Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(4): 6652–6665.
DOI: 10.3934/mbe.2023286
Received: 15 October 2022
Revised: 25 December 2022
Accepted: 08 January 2023
Published: 03 February 2023

Research article

An adaptive feature selection algorithm based on MDS with uncorrelated constraints for tumor gene data classification

Wenkui Zheng[1], Guangyao Zhang[2,*], Chunling Fu[3] and Bo Jin[2]

[1] School of Computer and Information Engineering, Henan University, Kaifeng 475004, China
[2] School of Artificial Intelligence , Henan University, Zhengzhou 450046, China
[3] School of Physics and Electronics, Henan University, Kaifeng 475004, China

* **Correspondence:** Email: zgy20200210@163.com.

Abstract: The developing of DNA microarray technology has made it possible to study the cancer in view of the genes. Since the correlation between the genes is unconsidered, current unsupervised feature selection models may select lots of the redundant genes during the feature selecting due to the over focusing on genes with similar attribute. which may deteriorate the clustering performance of the model. To tackle this problem, we propose an adaptive feature selection model here in which reconstructed coefficient matrix with additional constraint is introduced to transform original data of high dimensional space into a low-dimensional space meanwhile to prevent over focusing on genes with similar attribute. Moreover, Alternative Optimization (AO) is also proposed to handle the nonconvex optimization induced by solving the proposed model. The experimental results on four different cancer datasets show that the proposed model is superior to existing models in the aspects such as clustering accuracy and sparsity of selected genes.

Keywords: unsupervised feature selection; gene expression data; cancer classification; uncorrelated constraint; structure learning

1. Introduction

Cancer is considered to be the greatest threat against human health and a significant barrier for life expectancy increasing in the world. An early diagnosis is significant for differentiation of the characteristics of the tumor and enhancing the quality of life of the patients. At present, there is much data that can be used to assist cancer early diagnosis, in which gene expression data obtained by DNA microarray technology is considered to be the most important data. However, directly using raw gene expression data to classify cancer not only induces high storage costs and huge computational burden but also leads to poor clustering or classification accuracy due to the high dimensionality and

redundancy of the raw gene expression data. Therefore, extracting the key genes related to tumors from the large number of genes in the raw data, i.e., feature selection, is one of the noteworthy research areas for cancer classification.

Using the feature selection method can extract key genes strongly relevant to the certain cancer. Key genes extracted from the raw gene data should have properties such as low redundancy, low dimensionality and high classification power. Feature selection methods can be divided into three categories: filtering method, wrapper method and embedded method. Filtering methods score genes according to rules; for example: divergence or correlation, to select genes first, and then employ the selected genes to train filter [1, 2]. While easy to be fulfilled, filtering methods may lose the structural information concealed in the raw data since correlation between genes are unconsidered. Wrapper methods (e.g., RFE-SVM [3]) use the classifier to evaluate and update the selected genes iteratively. Although it can achieve higher accuracy rates comparing with filtering methods, wrapper methods suffer heavy computational burdens.

The embedded method selects the genes by following the objective function of a certain model, thus can retain the structural information concealed in the raw gene data. In addition, it's computational complexity is lower than that of the wrapper method. At present, the embedded methods can be divided into two categories: supervised feature selection (SFS) and unsupervised feature selection (UFS), according to whether labels are required. The SFS methods (e.g., Max-Relevance and Min-Redundancy (mRMR) [4], Regularized logistic regression (RLR) [5], etc.) employ labels to select genes with discrimination capability. Obviously, SFS methods is unsuitable in the scene which data label is unavailable. Although labelling data can handle this problem, it will consume a lot of time and labors. To extract key genes from the unlabeled data, the UFS models including Spectral Feature Selection (SPEC), Unsupervised Discriminative Feature Selection (UDFS), non-negative discriminant feature selection algorithm (NDFS) and Joint Embedding Learning and Sparse Regression (JELSR) [6–9] are proposed. Although the above methods can reveal the local flow structure of the gene data, clustering performance of them is unsatisfied since overall structural information concealed in gene data is unconsidered by them [10]. To handle this problem, Jin et al. proposed the adaptive feature selection method (MDS-AUFS) [11], which first employs multidimensional deflation to transform high dimensional original data into a low dimensional space while preserving the global structure of the original gene data. Then, the sparse regression item and probabilistic neighborhood graph are introduced in the proposed model to sparse select genes and to preserve the local structure of the low dimensional data, respectively. However, since Jin's method tends to extract the highly correlated genes, the genes extracted by the method loses the diversity of the genes, which may deteriorate to the clustering ability of the model [12]. Although generalized irrelevance constraints is introduced by Li to preserve the diversity of the selected genes [13], how to preserve the global structure of the original data is still unconsidered in this work. Namely, Li's method is also faced the problem that the inherent structure of the original data is changed.

To address the above problem, inspired by works of Jin and Li, we propose a new unsupervised feature selection model here which is called adaptive unsupervised feature selection for multidimensional deformation based on uncorrelated constraints (UCMAFS). Specifically, the objective function of the proposed model consists of three items. The first item employs reconstructed coefficient matrix to map high-dimensional original genetic data into a low-dimensional space which is obtained by using multidimensional deflation method (MDS). The second item encourages

reconstructed coefficient matrix to be row sparsity, which may be suitable for selecting sparse features related to the tumor. The third item employs probability neighborhood matrix to encourage reconstructed coefficient matrix preserving local flow structure of the original data. Moreover, an uncorrelated constraint is imposed on the reconstructed coefficient matrix to preserve the diversity of the selected genes. The contributions of this paper are summarized as follows.

1) A new unsupervised feature selection model is proposed by integrating structure learning and sparse feature selection together.
2) The reconstructed coefficient matrix and its $\ell_{2,1}$-norm constraint are introduced in proposed model to map high dimension original data into a low dimensional space meanwhile to guarantee sparsity of the selected features.
3) Global and local structures of original data are preserved by using MDS method and probabilistic neighborhood matrices respectively in the proposed model.
4) An uncorrelated constraint is imposed on the reconstructed coefficient matrix of the proposed model to enhance the diversity of the selected genes.

The subsequent part of the paper is organized as follows, in Section 2, we present the UCMAUFS model in detail and give its optimization algorithm. Comparative experiment with some benchmark models on real cancer datasets are designed to demonstrate the effectiveness of the UCMAUFS in the Section 3. Conclusions are given in Section 4.

2. Adaptive feature selection algorithm based on multidimensional deflation with uncorrelated constraints imposed

In this paper, we use $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ...\boldsymbol{x}_n] \in R^{d\times n}$ to represent the gene expression data, where $\boldsymbol{x}_i \in R^d$ denotes the i-th sample and n is the number of samples. $\boldsymbol{I}_n \in R^{n\times n}$ is an identity matrix and $\mathbf{1}_n$ is an n-dimensional vector with all elements of 1. The trace of a matrix $\boldsymbol{A} = (a_{i,j}) \in R^{n\times n}$ is written as $Tr(\boldsymbol{A})$, and $\|\boldsymbol{A}\|_{2,1}$ represents the $\ell_{2,1}$-norm of the matrix $\boldsymbol{A}$:

$$\|\boldsymbol{A}\|_{2,1} = \sum_{i=1}^{n} (\sum_{j=1}^{n} a_{i,j}^2)^2 \tag{1}$$

2.1. Proposed objective function

Although UFS algorithm can select genes by revealing the inherent structural information of unlabelled data, it ignores the correlation information concealed in the genes, so highly correlated genes may coexist in the selected genes, which reduces the diversity of selected genes, thus deteriorate the clustering performance of the model. To preserve the inherent structure of the data and the diversity of the genes simultaneously, we propose the UCMAUFS model here, whose objective function is shown below.

$$\begin{aligned} &\min_{\boldsymbol{W},\boldsymbol{P}} \{\|\boldsymbol{W}^T\boldsymbol{X} - \boldsymbol{Y}\|_F^2 + \alpha\|\boldsymbol{W}\|_{2,1} + \beta \sum_{i,j}^{n} (\|\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{x}_j\|^2 p_{i,j} + \lambda p_{i,j}^2)\} \\ &s.t. \quad \boldsymbol{W}^T\boldsymbol{U}_p\boldsymbol{W} = \boldsymbol{I}, 0 \le p_{i,j} \le 1, \textstyle\sum_{j=1}^{n} p_{i,j} = 1 \end{aligned} \tag{2}$$

where α and β are regularization parameters to balance the fitting accuracy of adaptive structure learning and the sparsity of the feature selection coefficient matrix. $\boldsymbol{W} \in R^{d\times p}$ is the reconstructed coefficient matrix which transforms high dimensional original data $\boldsymbol{X}$ to the low-dimensional representation $\boldsymbol{Y}$, where $\boldsymbol{Y}$ is obtained by the MDS algorithm, thus the first item of the model can preserve the global structure of the feature space by employing the spatial structure invariant property of the MDS algorithm, while the $\ell_{2,1}$-norm term in the second item forces row of the $\boldsymbol{W}$ matrix to be sparse, giving the $\boldsymbol{W}$ matrix feature selection capability. And $p_{i,j}$ is a element of the probabilistic neighborhood matrix $\boldsymbol{P} \in R^{n\times n}$, whose values is the probabilitie that $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are neighbors of each other, which can be obtained by solving for the following equation:

$$\begin{aligned} &\min_{\boldsymbol{W},\boldsymbol{P}} \sum_{i,j}^{n} \left\{ (\|\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{x}_j\|^2 p_{i,j} + \lambda p_{i,j}^2) \right\} \\ &s.t. \quad 0 \le p_{i,j} \le 1, \textstyle\sum_{j=1}^{\mathrm{n}} p_{i,j} = 1 \end{aligned} \tag{3}$$

where λ is a regularization parameter, and the regularization term $\lambda p_{i,j}^2$ can effectively avoid the trivial solution.

In addition, the probabilistic neighborhood matrix is utilized to preserve the local flow structure concealed in the original data, therefore, a new uncorrelated constraint is proposed here according to the original uncorrelated constraint [13]:

$$\boldsymbol{W}^T(\boldsymbol{X}(\boldsymbol{I}_n + \beta\boldsymbol{L}_p)\boldsymbol{X}^T + \alpha\boldsymbol{M})\boldsymbol{W} = \boldsymbol{W}^T\boldsymbol{U}_p\boldsymbol{W} = \boldsymbol{I} \tag{4}$$

where $\boldsymbol{L}_p = \boldsymbol{D}_p - (\boldsymbol{P}+\boldsymbol{P}^T)/2$, is the Laplace matrix constructed by the probabilistic neighborhood matrix $\boldsymbol{P}$, $\boldsymbol{D}_p$ is a diagonal matrix with the i-th diagonal element which can be described to $\sum_{j=1}^{\mathrm{n}} (p_{i,j} + p_{j,i})/2$. The diagonal matrix $\boldsymbol{M} \in R^{d\times d}$ with i-th diagonal element described in the (5) is proposed to relax the uncorrelated constraint with respect to $\boldsymbol{W}$:

$$m_{i,i} = \frac{1}{2\sqrt{\|\boldsymbol{w}_i\|_2^2 + \varepsilon}} \tag{5}$$

where $\boldsymbol{w}_i$ is the i-th row of $\boldsymbol{W}$. We can find that the (4) allows the model to effectively reduce the redundancy gene selected, thus to find more discriminative genes, leads the classification ability of the model to improving.

2.2. Optimization algorithms

Since the non-convex optimization problem (2) is to minimize an objective function which is composed of two regularization terms subject to three constraints, it is hard to solve directly. Inspired by the optimization methods which are described in the literature [14, 15], we handle the (2) by using the Alternative Optimization (AO), i.e., by given one optimized matrix to update the other, thus an original coupled matrix optimization problem can be transformed into multiple sub-problems to handle iteratively. Details of the UCMAUFS model are written in the Algorithm 1 described as follow.

2.2.1. Given variables $\boldsymbol{W}$ update $\boldsymbol{P}$

Given the $\boldsymbol{W}$, the optimization problem (2) can be described as follow.

Algorithm 1 UCMAUFS

Input: gene expression data matrix $\boldsymbol{X}$, regularization parameter α and β, the clustering number q, number of selected genes s.

Output: The top s ranked genes.

Initialization:

1: Identity matrix to be set as the initialization matrix of the $\boldsymbol{W}$ and $\boldsymbol{M}$; λ is initialized to be 1;
2: A low-dimensional representation matrix $\boldsymbol{Y}$ of $\boldsymbol{X}$ are obtained by using the MDS algorithm;

Repeat:

1: Given variables $\boldsymbol{W}$ update $\boldsymbol{P}$ and λ (*see algorithm2*);
2: Calculate $\boldsymbol{L}_p = \boldsymbol{D}_p - (\boldsymbol{P} + \boldsymbol{P}^T)/2$;
3: Given variables $\boldsymbol{P}$ update $\boldsymbol{W}$ and $\boldsymbol{M}$ (*see algorithm3*);

Until Convergence

Sort all genes based on $\|\boldsymbol{w}_i\|_2$ in descending order. The top s ranked genes are selected.

$$\min_{\substack{0\le p_{i,j}\le 1\\ \sum_{j=1}^{n} p_{i,j}=1}} \sum_{i,j} \left\{ \|\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{x}_j\|_2^2 p_{i,j} + \lambda p_{i,j}^2 \right\} \tag{6}$$

Let $b_{i,j} = \frac{1}{2\lambda}\|\boldsymbol{W}^T\boldsymbol{x}_i - \boldsymbol{W}^T\boldsymbol{x}_j\|^2$, Define the square matrix $\boldsymbol{B} = [\boldsymbol{b}_1^T, \boldsymbol{b}_2^T, ..., \boldsymbol{b}_n^T]^T \in R^{n\times n}$, then (6) can be expressed as:

$$\min_{\substack{0\le p_{i,j}\le 1\\ \mathbf{1}_n\boldsymbol{p}_i=1}} \frac{1}{2}\|\boldsymbol{p}_i + \boldsymbol{b}_i\|^2 \quad \forall i \in \{1, 2, ..., n\} \tag{7}$$

The Lagrange function of problem (7) is described as:

$$L(\boldsymbol{p}_i, \mu, \boldsymbol{v}_i) = \frac{1}{2}\|\boldsymbol{p}_i + \boldsymbol{b}_i\|^2 - \mu(\mathbf{1}_n\boldsymbol{p}_i - 1) - \boldsymbol{v}_i^T\boldsymbol{p}_i \tag{8}$$

where μ and $\boldsymbol{v}_i$ are Lagrangian multipliers. According to the KTT condition, the optimal solution of problem (8) is

$$p_{i,j} = \max(-b_{i,j} + \mu, 0) \tag{9}$$

Sort the elements of each row of matrix $\boldsymbol{B}$ in descending order to form matrix $\boldsymbol{B}^* \in R^{n\times n}$. The value of k can be determined by the following inequality:

$$\begin{cases} B_{i,\bar{k}}^* + \mu > 0 & \text{for } \bar{k} = 1, ..., k \\ B_{i,\bar{k}}^* + \mu \le 0 & \text{for } \bar{k} = k+1, ..., n \end{cases} \tag{10}$$

Since the $\boldsymbol{p}_i$ should follow the constraint $\mathbf{1}_n\boldsymbol{p}_i = 1$, thus μ is further obtained by (11).

$$\mu = \frac{1}{k}(1 - \sum_{d=1}^{k} b_{i,d}^*) \tag{11}$$

Substitute the (11) into the (9) to obtain the optimal value of $\boldsymbol{P}$:

$$p_{i,j} = \max(-b_{i,j} - \frac{1}{k}(1 - \sum_{d=1}^{k} b^*_{i,d}), 0) \tag{12}$$

Similar to the approach proposed in the literature [14], we set the regularization parameter λ according to the number of nearest neighbors k:

$$\lambda = \frac{1}{2n} \sum_{i=1}^{n} (kb^*_{i,k+1} - \sum_{j=1} b_{i,j}) \tag{13}$$

Algorithm 2 Given variables $\boldsymbol{W}$ to update $\boldsymbol{P}$

Input: gene expression data matrix $\boldsymbol{X}$, low-dimensional representation matrix $\boldsymbol{Y}$, Reconstructed coefficient matrix $\boldsymbol{W}$, regularization parameter λ.

Repeat:

1: Obtain the matrix $\boldsymbol{B}$;
2: Determine the value of k by using the inequality (10);
3: Update $\boldsymbol{P}$ according to equation (12);
4: Update λ according to equation (13);

Until Convergence

Output: Probability matrix $\boldsymbol{P}$ and regularization parameter λ.

2.2.2. Given variables $\boldsymbol{P}$ update $\boldsymbol{W}$

Given $\boldsymbol{P}$, the optimization problem is described as

$$\min_{\boldsymbol{W}^T\boldsymbol{U}_P\boldsymbol{W}=\boldsymbol{I}} \{Tr(\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{W} - 2\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{Y}) + Tr(\boldsymbol{W}^T(\alpha\boldsymbol{M})\boldsymbol{W}) + Tr(\boldsymbol{W}^T\boldsymbol{X}(\beta\boldsymbol{L}_p)\boldsymbol{X}^T\boldsymbol{W}) \tag{14}$$

Given $\boldsymbol{W}^T\boldsymbol{U}_p\boldsymbol{W} = \boldsymbol{I}$ and $\boldsymbol{P}$, solving problem (14) is equivalent to solving the following problem.

$$\begin{aligned} &\max_{\boldsymbol{W}} Tr(\boldsymbol{W}^T\boldsymbol{X}\boldsymbol{Y}) \\ &s.t. \quad \boldsymbol{W}^T\boldsymbol{U}_P\boldsymbol{W} = \boldsymbol{I} \end{aligned} \tag{15}$$

Transform the (15) to the (16)

$$\max_{\boldsymbol{Q}^T\boldsymbol{Q}=\boldsymbol{I}} Tr(\boldsymbol{Q}^T\boldsymbol{C}) \tag{16}$$

Where:

$$\boldsymbol{Q} = (\boldsymbol{U}_p)^{\frac{1}{2}}\boldsymbol{W}, \quad \boldsymbol{C} = (\boldsymbol{U}_p)^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{Y} \tag{17}$$

By solving (16) and (17), we can find that $\boldsymbol{W}$ can be described as

$$\boldsymbol{W} = (\boldsymbol{U}_p)^{-\frac{1}{2}}\boldsymbol{Q} \tag{18}$$

In summary, Optimization problem (14) can be solved efficiently [16]. The details are described in Algorithm 3.

Algorithm 3 Given variables $\boldsymbol{P}$ to update $\boldsymbol{W}$

Input: gene expression data matrix $\boldsymbol{X}$, low-dimensional representation matrix $\boldsymbol{Y}$, regularization parameter α and β, Diagonal matrix $\boldsymbol{M}$; Laplace matrix $\boldsymbol{L}_p$.

Repeat:

1: Obtain $\boldsymbol{U}_p = \boldsymbol{X}(\boldsymbol{I}_n + \beta \boldsymbol{L}_p)\boldsymbol{X}^T + \alpha \boldsymbol{M}$;
2: Obtain the matrix $\boldsymbol{C}$ according to equation (17);
3: Obtain $\boldsymbol{U}$ and $\boldsymbol{V}$ by performing a tight singular value decomposition of $\boldsymbol{C}$;
4: Obtain $\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{V}^T$;
5: Update $\boldsymbol{W} = (\boldsymbol{U}_p)^{-\frac{1}{2}}\boldsymbol{Q}$;
6: Update $\boldsymbol{M}$;

Until Convergence

Output: Reconstructed coefficient matrix $\boldsymbol{W}$ and Diagonal matrix $\boldsymbol{M}$.

3. Experiment

3.1. Experimental data

Comparative experiment on the four publicly available cancer gene datasets, including a lung cancer dataset, a colon cancer dataset, a lymphoma dataset and a glioma dataset are employed to evaluate the performance of the UCMSUFS. All dataset were downloaded from https://jundongl.github.io/scikit-feature/datasets.html (accessed on 1 May 2021), and details of the datasets are shown in Table 1.

Table 1. Details of the datasets.

Dataset	No. of samples	No. of genes	No. of Classes
Lung Cancer	203	3312	5
Colon Cancer	62	2000	2
Lymphoma	96	4026	9
Glioma	50	4434	4

3.2. Experimental methods and evaluation metrics

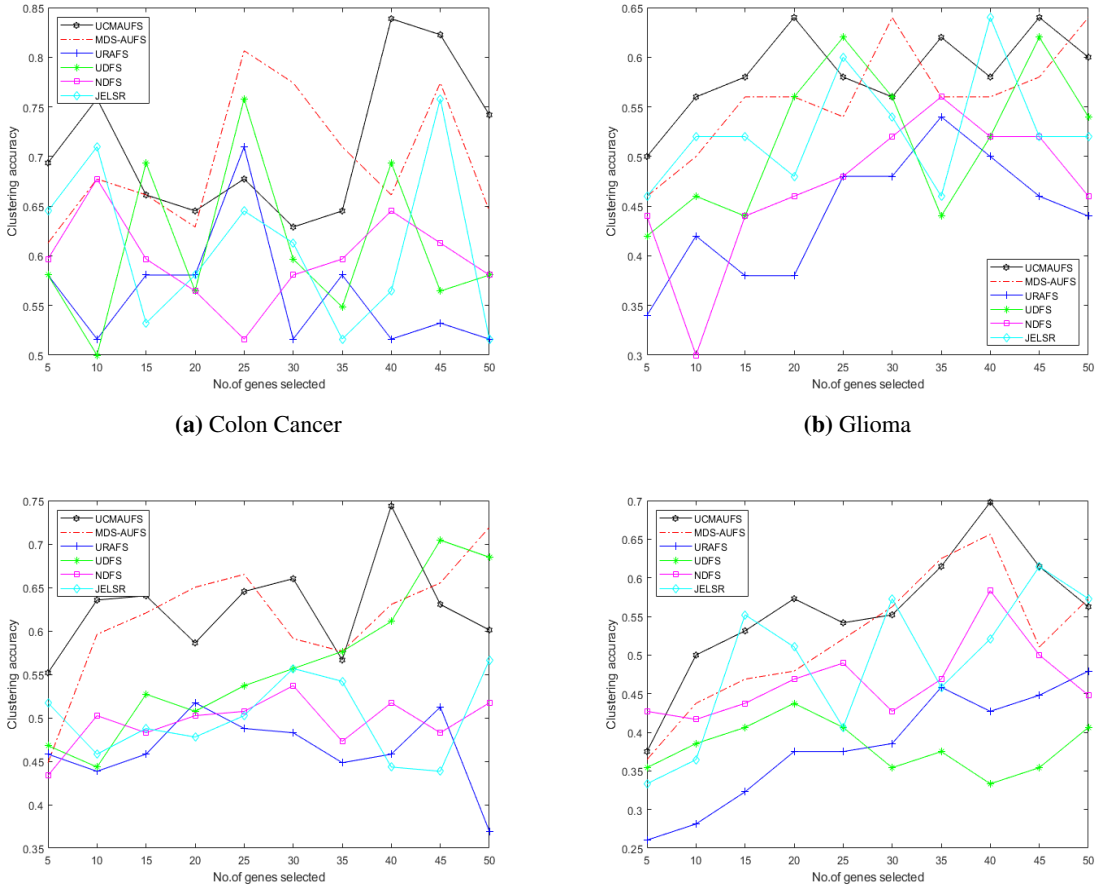
We use the K-means clustering algorithm [17] to evaluate the clustering accuracy (ACC) of the new unsupervised feature selection model (2) proposed in this paper. Specifically, the key genes are extracted by using UCMAUFS model and other comparison models, respectively, and then k-means algorithm are employed to cluster those extracted genes. The accuracy of the clustering is obtained by using (19) finally.

$$ACC = 1/n \sum_{i=1}^{n} \delta(map(c_i), C_i) \tag{19}$$

Where c_i is the cluster label of the i-th sample $\boldsymbol{x}_i$ and C_i is the actual label of $\boldsymbol{x}_i$. $\delta(.)$ is the δ-function and function $map(.)$ is the optimal mapping function [18], which projects each cluster label into the actual label by using the Kuhn-Munkres algorithm. Obviously, larger ACC values indicate higher clustering performance.

3.3. Analysis of experimental results

Comparative experiment with some classical feature selection models including MDS-AUFS [11], URAFS [13], UDFS [7], NDFS [8] and JELSR [9] are employed to evaluate the performance of proposed UCMSUFS model. To smooth ACC fluctuating induced by initial points choosing in the clustering algorithm, we repeat the clustering algorithm 20 times in each experiment and take the average of ACC. In different dataset, the average of ACC versus number of genes selected of different models are plotted in the Figure 1.



(a) Colon Cancer

(b) Glioma

(c) Lung Cancer

(d) Lymphoma

Figure 1. ACC performance of the six models on the four gene expression datasets. (a) Colon Cancer; (b) Glioma; (c) Lung Cancer; (d) Lymphoma.

It can be seen that the beginning, the clustering accuracy of all models on the four data sets almost trend to increasing with the number of selected genes when the number of selected genes is small. However, clustering accuracy of all models on the four data sets show decreasing with the number of selected genes after the clustering accuracy achieves its maximum. It can be inferred that all models perform well due to the reasonable number of selected genes, while further increases in the selected genes may induce redundant genes into models which may weaken the performance of the models.

Moreover, we can find that proposed model shows the almost best clustering performance on all datasets with no more than 40 genes selected. For example, UCMAUFS achieves a maximum clustering accuracy of 83.37% on the colon cancer dataset, much higher than that of MDS-AUFS with 80.65% clustering accuracy and so on. Similarly, on the lymphoma dataset in Figure 1(d), UCMAUFS also shows the best clustering accuracy in all models. Although UCMAUFS, MDS-AUFS and JELSR show the similar clustering accuracy on the glioma dataset, UCMAUFS achieves the highest clustering accuracy with only 20 genes selected (Figure 1(b)) which indicates that UCMAUFS has the best ability of sparse feature selecting. Noteworthy, in Figure 1(c), the MDS-AUFS model shows the consistent increases of clustering accuracy with number of selected genes increasing, even close to the maximum clustering accuracy of UCMAUFS at the 50 genes selected.

Average and maximum clustering accuracy of each model are respectively described in Figure 2 and Table 2 to clearly exhibit the performance of the proposed model. In addition, the best and second best results are written in bold and underlined text respectively in the Table 2. We can find that proposed UCMAUFS model exhibits apparently best performance in all models.

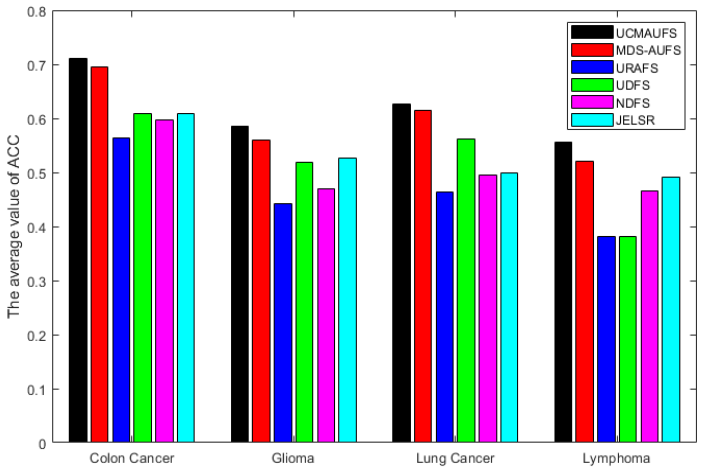


Figure 2. The average ACC of all models on four gene expression datasets.

Table 2. Maximum clusering accuracy of all modles on four gene expression datasets.

Dataset \ Method	URAFS	UDFS	NDFS	JELSR	MDS-AUFS	UCMAUFS
Colon Cancer	0.7097	0.7581	0.6774	0.7581	<u>0.8065</u>	**0.8387**
Glioma	0.5400	0.6200	0.5600	**<u>0.6400</u>**	**<u>0.6400</u>**	**<u>0.6400</u>**
Lung Cancer	0.5172	0.7044	0.5369	0.5665	<u>0.7192</u>	**0.7438**
Lymphoma	0.4792	0.4375	0.5833	0.6146	<u>0.6563</u>	**0.6979**

3.4. Convergence and computational complexity

Computational complexity of UCMAUFS and others comparison models are analyzed and written in Table 3. According to Algorithm 1, the computational complexity of UCMAUFS mainly determined by the reconstructed coefficient matrix $\boldsymbol{W}$ and the probability neighborhood matrix $\boldsymbol{P}$ updating. Since the computation complexity of updating coefficient matrix $\boldsymbol{W}$ and the probability neighborhood matrix $\boldsymbol{P}$ are described as $O(n^3)$ and $O(d^3)$, respectively. Thus, we can conclude that the computational complexity of UCMAUFS is $O(n^3 + d^3)$.

Table 3. Comparison of the computational complexity.

Method	Computational Complexity
URAFS	$O(d^3 + n^2d + n^2)$
UDFS	$O(d^3 + n^2c)$
NDFS	$O(d^3 + n^2c)$
JELSR	$O(d^3 + n^2q)$
MDS-AUFS	$O(min(n, d)^3 + n^2q)$
UCMAUFS	$O(n^3 + d^3)$

Convergences versus number of iterations of UCMAFUS model on four datasets are plotted in the Figure 3. It is clearly that UCMAFUS converges in all gene datasets quickly, almost achieving steady-state within 15 iterations.

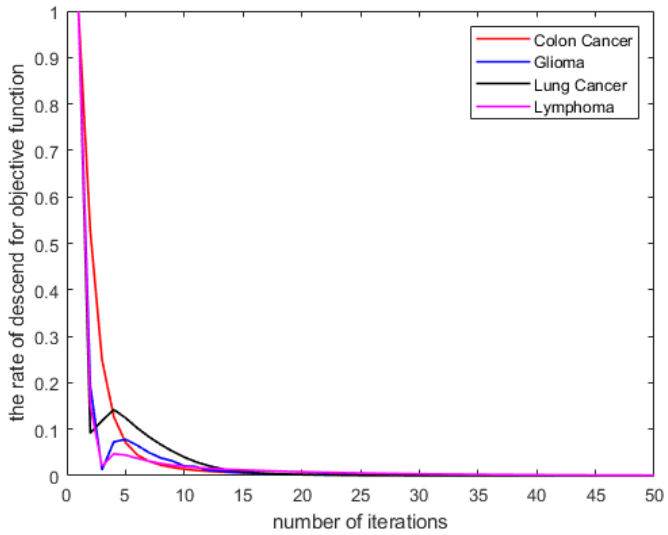
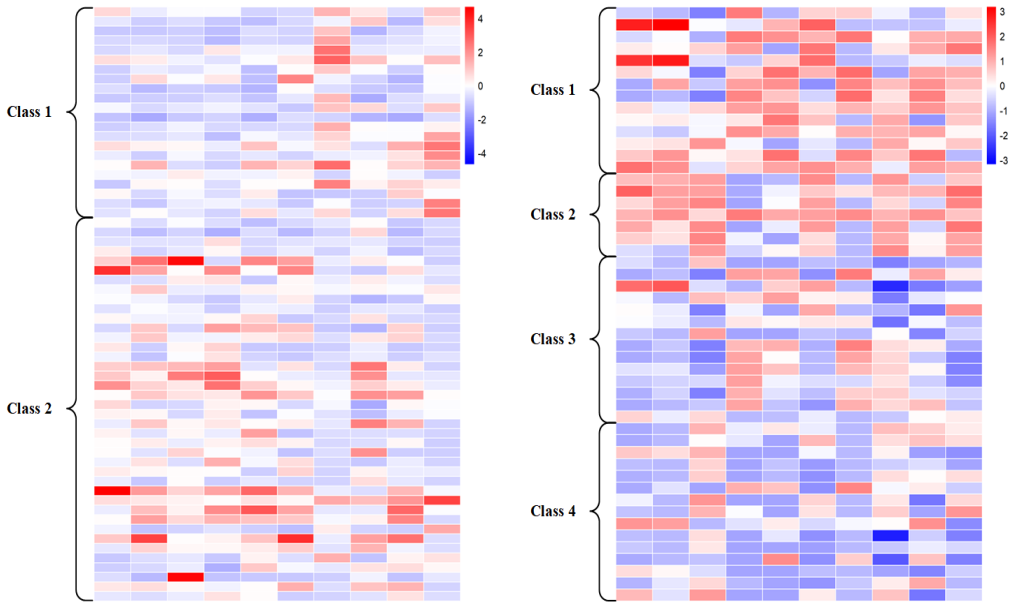


Figure 3. The convergence of UCMAUFS on four gene expression datasets.

3.5. Heatmap of the 10 Top Ranked Genes

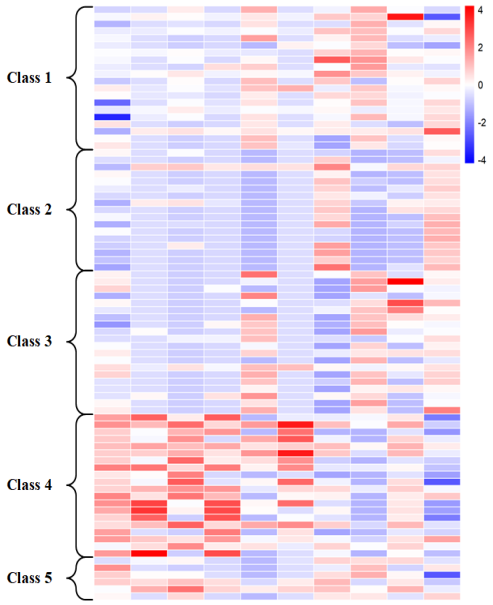
The heatmap is employed to further demonstrate the clustering performance of the UCMAUFS model. Specifically, the heatmap of the top ten genes selected by using UCMAUFS model on the four gene datasets are shown in Figure 4, where redder color block indicates higher expression level of

the corresponding gene while bluer color block indicates lower expression level of the corresponding gene. It can be seen that in the binary dataset Colon, the selected genes of different categories have different gene expression profiles. In other multi-categorical datasets, using a single gene may hard to distinguish different category, however, we can find that, overall, the gene expression profiles of different categories appear apparent distinguish.

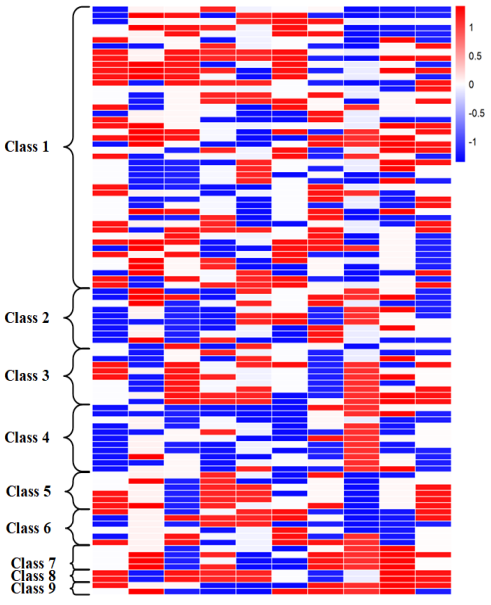


(a) Colon Cancer

(b) Glioma



(c) Lung Cancer



(d) Lymphoma

Figure 4. Heatmap of the 10 top ranked genes of 4 gene expression data. (a) Colon Cancer; (b) Glioma; (c) Lung Cancer; (d) Lymphoma.

To deeply discuss the clustering performance of the proposed model, the T-score and its P-value of the selected top ten genes in the Colon Cancer dataset and the F-score and its P-value of the selected top ten genes in the others datasets are presented in Table 4. It is clearly that seven of the top ten genes in the Colon Cancer dataset have T-score values greater than 2, meanwhile their P-value values are less than 0.05. While the top ten genes in the other three datasets all have large F-scores and its P-values are less than 0.05. This indicates that most of the selected genes of proposed model have good clustering performance.

Table 4. T-score or F-score of the top ten genes in the four datasets.

Colon Cancer		Glioma		Lung Cancer		Lymphoma	
T-score	P-value	F-score	P-value	F-score	P-value	F-score	P-value
2.637	1.06×10^{-2}	4.114	1.10×10^{-2}	18.47	6.49×10^{-13}	6.687	8.55×10^{-7}
2.905	5.16×10^{-3}	5.308	3.00×10^{-3}	29.86	1.94×10^{-19}	2.750	9.00×10^{-3}
2.475	1.62×10^{-2}	8.718	1.09×10^{-4}	24.92	1.01×10^{-16}	6.732	7.73×10^{-7}
4.370	7.39×10^{-4}	9.323	6.30×10^{-5}	56.37	1.11×10^{-31}	4.695	8.40×10^{-5}
2.366	2.13×10^{-2}	5.284	3.00×10^{-3}	10.55	9.02×10^{-8}	3.145	4.00×10^{-3}
1.602	1.14×10^{-1}	9.042	8.10×10^{-5}	17.64	2.14×10^{-12}	4.285	2.23×10^{-4}
4.556	6.50×10^{-7}	3.674	1.90×10^{-2}	6.649	4.90×10^{-5}	3.562	1.00×10^{-3}
1.917	6.00×10^{-2}	4.928	5.00×10^{-3}	23.09	1.13×10^{-15}	9.744	1.46×10^{-9}
1.484	1.43×10^{-1}	3.093	3.60×10^{-2}	3.761	6.00×10^{-3}	3.815	6.98×10^{-4}
3.886	8.70×10^{-5}	13.46	2.00×10^{-6}	10.45	1.07×10^{-7}	2.757	9.00×10^{-3}

4. Conclusions

In this paper, a new sparse embedded model is proposed to sparse selects genes with high relevant to cancer meanwhile to preserve the diversity of selected genes. By handling this new model, we propose an AO algorithm called UCMAUFS here. The comparative experiments with others models on four different cancer datasets shows that UCMAUFS can achieve the maximum clustering accuracy with the least selected genes which indicates UCMAUFS can obtain higher clustering accuracy by filtering out redundant genes.

Conflict of interest

The authors declare there is no conflict of interest.

References

1. S. M. Kopka, A. D. Long, E. T. Ito, L. Tolleri, M. M. Riehle, E. S. Paegle, et al., Global gene expression profiling in Escherichia coli K12: The effects of integration host factor, *J. Biol. Chem.*, **275** (2000), 29672–29684. https://doi.org/10.1074/jbc.M213060200

2. M. Berta, J. M. Renes, M. M. Wilde, Identifying the information gain of a quantum measurement, *IEEE Trans. Inform. Theory*, **60** (2014), 7987–8006. https://doi.org/10.1109/TIT.2014.2365207

3. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesiroy, et al., Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, *Science*, **286** (1999), 531–537. https://doi.org/10.1126/science.286.5439.531

4. H. Peng, F. Long, C. Ding, Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy, *IEEE Trans. Pattern Anal. Mach. Intell,*, **27** (2005), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159

5. L. Y. Li, Z. P. Liu, Biomarker discovery for predicting spontaneous preterm birth from gene expression data by regularized logistic regression, *Comput. Struct. Biotechnol. J.*, **18** (2020), 3434–3446. https://doi.org/10.1016/j.csbj.2020.10.028

6. Z. Zhao, H. Liu, Spectral feature selection for supervised and un-supervised Learning, in *Proceedings of the 24th international conference on Machine learning*, **227** (2007), 1151–1157. https://doi.org/10.1145/1273496.1273641

7. Y. Yang, H. T. Shen, Z. Ma, Z. Huang, X. Zhou, L_2, 1-Norm regularized discrimiNative feature selection for unsupervised learning, in *Proceedings of the 22nd International joint Conference on Artificial Intelligence*, (2011), 1589–1594. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-267

8. Z. Li, Y. Yang, J. Liu, X. Zhou, H. Lu, Unsupervised feature selection using nonnegative spectral analysis, in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Interlligence*, **26** (2012), 1026–1032. https://doi.org/10.1609/aaai.v26i1.8289

9. C. P. Hou, F. P. Nie, D. Y. Yi, Y. Wu, Feature selection via joint embedding Learning and sparse regression, in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, (2011), 1324–1229. https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-224

10. L. Du, Y. D. Shen, Unsupervised feature selection with adaptive structure learning, in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2015), 209–218. https://doi.org/10.1145/2783258.2783345

11. B. Jin, C. L. Fu, Y. Jin, W. Yang, S. B. Li, G. Y. Zhang, et al., An adaptive unsupervised feature selection algorithm based on MDS for tumor gene data classification, *Sensors*, **21** (2021), 3627. https://doi.org/10.3390/s21113627

12. X. Y. Xu, X. Wu, F. L. Wei, W. Zhong, F. P. Nie, A general framework for feature selection under orthogonal regression with global redundancy minimization, *IEEE Trans. Knowl. Data Eng.*, **34** (2021), 5056–5069. https://doi.org/10.1109/TKDE.2021.3059523

13. L. X. Li, H. Zhang, R. Zhang, Y. Liu, Generalized uncorrelated regression with adaptive graph for unsupervised feature selection, *IEEE Trans. Neural Networks Learn. Syst.*, **30** (2019), 1587–1595. https://doi.org/10.1109/TNNLS.2018.2868847

14. M. Yang, L. Zhang, X. C. Feng, D. Zhang, Sparse representation based fisher discrimination dictionary learning for image classification, *International Journal of Computer Vision*, **109** (2014), 209–232. https://doi.org/10.1007/s11263-014-0722-8

15. S. L. Peng, Y. Yang, W. Liu, F. Li, X. K. Liao, Discriminant projection shared dictionary learning for classification of tumors using gene expression data, *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, **18** (2021), 1464–1473. https://doi.org/10.1109/TCBB.2019.2950209
16. J. Huang, F. P. Nie, H. Huang, C. Ding, Robust manifold nonnegative matrix factorization, *ACM Trans. Knowl. Discovery Data*, **8** (2014), 1–21. https://doi.org/10.1145/2601434
17. R. Zhang, X. L. Li, Unsupervised feature selection via data reconstruction and side information, *IEEE Trans. Image Process.*, **29** (2020), 8097–8106. https://doi.org/10.1109/TIP.2020.3011253
18. A. Strehl, J. Ghosh, Cluster ensembles-A knowledge reuse framework for combining multiple partitions, *J. Mach. Learn. Res.*, **3** (2020), 583–617. https://doi.org/10.1162/153244303321897735

AIMS Press