressed as it is subject to image quality variability, content diversity, and deformability of the depicted anatomic structures. In this paper, we propose a novel approach to the discovery of consolidation patterns in chest radiographs. The proposed approach is based on non-negative matrix factorization (NMF) of statistical intensity signatures characterizing the densities of the depicted anatomic structures. Its experimental evaluation demonstrates its capability to recover semantically meaningful information from chest radiographs of patients with bacterial pulmonary infections. Moreover, the results reveal its comparative advantage over the baseline fuzzy C-means clustering approach.

1 Introduction

Artificial intelligence and data mining applications in medicine are increasingly becoming popular tools as the utilization of digital media meets the everyday clinical practice. Computer-aided diagnosis, intelligent information retrieval and knowledge discovery are some of the associated research directions, which can prove valuable to patient safety and quality of healthcare [1].

Health care–related infections comprise a major threat to patient safety. They are encountered in hospitals or health care facilities and they are usually reported as adverse events associated with medical procedures. Most commonly include pulmonary, urinary tract, skin, and soft tissue infections of bacterial origin [2].

The early detection of such infections as well as the choice of the appropriate antibiotic treatment can be life-saving especially for the critically ill patients. To this end, a computational approach that would be capable of automatically discovering patterns of infections and antibiotic prescription from patients' health records would constitute a valuable tool to the community [3].

Patients' health records may include both structured and unstructured data, digital signals and images. In the case of chest patients, chest radiographs provide substantial indications on the presence of a pulmonary infection. The most common radiographic manifestation of bacterial pulmonary infections is foci of consolidation. These are visible as bright shadows interfering with the interior lung intensities, which include intensities the lung parenchyma and intensities of superimposed structures of the thoracic cavity such as the ribs and the mediastinum. The diversity and the complexity of the visual content of the lung fields as well as the quality variability induced by the variable parameters of the radiation exposure, make its medical interpretation a challenging task. This task has motivated many researchers to develop computational methods for automatic lung field detection and analysis [4][5].

Current lung field analysis methods include size measurements of structures of the thoracic cavity [6], detection of the ribs [7], lung nodule detection [8], whereas fewer methods have been proposed for mining radiographic patterns associated with the presence of pulmonary infections [9]. Mining patterns of pneumonia and severe acute respiratory syndrome (SARS) has also been in the scope of contemporary research. In [10] a supervised approach using intensity-histograms and second-order statistical features has been proposed for mining pneumonia and SARS patterns, whereas most recently, the use of wavelet-based features have been proved useful for the detection radiographic patterns of childhood pneumonia under a supervised classification framework [11].

In contrast to the former methods this paper proposes an unsupervised approach to the discovery of consolidation patterns associated with bacterial pulmonary infections. Such an approach does not take into account any information extracted from previous images, thus avoiding the need for feature normalization between images. We use statistical intensity signatures characterizing the densities of the several anatomic structures depicted in chest radiographs to recover semantically meaningful information regarding the consolidation patterns. This is achieved by a clustering approach based on Non-negative Matrix Factorization (NMF) that involves cluster merging. The results obtained are compared with those obtained with the fuzzy C-means (FCM) clustering approach [12].

The rest of this paper consists of three sections. Section 2 provides a description of the proposed methodology, section 3 presents the results of its experimental application on a set of high-resolution chest radiographs, and section 4 summarizes the conclusions that can be derived from this study.

2 Methodology

Non-negative Matrix Factorization (NMF) was introduced by Paatero and Tapper [13] as a way to find a non-negative reduced representation of non-negative data, but it has gained popularity by the works of Lee and Seung [14, 15]. In contrast to other methods such as Principal Component Analysis (PCA), NMF allows only additive combinations of non-negative data, leading to a representation that is more intuitive and closer to the human perception.

Given a $m \times n$ non-negative matrix $\mathbf{V}$ and a reduced rank r $(r < \min(m,n))$, the non-negative matrix factorization problem lies in finding two non-negative factors $\mathbf{W}$ and $\mathbf{H}$ of $\overline{\mathbf{V}}$ such that:

$$\mathbf{V} \approx \overline{\mathbf{V}} = \mathbf{W} \times \mathbf{H} \tag{1}$$

where $\mathbf{W} \in \Re^{m \times r}$ and $\mathbf{H} \in \Re^{r \times n}$.

We may think of $\mathbf{W}$ as the matrix containing the NMF basis and $\mathbf{H}$ as the matrix containing the non-negative coefficients (or encodings) that exhibit a one-to-one correspondence with the data that consists $\mathbf{V}$. To quantify the distance between the data matrix $\mathbf{V}$ and the model matrix $\overline{\mathbf{V}}$ we used the Frobenius norm:

$$\min_{\mathbf{W},\mathbf{H}} \left\| \mathbf{V} - \mathbf{W}\mathbf{H} \right\|_F^2 \ , \ \mathbf{W}, \mathbf{H} \geq \mathbf{0} \tag{2a}$$

This optimization problem is solved using the following multiplicative update rules:

$$\mathbf{W}_{ir} \leftarrow \mathbf{W}_{ir} \frac{\left(\mathbf{V} \times \mathbf{W}^{\mathrm{T}}\right)_{ir}}{\left(\mathbf{W} \times \mathbf{H} \times \mathbf{H}^{\mathrm{T}}\right)_{ir}} \tag{2b}$$

$$\mathbf{H}_{rj} \leftarrow \mathbf{H}_{rj} \frac{\left(\mathbf{W}^{\mathrm{T}} \times \mathbf{V}\right)_{rj}}{\left(\mathbf{W}^{\mathrm{T}} \times \mathbf{W} \times \mathbf{H}\right)_{rj}} \tag{2c}$$

NMF can be considered as an alternative clustering technique [16,18-20] since given a normalized solution ($\widetilde{\mathbf{W}}$, $\widetilde{\mathbf{H}}$) of NMF, $\widetilde{\mathbf{H}}^{\mathrm{T}}$ can be interpreted as the cluster posterior and thus $\widetilde{\mathbf{H}}_{jr}$ represents the posterior probability that $\mathbf{v}_j$ belongs to the r-th cluster [20]. We can express the columns of $\mathbf{W}$ and $\mathbf{H}^{\mathrm{T}}$ as: $\mathbf{W} = \left(\mathbf{w}_1,...,\mathbf{w}_r\right)$, $\mathbf{H}^{\mathrm{T}} = \left(\mathbf{h}_1,...,\mathbf{h}_r\right)$ where in clustering terms $\mathbf{w}_r$ can be interpreted as the centroid of the r-th cluster and the $\mathbf{h}_r$ as the posterior probability of the r-th cluster. By normalizing the $\mathbf{W}$ and $\mathbf{H}^{\mathrm{T}}$ column-wisely we have the normalized columns are $\widetilde{\mathbf{W}} = \left(\widetilde{\mathbf{w}}_1,...,\widetilde{\mathbf{w}}_r\right)$ and $\widetilde{\mathbf{H}}^{\mathrm{T}} = \left(\widetilde{\mathbf{h}}_1,...,\widetilde{\mathbf{h}}_r\right)$ and the normalization:

$$\mathbf{WH} = \tilde{\mathbf{W}}\mathbf{S}\tilde{\mathbf{H}} \tag{3}$$

where

$$\tilde{\mathbf{W}} = \mathbf{W}\mathbf{D}_W^{-1}, \tag{4a}$$

$$\tilde{\mathbf{H}}^{\mathrm{T}} = \mathbf{H}^{\mathrm{T}}\mathbf{D}_H^{-1} \tag{4b}$$

$$\mathbf{S} = \mathbf{D}_W\mathbf{D}_H \tag{4c}$$

where $\mathbf{D}_W$ and $\mathbf{D}_H$ are diagonal matrices with diagonal elements be in the L_p-norm:

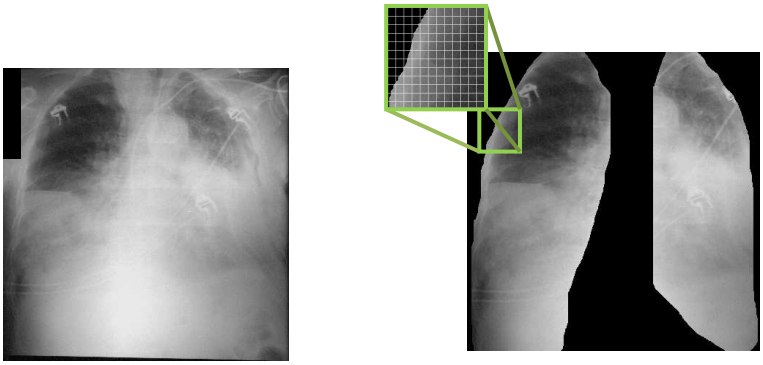
$$(\mathbf{D}_C)_{rr} = \left\|\tilde{\mathbf{c}}_r\right\|_p, (\mathbf{D}_H)_{rr} = \left\|\tilde{\mathbf{h}}_r\right\|_p \tag{5}$$

For the Euclidean distance case (L_2-norm) $\left\|\tilde{\mathbf{c}}_r\right\|_2 = 1$, $\left\|\tilde{\mathbf{h}}_r\right\|_2 = 1$ and due to the non-negativity of the data, this is just the condition that columns sum to 1. Thus, $\mathbf{D}_W$ contains the column sums of **W,** and $\mathbf{D}_H$ contains the column sums of $\mathbf{H}^{\mathrm{T}}$.

In this paper, we apply NMF as a clustering technique to extract consolidation patterns from radiographic images. A radiographic image is divided into a set of non-overlapping sub-images which are subsequently clustered into an even number of *r* clusters. Some of these clusters will correspond to patterns of normal lung parenchyma and the rest will correspond to consolidation patterns. The sub-images are represented by intensity histogram signatures characterizing the densities of the anatomic structures depicted in a chest radiograph [21]. Finally, the *r* clusters are dyadically merged down to two clusters based on the similarity of their centroids. Considering that the consolidations are dense foci in the lungs which are normally filled with air, we assume that the cluster with the lower intensity centroid corresponds to the patterns of the normal lung field parenchyma, and that the cluster with the higher intensity centroid corresponds to the consolidation patterns.

3 Results and Discussion

For the evaluation of the proposed approach, we used a collection of chest radiographs from twenty patients. The radiographic images were 8-bit grayscale with a size of 2816×2112 pixels. The lung fields were isolated using the methodology proposed in [5] and were divided into 32×32 sub-images. From each sub-image an intensity histogram signature was calculated so as to build the data matrix that was used as input to the NMF. A representative chest radiograph along with the isolated lung fields are illustrated in Fig. 1.



(a) (b)

Fig. 1. (a) A chest radiographic image, (b) the lung fields isolated from (a). The magnified area illustrates the sub-images considered. Consolidation areas are visible as bright shadows within the lung fields.

Each column in the initial data matrix, **V**, corresponds to the histogram information of each window. The resulting non-negative matrices, **W** and **H**, represent the feature bases and their membership probabilities accordingly. Both **W** and **H** were normalized by following the procedure described in the methodology section. Such a normalization allows to easily compare the bases with the initial feature vectors on the one hand, while on the other hand it leads to an easier interpretation of the probability of a signature to belong to a certain cluster or category.

To evaluate the performance of the proposed approach we applied the proposed as well as the conventional NMF-based approach on each radiographic image. Prior to the application of the algorithms, images were annotated by an expert so as to provide us with the necessary ground truth information. The results obtained are compared with the performance obtained with the fuzzy c-means (FCM) algorithm which is considered as a baseline method [12].

The performance measures considered in this study are: sensitivity, specificity and accuracy [23],

$$Sensitivity = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

$$Specificity = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \tag{7}$$

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{8}$$

where TP (true positive), TN (true negative), FP (false positive) and FN (false negative) are estimated as follows:

$$\begin{aligned} &TP=GTP \cap PCLA,\ TN=GTN \cap NCLA, \\ &FP=GTN \cap PCLA,\ FN=GTP \cap NCLA \end{aligned} \tag{9}$$

where PCLA (positive cluster lung area) is the area corresponding to the patterns considered as consolidations, NCLA (negative cluster lung area) is the area corresponding to the patterns considered as normal lung parenchyma, and GTP and GTN are the ground truth areas of consolidations and normal lung parenchyma, respectively.

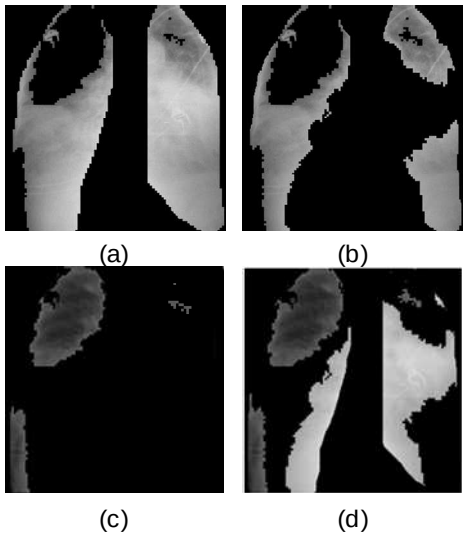


Fig. 2. Mining patterns of infections with the conventional NMF (left) and FCM (right) approaches using two clusters. (a) NMF first cluster, (b) FCM first cluster, (c) NMF second cluster, (d) FCM second cluster.

The formation of the clusters from the dataset derived from the image in Fig.1 is illustrated in Fig.2 for the 2 clusters case and for both NMF (on the left) and FCM (on the right). Figure shows that NMF achieves better separation of the consolidated areas (top left image), in contrast to FCM that fails to separate the consolidated areas from the normal ones. However, the separation of the consolidated from the normal areas is not always feasible using two clusters. An example is provided in Fig. 3 where the clustering of the lung fields in Fig.3(a) in two clusters results in an accuracy that does not exceed 40%. To cope with this problem, clustering in more than two clusters followed by a cluster merging scheme is proposed.

According to this approach the image signatures are initially clustered into an even number of clusters. Considering that the NMF bases actually represent the cluster centroids, the clusters are dyadically merged down to two based on the similarity of their centroids. Since the signatures are intensity histograms the similarity is evaluated by the histogram intersection metric [22].

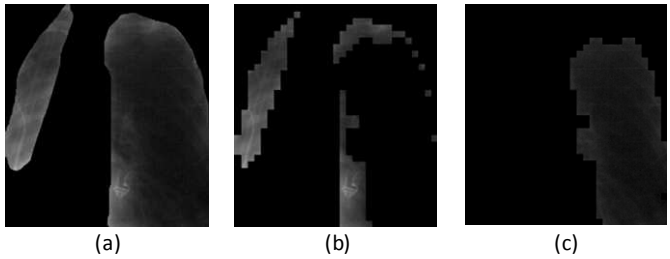


Fig. 3. Mining patterns of infections with the conventional NMF clustering approach using two clusters. (a) The lung fields to be clustered, (b) first cluster (consolidation areas), c) second cluster (normal lung field parenchyma).

An example of the application of this merging procedure is illustrated in Fig.4. The feature vectors (NMF bases) of the four initial clusters are graphically depicted above the upper row of images, which illustrates the image regions assigned to each cluster. The four clusters are subsequently merged down to two clusters, which are visualized in the last row of images in the figure. The consolidated areas are spotted in clusters 1 and 2. It is obvious that the proposed merging scheme achieves a better separation of the consolidation areas in contrast to the conventional clustering in two target clusters, which is quantified to an 87% of accuracy.

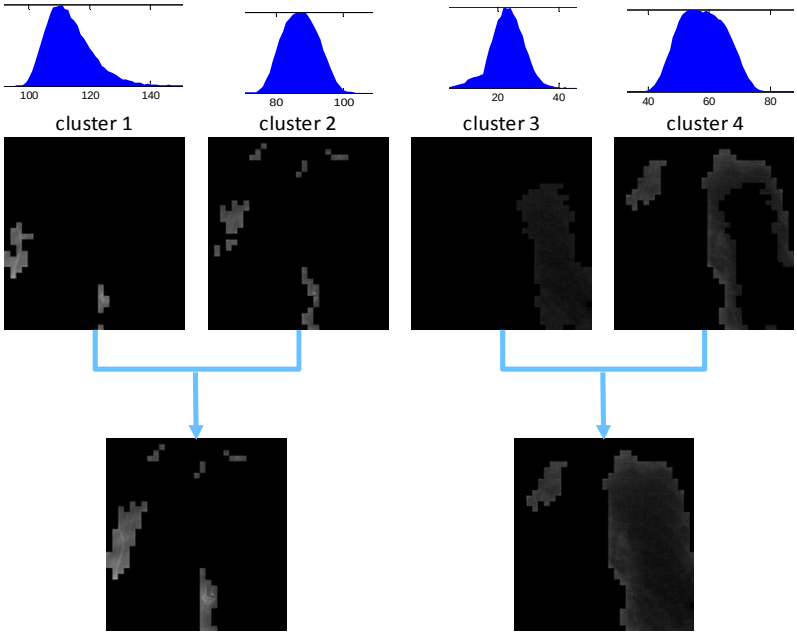


Fig. 4. Mining patterns of infections with the proposed approach. The resulting NMF bases after clustering the dataset derived from Fig.3(a) in 4 clusters (top row), the formation of the 4 clusters (top row) and the resulting merged clusters (bottom row).

The average results estimated from the application of the proposed approach on the whole dataset are summarized in Fig. 5. The average accuracy achieved by the NMF followed by cluster merging is 75%, whereas the accuracy achieved by the direct NMF clustering into two clusters is significantly lower reaching only 35%. It can be noticed that the average accuracy obtained with the FCM is poorer. Though, its sensitivity is much higher than the one obtained with the NMF after cluster merging, NMF provides much higher specificity and accuracy leading to an overall better performance. As it is illustrated in the figure, the accuracy obtained with the FCM is about 29% in the direct clustering case and 61% for the cluster merging case. Comparing the results of the proposed approach with the results of the FCM clustering with and without cluster merging as illustrated in Fig. 5, it becomes evident that the proposed approach is more suitable than the FCM for the particular clustering task.

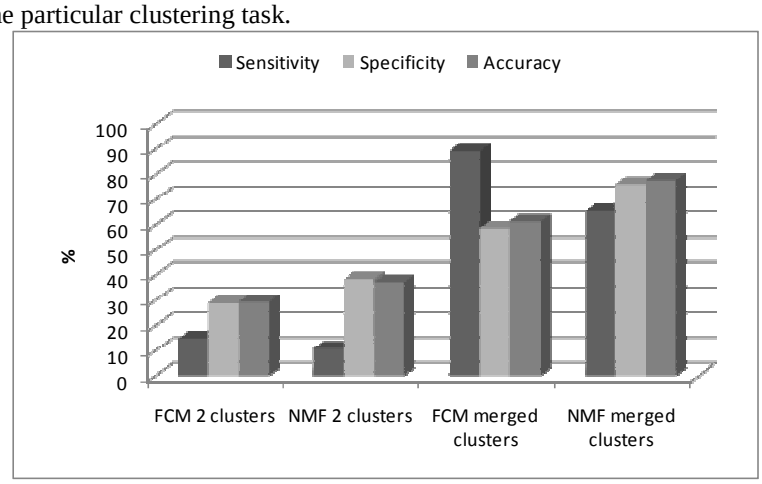


Fig. 5. Performance of the proposed cluster merging method in terms of sensitivity, specificity and accuracy.

4 Conclusions and Future Work

This study presented a novel approach to the discovery of patterns of bacterial pulmonary infections. The proposed approach is based on non-negative matrix factorization of statistical intensity signatures followed by a cluster merging scheme. The proposed approach was experimentally evaluated on radiographic images of patients with bacterial infection manifested as foci of consolidation. The experimental evaluation of the proposed technique demonstrates the superiority of the proposed NMF-based algorithm over the conventional NMF clustering scheme and the standard FCM for non-negative image data.

Currently the improvement of the proposed methodology is considered and our effort is made towards the development of an intelligent system for discovery

and assessment of pulmonary infections from radiographic images. Our future work involves further experimentation with the proposed and alternative cluster merging schemes, comparisons with state of the art unsupervised and supervised approaches, and utilization of various image features.

References

1. Irene M. Mullins, Mir S. Siadaty, Jason Lyman, Ken Scully, Carleton T. Garrett, W. Greg Miller, Rudy Muller, Barry Robson, Chid Apte, Sholom Weiss, Isidore Rigoutsos, Daniel Platt, Simona Cohen, William A. Knaus, Data mining and clinical data repositories: Insights from a 667,000 patient data set, Computers in Biology and MedicineVolume 36, Issue 12, , December 2006, Pages 1351-1377
2. D.L. Smith, J. Dushof, E.N. Perencevich, A.D. Harris, S.A. Levin, "Persistent Colonization and the spread of antibiotic resistance in nosocomial pathogens: Resistance is a regional problem," PNAS, vol. 101, no. 10, pp. 3709-3714, Mar. 2004
3. C. Lovis, D. Colaert, V.N. Stroetmann, "DebugIT for Patient Safety - Improving the Treatment with Antibiotics through Multimedia Data Mining of Heterogeneous Clinical Data," Stud Health Technol. Inform., vol. 136, 641-646, 2008
4. B.V. Ginneken, B.T.H. Romeny, and M.A. Viergever, "Computer-Aided Diagnosis in Chest Radiography: A Survey," IEEE Transactions Medical Imaging, vol. 20, no. 12, pp. 1228-1241, Dec. 2001
5. D.K. Iakovidis, and G. Papamichalis, "Automatic Segmentation of the Lung Fields in Portable Chest Radiographs Based on Bézier Interpolation of Salient Control Points," in Proceedings IEEE International Conference on Imaging Systems and Techniques, Chania, Greece, 2008, pp. 82-87
6. I.C. Mehta, Z.J. Khan, and R.R. Khotpa, Volumetric Measurement of Heart Using PA and Lateral View of Chest Radiograph, S. Manandhar et al. (Eds.): AACC 2004, LNCS 3285, pp. 34–40, 2004
7. M. Loog, B.van Ginneken: Segmentation of the posterior ribs in chest radiographs using iterated contextual pixel classification. IEEE Transactions Medical Imaging 25(5): 602-611 (2006)
8. Giuseppe Coppini, Stefano Diciotti, Massimo Falchini, N. Villari, Guido Valli: Neural networks for computer-aided diagnosis: detection of lung nodules in chest radiograms. IEEE Transactions on Information Technology in Biomedicine 7(4): 344-357 (2003)
9. B.V. Ginneken, S. Katsuragawa, B.T.H. Romeny, K. Doi, and M.A. Viergever, "Automatic Detection of Abnormalities in Chest Radiographs Using Local Texture Analysis", IEEE Transactions Medical Imaging, vol. 21, no. 2, pp. 139-149, Feb. 2002
10. X. Xie, X. Li, S. Wan, and Y. Gong, Mining X-Ray Images of SARS Patients, G.J. Williams and S.J. Simoff (Eds.): Data Mining, LNAI 3755, pp. 282–294, 2006
11. L.L.G. Oliveiraa, S. Almeida e Silvaa, L.H. Vilela Ribeirob, R. Maurício de Oliveiraa, C.J. Coelhoc and A.L.S.S. Andrade, "Computer-Aided Diagnosis in Chest Radiography for Detection of Childhood Pneumonia", International Journal of Medical Informatics, vol. 77, no. 8, pp. 555-564, 2007J.

12. C. Bezdek, J. Keller, R. Krisnapuram, and N.R. Pal. Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Kluwer Academic Publishers, 1999
13. Paatero and U. Tapper. Positive matrix factorization: a nonnegative factor model with optimal utilization of error estimates of data values. Environmetrics, 5(1):111–126, 1994
14. D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," Advanced Neural Information Processing Systems, 13, 2000, pp. 556–562
15. Chris D, XiaoFeng H, Horst D.S. On the equivalence of nonnegative matrix factorization and spectral clustering, Proceedings SIAM International Conference on Data Mining (SDM'05), 2005: 606–610
16. C. Ding, X. He, H.D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," Proceedings SIAM International Conference on Data Mining, Newport Beach, CA, April 2005, pp. 606–610
17. Xiong H.L, Chen X.W. Kernel-based distance metric learning for microarray data classification, BMC Bioinformatics, 2006, 7
18. Xu W, Liu X, Gong Y. Document clustering based on non-negative matrix factorization, Proceedings ACM Conference Research Development in Information Retrieval, 2003: 267–273
19. Yuan G, George C. Improving molecular cancer class discovery through sparse non-negative matrix factorization, Bioinformatics, 2005, vol 21, no.21:3970–3975
20. Ding, C., Li, T., Peng, W. On the equivalence between Non-negative Matrix Factorization and Probabilistic Latent Semantic Indexing (2008) Computational Statistics and Data Analysis, 52 (8), pp. 3913-3927.
21. Novelline, R.A. (1997) Squires's Fundamentals of Radiology. Cambridge: Harvard University Press
22. M.J. Swain, D.H. Ballard, Color Indexing. Int. J. Computer Vision, Vol. 7, No. 1, pp. 11–32, Nov. 1991
23. Han, J., Kamber, M., 2001. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers.