RESEARCH ARTICLE

Semi-Supervised Projective Non-Negative Matrix Factorization for Cancer Classification

Xiang Zhang[1,2], **Naiyang Guan**[1,2]*, **Zhilong Jia**[3], **Xiaogang Qiu**[4], **Zhigang Luo**[1,2]*

1 College of Computer, National University of Defense Technology, Changsha 410073, China, **2** National Laboratory for Parallel and Distributed Processing, National University of Defense Technology, Changsha 410073, China, **3** Department of Chemistry and Biology, College of Science, National University of Defense Technology, Changsha, Hunan, China, **4** College of Information System and Management, National University of Defense Technology, Changsha, Hunan, 410073 China

* ny_guan@nudt.edu.cn (NG); zgluo@nudt.edu.cn (ZL)

Abstract

Advances in DNA microarray technologies have made gene expression profiles a significant candidate in identifying different types of cancers. Traditional learning-based cancer identification methods utilize labeled samples to train a classifier, but they are inconvenient for practical application because labels are quite expensive in the clinical cancer research community. This paper proposes a semi-supervised projective non-negative matrix factorization method (Semi-PNMF) to learn an effective classifier from both labeled and unlabeled samples, thus boosting subsequent cancer classification performance. In particular, Semi-PNMF jointly learns a non-negative subspace from concatenated labeled and unlabeled samples and indicates classes by the positions of the maximum entries of their coefficients. Because Semi-PNMF incorporates statistical information from the large volume of unlabeled samples in the learned subspace, it can learn more representative subspaces and boost classification performance. We developed a multiplicative update rule (MUR) to optimize Semi-PNMF and proved its convergence. The experimental results of cancer classification for two multiclass cancer gene expression profile datasets show that Semi-PNMF outperforms the representative methods.

Citation: Zhang X, Guan N, Jia Z, Qiu X, Luo Z (2015) Semi-Supervised Projective Non-Negative Matrix Factorization for Cancer Classification. PLoS ONE 10(9): e0138814. doi:10.1371/journal.pone.0138814

Editor: Ramin Homayouni, University of Memphis, UNITED STATES

Received: March 4, 2015

Accepted: September 3, 2015

Published: September 22, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. All relevant data are available on the Zenodo repository, and the access URL is https://zenodo.org/record/21712.

Funding: This work was supported by Ministry of Education of the People's Republic of China Grant Number 20134307110017 (www.moe.edu.cn), PI: Zhigang Luo and National Natural Science Foundation of China Grant Number 91024030/G03 (http://www.nsfc.gov.cn/), PI: Xiaogang Qiu.

Introduction

In cancer prognosis and treatment, it is crucial to identify different cancer types and subtypes. Traditional methods often rely on similar morphological appearances but easily induce different clinical courses and responses to therapy because of subjective interpretations and personal experience. This usually results in diagnostic confusion. Fortunately, the emergence of the DNA microarray technique removes this barrier in an objective and systematic manner and has showed great potential in outcome prediction of cancer types in genome-wide scales [1–11].

Numerous learning methods have been developed for cancer classification based on gene expression profiles [1–3]. For instance, Golub *et al.* [1] used a weighted voting scheme for the molecular classification of acute leukemia. Nguyen *et al.* [3] incorporated partial least squares

Competing Interests: The authors have declared that no competing interests exist.

(PLS) into the logistic discrimination and quadratic discriminant analysis for tumor classification. However, these methods are not convenient for practical applications because the labeled samples are quite expensive in the clinical cancer research community. To overcome this deficiency, Xu *et al.* [12] used the semi-supervised Ellipsoid ARTMAP (ssEAM) method for cancer classification. Shi *et al.* [13] utilized the semi-supervised method termed low density separation (LDS, [14]) to classify different types of cancers. Moreover, Maulik *et al.* [15] investigated the effectiveness of transductive SVM (TSVM, [16]) in cancer classification. Nevertheless, these algorithmic challenges involve the curse of dimensionality, which indicates that the overwhelming number of measures for gene expression levels contrast with the small number of samples.

This problem often calls for dimension reduction techniques. This paper focuses on non-negative matrix factorization (NMF, [17, 18]) because it is a flexible framework for conducting dimension reduction and performing classification and clustering tasks [19–26]. NMF decomposes a data matrix into the product of two non-negative factors. Due to its effectiveness, NMF and its variants have been applied to analyzing large-scale gene expression datasets [27–29], cancer classification [30, 31] and new class discovery [30]. Brunet *et al.* [31] originally adopted NMF to uncover molecular meta-patterns by clustering samples of leukemia, medulloblastoma and central nervous system tumors, and indicating that NMF outperforms both hierarchy clustering (HC) and self-organizing map (SOM). However, NMF does not explicitly guarantee the sparseness of the decomposition and violates the uniqueness property. Recent works [32] show that this often degrades the clustering performance. To address this issue, Li *et al.* [32] proposed local NMF (LNMF) to overcome this deficiency by imposing the sparse constraints over the decomposition. Hoyer *et al.* proposed sparse NMF (SNMF, [33]) to enforce sparseness in NMF by penalizing the number of non-zero entries of the coefficients rather than the sum of the entries. Furthermore, Gao *et al.* [34] utilized SNMF to identify the meta-patterns of various cancers for identifying different types of tumors.

Because the aforementioned methods follow regularization theory, they are jointly non-convex and are difficult to optimize. Unlike the above methods, Yuan *et al.* [35] developed the projective NMF (PNMF) to induce parts-based representation by implicitly imposing the orthogonal constraint over the basis. However, because these methods are unsupervised learning methods that do not take into account labels, their performance in cancer classification can be further improved. In this paper, we propose a semi-supervised projective NMF method (Semi-PNMF) that utilizes both labeled and unlabeled samples to boost classification performance. Particularly, Semi-PNMF learns a non-negative subspace from concatenated labeled and unlabeled samples and predicts classes by the index of the largest entries of their coefficients. Benefiting from the unlabeled data, Semi-PNMF can learn more representative subspaces, which are beneficial for classification tasks. We explored a multiplicative update rule (MUR) to solve Semi-PNMF and proved its convergence. The experimental results of cancer identification for multiclass cancer gene expression profile datasets including GCM [8] and Acute Leukemia [36] datasets show that Semi-PNMF outperforms the representative methods in terms of quantity.

Materials and Methods

Semi-supervised Projective Nonnegative Matrix Factorization

Projective non-negative matrix factorization (PNMF) learns a non-negative projection matrix to project high-dimensional data into the lower-dimensional subspace. Because it can learn parts-based representation, PNMF has been widely applied in pattern recognition [21, 26, 35, 37]. Here, we introduce the other representation form of PNMF that learns the lower-dimensional coefficients of samples to approximate the class indicator for clustering. This is based on the assumption that the basis lies in the subspace spanned by the original samples. Given the data matrix $V = [v_1,\cdots,v_n]^T \in R^{n\times m}$, where n denotes the number of samples and m their dimensionality, PNMF learns the coefficients $H \in R^{n\times r}$ to represent original samples, i.e.,

$$\min_{H\geq 0} \quad \| V - HH^T V \|_F^2, \tag{1}$$

where $\|\bullet\|_F$ denotes the matrix Frobenius norm and r the number of clusters.

As in objective (1), it is non-trivial to analyze the convergence in theory because Eq (1) contains a fourth-order term. To remove such a high order term, we first introduce an auxiliary variable, i.e., the cluster centroids, and the equality constraint into Eq (1). Thus, we can obtain

$$\begin{aligned} &\min_{H\geq 0} \quad \| V - HW \|_F^2, \\ &s.t., W = H^T V. \end{aligned} \tag{2}$$

The objective is very similar to BPNMF [26], but we cannot directly apply the optimization algorithm of BPNMF to optimize it especially when additional constraints such as the sparseness constraint and Laplacian regularization are imposed over the coefficients, as these constraints easily induce PNMF to produce the trivial solution. To avoid such a drawback, we propose a semi-supervised PNMF method (Semi-PNMF) by recasting Eq (2) as

$$\min_{H,W\geq 0} \quad \frac{1}{2}\| V - HW \|_F^2 + \frac{\alpha}{2}\| W - H^T V \|_F^2, \tag{3}$$

where $\alpha \geq 0$ is a regularization constant and W denotes the non-negative cluster centroid. Model (3) significantly differs from BPNMF because Eq (3) favors the representative capacity of the cluster centroids, while BPNMF focuses on the orthogonality of the non-negative subspace. Thus, Eq (3) induces the sparse coefficients, while BPNMF produces the sparse basis.

According to Eq (3), we can incorporate the local coordinate constraint [38] to improve the representative power of the basis, meanwhile further inducing the sparse coefficients to be true classes. Thus, we recast Eq (3) as the following regularization form:

$$\min_{H,W\geq 0} \quad \frac{1}{2}\| V - HW \|_F^2 + \frac{\alpha}{2}\| W - H^T V \|_F^2 + \frac{\beta}{2}\sum_{i=1}^{n}\sum_{j=1}^{r}|H^{ij}| \| V^i - W^j \|_2^2, \tag{4}$$

where β trades off the local coordinate regularization and H^{ij} denotes the i-the row and j-th column element of coefficients H, W^j and V^i, signifying the i-th and j-th row vector of W and V, respectively.

To make full use of partial labeled samples, we propagate the labels of labeled samples to unlabeled ones by minimizing the distance between their coefficients and the corresponding class indicator. Particularly, we require the coefficients of labeled samples to be equivalent with the corresponding class indicator. Consider the first d examples labeled and the rest unlabeled; the data matrix V can be divided into two parts, i.e., $V = [V_L^T, V_U^T]^T$. Then, we can obtain the objective function of Semi-PNMF as follows:

$$\min_{W,H_U\geq 0} \quad J = \frac{1}{2}\left\| \begin{bmatrix} V_L \\ V_U \end{bmatrix} - \begin{bmatrix} Q \\ H_U \end{bmatrix} W \right\|_F^2 + \frac{\alpha}{2}\| W - H_U^T V_U \|_F^2 + \frac{\beta}{2}\sum_{i=1}^{n_U}\sum_{j=1}^{r}|H_U^{ij}| \| V_U^i - W^j \|_2^2, \tag{5}$$

where Q denotes the partial label matrix wherein $Q_{ij} = 1$ if v_i belongs to the j-th class; otherwise, $Q_{ij} = 0$. Both H_U and n_U denote the coefficients and number of the unlabeled samples, respectively.

Interestingly, Semi-PNMF has two distinct aspects. First, it replaces the learned coefficients of the labeled samples with the corresponding class indicator. The constraint is so strong that the learned basis completely biases the labeled samples. This might induce the trivial solution to the coefficients of the unlabeled samples. Second, Semi-PNMF completely ignores the representation contribution of the labeled samples. It is so unintelligible that the learned basis only favors the unlabeled samples. It appeared that both aspects contradict each other, but intrinsically, they mutually complement each other in our Semi-PNMF. In essence, the first aspect corresponds to supervised learning, which generates the reasonable solution yet does not ensure it is consistent with the underlying data distribution, while the second one considers data distribution but cannot yield the reasonable solution. Thus, the combination of both aspects can mutually complement each other. Semi-PNMF learns the shared basis by the labeled and unlabeled instances, meanwhile inducing similar instances to have a similar representation, i.e., the coefficients. Because we impose the restriction that coefficients of the labeled samples be their labels as well as the local coordinate constraint over the basis and coefficients, the unlabeled sample coefficients are implicitly as sparse as the label vectors. In this way, Semi-PNMF effectively propagates the labels of labeled samples to the unlabeled ones. Consequently, in cancer classification, it is reasonable that, for each unlabeled sample, we choose the index of the largest entry of its coefficient to predict the classes of this sample once objective (5) yields their coefficients. The above intuition can be further verified by the toy example given in Figs 1 and 2.

Optimization Algorithm

It is difficult to optimize Eq (5) because it is jointly non-convex with respect to both W and H. Fortunately, it is convex with respect to W and H, respectively. Thus, we can establish the following theorem:

Theorem 1: The objective function (5) is non-increasing under the following multiplicative update rules:

$$W = W \otimes \frac{Q^T V_L + (1+\alpha+\beta) H_U^T V_U}{Q^T Q W + H_U^T H_U W + \alpha W + \beta F_U W}, \tag{6}$$

, and

$$H_U = H_U \otimes \frac{(1+\alpha+\beta) V_U W^T}{H_U W W^T + \alpha V_U V_U^T H_U + \beta/2(A+B)}, \tag{7}$$

where $\otimes$ denotes the element-wise product operator, $F_U = diag(sum(H_U))$, $A = [a,\cdots,a]$ wherein $a = diag(V_U V_U^T)$, and $B = [b,\cdots,b]$, wherein $b = diag(WW^T)$.

Proof. According to Eq (5), we can obtain the objective with respect to W as follows:

$$\begin{aligned} J(W) = {} & \frac{1}{2} Tr(-2V_L W^T Q^T + QWW^T Q^T) + \frac{1}{2} Tr(-2V_U W^T H_U^T + H_U W W^T H_U^T) \\ & + \frac{\alpha}{2} Tr(WW^T - 2WV_U^T H_U) + \frac{\beta}{2} Tr(\sum_{i=1}^{n_U} (V_U^i)^T 1^T \Lambda_i 1 V_U^i - 2W^T H_U^T V_U + W^T F_U W) \end{aligned}, \tag{8}$$

where Λ_U^i denotes the diagonal matrix whose diagonal elements are the i-th row vector values of V_U.

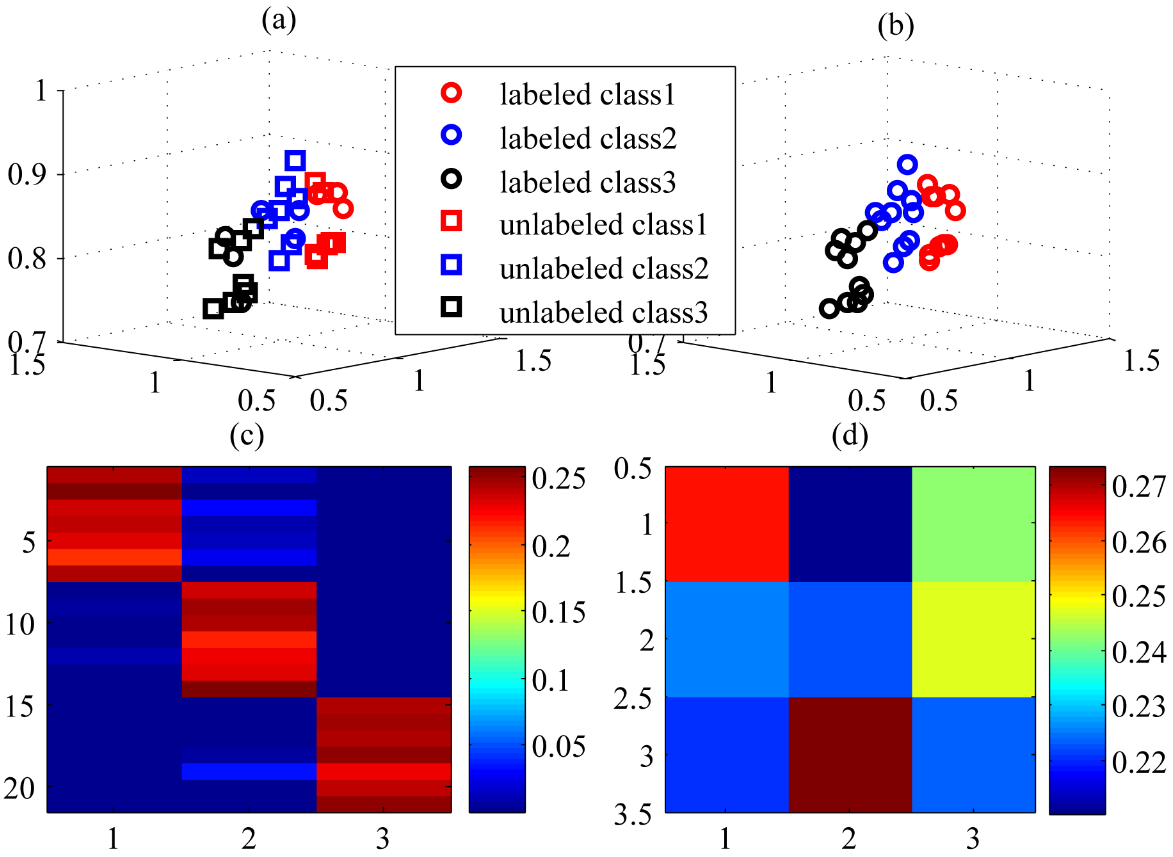


Fig 1. The toy example illustrating (a) the synthetic 3D original data including the labeled and unlabeled samples and the ground-truth labels, (b) the labeled results of unlabeled samples, (c) the learned coefficients of the unlabeled samples, and (d) the learned basis by Semi-PNMF. In Fig (a), both the square and circle markers signify the unlabeled and labeled samples, respectively, and three different colors stand for three different categories. Fig (b) shows that the unlabeled samples are marked as the ground-truth markers and colors. Figs (c) and (d) shows the coefficients and basis learned by Semi-PNMF, respectively. The index of maximum value of the coefficient for an unlabeled sample appears in red and indicates its class.

doi:10.1371/journal.pone.0138814.g001

By Eq (8), we can define the auxiliary function of $J(W)$ as

$$\begin{aligned} G(W,W') = & -(1+\alpha)Tr(WV_U^T H_U) - Tr(WV_L^T Q) \\ & + \frac{1}{2}\sum_{ij}\frac{(Q^T Q W')_{ij}}{W'_{ij}}W_{ij}^2 + \frac{1}{2}\sum_{ij}\frac{(H_U^T H_U W')_{ij}}{W'_{ij}}W_{ij}^2 + \frac{\alpha}{2}Tr(WW^T) \\ & - \beta Tr(W^T H_U^T V_U) + \frac{\beta}{2}\sum_{ij}\frac{(F_U W')_{ij}}{W'_{ij}}W_{ij}^2. \end{aligned} \tag{9}$$

Obviously, objective (9) has

$$G(W,W') \geq J(W) = G(W,W). \tag{10}$$

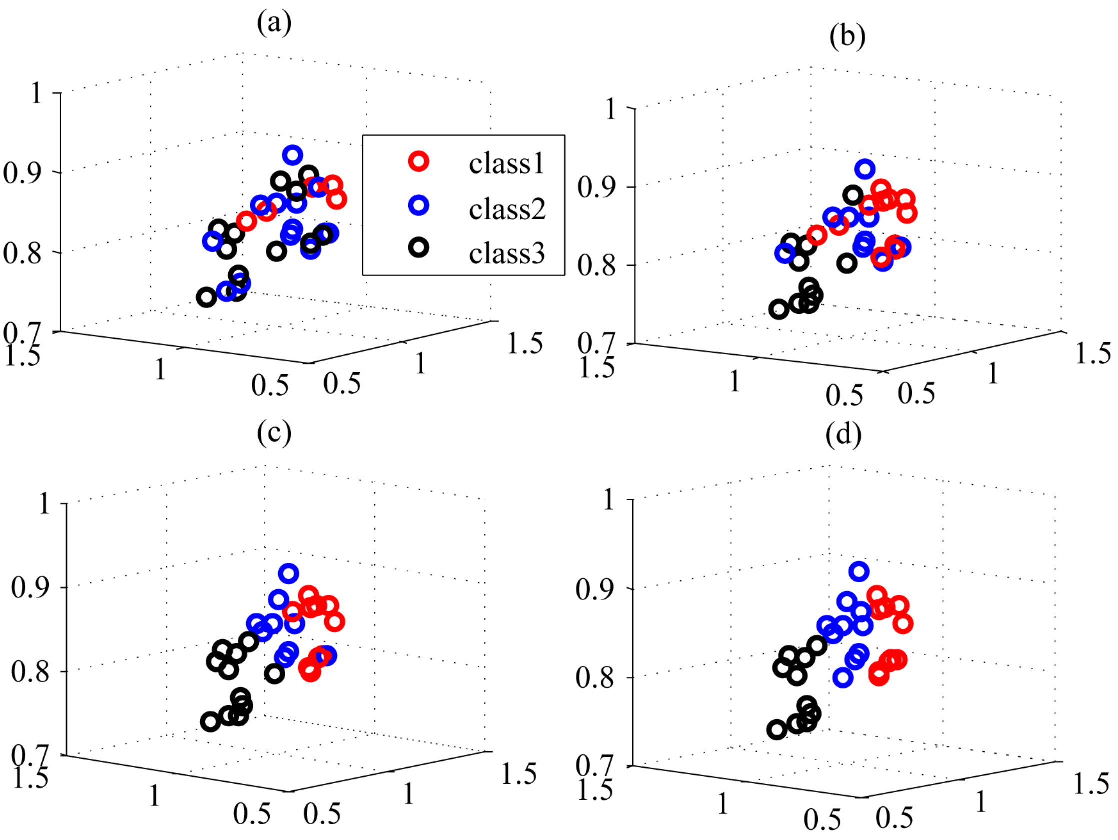


Fig 2. The toy example illustrating the labeling results obtained from the coefficients when the propagation procedure arrives at the (a) initialization stage, (b) 50-th iteration round, (c) 300-th iteration round, and (d) the resultant convergence (1500-th iteration round), respectively.

doi:10.1371/journal.pone.0138814.g002

We can obtain the derivative of Eq (9) as follows:

$$\begin{aligned}\frac{\partial G(W,W')}{\partial W_{ij}} &= -((1+\alpha)H_U^T V_U - Q^T V_L)_{ij} + \frac{(Q^T Q W')_{ij}}{W'_{ij}} W_{ij} \\ &+ \frac{(H_U^T H_U W')_{ij}}{W'_{ij}} W_{ij} + \alpha W_{ij} - \beta(H_U^T V_U)_{ij} + \beta\frac{(F_U W')_{ij}}{W'_{ij}} W_{ij},\end{aligned} \tag{11}$$

Based on Eq (11), we have

$$W_{ij} = W'_{ij}\frac{(Q^T V_L + (1+\alpha+\beta)H_U^T V_U)_{ij}}{(Q^T Q W' + H_U^T H_U W' + \alpha W' + \beta F_U W')_{ij}}. \tag{12}$$

By simple algebra, the formula (6) can be deduced from Eq (12). Likewise, we can obtain the auxiliary function of $J(H_U)$ as follows:

$$\begin{aligned} G(H_U, H'_U) &= -(1+\alpha)Tr(WV_U^T H_U) \\ &\quad + \frac{1}{2}\sum_{ij}\frac{(H'_U WW^T)_{ij}}{(H'_U)_{ij}}(H_U)_{ij}^2 + \frac{\alpha}{2}\sum_{ij}\frac{(V_U V_U^T H'_U)_{ij}}{(H'_U)_{ij}}(H_U)_{ij}^2 \\ &\quad + \frac{\beta}{2}Tr(H_U^T A - 2W^T H_U^T V_U + BH_U^T), \end{aligned} \tag{13}$$

Setting $\frac{\partial G(H_U, H'_U)}{\partial (H_U)_{ij}} = 0$, we have

$$(H_U)_{ij} = (H'_U)_{ij}\frac{(1+\alpha+\beta)(V_U W^T)_{ij}}{(H'_U WW^T + \alpha V_U V_U^T H'_U + \beta/2(A+B))_{ij}}, \tag{14}$$

Thus, according to Eq (14), we also obtain the update rule (7) for H_U.

Moreover, according to Eqs (10), (12) and (14), we have

$$J(W^{t+1}, H_U^{t+1}) \le J(W^{t+1}, H_U^t) \le J(W^t, H_U^t). \tag{15}$$

Based on Eq (15), these update rules always guarantee that the objective function monotonically decreases. Thus, this completes the proof. ■

According to the above theorem, we summarize the multiplicative update rule (MUR) for Semi-PNMF in **Algorithm 1**.

```
Algorithm 1 MUR for Semi-PNMF
Input: Examples $V \in R^{m \times n}$, penalty parameter $\alpha$, partial label matrix $Q$.
Output: $H_U$.
  1: Randomly initialize $W^0$ and $H_U^0$, and $l = 0$.
  2: repeat
  3:   Update $W_{l+1}$ according to Eq (6).
  4:   Calculate $H_U^{l+1}$ according to Eq (7).
  5:   $l \leftarrow l+1$.
  6: until { Stopping criterion $\frac{\|J^{l+1} - J^l\|_F}{\|J^l\|_F} < \varepsilon$ is satisfied.}
  7: $H_U = H_U^l$.
```

To reduce the time overhead, **Algorithm 1** utilizes the objective relative error as the stopping criterion; in addition, set ε to 10^{-7} in our experiments. The main time cost of **Algorithm 1** lies in line 3 and line 4. Their time complexities are $O(r^2n+mrn+r^2m+rm)$ and $O(mr(n-d)+r^2m+rm+r^2+r^2(n-d))$, respectively. Thus, the total time complexity of **Algorithm 1** is $O(r^2n+mrn+mr(n-d)+mrd+r^2m+rm+r^2+r^2(n-d))$.

Results

This section conducts a series of experiments on both synthetic and real-world datasets to verify the method proposed in this paper.

Synthetic Dataset

This section generates a small synthetic dataset to clarify the mechanism of Semi-PNMF. The synthetic dataset consists of three categories constructed by the following random samples:

$$y_1 = [1, 0.8, 0.8]^T + 0.1x,$$

$$y_2 = [0.8, 0.8, 0.8]^T + 0.1x,$$

and

$$y_3 = [0.8, 1, 0.7]^T + 0.15x,$$

where $x \in R^3$, and each of its entry is sampled from the standard uniform distribution $U(0,1)$. For each category, we randomly generated 10 samples, within which three samples were selected as labeled samples and the rest as unlabeled ones. Therefore, the synthetic dataset contains 30 samples in total. For clear illustration, three categories are marked as three different colors, and the labeled and unlabeled samples are distinguished by two shapes.

Fig 1(a) and 1(b) shows the ground truth and resultant labeled results of the unlabeled samples by Semi-PNMF, respectively, while Fig 1(c) and 1(d) displays the learned coefficients of the unlabeled samples and basis. In Fig 1(d), each row of the learned basis has different colors, implying that the basis stands for the centroids of different categories and owns the discriminative representation ability. According to Fig 1(c), each row of the learned coefficients is the lower-dimensional coefficient of the corresponding unlabeled sample. The larger the entry of the coefficient is, the darker its color is. As shown in Fig 1(c), the maximum entry of the coefficient largely exceeds the other entries. All maximum entries make the coefficients take up the diagonal form and imply the cluster memberships of all the samples. Thus, it is reasonable to select the index of the maximum entry of the coefficient as the classes of an unlabeled sample. This verifies our previous intuition. Since all samples shares the common basis, their coefficients become close to each other if they have the same labels. We impose the restriction that the coefficients of labeled samples be equivalent to their label vectors, and thus this also induces

Table 1. Description of the GCM dataset.

Cancer Types	Number of Samples
Breast adenocarcinoma (BR)	12
Prostate adenocarcinoma (PR)	14
Lung adenocarcinoma (LU)	12
Colorectal adenocarcinoma (CO)	12
Lymphoma (LY)	22
Bladder transitional cell carcinoma (BL)	11
Melanoma (ML)	10
Uterus adenocarcinoma (UT)	10
Leukemia (LE)	30
Renal cell carcinoma (RE)	11
Pancreas adenocarcinoma (PA)	11
Ovarlan adenocarcinoma (OV)	12
Pleural mesothelioma (MS)	11
Central nervous system (CNS)	20
Total	198

doi:10.1371/journal.pone.0138814.t001

the coefficients of the unlabeled to be close to their label vectors. In this way, Semi-PNMF can propagate the labels of the labeled samples to the unlabeled ones. The propagation procedure is illustrated in Fig 2.

GCM Dataset

This experiment merely compares traditional semi-supervised learning methods including low density separation (LDS, [14]), transductive SVM (TSVM, [16]), constrained NMF (CNMF, [24]), soft-constrained NMF (SCNMF, [25]) and Semi-PNMF by separating different types of cancers on the GCM dataset. The GCM dataset [8] contains the expression profiles of 218 tumor samples representing 14 common human cancer classes. It is available on the public website: http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi, and can also be downloaded from the website: https://zenodo.org/record/21712. According to [8], we combine the training and testing set of this gene expression data into a dataset for cancer classification. Thus, the combined dataset contains 198 samples with 16,063 genes. Table 1 gives a brief description of this dataset. To remove very low noisy values and saturation effects of very high values, we bound the gene expression data into a specific box constraint ranging from 20 to 16,000 units and then exclude those genes whose ratios and absolute variations across samples are under 5 and 500, respectively. Consequently, the resultant expression profile dataset contains the 11,370 genes passing. We compare the effectiveness of Semi-PNMF with LDS, TSVM, CNMF and SCNMF under varying configurations. Both CNMF and SCNMF involve no parameter tuning. For Semi-PNMF, we set two parameters $\alpha = 2$, and $\beta = 0.0001$, respectively. Because these representative methods enable convergence within 1,500 iteration rounds, we set the maximum number of loops to 1,500. For LDS and TSVM, we adopt the parameter settings provided in the source code to obtain the classification results.

We evaluate the cancer classification by the cross-validation over the whole dataset. This process selects one sample as the unlabeled sample and, meanwhile, learns the prediction model on all the samples for cancer diagnosis. For the unlabeled sample, we choose the index of the largest value of the resultant consensus matrix to predict the classes of this sample. As shown in Figs 3 to 7, the confusion matrix of the predicted results of Semi-PNMF, CNMF, SCNMF, LDS and TSVM are reported in detail. Each column denotes how many the unlabeled samples are assigned to each cancer, while each row signifies the number of the unlabeled samples affiliated to the real tumor type. Each color not only represents a specific cancer type but also highlights the correct prediction results, i.e., the diagonal elements of the confusion matrix.

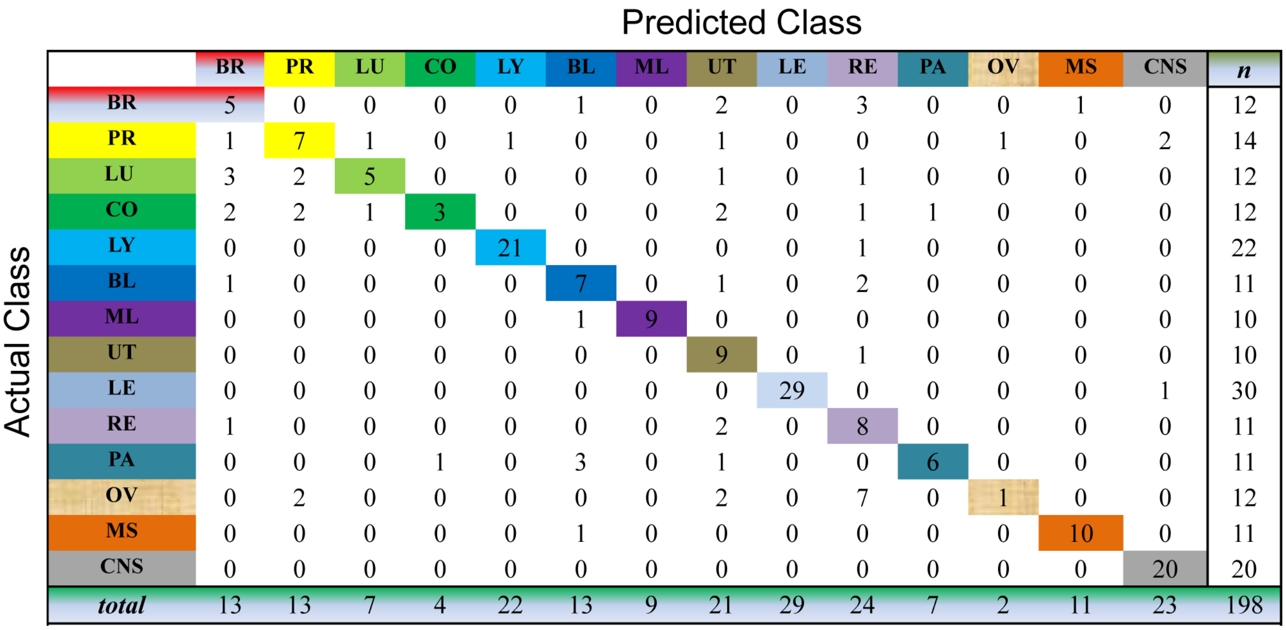
Figs 3 to 7 imply that Semi-PNMF can identify different tumor types more accurately than the representative methods. For example, when working with two labeled samples from each tumor type, Semi-PNMF achieves 70.71% classification accuracy and exceeds LDS, TSVM, SCNMF, and CNMF by 10.6%, 21.72%, 21.72%, and 32.3%, respectively. Moreover, Table 2 further implies the effectiveness of Semi-PNMF compared with CNMF, SCNMF, TSVM, and LDS in terms of both sensitivity and specificity. For completeness, we list their definitions as follows:

$$sensitivity = \frac{TP}{TP + FN}, \tag{16}$$

and

$$specificity = \frac{TN}{TN + FP}, \tag{17}$$

where TP, TN, FP, and FN denote the number of true positive, true negative, false positive and false negative samples, respectively.

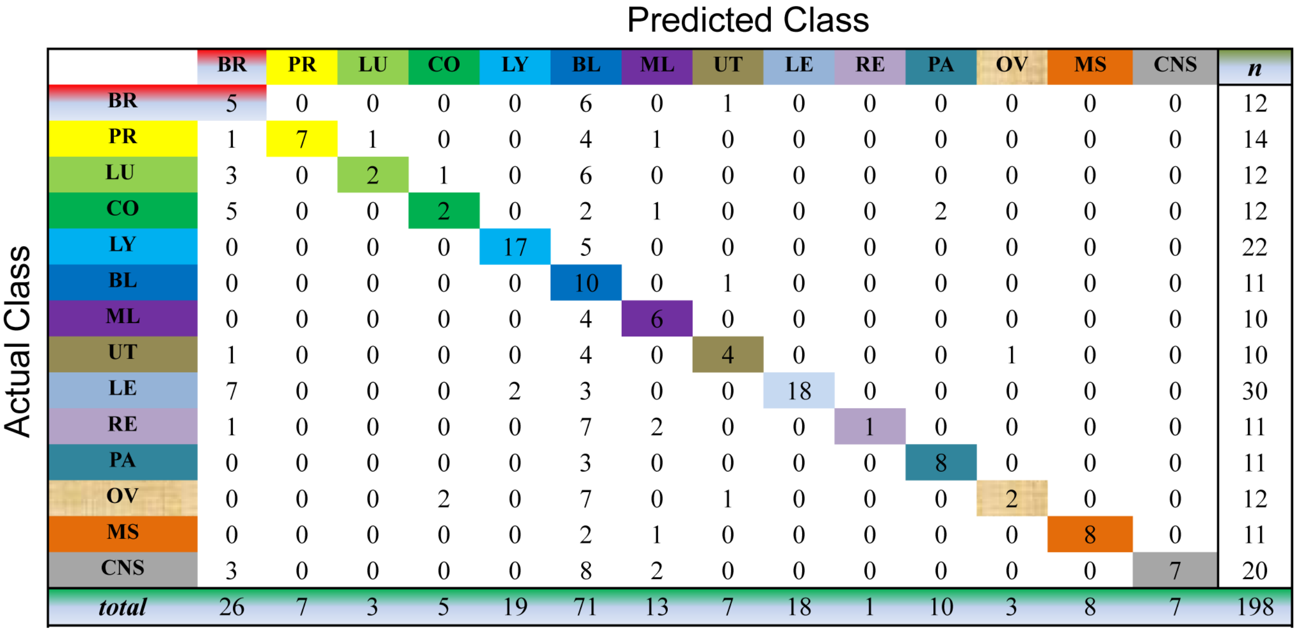

Predicted Class
Actual Class

	BR	PR	LU	CO	LY	BL	ML	UT	LE	RE	PA	OV	MS	CNS	*n*
BR	5	0	0	0	0	1	0	2	0	3	0	0	1	0	12
PR	1	7	1	0	1	0	0	1	0	0	0	1	0	2	14
LU	3	2	5	0	0	0	0	1	0	1	0	0	0	0	12
CO	2	2	1	3	0	0	0	2	0	1	1	0	0	0	12
LY	0	0	0	0	21	0	0	0	0	1	0	0	0	0	22
BL	1	0	0	0	0	7	0	1	0	2	0	0	0	0	11
ML	0	0	0	0	0	1	9	0	0	0	0	0	0	0	10
UT	0	0	0	0	0	0	0	9	0	1	0	0	0	0	10
LE	0	0	0	0	0	0	0	0	29	0	0	0	0	1	30
RE	1	0	0	0	0	0	0	2	0	8	0	0	0	0	11
PA	0	0	0	1	0	3	0	1	0	0	6	0	0	0	11
OV	0	2	0	0	0	0	0	2	0	7	0	1	0	0	12
MS	0	0	0	0	0	1	0	0	0	0	0	0	10	0	11
CNS	0	0	0	0	0	0	0	0	0	0	0	0	0	20	20
total	13	13	7	4	22	13	9	21	29	24	7	2	11	23	198

Fig 3. Confusion matrix of prediction results using Semi-PNMF, which achieves a total accuracy of 70.71%. Matrix delineates distribution of actual compared with predicted class membership for multiclass cancer prediction on the GCM dataset.

doi:10.1371/journal.pone.0138814.g003

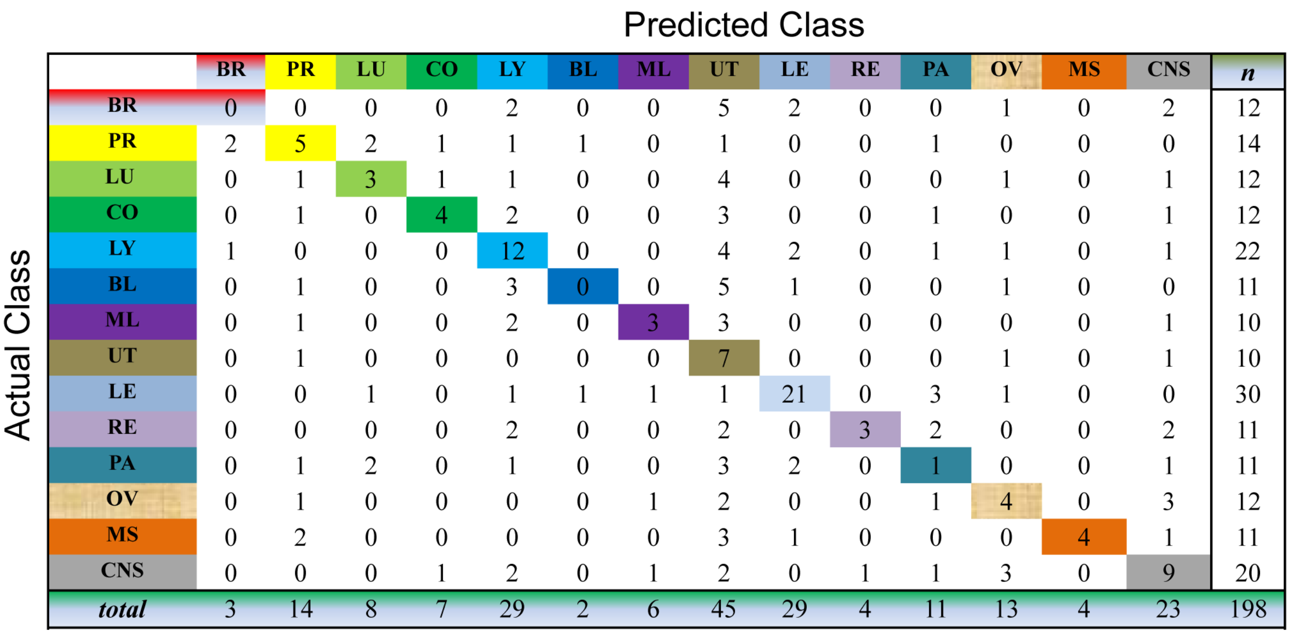
The number of labeled examples is an important factor affecting the performance of semi-supervised learning methods. Hence, it is very necessary to observe the classification accuracy of Semi-PNMF under different numbers (1–6) of labeled samples in each class. Here, we randomly select different numbers of examples from each class as labeled examples and regard the


Predicted Class
Actual Class

	BR	PR	LU	CO	LY	BL	ML	UT	LE	RE	PA	OV	MS	CNS	*n*
BR	5	0	0	0	0	6	0	1	0	0	0	0	0	0	12
PR	1	7	1	0	0	4	1	0	0	0	0	0	0	0	14
LU	3	0	2	1	0	6	0	0	0	0	0	0	0	0	12
CO	5	0	0	2	0	2	1	0	0	0	2	0	0	0	12
LY	0	0	0	0	17	5	0	0	0	0	0	0	0	0	22
BL	0	0	0	0	0	10	0	1	0	0	0	0	0	0	11
ML	0	0	0	0	0	4	6	0	0	0	0	0	0	0	10
UT	1	0	0	0	0	4	0	4	0	0	0	1	0	0	10
LE	7	0	0	0	2	3	0	0	18	0	0	0	0	0	30
RE	1	0	0	0	0	7	2	0	0	1	0	0	0	0	11
PA	0	0	0	0	0	3	0	0	0	0	8	0	0	0	11
OV	0	0	0	2	0	7	0	1	0	0	0	2	0	0	12
MS	0	0	0	0	0	2	1	0	0	0	0	0	8	0	11
CNS	3	0	0	0	0	8	2	0	0	0	0	0	0	7	20
total	26	7	3	5	19	71	13	7	18	1	10	3	8	7	198

Fig 4. Confusion matrix of prediction results using SCNMF, which achieves a total accuracy of 48.99%. Matrix delineates distribution of actual compared with predicted class membership for multiclass cancer prediction on the GCM dataset.

doi:10.1371/journal.pone.0138814.g004


Predicted Class
Actual Class

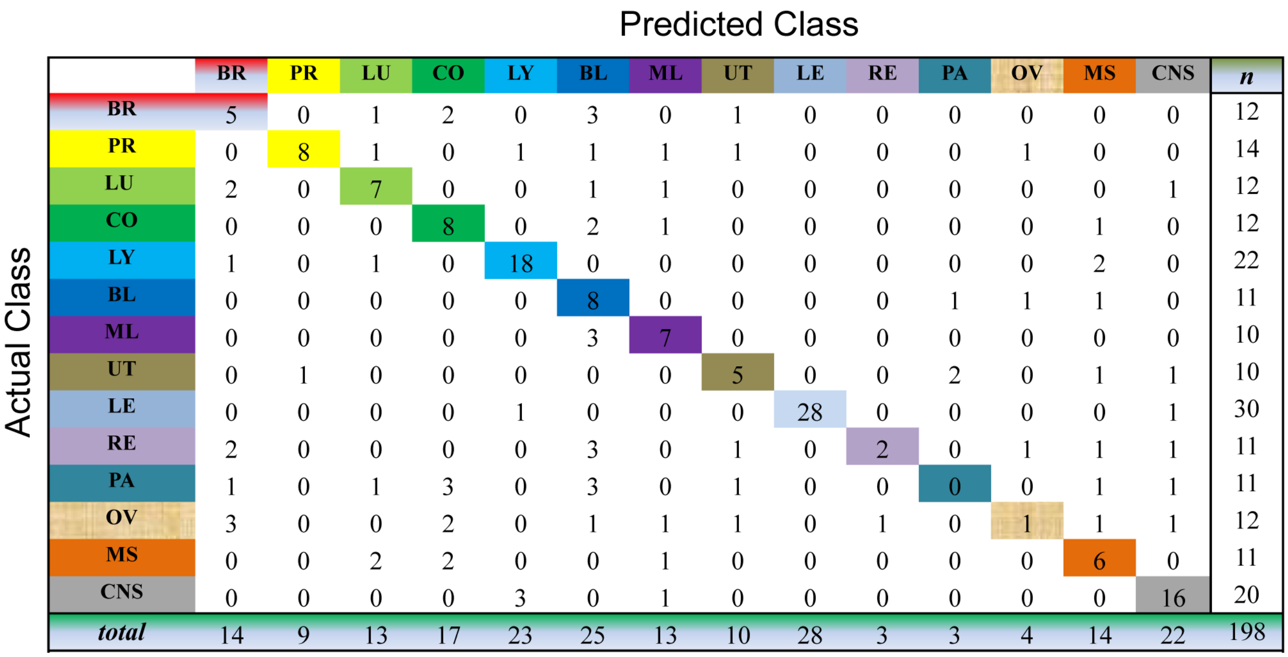
	BR	PR	LU	CO	LY	BL	ML	UT	LE	RE	PA	OV	MS	CNS	*n*
BR	0	0	0	0	2	0	0	5	2	0	0	1	0	2	12
PR	2	5	2	1	1	1	0	1	0	0	1	0	0	0	14
LU	0	1	3	1	1	0	0	4	0	0	0	1	0	1	12
CO	0	1	0	4	2	0	0	3	0	0	1	0	0	1	12
LY	1	0	0	0	12	0	0	4	2	0	1	1	0	1	22
BL	0	1	0	0	3	0	0	5	1	0	0	1	0	0	11
ML	0	1	0	0	2	0	3	3	0	0	0	0	0	1	10
UT	0	1	0	0	0	0	0	7	0	0	0	1	0	1	10
LE	0	0	1	0	1	1	1	1	21	0	3	1	0	0	30
RE	0	0	0	0	2	0	0	2	0	3	2	0	0	2	11
PA	0	1	2	0	1	0	0	3	2	0	1	0	0	1	11
OV	0	1	0	0	0	0	1	2	0	0	1	4	0	3	12
MS	0	2	0	0	0	0	0	3	1	0	0	0	4	1	11
CNS	0	0	0	1	2	0	1	2	0	1	1	3	0	9	20
total	3	14	8	7	29	2	6	45	29	4	11	13	4	23	198

Fig 5. Confusion matrix of prediction results using CNMF, which achieves a total accuracy of 38.4%. Matrix delineates distribution of actual compared with predicted class membership for multiclass cancer prediction on the GCM dataset.

doi:10.1371/journal.pone.0138814.g005

rest as unlabeled. For fair comparison, we independently conduct 100 individual experiment trails to remove the effect of randomness.

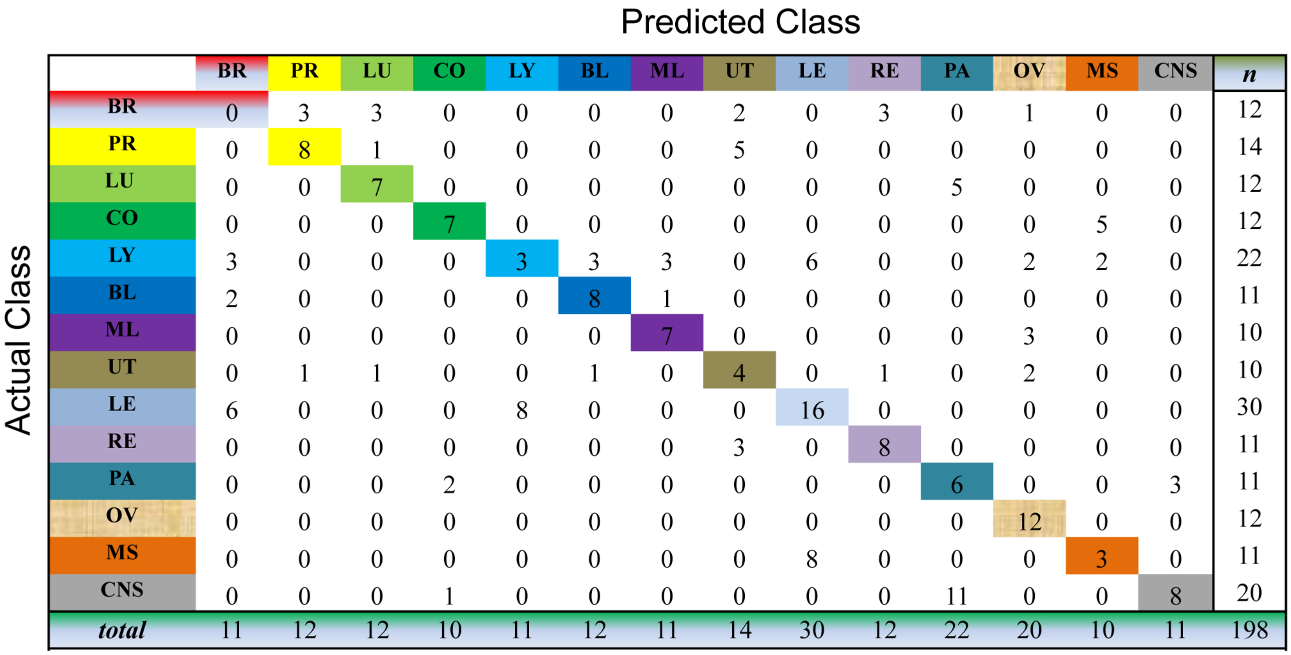
Fig 8 compares the average accuracy of CNMF, SCNMF, TSVM, LDS, and Semi-PNMF under different numbers of labeled samples for each class. It also shows that Semi-PNMF achieves the highest accuracy and takes on an increasing tendency with the rise in the number of labeled samples.


Predicted Class
Actual Class

	BR	PR	LU	CO	LY	BL	ML	UT	LE	RE	PA	OV	MS	CNS	*n*
BR	5	0	1	2	0	3	0	1	0	0	0	0	0	0	12
PR	0	8	1	0	1	1	1	1	0	0	0	1	0	0	14
LU	2	0	7	0	0	1	1	0	0	0	0	0	0	1	12
CO	0	0	0	8	0	2	1	0	0	0	0	0	1	0	12
LY	1	0	1	0	18	0	0	0	0	0	0	0	2	0	22
BL	0	0	0	0	0	8	0	0	0	0	1	1	1	0	11
ML	0	0	0	0	0	3	7	0	0	0	0	0	0	0	10
UT	0	1	0	0	0	0	0	5	0	0	2	0	1	1	10
LE	0	0	0	0	1	0	0	0	28	0	0	0	0	1	30
RE	2	0	0	0	0	3	0	1	0	2	0	1	1	1	11
PA	1	0	1	3	0	3	0	1	0	0	0	0	1	1	11
OV	3	0	0	2	0	1	1	1	0	1	0	1	1	1	12
MS	0	0	2	2	0	0	1	0	0	0	0	0	6	0	11
CNS	0	0	0	0	3	0	1	0	0	0	0	0	0	16	20
total	14	9	13	17	23	25	13	10	28	3	3	4	14	22	198

Fig 6. Confusion matrix of prediction results using LDS, which achieves a total accuracy of 60.1%. Matrix delineates distribution of actual compared with predicted class membership for multiclass cancer prediction on the GCM dataset.

doi:10.1371/journal.pone.0138814.g006



Predicted Class

Actual Class

	BR	PR	LU	CO	LY	BL	ML	UT	LE	RE	PA	OV	MS	CNS	*n*
BR	0	3	3	0	0	0	0	2	0	3	0	1	0	0	12
PR	0	8	1	0	0	0	0	5	0	0	0	0	0	0	14
LU	0	0	7	0	0	0	0	0	0	0	5	0	0	0	12
CO	0	0	0	7	0	0	0	0	0	0	0	0	5	0	12
LY	3	0	0	0	3	3	3	0	6	0	0	2	2	0	22
BL	2	0	0	0	0	8	1	0	0	0	0	0	0	0	11
ML	0	0	0	0	0	0	7	0	0	0	0	3	0	0	10
UT	0	1	1	0	0	1	0	4	0	1	0	2	0	0	10
LE	6	0	0	0	8	0	0	0	16	0	0	0	0	0	30
RE	0	0	0	0	0	0	0	3	0	8	0	0	0	0	11
PA	0	0	0	2	0	0	0	0	0	0	6	0	0	3	11
OV	0	0	0	0	0	0	0	0	0	0	0	12	0	0	12
MS	0	0	0	0	0	0	0	0	8	0	0	0	3	0	11
CNS	0	0	0	1	0	0	0	0	0	0	11	0	0	8	20
total	11	12	12	10	11	12	11	14	30	12	22	20	10	11	198

Fig 7. Confusion matrix of prediction results using TSVM, which achieves a total accuracy of 48.99%. Matrix delineates distribution of actual compared with predicted class membership for multiclass cancer prediction on the GCM dataset.

doi:10.1371/journal.pone.0138814.g007

Acute Leukemia Dataset

We also conduct a cancer classification experiment to verify the classification performance of Semi-PNMF compared with low density separation (LDS, [14]), transductive SVM (TSVM, [16]), constrained NMF (CNMF, [24]), and soft-constrained NMF (SCNMF, [25]) on another popular dataset, i.e., the Acute Leukemia dataset [36]. This dataset comes from Gene

Table 2. Sensitivity and Specificity of the compared methods over 14 cancer subtypes on the GCM dataset.

	Sensitivity					Specificity				
	CNMF	SCNMF	TSVM	LDS	Semi-PNMF	CNMF	SCNMF	TSVM	LDS	Semi-PNMF
BR	0	**0.42**	0	**0.42**	**0.42**	**0.98**	0.89	0.94	0.95	0.96
PR	0.36	0.5	**0.57**	**0.57**	0.5	0.95	**1**	0.98	0.99	0.98
LU	0.25	0.17	**0.58**	**0.58**	0.42	0.97	**0.99**	0.97	0.97	**0.99**
CO	0.33	0.17	0.58	**0.67**	0.25	0.98	0.98	0.98	0.95	**0.99**
LY	0.55	0.77	0.14	0.82	**0.95**	0.9	**0.99**	0.95	0.97	**0.99**
BL	0	**0.91**	0.73	0.73	0.64	**0.99**	0.67	0.98	0.91	0.97
ML	0.3	0.6	0.7	0.7	**0.9**	0.98	0.96	0.98	0.97	**1**
UT	0.7	0.4	0.4	0.5	**0.9**	0.79	**0.98**	0.95	0.97	0.94
LE	0.7	0.6	0.53	0.93	**0.97**	0.95	**1**	0.92	**1**	**1**
RE	0.27	0.09	0.73	0.18	**0.73**	0.99	**1**	0.98	0.99	0.91
PA	0.09	**0.73**	0.55	0	0.55	0.95	**0.99**	0.91	0.98	**0.99**
OV	0.33	0.17	**1**	0.083	0.08	0.95	**0.99**	0.96	0.98	**0.99**
MS	0.36	0.73	0.27	0.55	**0.91**	**1**	**1**	0.96	0.96	0.99
CNS	0.45	0.35	0.4	0.8	**1**	0.92	**1**	0.98	0.97	0.98
Avg.	0.34	0.47	0.51	0.538	**0.66**	0.95	0.96	0.96	0.97	**0.98**

doi:10.1371/journal.pone.0138814.t002

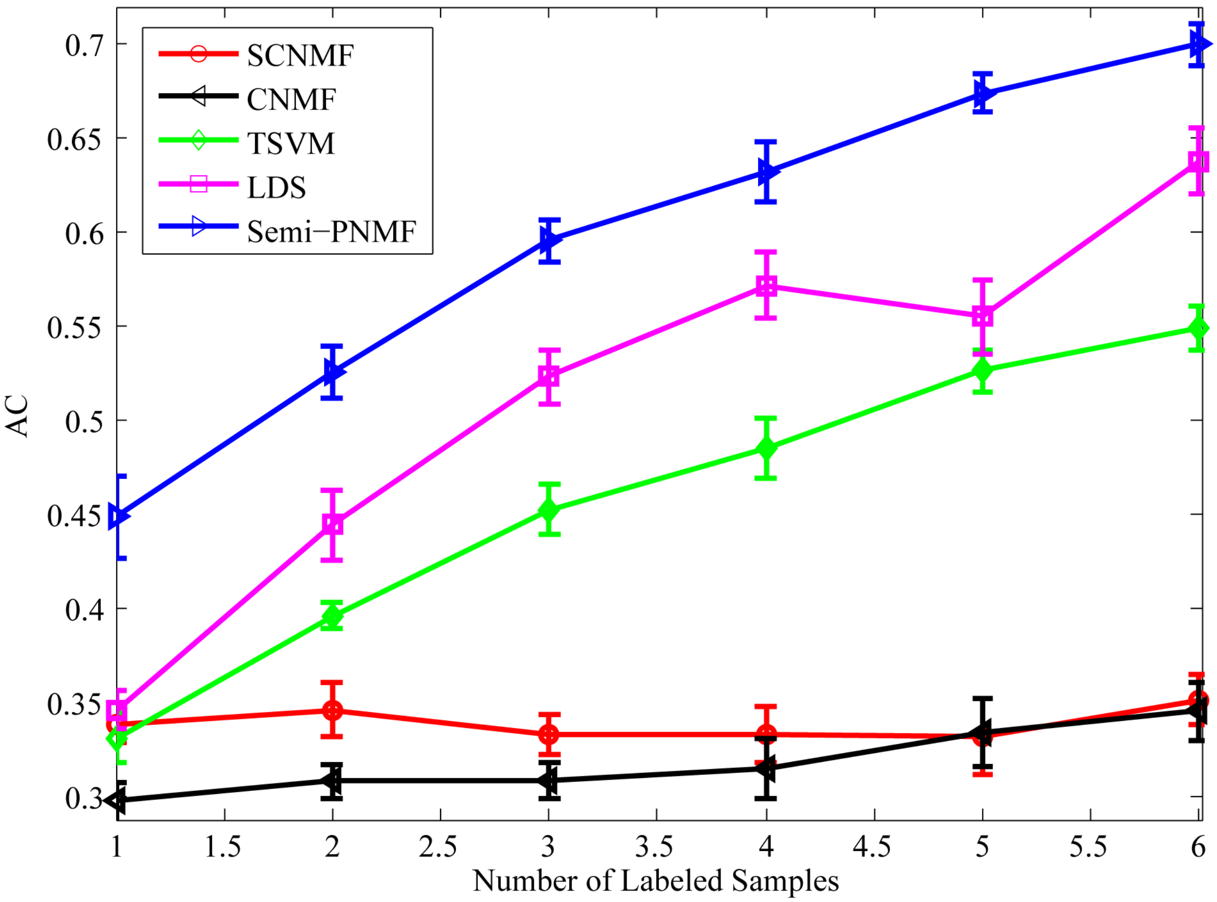


Fig 8. Average accuracies versus different numbers (1–6) of the labeled samples for each class of the GCM dataset.

doi:10.1371/journal.pone.0138814.g008

Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE13159), and can also be downloaded from the website: https://zenodo.org/record/21712. We replace the unavailable entries of this dataset with the average values of their *k*-nearest neighbor elements. This dataset consists of 2,096 samples along with 54,675 probes in total. This dataset contains different cancer subtypes of the acute leukemia and thus is not suited for cancer classification in contrast with the GCM dataset. Table 3 gives a brief description of this dataset. Then, we feed this dataset to all the compared methods.

For Semi-PNMF, we set two parameters $\alpha = 0.2$, and $\beta = 0.01$. For the traditional semi-supervised learning methods, we adopt the same configurations as the above subsection. The cross-validation process of the above subsection is repeatedly conducted to evaluate the compared methods on this dataset. As shown in Figs 9 to 13, the confusion matrix of the predicted results of Semi-PNMF, CNMF, SCNMF, LDS and TSVM are reported in detail. Each column denotes how many unlabeled samples are assigned to each cancer subtype, while each row signifies the number of unlabeled samples affiliated to the real tumor subtype. Each color not only represents a specific cancer subtype but also highlights the correct prediction results, i.e., the diagonal elements of the confusion matrix.

Figs 9 to 13 imply that Semi-PNMF can identify different tumor types more accurately than the representative methods. Semi-PNMF achieves the highest total classification accuracy

Table 3. Description of the Acute Leukemia dataset.

Cancer Types	Number of Samples
Mature B-ALL with t(8;14)	13
Pro-B-ALL with t(11q23)/MLL	70
c-ALL/pre-B-ALL with t(9;22)	122
T-ALL	174
ALL with t(12;21)	58
ALL with t(1;19)	36
ALL with hyperdiploid karyotype	40
c-ALL/pre-B-ALL without t(9;22)	237
AML with t(8;21)	40
AML with t(15;17)	37
AML with inv(16)/t(16;16)	28
AML with t(11q23)/MLL	38
AML with normal karyotype+other abnormalities	351
AML complex aberrant karyotype	48
CLL	448
CML	76
MDS	206
Non-leukemia and healthy bone marrow	74
Total	2,096

Abbreviations: B-ALL, B-cell acute lymphoblastic leukemia; MLL, myeloid/lymphoid or mixed-lineage leukemia; pre, precursor; c-ALL, childhood acute lymphoblastic leukemia; T-ALL, T-cell acute lymphoblastic leukemia; ALL, acute lymphoblastic leukemia; AML, acute myeloid leukemia; CLL, chronic lymphocytic leukemia; CML, chronic myelogenous leukemia; MDS, myelodysplastic syndrome.

doi:10.1371/journal.pone.0138814.t003

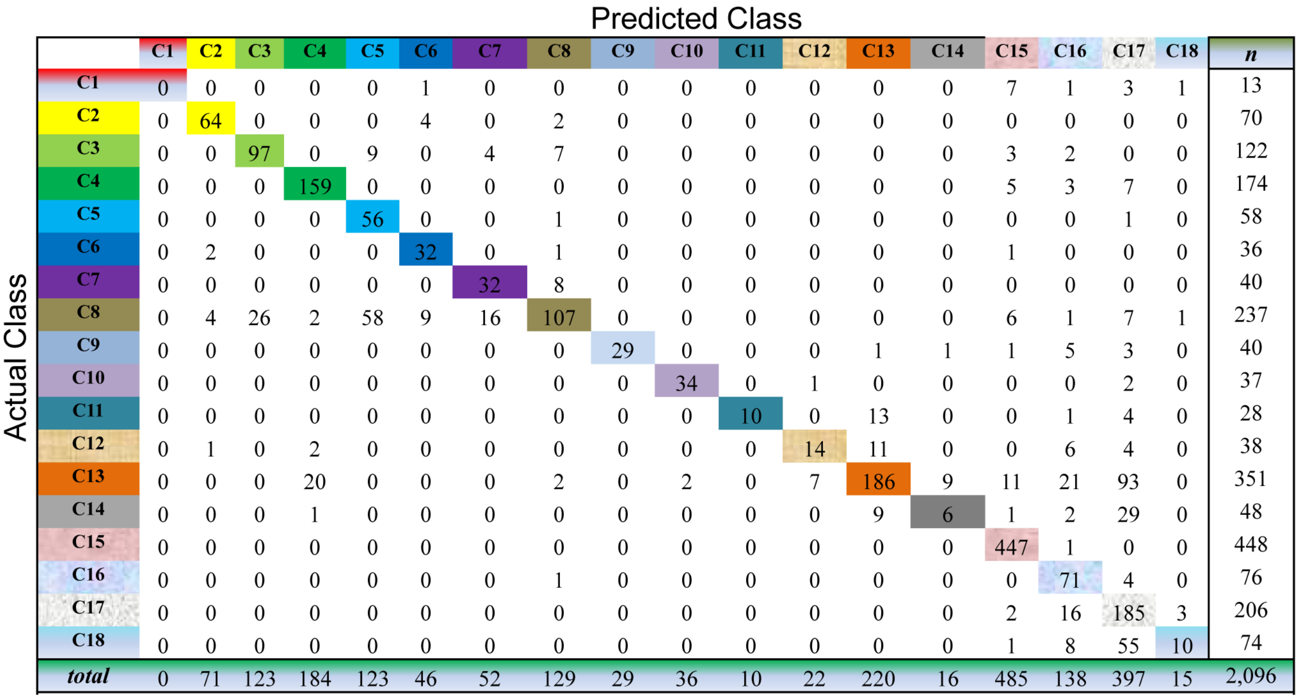
compared with CNMF, SCNMF, TSVM and LDS in terms of the prediction results in the confusion matrix. Moreover, Table 4 also indicates that Semi-PNMF consistently outperforms the compared methods on eighteen cancer subtypes in terms of both sensitivity and specificity. In summary, these results suggest the effectiveness of Semi-PNMF in cancer classification.

The number of the labeled samples is an important factor affecting the performance of semi-supervised learning methods. Hence, it is very necessary to observe the classification accuracy of Semi-PNMF under different numbers (1–6) of labeled samples in each class. Here, we randomly select different numbers of examples from each class as labeled examples and regard the rest as unlabeled. Then, we independently conduct 10 individual experiment trails to remove the effect of randomness.

Fig 14 compares the average accuracy of CNMF, SCNMF, TSVM, LDS, and Semi-PNMF under different numbers of labeled samples for each class. It also shows that Semi-PNMF achieves the highest accuracy and has an increasing tendency with the rise in the number of labeled samples.

Discussion

This paper proposes the semi-supervised PNMF method (Semi-PNMF), which incorporates two types of constraints as well as the auxiliary basis to boost PNMF. Particularly, Semi-PNMF utilizes the linear combination of examples to approximate the cluster centroids such that the cluster centroids have more powerful representative ability. To effectively indicate the classes of unlabeled samples, Semi-PNMF enforces the coefficients of labeled samples to approach



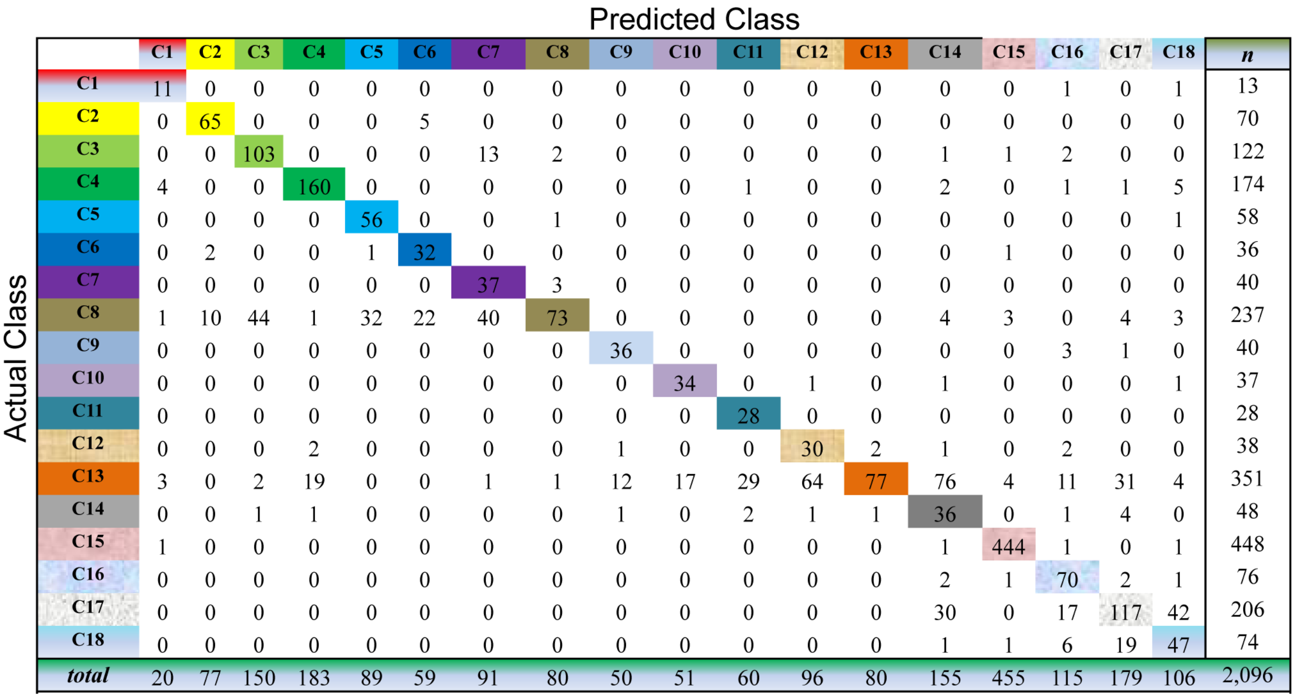
Predicted Class

Actual Class

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	*n*
C1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	7	1	3	1	13
C2	0	64	0	0	0	4	0	2	0	0	0	0	0	0	0	0	0	0	70
C3	0	0	97	0	9	0	4	7	0	0	0	0	0	0	3	2	0	0	122
C4	0	0	0	159	0	0	0	0	0	0	0	0	0	0	5	3	7	0	174
C5	0	0	0	0	56	0	0	1	0	0	0	0	0	0	0	0	1	0	58
C6	0	2	0	0	0	32	0	1	0	0	0	0	0	0	1	0	0	0	36
C7	0	0	0	0	0	0	32	8	0	0	0	0	0	0	0	0	0	0	40
C8	0	4	26	2	58	9	16	107	0	0	0	0	0	0	6	1	7	1	237
C9	0	0	0	0	0	0	0	0	29	0	0	0	1	1	1	5	3	0	40
C10	0	0	0	0	0	0	0	0	0	34	0	1	0	0	0	0	2	0	37
C11	0	0	0	0	0	0	0	0	0	0	10	0	13	0	0	1	4	0	28
C12	0	1	0	2	0	0	0	0	0	0	0	14	11	0	0	6	4	0	38
C13	0	0	0	20	0	0	0	2	0	2	0	7	186	9	11	21	93	0	351
C14	0	0	0	1	0	0	0	0	0	0	0	0	9	6	1	2	29	0	48
C15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	447	1	0	0	448
C16	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	71	4	0	76
C17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	16	185	3	206
C18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	8	55	10	74
total	0	71	123	184	123	46	52	129	29	36	10	22	220	16	485	138	397	15	2,096

Fig 9. Confusion matrix of prediction results using Semi-PNMF, which achieves a total accuracy of 73.43%. Matrix delineates distribution of actual compared with predicted class membership for multiclass cancer prediction on the Acute Leukemia dataset.

doi:10.1371/journal.pone.0138814.g009



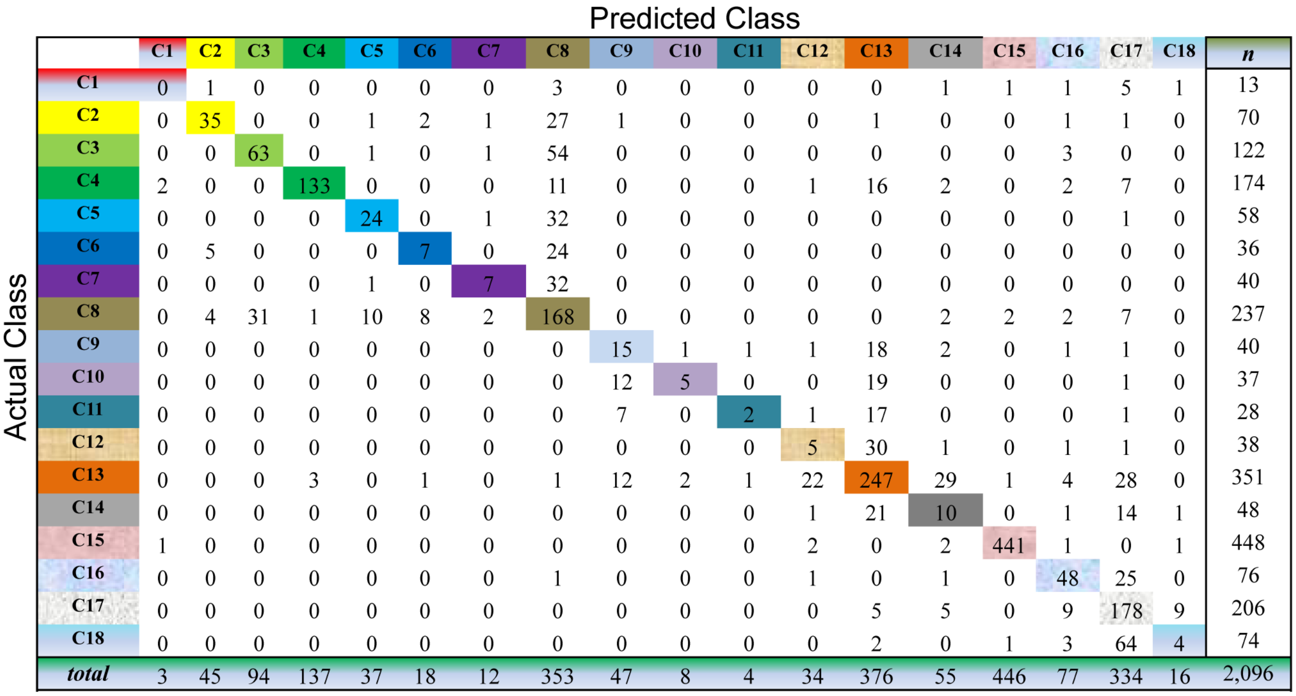
Predicted Class

Actual Class

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	*n*
C1	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	13
C2	0	65	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	70
C3	0	0	103	0	0	0	13	2	0	0	0	0	0	1	1	2	0	0	122
C4	4	0	0	160	0	0	0	0	0	0	1	0	0	2	0	1	1	5	174
C5	0	0	0	0	56	0	0	1	0	0	0	0	0	0	0	0	0	1	58
C6	0	2	0	0	1	32	0	0	0	0	0	0	0	0	1	0	0	0	36
C7	0	0	0	0	0	0	37	3	0	0	0	0	0	0	0	0	0	0	40
C8	1	10	44	1	32	22	40	73	0	0	0	0	0	4	3	0	4	3	237
C9	0	0	0	0	0	0	0	0	36	0	0	0	0	0	0	3	1	0	40
C10	0	0	0	0	0	0	0	0	0	34	0	1	0	1	0	0	0	1	37
C11	0	0	0	0	0	0	0	0	0	0	28	0	0	0	0	0	0	0	28
C12	0	0	0	2	0	0	0	0	1	0	0	30	2	1	0	2	0	0	38
C13	3	0	2	19	0	0	1	1	12	17	29	64	77	76	4	11	31	4	351
C14	0	0	1	1	0	0	0	0	1	0	2	1	1	36	0	1	4	0	48
C15	1	0	0	0	0	0	0	0	0	0	0	0	0	1	444	1	0	1	448
C16	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	70	2	1	76
C17	0	0	0	0	0	0	0	0	0	0	0	0	0	30	0	17	117	42	206
C18	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	6	19	47	74
total	20	77	150	183	89	59	91	80	50	51	60	96	80	155	455	115	179	106	2,096

Fig 10. Confusion matrix of prediction results using SCNMF, which achieves a total accuracy of 69.47%. Matrix delineates distribution of actual compared with predicted class membership for multiclass cancer prediction on the Acute Leukemia dataset.

doi:10.1371/journal.pone.0138814.g010

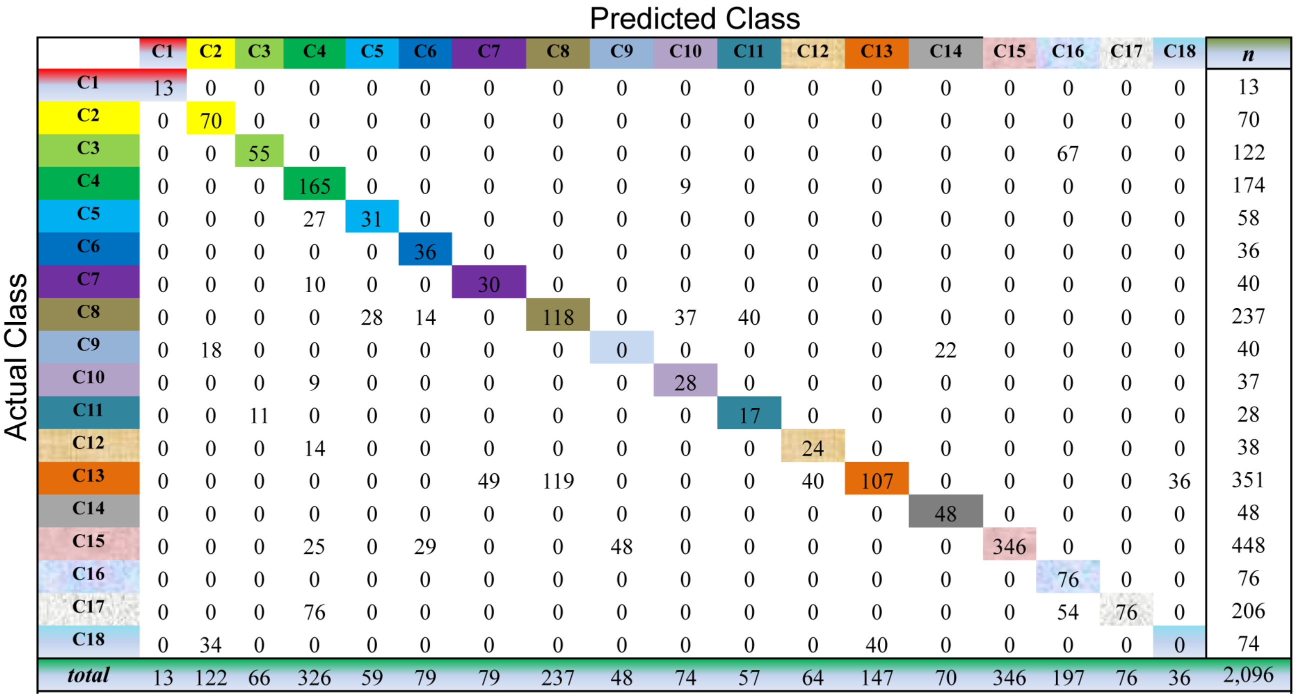


Predicted Class / Actual Class

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	*n*
C1	0	1	0	0	0	0	0	3	0	0	0	0	0	1	1	1	5	1	13
C2	0	35	0	0	1	2	1	27	1	0	0	0	1	0	0	1	1	0	70
C3	0	0	63	0	1	0	1	54	0	0	0	0	0	0	0	3	0	0	122
C4	2	0	0	133	0	0	0	11	0	0	0	1	16	2	0	2	7	0	174
C5	0	0	0	0	24	0	1	32	0	0	0	0	0	0	0	0	1	0	58
C6	0	5	0	0	0	7	0	24	0	0	0	0	0	0	0	0	0	0	36
C7	0	0	0	0	1	0	7	32	0	0	0	0	0	0	0	0	0	0	40
C8	0	4	31	1	10	8	2	168	0	0	0	0	0	2	2	2	7	0	237
C9	0	0	0	0	0	0	0	0	15	1	1	1	18	2	0	1	1	0	40
C10	0	0	0	0	0	0	0	0	12	5	0	0	19	0	0	0	1	0	37
C11	0	0	0	0	0	0	0	0	7	0	2	1	17	0	0	0	1	0	28
C12	0	0	0	0	0	0	0	0	0	0	0	5	30	1	0	1	1	0	38
C13	0	0	0	3	0	1	0	1	12	2	1	22	247	29	1	4	28	0	351
C14	0	0	0	0	0	0	0	0	0	0	0	1	21	10	0	1	14	1	48
C15	1	0	0	0	0	0	0	0	0	0	0	2	0	2	441	1	0	1	448
C16	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	48	25	0	76
C17	0	0	0	0	0	0	0	0	0	0	0	0	5	5	0	9	178	9	206
C18	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	3	64	4	74
total	3	45	94	137	37	18	12	353	47	8	4	34	376	55	446	77	334	16	2,096

Fig 11. Confusion matrix of prediction results using CNMF, which achieves a total accuracy of 66.41%. Matrix delineates distribution of actual compared with predicted class membership for multiclass cancer prediction on the Acute Leukemia dataset.

doi:10.1371/journal.pone.0138814.g011

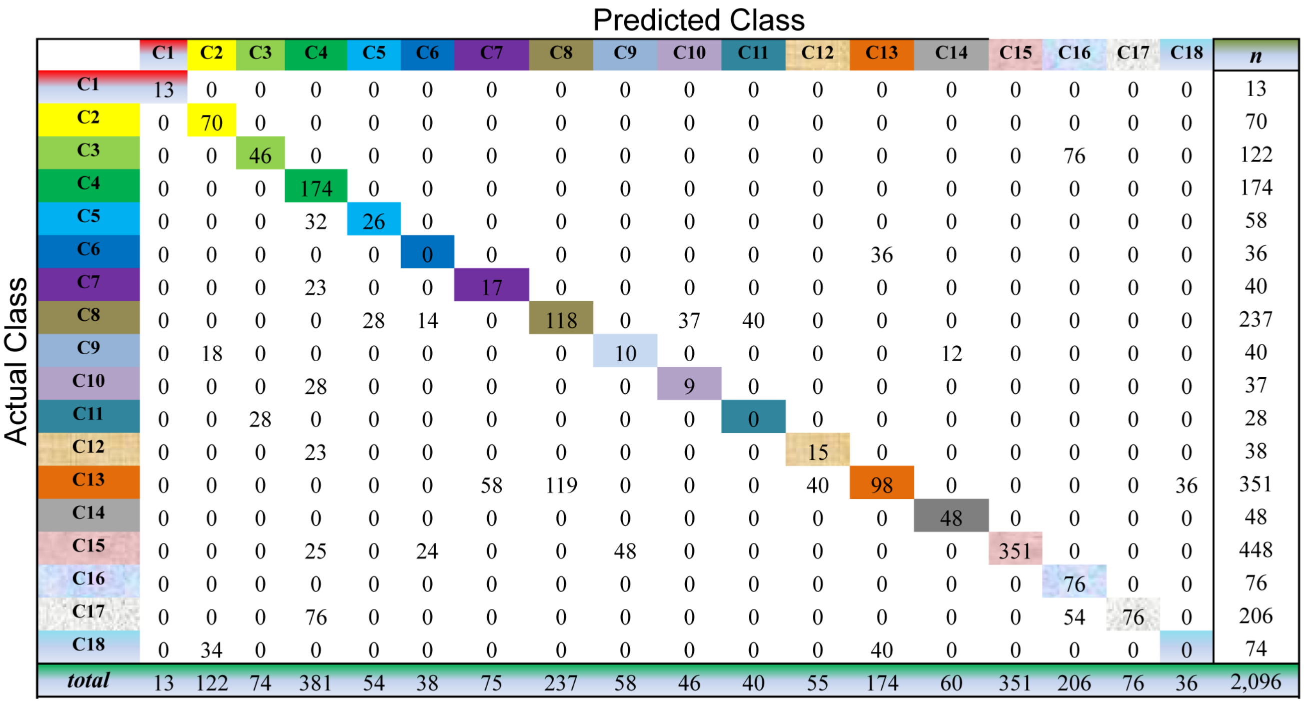


Predicted Class / Actual Class

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	*n*
C1	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13
C2	0	70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	70
C3	0	0	55	0	0	0	0	0	0	0	0	0	0	0	0	67	0	0	122
C4	0	0	0	165	0	0	0	0	0	9	0	0	0	0	0	0	0	0	174
C5	0	0	0	27	31	0	0	0	0	0	0	0	0	0	0	0	0	0	58
C6	0	0	0	0	0	36	0	0	0	0	0	0	0	0	0	0	0	0	36
C7	0	0	0	10	0	0	30	0	0	0	0	0	0	0	0	0	0	0	40
C8	0	0	0	0	28	14	0	118	0	37	40	0	0	0	0	0	0	0	237
C9	0	18	0	0	0	0	0	0	0	0	0	0	0	22	0	0	0	0	40
C10	0	0	0	9	0	0	0	0	0	28	0	0	0	0	0	0	0	0	37
C11	0	0	11	0	0	0	0	0	0	0	17	0	0	0	0	0	0	0	28
C12	0	0	0	14	0	0	0	0	0	0	0	24	0	0	0	0	0	0	38
C13	0	0	0	0	0	0	49	119	0	0	0	40	107	0	0	0	0	36	351
C14	0	0	0	0	0	0	0	0	0	0	0	0	0	48	0	0	0	0	48
C15	0	0	0	25	0	29	0	0	48	0	0	0	0	0	346	0	0	0	448
C16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	76	0	0	76
C17	0	0	0	76	0	0	0	0	0	0	0	0	0	0	0	54	76	0	206
C18	0	34	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	74
total	13	122	66	326	59	79	79	237	48	74	57	64	147	70	346	197	76	36	2,096

Fig 12. Confusion matrix of prediction results using LDS, which achieves a total accuracy of 59.16%. Matrix delineates distribution of actual compared with predicted class membership for multiclass cancer prediction on the Acute Leukemia dataset.

doi:10.1371/journal.pone.0138814.g012



Predicted Class

Actual Class

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	*n*
C1	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13
C2	0	70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	70
C3	0	0	46	0	0	0	0	0	0	0	0	0	0	0	0	76	0	0	122
C4	0	0	0	174	0	0	0	0	0	0	0	0	0	0	0	0	0	0	174
C5	0	0	0	32	26	0	0	0	0	0	0	0	0	0	0	0	0	0	58
C6	0	0	0	0	0	0	0	0	0	0	0	0	36	0	0	0	0	0	36
C7	0	0	0	23	0	0	17	0	0	0	0	0	0	0	0	0	0	0	40
C8	0	0	0	0	28	14	0	118	0	37	40	0	0	0	0	0	0	0	237
C9	0	18	0	0	0	0	0	0	10	0	0	0	0	12	0	0	0	0	40
C10	0	0	0	28	0	0	0	0	0	9	0	0	0	0	0	0	0	0	37
C11	0	0	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	28
C12	0	0	0	23	0	0	0	0	0	0	0	15	0	0	0	0	0	0	38
C13	0	0	0	0	0	0	58	119	0	0	0	40	98	0	0	0	0	36	351
C14	0	0	0	0	0	0	0	0	0	0	0	0	0	48	0	0	0	0	48
C15	0	0	0	25	0	24	0	0	48	0	0	0	0	0	351	0	0	0	448
C16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	76	0	0	76
C17	0	0	0	76	0	0	0	0	0	0	0	0	0	0	0	54	76	0	206
C18	0	34	0	0	0	0	0	0	0	0	0	0	40	0	0	0	0	0	74
total	13	122	74	381	54	38	75	237	58	46	40	55	174	60	351	206	76	36	2,096

Fig 13. Confusion matrix of prediction results using TSVM, which achieves a total accuracy of 54.72%. Matrix delineates distribution of actual compared with predicted class membership for multiclass cancer prediction on the Acute Leukemia dataset.

doi:10.1371/journal.pone.0138814.g013

Table 4. Sensitivity and Specificity of the compared methods over 18 cancer subtypes on the Acute Leukemia dataset.

	Sensitivity					Specificity				
	CNMF	SCNMF	TSVM	LDS	Semi-PNMF	CNMF	SCNMF	TSVM	LDS	Semi-PNMF
C1	0	0.85	**1**	**1**	0	**1**	**1**	**1**	**1**	**1**
C2	0.5	0.93	**1**	**1**	0.91	**1**	0.99	0.97	0.97	**1**
C3	0.52	**0.84**	0.38	0.45	0.79	0.98	0.98	**0.99**	**0.99**	**0.99**
C4	0.76	0.92	**1**	0.95	0.91	**1**	0.99	0.89	0.92	0.97
C5	0.41	**0.97**	0.45	0.53	**0.97**	**0.99**	0.98	**0.99**	**0.99**	**0.99**
C6	0.19	0.89	0	**1**	0.89	**0.99**	**0.99**	0.98	0.98	**0.99**
C7	0.18	**0.93**	0.43	0.75	0.8	**1**	0.97	0.97	0.97	0.99
C8	**0.71**	0.31	0.5	0.5	0.45	0.9	1	0.94	0.94	**1**
C9	0.38	**0.9**	0.25	0	0.73	0.98	0.99	0.98	0.98	**1**
C10	0.14	**0.92**	0.24	0.76	**0.92**	**1**	0.99	0.98	0.98	**1**
C11	0.07	**1**	0	0.61	0.36	**1**	0.98	0.98	0.98	**1**
C12	0.13	**0.79**	0.39	0.63	0.37	**0.99**	0.98	0.98	0.98	0.98
C13	**0.7**	0.22	0.28	0.3	0.53	0.93	1	0.96	0.98	**1**
C14	0.21	0.75	**1**	**1**	0.13	0.98	0.94	**0.99**	**0.99**	0.98
C15	0.98	0.99	0.78	0.77	**1**	**1**	0.99	**1**	**1**	0.97
C16	0.63	0.92	**1**	**1**	0.93	0.99	0.98	0.94	0.94	0.97
C17	0.86	0.57	0.37	0.37	**0.9**	0.92	0.97	**1**	**1**	0.89
C18	0.05	**0.64**	0	0	0.14	0.99	0.97	0.98	0.98	**1**
Avg.	0.41	**0.8**	0.5	0.646	0.65	0.98	0.9828	0.9733	0.9761	**0.9844**

Each row indicates the specific cancer sub-style corresponding to each row of Table 3.

doi:10.1371/journal.pone.0138814.t004

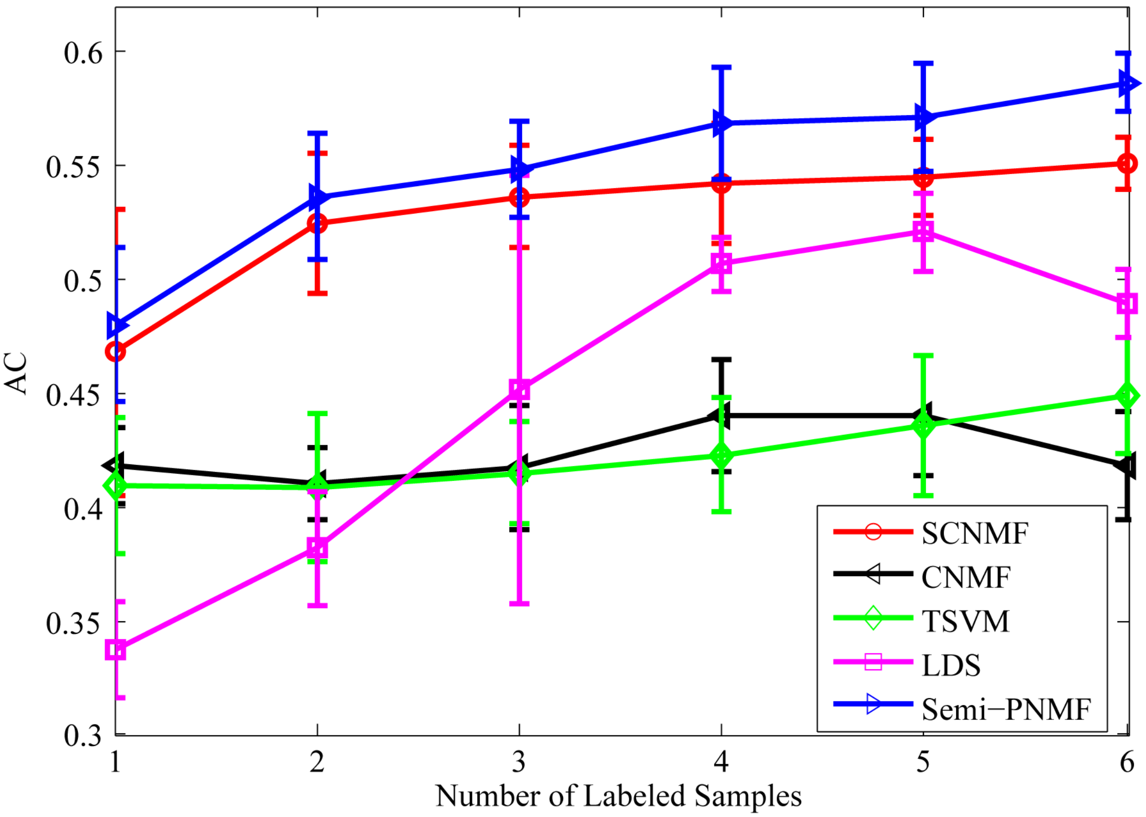


Fig 14. Average accuracies versus different numbers (1–6) of the labeled samples for each class on the Acute Leukemia dataset.

doi:10.1371/journal.pone.0138814.g014

their labels, meanwhile representing the unlabeled samples using the identical cluster centroid. To optimize Semi-PNMF, we devised the multiplicative update rule (MUR) to establish the convergence guarantee. Experiments of cancer classification on two real-world datasets show that Semi-PNMF outperforms the representative methods in terms of quantity.

Recently, Bayesian methods that incorporate both sparsity and a large number of covariates in the model have been extensively used for parameter estimation and classification in data sets compared to small sample sizes such as gene expression data [39–41]. They also improve model accuracy by introducing a slight bias in the model [40]. In future works, we can borrow from the merits of Bayesian methods to further improve the classification performance of Semi-PNMF for a large-scale dataset. Semi-PNMF has provided a flexible framework for learning methods in cancer data processing and can be utilized in other applications such as cancer recurrence [42, 43].

Author Contributions

Conceived and designed the experiments: XZ NG. Performed the experiments: XZ ZJ. Analyzed the data: XZ NG. Contributed reagents/materials/analysis tools: XZ ZJ NG XQ ZL. Wrote the paper: XZ NG.

References

1. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science. 1999; 286 (5439):531–537. doi: 10.1126/science.286.5439.531 PMID: 10521349

2. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, et al. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proceedings of the National Academy of Sciences. 1999; 96(12):6745–6750. doi: 10.1073/pnas.96.12.6745

3. Nguyen DV, Rocke DM. Multi-class cancer classification via partial least squares with gene expression profiles. Bioinformatics. 2002; 18(9):1216–1226. doi: 10.1093/bioinformatics/18.9.1216 PMID: 12217913

4. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000; 403(6769):503–511. doi: 10.1038/35000501 PMID: 10676951

5. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000; 406(6797):747–752. doi: 10.1038/35021093 PMID: 10963602

6. Network CGA, et al. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490 (7418):61–70. doi: 10.1038/nature11412

7. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proceedings of the National Academy of Sciences. 1999; 96(6):2907–2912. doi: 10.1073/pnas.96.6.2907

8. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. Proceedings of the National Academy of Sciences. 2001; 98(26):15149–15154. doi: 10.1073/pnas.211566398

9. Bicciato S, Luchini A, Di Bello C. PCA disjoint models for multiclass cancer analysis using gene expression data. Bioinformatics. 2003; 19(5):571–578. doi: 10.1093/bioinformatics/btg051 PMID: 12651714

10. Tan Y, Shi L, Tong W, Hwang GG, Wang C. Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models. Computational Biology and Chemistry. 2004; 28(3):235–243. doi: 10.1016/j.compbiolchem.2004.05.002 PMID: 15261154

11. Tan Y, Shi L, Tong W, Wang C. Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data. Nucleic acids research. 2005; 33(1):56–65. doi: 10.1093/nar/gki144 PMID: 15640445

12. Rui, X, Anagnostopoulos, G. Multiclass Cancer Classification Using Semisupervised Ellipsoid ARTMAP and Particle Swarm Optimization with Gene Expression Data. 2007;.

13. Shi M, Zhang B. Semi-supervised learning improves gene expression-based prediction of cancer recurrence. Bioinformatics. 2011; 27(21):3017–3023. doi: 10.1093/bioinformatics/btr502 PMID: 21893520

14. Chapelle O, Zien A. Semi-supervised classification by low density separation. In: Proceedings of the 10-th International Workshop on Artificial Intelligence and Statistics, 2005. p. 57–64.

15. Maulik U, Mukhopadhyay A, Chakraborty D. Gene-expression-based cancer subtypes prediction through feature selection and transductive SVM. IEEE Transactions on Biomedical Engineering. 2013; 60(4):1111–1117. doi: 10.1109/TBME.2012.2225622 PMID: 23095982

16. Vapnik VN, Vapnik V. Statistical learning theory. Wiley New York; 1998.

17. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. Nature. 1999; 401(6755):788–791. doi: 10.1038/44565 PMID: 10548103

18. Lee DD, Seung HS. Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems; 2001. p. 556–562.

19. Guan N, Tao D, Luo Z, Yuan B. NeNMF: an optimal gradient method for nonnegative matrix factorization. IEEE Transactions on Signal Processing. 2012; 60(6):2882–2898. doi: 10.1109/TSP.2012.2190406

20. Cho Y, Saul LK. Nonnegative Matrix Factorization for Semi-supervised Dimensionality Reduction. arXiv preprint arXiv:11123714. 2011.

21. Yang Z, Oja E. Linear and nonlinear projective nonnegative matrix factorization. IEEE Transactions on Neural Networks. 2010; 21(5):734–749. doi: 10.1109/TNN.2010.2041361 PMID: 20350841

22. Guan N, Huang X, Lan L, Luo Z, Zhang X. Graph based semi-supervised non-negative matrix factorization for document clustering. In: 2012 IEEE 11th International Conference on Machine Learning and Applications (ICMLA); 2012. p.404–408.

23. Shahnaz F, Berry MW, Pauca VP, Plemmons RJ. Document clustering using nonnegative matrix factorization. Information Processing & Management. 2006; 42(2):373–386. doi: 10.1016/j.ipm.2004.11.005

24. Liu H, Wu Z, Li X, Cai D, Huang TS. Constrained nonnegative matrix factorization for image representation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012; 34(7):1299–1311. doi: 10.1109/TPAMI.2011.217

25. Lan L, Guan N, Zhang X, Tao D, Luo Z. Soft-constrained nonnegative matrix factorization via normalization. In: 2014 IEEE International Joint Conference on Neural Networks (IJCNN); 2014. p. 3025–3030.

26. Zhang X, Guan N, Lan L, Tao D, Luo Z. Box-constrained projective nonnegative matrix factorization via augmented Lagrangian method. In: 2014 IEEE International Joint Conference on Neural Networks (IJCNN); 2014. p. 1900–1906.

27. Lee CM, Mudaliar MA, Haggart D, Wolf CR, Miele G, Vass JK, et al. Simultaneous non-negative matrix factorization for multiple large scale gene expression datasets in toxicology. PloS ONE. 2012; 7(12): e48238. doi: 10.1371/journal.pone.0048238 PMID: 23272042

28. Taslaman L, Nilsson B. A framework for regularized non-negative matrix factorization, with application to the analysis of gene expression data. PloS ONE. 2012; 7(11):e46331. doi: 10.1371/journal.pone.0046331 PMID: 23133590

29. Zhang S, Li Q, Liu J, Zhou XJ. A novel computational framework for simultaneous integration of multiple types of genomic data to identify microRNA-gene regulatory modules. Bioinformatics. 2011; 27(13): i401–i409. doi: 10.1093/bioinformatics/btr206 PMID: 21685098

30. Fogel P, Young SS, Hawkins DM, Ledirac N. Inferential, robust non-negative matrix factorization analysis of microarray data. Bioinformatics. 2007; 23(1):44–49. doi: 10.1093/bioinformatics/btl550 PMID: 17092989

31. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the national academy of sciences. 2004; 101(12):4164–4169. doi: 10.1073/pnas.0308531101

32. Li SZ, Hou X, Zhang H, Cheng Q. Learning spatially localized, parts-based representation. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition; 2001. p. 207–212.

33. Hoyer PO. Non-negative matrix factorization with sparseness constraints. The Journal of Machine Learning Research. 2004; 5:1457–1469.

34. Gao Y, Church G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. Bioinformatics. 2005; 21(21):3970–3975. doi: 10.1093/bioinformatics/bti653 PMID: 16244221

35. Yuan Z, Oja E. Projective nonnegative matrix factorization for image compression and feature extraction. In: Image Analysis. Springer; 2005. p. 333–342.

36. Staunton JE, Slonim DK, Coller HA, Tamayo P, Angelo MJ, Park J, et al. Chemosensitivity prediction by transcriptional profiling. Proceedings of the National Academy of Sciences. 2001; 98(19):10787–10792. doi: 10.1073/pnas.191368598

37. Zhang H, Yang Z, Oja E. Adaptive multiplicative updates for projective nonnegative matrix factorization. In: Neural Information Processing. Springer; 2012. p. 277–284.

38. Yu K, Zhang T, Gong Y. Nonlinear Learning using Local Coordinate Coding. In: Advances in Neural Information Processing Systems. vol. 22; 2009. p. 2223–2231.

39. Kyounghwa B, Mallick B. Gene Selection Using a Two-level Hierarchical Bayesian Model. Bioinformatics. 2004; 20(18):3423–3430. doi: 10.1093/bioinformatics/bth419

40. Behrouz M, Deng L, Homayouni R. Application of Sparse Bayesian Generalized Linear Model to Gene Expression Data for Classification of Prostate Cancer Subtypes. Bioinformatics. 2014; 20(18):3423–3430.

41. Li J, Das K, Fu G, Li R, Wu R. The Bayesian Lasso for Genome-wide Association Studies. Bioinformatics. 2011; 27(4):516–523. doi: 10.1093/bioinformatics/btq688 PMID: 21156729

42. Shi M, Zhang B. Semi-supervised Learning Improves Gene Expression-based Prediction of Cancer Recurrence. Bioinformatics. 2011; 27(21):3017–3023. doi: 10.1093/bioinformatics/btr502 PMID: 21893520

43. Hofree M, Shen J, Carter H, Gross A, Ideker T. Network-based Stratification of Tumor Mutations. Nature methods. 2013; 10(11):1108–1115. doi: 10.1038/nmeth.2651 PMID: 24037242