Unsupervised feature selection via two-way ordering in gene expression analysis

Chris H.Q. Ding

NERSC Division, Lawrence Berkeley National Laboratory, University of California, Berkeley, CA 94720, USA

Received on August 7, 2002; revised on October 31, 2002; January 17, 2003; accepted on January 22, 2003

ABSTRACT

Motivation: Selection of genes most relevant and informative for certain phenotypes is an important aspect in gene expression analysis. Most current methods select genes based on known phenotype information. However, certain set of genes may correspond to new phenotypes which are yet unknown, and it is important to develop novel effective selection methods for their discovery without using any prior phenotype information.

Results: We propose and study a new method to select relevant genes based on their similarity information only. The method relies on a mechanism for discarding irrelevant genes. A two-way ordering of gene expression data can force irrelevant genes towards the middle in the ordering and thus can be discarded. Mechanisms based on variance and principal component analysis are also studied. When applied to expression profiles of colon cancer and leukemia, the unsupervised method outperforms the baseline algorithm that simply uses all genes, and it also selects relevant genes close to those selected using supervised methods.

Supplement: More results and software are online: http://www.nersc.gov/~cding/2way.

Contact: chqding@lbl.gov

1 INTRODUCTION

High density DNA microarray technology can simultaneously monitor the expression levels of thousands of genes, reflecting the state of the cell with different protein and mRNA compositions. Thus gene expression profiles of samples corresponding to different pathological states of the same tissue provide molecular rather than morphological signature of the malignancy (Alon et al., 1999; Golub et al., 1999; Alizadeh et al., 2000).

When analyzing expression profiles using machine learning methods, an important issue is feature (gene) selection for the target phenotypes. Of the thousands of genes, only a small number of them show strong correlation with a certain phenotype. For instance, for a two-way cancer/non-cancer diagnose, 50 such informative genes are usually sufficient (Golub et al., 1999).

There are two general types of feature selection methods: filters and wrappers (John et al., 1994). Filter type methods are essentially data pre-processing or data filtering methods. Features are selected based on their relevance to certain classes. These features therefore have more discriminative power. Filtering methods range from simple methods such as information gain, statistical tests (t-test) (Alon et al., 1999; Golub et al., 1999; Ding, 2002) to more sophisticated methods such as Markov blanket based on conditional independence (Koller and Sahami, 1996). The bias in the feature selection does not correlate with the bias in a learning method, therefore they have better generalization property. In wrapper methods, feature selection is ‘wrapped around’ a particular learning algorithm: the usefulness of features is determined by the estimated accuracy of the learning algorithm. One can often obtain a very small subset of features with relatively high accuracy (Li and Yang, 2000; Xiong et al., 2001), because the characteristics in the feature set correlate strongly with those of the learning algorithm.

All above feature selection methods are supervised in nature: they depend on known class information. This presents a problem for clustering analysis for discovering new phenotypes or subtypes: to uncover new phenotypes we need a subset of most relevant features while the selection of these features depends on a priori knowledge of cluster structure that we seek in the first place.

In this paper, we study unsupervised feature selection, i.e., selecting genes without prior phenotype knowledge. One approach to this chicken-and-egg problem is iterative feature filtering (Xing and Karp, 2001): given an initial clustering, one can use supervised methods to select relevant features which are used in turn to obtain an improved clustering. A difficulty with this approach is that the features selected based on the initial clustering are often biased and iterative procedure will not necessarily produce a better clustering (see Tables 2 and 4). In other words, the iterative procedure is easily stuck in a local
fixed point in the high dimensional parameter space.

Is there a way to select relevant features without knowing the correct cluster structure? In this paper, we propose a novel approach to this chicken-and-egg problem. Rather than selecting ‘relevant’ features, we attempt to identify ‘irrelevant’ or non-discriminant features. Discarding these irrelevant features we use the remaining features to obtain improved clustering results, as compared to the clustering results obtained using all features.

We propose a mechanism that can effectively identify non-discriminant features. The mechanism depends on a two-way ordering of the gene expression profiles based on an optimal similarity criterion: genes (rows of the gene expression matrix \( G \)) and tissue samples (columns) are simultaneously re-ordered such that genes (samples) adjacent in the order are similar and genes (samples) far-away from each other are dissimilar. With this two-way ordering, nondiscriminant genes (genes having large similarity to both clusters) will locate near the middle and can be discarded.

Note that the two-way ordering of gene expression profiles has its own merit. In hierarchical clustering, clusters and the members of clusters on the binary tree are often displayed in linear order. Biological and clinical studies are often performed in the context of this linear ordering. In (Eisen et al., 1998), leaf nodes are ordered based on average expression levels and visible structures. In self-organizing maps (Tamayo et al., 1999), clusters are organized as a two-dimensional topological mesh. In (Alon et al., 1999), similarity between nodes and their parent’s siblings are utilized to order leaf nodes. A more quantitative optimal ordering method is proposed by Bar-Joseph et al. (2001) to insures adjacent nodes to have large similarities, which is further improved (Ding, 2002) to ensure that nodes far-away along the order have small similarities. The one-way ordering algorithm in (Ding, 2002) applies only to one-type nodes, either based on gene-gene similarity or based on tissue–tissue similarities. It is extended here to two-way simultaneous ordering of both genes and tissue samples, based on a bipartite graph model.

Our method has similar flavor as the gene-shaving method (Hastie et al., 2000), where genes are clustered via a gradual process of discarding genes with small variance as measured by its first principal component. We can also incorporate variance or PCA as optional mechanisms to discard irrelevant genes.

## 2 UNSUPERVISED FEATURE SELECTION

The UFS algorithm includes (i) identify and discard irrelevant genes; (ii) using remaining genes to perform an initial clustering; (iii) based on the cluster structure obtained in (ii) to select final set of genes using supervised method. Step (i) is the critical part of the algorithm. Steps (ii–iii) are largely for assessing the validity of step (i), where the final set of selected genes in UFS are compared with those selected via supervised method.

The complete UFS algorithm is the following.

1. Pre-processing gene expression data \((m\text{ genes and } n\text{ tissue samples})\). Set the number of initial features \(m_i\) and the number of final features \(m_f\).

2. Identifying and discarding \(m - m_i\) irrelevant genes using:
   - (a) TWO-way-ordering: details in Section 3.
   - (b) Variance: details in Section 2.1.
   - (c) PCA: details in Section 2.1.

3. Based on the \(m_i\) initial features, compute Pearson correlation between tissue samples and construct similarity matrix. Cluster tissues using min–max cut algorithm (Section 4). This gives initial clustering.

4. Based on the initial clustering, use supervised feature selection method (we use t-test) to select \(m_f\) features from entire \(m\) feature set.

### 2.1 Initial feature selection

**Two-way ordering** The main contribution of the paper is to introduce the two-way ordering for bipartite graphs and use it to discard irrelevant features. The details of this new method are described in Section 3. We call this method unsupervised feature selection via two-way ordering (UFS-2way).

**Variance and PCA** We also propose two alternative methods for initial feature selection, motivated by the ideas in (Hastie et al., 2000). The main idea here is that genes with large variance across tissue samples are interesting features, because they explain most of the total variances, a central theme in statistics. Therefore one method is to choose \(m_i\) genes with the \(m_i\) largest variances. We call this method UFS-variance.

PCA is a widely used dimension reduction method in statistics and machine learning. Usually, the first few principal components span the dimensions of the largest variances and are retained. Here we use PCA, but in an unconventional way. Given the gene expression matrix \(G = (g_{ij})\), we compute the first principal component, \(u_1\). Since \(u_1\) is a linear combination of genes, the magnitude of \(u_1(i)\) is indicative of the variance of gene \(i\). We discard \(m - m_i\) genes with the smallest (magnitude) elements in \(u_1\) and retain the rest \(m_i\) genes as initial selected genes. We call this method UFS-PCA.

### 2.2 Graph model for gene expressions

It is convenient to use a graph model for gene expression analysis. For class discovery or clustering, a tissue sample
is represented by a node in a graph and the similarity between two tissue samples \( i, j \) becomes the weight \( w_{ij} \) on the edge between them. We use Pearson correlation \( c(i, j) \) and define the pairwise similarity as \( u_{ij} = \exp(c(i, j)/c) \), where \( c \) is an average correlation. The matrix \( W = (w_{ij}) \) defines an undirected similarity graph. Clustering and node ordering uses the graph model.

### 2.3 Initial clustering with MinMaxCut

Once the initial features are selected (or more precisely, those non-discriminant features are discarded), we perform a cluster analysis on the expression data with these initial features. The cluster structure obtained is the initial clustering. It is important that the quality of this clustering is good. There are several clustering methods previously employed in gene expressions are hierarchical clustering (Eisen et al., 1998), self-organized maps (Tamayo et al., 1999), deterministic annealing (Alon et al., 1999), and graph partitioning methods (Ben-Dor et al., 1999; Sharan and Shamir, 2000; Xing and Karp, 2001).

We use a clustering algorithm which is based on a min-max clustering principle: clustering is a grouping of data such that similarities between different clusters are minimized while similarities within same clusters are maximized (Ding et al., 2001; Ding, 2002).

Given a weighted graph \( G \) with weight matrix \( W = (w_{ij}) \) that measures the similarity between nodes \( i, j \), we wish to partition it into two subgraphs \( A, B \). The similarities between \( A \) and \( B \) is the sum of weights between them, \( s(A, B) = \sum_{i \in A, j \in B} w_{ij} \). The self-similarity within cluster \( A \) is \( s(A, A) \), the sum of all edge weights within \( A \). The self-similarity within cluster \( B \) is \( s(B, B) \). Based on the min–max clustering principle, MinMaxCut minimizes the objective function,

\[
J_{\text{MMC}} = \frac{s(A, B)}{s(A, A)} + \frac{s(A, B)}{s(B, B)}. \tag{1}
\]

An efficient solution is the following. (i) Compute the second lowest eigenvector \( q_2 \) of \( (D - W)q = \xi Dq \), and sort \( q_2 \) to establish a linear order. (ii) A cutpoint \( i_{\text{cut}} \) cuts the graph is into two subgraphs (clusters): \( A = \{q(i) \mid i \leq i_{\text{cut}}\}, \; B = \{q(i) \mid i > i_{\text{cut}}\} \) We search for the optimal cutpoint \( i_{\text{cut}} \) along the linear order, such that \( J_{\text{MMC}}(A, B) \) is minimized. (iii) Perform refinement on \( A \) and \( B \) based on \( J_{\text{MMC}} \).

By construction, \( J_{\text{MMC}} \) is a measure of the quality of clustering on a dataset. For well separated clusters, \( s(A, B) \) is small, so is \( J_{\text{MMC}} \). When clusters overlap significantly, \( s(A, B) \) becomes comparable to within-cluster self-similarity and thus \( J_{\text{MMC}} \sim 2 \). Thus the smaller \( J_{\text{MMC}} \) the better quality of the clustering. Irrelevant genes contribute mostly to overlap \( s(A, B) \). Their removal will enhance the cluster separation and therefore reduce \( J_{\text{MMC}} \).

### 2.4 Final feature selection

Given the cluster structure obtained in the initial clustering, we now use supervised method to select the final features. The final features can be compared with those obtained by supervised method, as a quality assessment of this unsupervised feature selection method. Here we select features from all the original features. using statistic test (t-test). For a gene \( g = (g_1, g_2, \ldots, g_n) \) with expression across \( n \) samples, \( t \)-value is defined as

\[
t = \frac{\bar{g}_1 - \bar{g}_2}{\sigma}, \quad \sigma^2 = \frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n - 2} \tag{2}
\]

where \( \bar{g}_k \) and \( \sigma_k^2 \) are the mean and variance of gene expressions within class \( C_k \) (of size \( n_k \)). We select final \( m_f \) genes with \( m_f \) largest \( t \)-values. Information gain is also shown to be effective, but it requires a discretization procedure to turn continuous values into probabilities.

### 3 NODE ORDERING

#### 3.1 One-way ordering for undirected graphs

The objective of node ordering is to insure that (i) adjacent nodes are most similar; (ii) near neighbor nodes are less similar; (iii) nodes far away are least similar. The pairwise similarity \( w_{ij} \) between two nodes \( i, j \) are the edge weights on the undirected graph (see Section 2.1). One appropriate, distance-sensitive, objective function is the following (Ding, 2002):

\[
\min_{\pi} J_d(\pi), \quad J_d(\pi) = \sum_{\ell=1}^{n-1} c^2 J_{d=\ell}(\pi) \tag{3}
\]

where \( J_{d=\ell}(\pi) = \sum_{i,j=1}^{n-\ell} w_{\pi_i,\pi_{i+\ell}} \) and the index permutation \( \pi = (\pi_1, \ldots, \pi_n) \) defines the order. Here large distance similarities are penalized with larger weights to ensure that the larger the distance between a pair of nodes, the less similar these two nodes are. We may rewrite the distance-sensitive objective function \( J_d \) as

\[
J_d(\pi) = 4w + \sum_{\ell=3}^{n-1} (\ell^2 - 4) J_{d=\ell}(\pi) - 3J_{d=1}(\pi),
\]

where \( w = \sum_{i,j} w_{ij} \) is the total weight of the graph which is a constant. Therefore, minimizing \( J_d \) is equivalent to simultaneously minimizing large distance similarities and maximizing adjacent similarities \( J_{d=1} \) (which is the objective function used in Bar-Joseph et al. (2001)).

An \( O(n^2) \) approximate algorithm to compute the permutation that minimizes \( J_d \) is the following. Compute the second lowest eigenvector \( p_2 \) of \( (D - W)p = \lambda Dp \). Sort \( p_2 \) to increasing order. This sorting induces an index permutation which is the desired solution. With appropriate modification, this algorithm can also produce an optimal ordering that preserve cluster structure. More details and experimental results are in Ding (2002).
3.2 Two-way ordering for bipartite graphs

In one-way ordering, genes contribute to the definition of similarity between tissue samples; but they do not show up directly in the ordering consideration. Gene expression data as rectangle gene–tissue association matrix $B = (b_{ij})$ can be modeled as a weighted bipartite graph. Each gene is a g-type node and each tissue sample is a s-type node.

In a two-way ordering of the bipartite graph, both tissues and genes show up directly in the ordering, therefore providing mechanism to study the relationship among genes, while they contribute to the similarity between tissues. Our goal here is to simultaneously order the genes and tissue samples with the same objective as undirected tissues. Our goal here is to simultaneously order the genes and tissue samples with the same objective as undirected graphs in Section 3.1. We solve this two-way ordering problem using the same approach there.

Let $g = (g_1, \ldots, g_m)^T$ represent an index permutation of genes, and $s = (s_1, \ldots, s_n)^T$ represent an index permutation of tissue samples. Putting both type of nodes together, we write $p = (g, s)^T$. The symmetric weighted adjacency matrix $W$ for the bipartite graph is

$$
W = \begin{pmatrix}
0 & B \\
B^T & 0
\end{pmatrix}.
$$

The diagonal matrix of node degrees is $D = \text{diag}(D_g, D_s)$, where $D_g = \text{diag} (b_{gn})$, $D_s = \text{diag} (b_{sm})$, and $e = (1, \ldots, 1)^T$ with appropriate size. Here $g$ and $s$ are represented as

$$
z = \begin{pmatrix}
u \\
v
\end{pmatrix} = D^{1/2} p = \begin{pmatrix}D_g^{1/2} g \\
D_s^{1/2} s
\end{pmatrix}.
$$

Putting $W, D, z$ into $(D - W)p = \lambda Dp$ as in Section 3.1, we obtain

$$
\hat{B}v = \zeta u, \quad \hat{B}^Tv = \zeta u, \quad \hat{B} = D_g^{-1/2} B D_s^{-1/2},
$$

where $\zeta = 1 - \lambda$. The solutions to these two equations are the singular value decomposition (SVD) of $\hat{B}$: $\hat{B} = \sum_{k=1}^K u_k \hat{\xi}_k v_k^T$. (This SVD solution is first obtained in slightly different form for clustering (Zha et al., 2001)). Therefore we compute the second principle components $u_2, v_2$, and obtain $g_2 = D_g^{-1/2} u_2$ and $s_2 = D_s^{-1/2} v_2$. We sort $g_2$ to increasing order and use the induced permutation to order genes. At same time, we sort $s_2$ to increasing order and use the induced permutation to order tissue samples. We can show that the two-way ordering leads to the correspondence analysis (Greenacre, 1984).

One advantage of this approach is the computational efficiency. Typically the number of genes $m \sim 5000$ while the number of samples $n \sim 100$. Thus the $n \times n$ matrix $\hat{B}^T \hat{B}$ can be easily solved to obtain $v_2$. Afterwards, we obtain $u_2$ through the relation $u_2 = \hat{B}v_2/\lambda_2$. This way, we never need to solve the much larger $m \times m$ matrix $B \hat{B}^T$.

3.3 Discarding irrelevant genes

The primary advantage of two-way ordering is to locate and discard irrelevant genes. To be precise, we define a gene to be irrelevant if the gene has high associations with both clusters. Irrelevant genes have no discriminative power in both clustering and classification (prediction). The presence of large number of irrelevant genes mix clusters together.

Figure 1 illustrates the most important feature of the two-way ordering. In this synthetic dataset, there are two clusters. We added four genes (rows, in red) that are non-discriminant in regarding to the cluster structure. After two-way ordering, the four genes move to the bordering area between the two clusters.

The reason that irrelevant genes move toward to bordering area can be explained by the design of the ordering objective function: (i) the larger the distance between two
Unsupervised feature selection

genes, the less similar they are. This dictates that genes of different clusters are far away from each other. Thus the two clusters should be on two ends: either one up and one down, or one left and one right (see Fig. 2). (ii) Each gene cluster has high association with one tissue sample cluster. This implies genes near both ends have high association with either tissue sample cluster, therefore are more discriminative. (iii) Irrelevant genes have large associations with both clusters; therefore, they tend to move to the bordering area.

In unsupervised gene selection, we do not know a priori the cluster structures and therefore the bordering area is also unknown. But by (i)–(iii) above, we know that irrelevant genes should be somewhere near the middle, together with those less discriminative genes. If we discard, say, large part of genes in the middle, that should eliminate irrelevant genes and enhance the cluster structure. Note that our goal here is not to clearly identify those irrelevant genes (in fact we can not achieve this without knowing the cluster structure). Rather, we seek to discard a large number of genes which are either irrelevant or not very discriminative, utilizing the properties of the two-way ordering. These observations form the basis for our unsupervised feature selection algorithm.

4 ANALYSIS OF COLON CANCER DATA

We apply UFS to the expression profiles of colon tumor tissues (Alon et al., 1999). There are 40 cancerous samples and 22 normal samples. Each RNA sample was hybridized to oligonucleotide microarray complementary to about 6500 genes and ESTs. The expression data of 2000 genes are provided for analysis. The data are the raw intensities directly obtained from microarray (not log-transformed, nor ratio over control expression as in cDNA microarray). To see the effects of feature selection, we apply the min–max cut clustering to the data, using either all 2000 genes or only the 50 genes selected via t-test based on known class information. Clustering results are specified by a contingency table $T = (t_{ij})$, and the Q-accuracy (Rost and Sander, 1993; Ding and Dubchak, 2001) is defined as $\sum_i t_{ii}/N$. The clustering results has accuracy of 0.71 when all 2000 genes are used. When using the 50 selected genes, clustering accuracy increases to 0.903.

Now we apply UFS to this dataset, setting $B = G$ in Equation (4). The three mechanisms for discarding irrelevant genes are shown in Figure 2. Genes are ordered according to the two-way order (Section 3.2). We use t-value of Equation (2) as a measure of relevance. One sees that genes near the two ends have large t-values while genes near the middle have small t-values, although noise does exist. This confirms our theoretical analysis in Section 3.3.

**Table 1.** Clustering results using genes selected by UFS with two-way ordering for colon cancer/normal classes

<table>
<thead>
<tr>
<th>$m_i$ genes</th>
<th>accuracy</th>
<th>$J_{\text{MMC}}$</th>
<th>$m_f$</th>
<th>accuracy</th>
<th>$J_{\text{MMC}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>0.710</td>
<td>1.61</td>
<td>50</td>
<td>0.742</td>
<td>0.46</td>
</tr>
<tr>
<td>800</td>
<td>0.839</td>
<td>0.70</td>
<td>50</td>
<td>0.839</td>
<td>0.20</td>
</tr>
<tr>
<td>400</td>
<td>0.855</td>
<td>0.53</td>
<td>50</td>
<td>0.839</td>
<td>0.14</td>
</tr>
<tr>
<td>200</td>
<td>0.871</td>
<td>0.32</td>
<td>50</td>
<td>0.871</td>
<td>0.14</td>
</tr>
</tbody>
</table>

For example, $m_i = 400$ are selected using UFS out of all genes. Initial clustering is performed on these 400 genes, with accuracy = 0.855. Based on this initial clustering, $m_f = 50$ genes are selected using t-test. Final clustering is perform on this 50 genes, with accuracy = 0.839.

**Table 2.** Accuracy of the clustering using genes selected by UFS with two-way ordering, PCA and variance

<table>
<thead>
<tr>
<th>Genes</th>
<th>Unsupervised 2-way order</th>
<th>PCA</th>
<th>Variance</th>
<th>Supervised t-stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>800</td>
<td>0.839</td>
<td>0.613</td>
<td>0.629</td>
<td>0.887</td>
</tr>
<tr>
<td>400</td>
<td>0.855</td>
<td>0.532</td>
<td>0.629</td>
<td>0.887</td>
</tr>
<tr>
<td>200</td>
<td>0.871</td>
<td>0.516</td>
<td>0.629</td>
<td>0.871</td>
</tr>
</tbody>
</table>

Accuracy of the clustering using genes selected by supervised method is also listed. For colon/normal classes.

**Fig. 2.** t-value (top), magnitude of first PCA component (middle) and variance (bottom) of 2000 genes based on the two-way ordering. x-axis are gene numbers.
Applying UFS-2way-ordering, with initial number of features \( m_i = 800, 400, 200 \), the results of initial and final clustering are listed in Table 1. The clustering results of \( m_i = 2000 \) serve as the baseline algorithm. From these initial clustering results, UFS outperform the baseline algorithm substantially.

In Table 1, we also give \( J_{MMC} \) values for each clustering as a measure of quality of clustering. For the initial clusterings, as more irrelevant features are discarded, \( J_{MMC} \) decrease steadily, from 1.61 for 2000 genes to 0.32 for 200 genes, indicating more separated cluster structures.

The 50 genes selected by UFS and the 50 genes selected using supervised \( t \)-test are shown in Supplementary materials. An interesting question is how much they overlap? There are 40 genes appear in both final gene sets. This indicates the unsupervised method capture essential knowledge that human expertise has on the class structure. This also indicates that the class structure are reasonably reflected by the pairwise similarity defined via Pearson correlation (see Section 2.1).

We also studied PCA and variance as irrelevant feature removal mechanisms (Section 2.2). In Figure 2, the 1st PCA component and variance for each gene are shown in the order obtained by two-way ordering. One can see that there are strong correlations between PCA and variance, but they have little correlations with \( t \)-value. Using UFS with PCA or variance to select initial 800, 400 and 200 genes, the corresponding initial clustering results are shown in Table 2. UFS-PCA and UFS-variance generally performed poorer than the baseline algorithm with all 2000 genes, and also poorer than UFS-two-way-ordering.

5 ANALYSIS OF LEUKEMIA SUBTYPES

We next apply UFS to the leukemia dataset of Golub et al. (1999). Here we use the training dataset of 7070 gene expressions of 38 tumor tissue samples from Affymetrix oligonucleotide microarray. The target class here are the two phenotypes of the cancer: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML).

Intensity readings from oligonucleotide gene chips are generally non-negative, such as the colon cancer data in Section 4. This is necessary for the two-way ordering of bipartite graphs, since the edge weights must be non-negative. However, due to a software problem in Affymetrix microarray, the intensity readings have many negative values (this problem is resolved in latest Affymetrix products). To apply UFS to this dataset, we pre-process the leukemia data as

\[
b_{ij} = g_{ij} + 1000; \quad 0 \leq b_{ij} \leq 5000.
\]

i.e. we shift all data up by 1000; values below 0 are set to 0 and values above 5000 is set to 5000. This kind of pre-processing (also done in Golub et al., 1999) mainly serve the purpose of removing outliers. Note that shifting by a constant will not change Pearson correlation, \( t \)-test, and many other linear properties. This pre-processing improves the clustering. For examples, using all 7070 genes, the accuracy of clustering is 0.763, comparing to 0.736 obtained without any pre-processing (Ding, 2002). Using 50 genes selected based on supervised \( t \)-test, the clustering accuracy increases to 1.0. Supervised feature selection biases the clustering results toward the pre-defined class labels.

We apply UFS to these gene expressions (see figures in the Supplement). Using UFS with two-way ordering, the results of initial clustering and final clustering are listed in Table 3. Using the baseline, i.e. with all 7070 genes, the accuracy of initial clustering is 0.763. Using \( m_i = 1000 \) genes selected by UFS, the accuracy increases to 0.868. Based on this initial clustering, 50 genes are selected and cluster results based on these 50 genes have accuracy of 1. However, for \( m_i = 400, 200 \), the accuracy of clustering do not improve.

Clustering results using UFS with two-way ordering, PCA and variance are given in Table 4. UFS with PCA appear to be the best. The reason is that 1st PCA component shows high correlation with \( t \)-value, thus picking up genes from both ends are almost like picking genes according to supervised \( t \)-value (see figure in Supplement). On the other hand, the variance appears to have low correlations with either \( t \)-value or PCA; its clustering accuracy varies widely from 0.5 to 0.921.

These results and those from earlier studies (Golub et al., 1999; Xing and Karp, 2001; Ding, 2002) indicate that the leukemia phenotype structure appears to be reasonably well separated. Thus the clusters obtained using unsupervised feature selection agree reasonable with pre-defined class labels. However, comparing Table 3 with Table 1, Table 4 with Table 2, indicates the level of agreement for leukemia data is lower than that for colon data: 0.737 versus 0.871. Thus the clusters obtained using unsupervised feature selection may have something new or different from pre-defined class labels. Golub et al. (1999), based on the results of self-organizing maps, suggest that ALL class should be split into T-lineage ALL and B-lineage ALL classes. Thus there are three clusters, not two, in the data. This complication could be one reason why the clusters obtained using unsupervised feature selection differ somewhat from the cluster labels based on two-class labeling. This point needs to be further explored.

We checked how sensitive the results depend on the data pre-processing Equation (7). If we double the ranges, i.e. set \( b_{ij} = g_{ij} + 2000; \quad 0 \leq b_{ij} \leq 10000 \), the results remain almost identical.

If we use the original data without any pre-processing, PCA, and variances change substantially, (see figure in
appears to be more sensitive to data pre-processing while also changed in non-negligible way. In other words, PCA have significant effects on the mean value such that PCA is the outliers outside the specified ranges in Equation (7) genes have different means and is thus shifted differently. This is because in PCA, the data is first centered; different abundance (the ratio of two positive intensities of the target genes and control genes) are measured (Eisen et al., 1998; Alizadeh et al., 2000). The non-negativity is automatically satisfied. This positive ratio should be used in the two-way ordering, although the logarithm of this ratio is often used in gene expression analysis.

In this paper, we concentrate on two-class problems. For many important clinical diagnose and prognosis, two-class problem remain the most important case. For multi-class problems, we suggest to follow the hierarchical clustering approach dealing with two-class problem one at a time. There are several applications of the unsupervised gene selection. One application is to identify marker genes which are discriminative or predictive for the desired phenotypes. Although tissue samples are labeled by human expertise, some of the labels could be occasionally mistaken or ambiguous. Applying the unsupervised approach to identify relevant genes and compare with those selected through supervised methods provides an important consistency check. More fruitful applications are in areas where the natural phenotypes are not clearly separated (at least initially); UFS can be used to help identify the phenotypes. The unbiased nature of UFS help to ensure that the phenotypes eventually identified are reasonably encoded or reflected in the data.

6 DISCUSSIONS

We have proposed and studied a new unsupervised feature selection method for unsupervised learning. The key idea is to use a two-way ordering mechanism to approximately identify irrelevant features and discard them. Theoretical basis is discussed in detail and applications on two datasets of gene expressions profiles show that the unsupervised method outperforms the baseline algorithm that simply uses all genes. We also studied using PCA and variance as criteria to discard irrelevant genes, with mixed results.

In analyses of gene expression datasets, we use the known cluster information as the basis to assess the viability of the results of the unsupervised method. It should be emphasized that the known cluster information based on human expertise are not necessarily fully reflected or encoded in the datasets. Thus the clusters obtained using unsupervised feature selection may differ somewhat from the human assigned labels. In both colon and leukemia datasets, clusters are reasonably well separated; This explains why clusters obtained using unsupervised feature selection agree reasonably well with human assigned labels. When clusters are not well-separated, clusters obtained using unsupervised feature selection will likely differ from human assigned labels. UFS methods are more useful in these datasets.

Our UFS method requires that gene expression data are non-negative (weights in bipartite graphs must be non-negative). When gene expression data are raw intensities this is always satisfied (see Section 5 for some clarification). In cDNA microarrays, the relative transcript abundance (the ratio of two positive intensities of the target genes and control genes) are measured (Eisen et al., 1998; Alizadeh et al., 2000). The non-negativity is automatically satisfied. This positive ratio should be used in the two-way ordering, although the logarithm of this ratio is often used in gene expression analysis.

In this paper, we concentrate on two-class problems. For many important clinical diagnose and prognosis, two-class problem remain the most important case. For multi-class problems, we suggest to follow the hierarchical clustering approach dealing with two-class problem one at a time. There are several applications of the unsupervised gene selection. One application is to identify marker genes which are discriminative or predictive for the desired phenotypes. Although tissue samples are labeled by human expertise, some of the labels could be occasionally mistaken or ambiguous. Applying the unsupervised approach to identify relevant genes and compare with those selected through supervised methods provides an important consistency check. More fruitful applications are in areas where the natural phenotypes are not clearly separated (at least initially); UFS can be used to help identify the phenotypes. The unbiased nature of UFS help to ensure that the phenotypes eventually identified are reasonably encoded or reflected in the data.

ACKNOWLEDGEMENTS

I thank Hongyuan Zha for discussions. This work is supported by U.S. Department of Energy (Office of Science, Office of Advanced Scientific Research/MISC Division and Office of Laboratory Policy and Infrastructure/LDRD) under contract DE-AC03-76SF00098.

REFERENCES


