

# CANCER MOLECULAR PATTERN DISCOVERY BY SUBSPACE CONSENSUS KERNEL CLASSIFICATION

Xiaoxu Han

*Department of Mathematics and Bioinformatics Program, Eastern Michigan University  
Ypsilanti, MI 48197, USA  
xiaoxu.han@emich.edu*

Cancer molecular pattern efficient discovery is essential in the molecular diagnostics. The characteristics of the gene/protein expression data are challenging traditional unsupervised classification algorithms. In this work, we describe a subspace consensus kernel clustering algorithm based on the projected gradient nonnegative matrix factorization (PG-NMF). The algorithm is a consensus kernel hierarchical clustering (CKHC) method in the subspace generated by the PG-NMF. It integrates convergence-soundness parts-based learning, subspace and kernel space clustering in the microarray and proteomics data classification. We first integrated subspace methods and kernel methods by following our framework of the input space, subspace and kernel space clustering. We demonstrate more effective classification results from our algorithm by comparison with those of the classic NMF, sparse-NMF classifications and supervised classifications (KNN and SVM) for the four benchmark cancer datasets. Our algorithm can generate a family of classification algorithms in machine learning by selecting different transforms to generate subspaces and different kernel clustering algorithms to cluster data.

## 1. INTRODUCTION

With the development of genomics and proteomics, Molecular diagnostics has appeared as a new tool to diagnose cancers. It picks a patient's tissues or blood samples and uses DNA microarray or mass spectrometry (MS) based proteomics techniques to generate their gene expressions or protein expressions. The gene/protein expressions reflect gene/protein activity patterns in different types of cancerous or precancerous cells. They are molecular patterns or molecular signatures of cancers. Different cancers will have different molecular patterns and the molecular patterns of a normal cell will be different from those of a cancer cell. Clinicians identify the potential biomarkers by analyzing the gene/protein patterns. However, robustly classifying cancer molecular patterns is still a challenge for clinicians and bioinformaticans.

Many classification methods from statistical and machine learning are proposed for cancer molecular pattern classification. These methods can be generally classified as supervised classification methods, such as k-nearest neighborhood (kNN), linear discriminant analysis (LDA), neural networks (NN), support vector machines (SVM);<sup>1-3</sup> unsupervised classification (clustering) methods, such as hierarchical clustering

(HC), self-organizing maps (SOM), principal component analysis (PCA); and their variants, such as particle swarm optimization support vector machines (PSO-SVM), kernel principal component analysis (KPCA) etc.<sup>4-7</sup> We are particularly interested in the unsupervised molecular pattern discovery algorithms, because they do not need or have prior knowledge about data. They also have potentials to explore the latent structure of data. However, the traditional clustering algorithms: HC and SOM were already proved unstable for gene and protein expression data although they are widely used in the cancer molecular pattern discovery community.<sup>4,8,15</sup>

Actually, the characteristics of gene and protein expression data are challenging the traditional unsupervised classification algorithms. These high dimensional data can be represented by an  $n \times m$  matrix after preprocessing. The row data in the matrix are the expression levels of a gene across different experiments or intensity values of a measured data point in different samples (observations) corresponding to an m/z ratio. The column data are the gene expression levels of a genome under an experiment or intensity values of all measured data points in a sample corresponding to m/z ratios. Usually,  $n \gg m$ ; that is, the number of variables



under a rank  $r$  by the multiplicative update algorithm, i.e. each observation is represented as the linear combination of bases by Eq. (1), where  $h_{ij}$  is the  $i$ -th element of the  $H_j$ , which is the prototype of the  $j$ -th observation  $X_j$  after feature selection. Second, clustering is conducted by the following query asked by each sample: ‘which basis has the largest expression level in my prototype? I will belong to the cluster associative with that basis’. For example, suppose  $h_{ij}$  is the largest value in  $H_j$ , then sample  $X_j$  will be assigned to the cluster  $i$  because the  $i^{\text{th}}$  basis has the largest expression level in its prototype  $H_j$ . The number of clusters is just the decomposition rank  $r$ . Finally, the rank leading to the most meaningful clustering is decided by a Monte Carlo based model selection mechanism by finding a rank with the maximum cophenetic correlation coefficient in the hierarchical clustering. The cophenetic correlation coefficient is a measure to evaluate the stability of hierarchical clustering. It is the correlation between the pairwise distance and linkage distance in the hierarchical clustering. A large cophenetic correlation coefficient value will indicate the high stability of a hierarchical clustering.

Brunet *et al* proved this method was superior to HC and SOM methods for three benchmark cancer datasets.<sup>15</sup> Inspired by this work, Gao and Church developed a sparse nonnegative matrix factorization to cluster the cancer samples by adding sparseness control in the basic NMF formulation (sparse-NMF).<sup>16, 17</sup> They demonstrated the sparse-NMF based clustering was superior to the basic NMF clustering method for the same datasets.

However, Brunet *et al*’s NMF based clustering method has following weak points. 1. The multiplicative update algorithm in the NMF lacks the convergence soundness. The model selection mechanism in the NMF clustering is expensive, because it requires to compute cophenetic correlation coefficients for the hierarchical clustering conducted at all possible ranks to decide the final optimal decomposition rank.

## 1.2. Contributions

In this study, we describe a subspace consensus kernel clustering technique based on the projected gradient nonnegative matrix factorization (PG-NMF), which was developed by Lin,<sup>14</sup> to conduct cancer molecular pattern classification for microarray and proteomics data. The projected gradient nonnegative matrix factorization (PG-

NMF) has sound convergence and converges faster than the basic NMF.<sup>14</sup> In addition, we present the ideas of input space, subspace and kernel space clustering before elaborating on our PG-NMF based classification method under the framework of subspace and kernel space clustering.

The idea of our method is to transform a gene/protein expression data set  $X \in \mathfrak{R}^n$  into a subspace  $S \subset \mathfrak{R}^n$  by using the PG-NMF algorithm. Then, a consensus kernel hierarchical clustering algorithm (CKHC) is developed to cluster the projections of a dataset  $X$  in the subspace  $S$  to infer the latent structure of the data. We have showed that the PG-NMF based subspace kernel clustering (PG-NMF-CKHC) is superior to the basic NMF, sparse-NMF clustering and supervised clustering (KNN and SVM) in the cancer molecular pattern discovery for four benchmark cancer datasets.

This paper is organized as follows. Section 2 presents the concepts of input space, subspace and kernel space clustering before introducing our PG-NMF based consensus kernel hierarchical clustering in the section 3. Section 4 shows the experimental results of our algorithm. Finally, we discuss the possible algorithm generalizations and draw conclusions.

## 2. INPUT SPACE, SUBSPACE AND KERNEL CLUSTERING

For a given data set  $X = (x_1, x_2, \dots, x_n)^T \in \mathfrak{R}^{n \times m}$ , clustering is to find an implicit classification function  $f: X \rightarrow \Gamma$  that maps each data sample  $x_i$ , to its target function value  $y_j$  (label) in a set  $\Gamma$  according to some dissimilarity metric ( $j = 1, 2, \dots, |\Gamma|$ ). Data samples with a same target function value (label) after classification will claim to share a same cluster.

We classify clustering as the input space, subspace and kernel space clustering according to where the implicit classification function  $f$  is computed. In the input space clustering, the implicit classification function  $f$  is computed in the input space  $\mathfrak{R}^{n \times m}$  of the dataset. Hierarchical clustering (HC), K-means clustering and expectation maximization (EM) clustering all belong to the input space clustering. In the kernel space clustering, the classification function  $f$  is computed in a kernel space  $\Omega$  of the input space, which is a high dimensional Hilbert space generated by a feature map function  $\Phi: X \rightarrow \Omega$ ,  $\dim(\Omega) \gg \dim(X)$ . That is, the clustering is conducted for the high



dissimilarity metrics in HC. The Euclidean distance between samples  $x_i$  and  $x_j$  in the kernel space can be which can be kernelized as:

$$d(\Phi(x_i), \Phi(x_j)) = (K_{ii} - 2K_{ij} + K_{jj})^{1/2} \quad (5)$$

where  $K_{ij} = K(x_i, x_j) = ((\Phi(x_i) \bullet \Phi(x_j)))$ .

In the kernelization of the correlation distance between samples  $x_i$  and  $x_j$ , we assume the mapped vectors  $\Phi(x_i), \Phi(x_j)$  are zero mean data in the kernel space  $\Omega$ , then the correlation distance between  $\Phi(x_i)$  and  $\Phi(x_j)$  can be formulated as the following inner product form in Eq. (6), where  $c_{ij} = c(\Phi(x_i), \Phi(x_j))$ .

$$c_{ij} = 1 - \frac{(\Phi(x_i) \bullet \Phi(x_j))}{(\Phi(x_i) \bullet \Phi(x_i))^{1/2} (\Phi(x_j) \bullet \Phi(x_j))^{1/2}} \quad (6)$$

However, we shall drop this assumption in the kernel space for more general practice. We use the expectation of all feature data to center each feature data,

$$\bar{\Phi}(x_i) = \Phi(x_i) - \frac{1}{m} \sum_{i=1}^m \Phi(x_i) \quad (7)$$

Then the corresponding correlation distance can be formulated as the similar form as in the Eq. (6). Let  $K'_{ij} = (\bar{\Phi}(x_i) \bullet \bar{\Phi}(x_j))$ , then we have the following result:

$$K'_{ij} = K_{ij} - \frac{1}{m} \sum_{n=1}^m K_{in} - \frac{1}{m} \sum_{l=1}^m K_{lj} + \frac{1}{m^2} \sum_{l=1}^m \sum_{n=1}^m K_{ln} \quad (8)$$

Since the kernel matrix  $K$  is a semi-positive definite matrix, summarizing previous results, we have the correlation distance in the kernel space between  $\Phi(x_i)$  and  $\Phi(x_j)$  can be computed as

$$c(\Phi(x_i), \Phi(x_j)) = \frac{(K'_{ii} K'_{jj})^{1/2} - K'_{ij}}{(K'_{ii} K'_{jj})^{1/2}} \quad (9)$$

The extension of the single, complete and average linkage in the kernel space is trivial but not for the centroid linkage. The centroid linkage between two clusters is defined as the Euclidean distance between the centroid of two clusters. We give the centroid linkage  $d_{rs}$  between the clusters  $C_r$  and  $C_s$  in the Eq. (10).

$$d_{rs} = \left( \frac{1}{|C_r|^2} \sum_{i,j=1}^{|C_r|} k_{ij}^{(r)} - \frac{1}{|C_r||C_s|} \sum_{i=1}^{|C_r|} \sum_{j=1}^{|C_s|} k_{ij}^{(r,s)} + \frac{1}{|C_s|^2} \sum_{i,j=1}^{|C_s|} k_{ij}^{(s)} \right)^{1/2} \quad (10)$$

Where  $x_i^{(r)}$  is the  $i^{\text{th}}$  sample in the cluster  $C_r$ ; The  $|C_r|, |C_s|$  are the number of samples in the clusters  $C_r$  and  $C_s$ ;  $k_{ij}^{(r)} = k(x_i^{(r)}, x_j^{(r)})$ ,  $k_{ij}^{(s)} = k(x_i^{(s)}, x_j^{(s)})$  and  $k_{ij}^{(r,s)} = k(x_i^{(r)}, x_j^{(s)})$ .

### 2.3. What's the ideal unsupervised classification algorithm for the high dimensional gene/protein expression data?

We believe that an ideal unsupervised classification or clustering algorithm for the high dimensional gene and protein data should satisfy following criteria. 1. Some feature selection methods ought to be applied to reduce data dimensions such that data are "clean and compact". 2. The feature selection method employed should have the part-base learning property to maintain the data locality well; that is, the feature selection method can conduct local feature selection. 3. Kernel tricks are desirable to be applied in the clustering of the data after feature selection to achieve better classification results in a kernel space.

According to the criteria, we give our subspace consensus kernel classification algorithm based on the projected gradient NMF (PG-NMF). The basic idea is to apply a convergent soundness local feature algorithm: PG-NMF to the gene/protein expression dataset  $X$ , which is equivalent to project the dataset  $X$  into the subspace  $S$  generated by the PG-NMF:  $X \sim WH$ , where  $W$  is the basis matrix generating the subspace. Then kernel hierarchical clustering is applied to column data the feature matrix  $H$ , which are the prototype data of the original data. Since the basis matrix and feature matrix are not unique in the NMF. We develop the consensus kernel hierarchical clustering algorithm (CKHC) to get the final classification.

### 3. PG-NMF SUBSPACE KERNEL HIERARCHICAL CLASSIFICATION

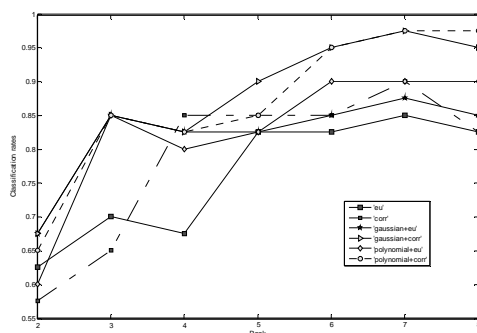
PG-NMF based subspace kernel classification is to conduct consensus kernel hierarchical clustering (CKHC) to each feature matrix  $H$  in a subspace  $S$  generated by the PG-NMF. The CKHC is an algorithm to run the kernel hierarchical clustering in a Monte Carlo simulation approach and compute the final classification by building a consensus tree. It consists of two general steps. 1. Build a consensus tree for the expression dataset  $X$  at each rank by conducting CKHC







linkage metric. Classification rates generally decrease after the rank 7 and the correlation distance generally performs better than the Euclidean distance in the classification.



**Fig. 7.** The classification rates of the PG-NMF-CKHC for this dataset: polynomial kernel + correlation distance reaches the best classification rate.

We also apply NMF and sparse-NMF classification for the proteomics data, although they were developed under the context of gene expression data. There are 8 samples misclassified from NMF clustering and 12 samples misclassified from the Sparse NMF clustering for our ovarian cancer dataset. Both algorithms indicate there are 2 clusters from their cophenetic coefficients. Since a proteomics dataset generally has much higher dimensionalities than a gene expression dataset, NMF and sparse NMF clustering have large time complexity for a proteomics dataset. For this dataset, NMF clustering takes >78 hours and sparse-NMF clustering takes >153 hours running under two PCs with 3.0 GHZ CPU and 504 RAM running under WIN-XP OS. It seems that NMF based clustering/classification mechanism can't work well in the context of the proteomics data.

#### 4.1. Comparing classification results from kNN, sparse-NMF and support vector machines (SVM)

We compare PG-NMF-CKHC for the four datasets (the leukemia, medulloblastoma, ovarian cancer dataset and a colon cancer dataset, which consists of 22 controls and 40 cancer data samples) with the classic NMF clustering, sparse-NMF clustering, and SVM and kNN

classifications. In kNN and SVM, We run classification 10 times under holdout cross-validation with 50% hold-out percentage for each case. We take the average classification rates as the final classification rates. In the SVM classification, we also use linear, polynomial and Gaussian kernel. We select the best final classification rate from three kernels as the final classification rate of SVM. In the leukemia data, we use SVM/kNN to classify ALL and AML types instead of all three types. Although the pathogenesis of medulloblastoma is not well established, we still compute the classification rates of this dataset based on the general assumption that samples are divided as 25 classic and 9 desmoplastic medulloblastomas, for the convenience of comparisons. Table 1 shows the classification rates for the four benchmark datasets from kNN, PG-NMF-CKHC, NMF, sparse-NMF and SVM classifications.

We have found that our algorithm is superior to the NMF, sparse-NMF and supervised SVM classification algorithms for these datasets; The NMF classification has better performance than SVM and kNN for three gene expression datasets. Sparse-NMF has averagely better performance than kNN for three gene expression datasets. However, the NMF and sparse-NMF can't compete with kNN and SVM for the proteomics data.

According to our classification results, it seems that sparseness constraint on the NMF may not always contribute to the improvement in the classifications for some datasets. Besides the ovarian dataset, for the medulloblastoma dataset, the classic NMF clustering seems to perform better in classifying desmoplastic medulloblastomas than the sparse-NMF clustering at rank 5, where both algorithms reaches the most robust reproducibility partitions. We also noticed the NMF and sparse-NMF clustering can not compete with SVM classification for the ovarian dataset. It is interesting to see that sparseness constraint may not lead to the better classification results for the colon cancer dataset. The classic NMF clustering reaches its largest cophenetic correlation coefficient at rank 2 (2 clusters) and its corresponding classification rate is 0.9355. However, the sparse NMF clustering reaches its largest cophenetic correlation coefficient at rank 4 (4 clusters) and its corresponding classification rate is 0.7581. It is possible due to the fact that the expression patterns of those dominant co-expressed genes such as, oncogenes, tumor suppressor genes are not extracted out in the sparse representation. This may also indicate that sparseness



## Acknowledgments

Author wants to thank the support from the New Faculty Research Award at Eastern Michigan University for this research.

## References

- Lilien, R. and Farid, H. Probabilistic Disease Classification of Expression-dependent Proteomic Data from Mass Spectrometry of Human Serum, *Journal of Computational Biology* 2003; **10** (6), 925-946.
- Golub, T. et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science* 1999; **286**: 531-537.
- Furey T., Cristianini N., Duffy N, Bednarski D., Schummer M. and Haussler D. Support vector machine classification and validation of cancer tissue samples using microarray expression data, *Bioinformatics* 2000; **16** (10): 906-914.
- Hautaniemi, S. , Yli-Harja, O., Jaakko Astola, J., Kauraniemi, P. et al. Analysis and Visualization of Gene Expression Microarray Data in Human Cancer Using Self-Organizing Maps, *Machine Learning* 2003; **52**: 45-66.
- Ressom, H., Varghese, R., Saha, D., Orvisky, R. et al. Analysis of mass spectral serum profiles for biomarker selection. *Bioinformatics* 2005; **21**: 4039-4045.
- Liu Z., Chen D. and Bensmail H. Gene expression data classification with Kernel principal component analysis. *J Biomed Biotechnol.* 2005 (2) 155-159.
- Eisen, M. et al. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* 1998; **95**: 14863-14868.
- Tamayo, P. et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA* 1999; **96**: 2907-2912.
- Bicciato, S. et al. PCA disjoint models for multiclass cancer analysis using gene expression data. *Bioinformatics* 2003; **19**: 571-578
- Wall, M., Andreas, R., Rocha, L. Singular value decomposition and principal component analysis. *A Practical Approach to Microarray Data Analysis*. Berrar, D., W. Dubitzky, W., Granzow, M. eds. Kluwer: Norwell, 2003; 91-109.
- Tan, Y., Shi, L., Tong, W., and Wang, C. Multi-class cancer classification by total principal component regression using microarray gene expression data. *Nucleic Acids Res.* 2005; **33**(1) 56-65.
- Zhang, X., Yap, Y., Wei, D., Chen, F. and Danchin, A. Molecular diagnosis of humancancer type by gene expression profiles and independent component analysis, *European Journal of Human Genetics* 2005; **1-9**: 1018-4813.
- Daniel D. Lee and H. Sebastian Seung.: Learning the parts of objects by non-negative matrix factorization. *Nature* 1999; **401**: 788-791.
- Lin, C. Projected gradient methods for non-negative matrix factorization, *Neural Computation* 2007; In Press.
- Brunet, J., Tamayo, P., Golub, T. and Mesirov., J. Molecular pattern discovery using matrix factorization. *Proc. Natl Acad. Sci. USA*, 2004, **101**,12: 4164-4169.
- Gao, Y. and Church, G. Improving molecular cancer class discovery through sparse nonnegative matrix factorization, *Bioinformatics* 2005; **21** (21):, 3970-3975.
- Patrik O. Hoyer: Non-negativematrix factorization with sparseness constraints. *Journal of Machine Learning Research* 2004, **5**: 1457-1469.
- Yeung, K. and Ruzso, W.: Principal Component Analysis for clustering gene expression data, *Bioinformatics*, 2001; **17** (9): 763-774.
- Lee, S. and Batzoglou, S. ICA-Based Clustering of Genes from Microarray Expression Data, *Neural Information Processing Systems(NIPS)* 2003.
- Vapik, V. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.
- Qin. J. et al.: Kernel hierarchical gene clustering from microarray expression data, *Bioinformatics*, 2003, **19** (16), 2097-2104.
- Mika, S., Rätsch, G., Weston, J., Schölkopf, B. and Müller, KR. Fisher discriminant analysis with kernels, *Neural Networks for Signal Processing IX*, 1999; 41-48.