The First International Symposium on Optimization and Systems Biology (OSB'07)
Beijing, China, August 8–10, 2007
Copyright © 2007 ORSC & APORC, pp. 242–249

Two Improvements of NMF Used for Tumor Clustering*

Zhong-Yuan Zhang† Xiang-Sun Zhang‡

Institute of Applied Mathematics, Academy of Mathematics and Systems Science
Chinese Academy of Sciences, Beijing 100080, China

Abstract Non-negative Matrix Factorization (NMF) is one of the promising methods used in data mining, such as clustering human tumor samples into different types or subtypes based on microarray technology. In this paper we briefly review this method, especially when it is used for tumor clustering, and present two small but effective improvements.

Keywords NMF; Microarray; Tumor classification.

1 Introduction

It has been observed that tumors that have similar histopathological appearance may follow significantly different clinical courses and show different responses to therapy, so based primarily on morphological appearance one may make an erroneous diagnosis. Microarray technology, as a mark of the advent of the systems biology, makes it possible to classify tumor samples based on gene expressions and thus has been widely used in systems biology and iatrology. Many methods from statistical and machine learning area have been applied for this purpose such as Hierarchical Clustering (HC, [9], [3], [22]), Self-Organizing Mapping (SOM, [26], [12]) for clustering and k-Nearest Neighbor (k-NN), Support Vector Machine (SVM, [11]) for classification. But the characteristics of gene expression data have presented new challenges for many traditional statistical and machine learning methods. First, gene expression data have very high dimensionality in feature (gene) space. On the contrary, the dimensionality of observation (sample) space is very low. In short, the abundant information we get is along 'the wrong dimension'. Finally, the high noise level of the data requires more robust methodology. Non-negative matrix factorization (NMF) is a rising methodologies to cope with these difficulties [29]. Many studies have shown that it outperforms other methods. In fact the last ten years have witnessed its boom in many fields such as bioinformatics ([5], [8], [10], [14]), physics ([25]), multimedia data ([6]), text mining ([20], [28]), etc. since it was first presented

*This work was partially supported by the National Natural Science Foundation of China under grant No.10631070, and the Ministry of Science and Technology, China, under grant No.2006CB503905.

†Email: zhzhyuan@amss.ac.cn.

‡Corresponding author. Email: zxs@amt.ac.cn.

in [19][16]. One of the most interesting applications of NMF is to cluster data, i.e. discovering patterns automatically from data. The NMF clustering property is studied in Ding et al. ([7]) who proved that NMF is equivalent to K-means clustering, one of the most popular clustering method. In this paper we briefly review the method from the nonlinear programming research point of view and present two small but effective improvements.

2 Methods

2.1 A brief review of NMF

Mathematically, Non-negative Matrix Factorization (NMF) can be described as follows: given an $n \times m$ matrix V composed of non-negative elements where $n \gg m$, our task is to factorize V into a non-negative matrix W of size $n \times r$ and another non-negative matrix H of size $r \times m$ such that $V \approx WH$. r is preassigned and should satisfy the principle $r < nm/(n+m)$. W and H can be explained variously in different fields, for specific purposes or even by different persons.

In short, the derived algorithm of NMF is as follows:

Step 1: Randomize W and H with positive numbers in [0, 1]. Select a cost function to be minimized.
Step 2: With W fixed, update H, then update W for the updated H. Iterate until the process converges.

The cost function is frequent $D_1(V,WH) = ||V - WH||_F^2$ or the generalized Kullback- Leibler divergence $D_2(V,WH) = \sum_{i,j}(V_{ij}logV_{ij}/(WH)_{ij} - V_{ij} + (WH)_{ij})$. When D_1 is used, the update formulae of H and W are

$$H_{au} := H_{au}\frac{(W^TV)_{au}}{(W^TWH)_{au}}, \tag{1}$$

$$W_{ia} := W_{ia}\frac{(VH^T)_{ia}}{(WHH^T)_{ia}}. \tag{2}$$

Otherwise, if D_2 is used, the corresponding formulae can be written as:

$$H_{au} := H_{au}\frac{\sum_i (W_{ia}V_{iu})/(WH)_{iu}}{\sum_k W_{ka}}, \tag{3}$$

$$W_{ia} := W_{ia}\frac{\sum_u H_{au}V_{iu}/(WH)_{iu}}{\sum_v H_{av}}. \tag{4}$$

All these expressions are obtained via the gradient decent method in nonlinear programming. We take (1) and (2) as an example to demonstrate the reasoning process.

Firstly, the derivative of the cost function $D_1(V,WH) = ||V - WH||_F^2$ with respect to H is:

$$\frac{\partial}{\partial h_{au}} D_1(V,WH) = -\sum_i (V_{iu} - (WH)_{iu}) W_{ia}.$$

Let the step size be $\alpha_{au} = H_{au}/W^T(WH))_{au}$, then

$$H_{au} = H_{au} - \alpha \frac{\partial}{\partial h_{au}} D_1(V,WH) = H_{au}(\frac{(W^TV)_{au}}{(W^TWH)_{au}}).$$

By reversing the roles of the W and H, one can easily get (2) as the update rule of W.

Local minimum is guaranteed. People who are interested in the theoretical aspect of NMF can get more information from [17].

NMF has been widely used in bioinformatics, especially in clustering tumor samples based on microarray experiments. Microarray is a new and developing technique and its data can be represented as a matrix A of size $n \times m$ whose rows contain the expression levels of the genes across m samples (m time points or m conditions). r is the class number of tumor samples, the cost function D_2 is selected because it has better numerical result than D_1 has. W, H are obtained using (3) and (4). Each column of W is defined as a metagene, so in fact, metagenes are the linear combination of the measured genes. The component of W denotes the weight of the corresponding gene in the metagene. Each row of H is viewed as the expression level of the metagene across different samples. The clustering method using NMF is based on a hypothesis, that is:

Hypothesis: *The metagenes should have similar expression patterns in the samples which belong to the same class.*

Under this hypothesis, samples can be clustered according to metagenes expression patterns, in other words, sample j is clustered into class i if h_{ij} is the largest value of the column i of H. This means that the metagene i is the most active in sample j. One can refer to [5] for more details.

2.2 Two improvements for NMF

As we can see, the step-size α is not selected through linear search, so it is not necessarily the best one. We multiply α by a scalar β, where $\beta \in (0,1]$, thus we can have more choices. Then the corresponding update rules become:

$$H_{au} := H_{au}(1 - \beta_2 + \frac{\beta_2 (W^TV)_{au}}{(W^TWH)_{au}}), \tag{1'}$$

$$W_{ia} := W_{ia}(1 - \beta_1 + \frac{\beta_1 (VH^T)_{ia}}{(WHH^T)_{ia}}). \tag{2'}$$

$$H_{au} := H_{au}(1-\beta_2+\beta_2\frac{\sum_i (W_{ia}V_{iu})/(WH)_{iu}}{\sum_k W_{ka}}), \tag{3'}$$

$$W_{ia} := W_{ia}(1-\beta_1+\beta_1\frac{\sum_u H_{au}V_{iu}/(WH)_{iu}}{\sum_v H_{av}}). \tag{4'}$$

Local minimum is guaranteed since the cost function $D_1(V,WH)$ and $D_2(V,WH)$, as $W's$ or $H's$, are convex, and the convergence in the case of $\beta_1 = 1, \beta_2 = 1$ has been proved[17][15]. Later numerical result shows that $\beta_1 = .5, \beta_2 = 1$ is a good choice.

Another disadvantage of NMF is that it is time-consuming which is mainly because of the high dimension of W. But as a matter of fact, in many computation cases we don't need to know W at all and people can easily observe that $V^TV = H^TW^TWH$, in other words, $K = H^TSH$ where $S = W^TW, K = V^TV$. the update rules for $D_1(K, H^TSH)$ is [7]:

$$S_{ik} := S_{ik}(1-\beta_3+\beta_3\frac{(HKH^T)_{ik}}{(HH^TSHH^T)_{ik}}), \tag{5}$$

$$H_{ik} := H_{ik}(1-\beta_4+\beta_4\frac{(SHK)_{ik}}{(SHH^TSH)_{ik}}). \tag{6}$$

Its effectiveness, especially for relatively small dataset, will be shown in the next section.

3 Application

3.1 Assess Standard

Purity has been widely used in data mining to assess the quality of clustering result which can be defined as follows:

Definition: Purity=$\sum_{i=1}^{K}\frac{n_iP(S_i)}{n}$ where K is the number of clusters, n is the number of data points (samples), n_i is the size of the i-th implanted class denoted by S_i, $P(S_i) = \frac{1}{n_i}max_j(n_i^j)$ where n_i^j is the number of samples of the i-th implanted class that are assigned to the j-th computed cluster.

As one can see, if the clustering result matches the implanted class structures exactly, the purity is one. In general, the purity measures the extent to which each cluster contains the samples from one of the implanted class, the larger the purity, the better the clustering result is.

3.2 Dataset

Six datasets are used to verify our improvements, the result shows that $\beta_1 = .5, \beta_2 = 1, \beta_3 = .5, \beta_4 = 1$ is strongly recommended.

ALL-AML

This dataset, as a golden standard in the cancer classification community, includes two types of human tumor-acute myelogenous leukemia (AML, 11 samples) and acute lymphoblastic leukemia (ALL, 27 samples). Also ALL can be divided into two subtypes-ALL-T (8 samples) and ALL-B (19 samples) [5].

Central Nervous System (CNS)

This dataset comes from [23] which consists of 34 samples: 10 classical medulloblastomas, 10 malignant, gliomas, 10 rhabdoids and 4 normals.

Lung cancer (LC)

This dataset, composed of 181 samples, is from [13] which is about malignant pleural mesothelioma (MPM, 31 samples) and adenocarcinoma (ADCA, 150 samples) of the lung .

Subtypes of Acute Lymphoblastic Leukemia

This dataset is including six prognostically important eukemia subtypes: T-ALL, E2A-PBX1, BCR-ABL, TEL-AML1, MLL, hyperdiploid>50 chromosomes. We select E2A-PBX1 (18 samples), MLL (14 samples), T-ALL (28 samples) as one test dataset, and E2A-PBX1 (18 samples), Hyperdiploid>50 (42 samples), T-ALL (28 samples), TEL-AML1 (52 samples) as another.

The original data contains about 12000 genes. In our experiment, the genes are ranked according to their coefficient of variation (i.e., standard deviation divided by the mean) and the top 8000 are selected.

All these data can be obtained directly from [4].

3.3 Result

The following six tables show the computational results, from which we can see that $\beta_1 = .5, \beta_2 = 1$ is consistently better. Another six tables to illustrate β_3, β_4 are omitted, where again $\beta_3 = .5, \beta_4 = 1$ is better, especially when the dataset is relatively small.

W	H	purity (%)
$\beta_1 = 1$	$\beta_2 = 1$	94.12
$\beta_1 = .5$	$\beta_2 = .5$	94.12
$\beta_1 = .5$	$\beta_2 = 1$	**97.06**
$\beta_1 = 1$	$\beta_2 = .5$	94.12

Table 1: CNS

W	H	purity (%)
$\beta_1 = 1$	$\beta_2 = 1$	94.74
$\beta_1 = .5$	$\beta_2 = .5$	94.74
$\beta_1 = .5$	$\beta_2 = 1$	**100**
$\beta_1 = 1$	$\beta_2 = .5$	94.74

Table 2: AML/ALL, k=2

W	H	purity (%)
$\beta_1 = 1$	$\beta_2 = 1$	94.74
$\beta_1 = .5$	$\beta_2 = .5$	94.74
$\beta_1 = .5$	$\beta_2 = 1$	**97.37**
$\beta_1 = 1$	$\beta_2 = .5$	94.74

Table 3: AML/ALL, k=3

W	H	purity (%)
$\beta_1 = 1$	$\beta_2 = 1$	93.92
$\beta_1 = .5$	$\beta_2 = .5$	92.62
$\beta_1 = .5$	$\beta_2 = 1$	**95.03**
$\beta_1 = 1$	$\beta_2 = .5$	90.61

Table 4: Lung Cancer

W	H	purity (%)
$\beta_1 = 1$	$\beta_2 = 1$	90
$\beta_1 = .5$	$\beta_2 = .5$	90
$\beta_1 = .5$	$\beta_2 = 1$	**91.67**
$\beta_1 = 1$	$\beta_2 = .5$	88.33

Table 5: subtypes, k=3

W	H	purity (%)
$\beta_1 = 1$	$\beta_2 = 1$	95.71
$\beta_1 = .5$	$\beta_2 = .5$	95.71
$\beta_1 = .5$	$\beta_2 = 1$	**96.43**
$\beta_1 = 1$	$\beta_2 = .5$	95.71

Table 6: subtypes,k=4

The reason why the result of $\beta_1 = 0.5, \beta_2 = 1$ is better than that of $\beta_1 = \beta_2 = 1$ is as follows: although the original NMF can maintain the positive property of W and H, this doesn't mean that the method can converge to the global optimal solution, in fact, only the local minimum is guaranteed. From the numerical test, we can see that $\beta_1 = 0.5, \beta_2 = 1$ is better.

As to the second improvement, we can explain it from the point of view of computational complexity: in each step, the computational complexity of (5) is of the order $m^2r + 3r^2m + r^3$ and that of (6) is of the order $2r^2m + mr^2$, while the order of equations (1') and (2') is $mnr + nr^2 + r^2m$ where $n \gg m$.

4 Discussion

Obviously, some information of V has been lost when we factorize V^TV to get H, but this is not serious when the data size is small.

Furthermore K can be viewed as a kernel matrix or comparability matrix, then we have extended NMF from the classification on the sample matrix to classification on the distance matrix. Thus it can be used in many fields, for example, the detection of community structure of network.

Acknowledgement

The authors are very grateful to Professor Chris Ding for his insightful suggestions.

References

[1] Alon U, Barkai N, Notterman D. A, et al. *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays*, Proc. Natl. Acad. Sci. USA, 1999, 96: 6745-6750.

[2] Aisner J.D. *Staging and natural history of pleural mesothelioma.* In J.Aisner, R. Arriagada, M. R. Green, N. Martini, and M. C. Perry (eds.), Comprehensive Textbook of Thoracic Oncology. Baltimore: Williams and Wiikims, 1996.

[3] Alizadeh A. A, Eisen M. B, Davis R. E, Ma C, Lossos I. S, Rosenwald A, Boldrick J. C, Sabet H, Tran T, Yu X, et al. *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling*, Nature 2000, 403, 503–511.

[4] Bio-medical Data Analysis [http://sdmc.lit.org.sg/GEDatasets/].

[5] Brunet J.P, Tamayo P, Golub T.R, MesirovJill P. *Metagenes and molecular pattern discovery using matrix factorization*, Proc. Natl. Acad. Sci. USA, 2004, 101:4164–4169.

[6] Cooper M, Foote J. *Summarizing video using non-negative similarity matrix factorization*, Proc. IEEE Workshop on Multimedia Signal Processing, 2002: 25-28.

[7] Chris D, XiaoFeng H, Horst D.S. *On the equivalence of nonnegative matrix factorization and spectral clsutering*, Proc. SIAM Int'l Conf. Data Mining (SDM'05), 2005: 606–610.

[8] Carmona-Saez P, Pascual-Marqui R.D, Tirado F, Carazo J.M, Pascual-Montano A. *Biclustering of gene expression data by non-smooth non-negative matrix factorization*, BMC Bioinformatics, 2006.

[9] Eisen M, Spellman P, Brown P, and Botstein D. *Cluster analysis and display of genome-wide expression patterns*, Proc. Natl. Acad. Sci. USA 1998, 95, 14863–14868.

[10] Fogel P, Young S.S, Hawkins D.M, Ledirac N, *Inferential, robust non-negative matrix factorization analysis of microarray data*, Bioinformatics, 2007, Vol.23. No.1, 44-49.

[11] Furey T.S, Cristianini N, DuffyN, Bednarski D.W, Schummer M, and Haussler D. *Support vector machine classification and validation of cancer tissue samples using microarray expression data*, Bioinformatics 2000, 16:906–914.

[12] Golub T.R, et al. *Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring*, Science 1999, 286: 531–537.

[13] Gordan J.G, Jenson R.V, Hsiao L, et al. *Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma*, Cancer Research, 2002, 62:4963–4967.

[14] Heger A, Holm L.*Sensitive pattern discovery with 'fuzzy' alignments of distantly related proteins*, Bioinformatics, 19, Suppl., i130–i137.

[15] Inderjit S. D, Suvrit S. *Generalized Nonnegative Matrix Approximations with Bregman Divergences*, Neural Information Processing Systems (NIPS), Vancouver, Canada, 2005.

[16] Lee D.D, Seung H. S. *Learning the parts of objects by non-negative matrix factorization*, Natrue 1999, 401: 788–791.

[17] Lee D.D, Seung H. S. *Algorithms for non-negative matrix factorization*, In Advances in Neural Information Processing Systems, 2001, vol 13, 556–562.

[18] Ordonez N.G. *The immunohistochemical diagnosis of epithelial mesothelioma*, Hum. Pathol., 1999, 30:313–323.

[19] Paatero P, Tapper U. *Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values*, Environmetrics, 1994, 5:111–126.

[20] Pauca V.P, Shahmaz F, Berry M.W, Plemmons R.J. *Text mining using non-negative matrix factorization*, Proc. SIAM Int'l conf on Data Mining, 2004:452–456.

[21] Philip M. K, Bruce T. *Subsystem identification through dimensionality reduction of large-scale gene expression data* , Genome Research, 2003, 13: 1706–1718.

[22] Perou C. M, Sorlie T, Eisen M. B, et al. *Molecular portraits of human breast tumors*, Nature 2000, 406, 747–752.

[23] Scott A. A, Jane E. S, Lewis B. S, et al. *Prediction of central nervous system embryonal tumor outcome based on gene expression*, Nature, 2002, 415:436–442.

[24] Scott A. A, Jane E. S, Lewis B. S, et al. *MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia*, Nature, 2002, 30:41–47.

[25] Stefan W, James C, Anne D. *Motivating Non-Negative Matrix Factorization*, Proceedings of the Eighth SIAM Conference on Applied Linear Algebra, Williamsburg, VA, July 15–19, 2003.

[26] Tamayo P, Slonim D, Mesirov J, Zhu Q, Dmitrovsky E, Lander E. S. and Golub T. R. *Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation*, Proc. Natl. Acad. Sci. USA 1999, 96, 2907–2912.

[27] Xiong H.L, Chen X.W. *Kernel-based distance metric learning for microarray data classification*, BMC Bioinformatics, 2006, 7.

[28] Xu W, Liu X, Gong Y. *Document clustering based on non-negative matrix factorization*, Proc.ACM Conf.Research Development in IR, 2003: 267–273.

[29] Yuan G, George C. *Improving molecular cancer class discovery through sparse non-negative matrix factorization*, Bioinformatics, 2005, vol 21, no.21:3970–3975.