Bioinformation
www.bioinformation.net

open access
Hypothesis

Supplementary Material:

Methdology :

Feature Selection:

Overall algorithmic steps of our feature subset selection method are illustrated in **Figure 1**. We denote as $C(N,M)$ a microarray data in a matrix form of N gene expression levels and M samples of experiment conditions. Let $C^i(N, 1)$ denote i-th column vector of $C(N,M)$. Referring to Figure 1, the details of the feature selection steps are as follows.

Step 1: First, column vectors $C^i(N, 1)$, $i = 1, 2,\ldots, M$, are created from $C(N,M)$. Each of these column vectors actually corresponds to one sample of gene expression data. Each column vector $C^i(N, 1)$ is fed into a difference operation (denoted as $\circ$) which computes the element-wise difference between $C^i(N, 1)$ and all the columns of $C(N,M)$. For each i, this operation outputs a difference matrix,

$$D^i(N,M) = C(N,M) \circ C^i(N, 1), \quad i = 1,2,..., M, \tag{1}$$

where each entry of $Di(N;M)$ is computed as the difference between two real numbers,

$$D^i(r, c) = C(r, c) - C^i(r, 1), \quad r = 1,2,\ldots,N, \quad c = 1,2,\ldots,M. \tag{2}$$

Let $D(N, j)$ denote j-th column of matrix $D(N,M)$. By definition, $D^i(N, j)$ contains the measure about how i-th sample $C^i(N, 1)$ differs from j-th sample $C^j(N, 1)$. One advantage of this vector-based scheme is that between-gene difference information is also kept in the column vectors together with between-sample differences. This is useful in examining how all the genes are correlated as will be shown shortly.

Step 2: Magnitude of elements of $D^i(N,M)$ can be used as a measure to determine how useful each gene (row) is in classification. Our strategy here is based on a simple idea: differences between two samples in the same class will be small for most genes, while two samples coming from different classes will show large differences for many genes. Our objective is to identify and select out those genes that behave according to the conjecture. In order to mark the genes which take big or small values in $D^i(N,M)$, we introduce upper threshold u and lower threshold l, which are set to 75-percentile and 25-percentile, respectively, of the values in $D^i(N,M)$. As the range of difference values could be varying depending on class, thresholds are determined using the values within a class and represented as u^w and l^w for class w.

Step 3: Marking of the values in $D^i(N,M)$ is carried out as follows. For an element $D^i(n,m)$ given, if i-th sample and m-th sample belong to the same class, then the absolute value of $D^i(n,m)$ would be expectedly small for gene n. Otherwise if they belong to different classes, then $D^i(n,m)$ would take a large value for gene n. In any case, if this expectation is met for gene n, we mark the gene by setting $I^i(n,m)$ to 1. This marking implies that n-th gene is useful for describing the (inverse) correlation between i-th to m-th sample in terms of classification. Also it should be noted that this naturally provides hints about where irrelevant features arise. However, final decision on usefulness of a gene in classification should be postponed until the gene proves to be useful for all the samples involved, which is taken care of in the next step.

Step 4: The objective of this step is to construct class-specific features from a set $\{ I^i(N,M), i = 1,2, \ldots, M\}$, which holds individual sample-based information. As we assume M samples are collected from W different classes, M columns can be decomposed into a partition of W blocks,

$$I^i(N, M) = I^i(N, M_1 + M_2 + ... + M_W), \tag{3}$$

where M_w refers to the number of samples in class w. With this scheme, for each w, we element-wise add $I^i(N,M)$'s within class w to get $S^w(N, M)$. For example, if we suppose M_2 consists of 3 samples indexed from 5 to 7, the features specific to class 2 are computed by:

$S^2(n,m) = I^5(n,m) + I^6(n,m) + I^7(n,m), \quad n = 1,2,...,N, \quad m = 1,2,...,M.$ Once we have constructed all $S^w(N,M)$'s, we identify the elements taking significant values by applying a threshold which is set to 90-percentile of the values in $S^w(N,M)$ for each class w. This threshold operation produces a binary matrix $F^w(N, M)$. It should be noted that $F^w(N, M)$ holds a useful measure for determining how much each gene n contributes to the classification of each sample m.

Step 5: After we collect the within-class features in $F^w(N, M)$, we then move on to selecting the most influencing genes. Selection of genes at this step is rather trivial, because all useful information has already gathered in $F^w(N, M)$. We just count the marked elements in $F^w(N, M)$ for each gene n, whose value is denoted by $F^w(n)$, and use it as a final measure to determine the usefulness of gene n in the classification task as described in **step 6**.

Clustering of Samples

Gene expression data are typically given without any information about the phenotype of genes within each class. In handling such case of lacking a priori knowledge of representative patterns, nonnegative matrix factorization (NMF) has proved to be successful in capturing biologically meaningful clusters in the unsupervised manner **[3, 12, 13]**. In contrast to holistic methods such as principle component analysis (PCA) and self-organizing map (SOM), NMF yields a sparse, parts-based decomposition of data without discarding the original interpretation of features [14]. Suppose gene expression data is represented as $N \times M$ nonnegative matrix A which is $C(N', M)$ after feature selection. The number N of genes is usually in the thousands. NMF method decomposes A into two nonnegative matrices, V of size $N \times \kappa$ and H of size $\kappa \times M$, so that $A \sim V H$. The rank κ of factorization defines the number of metagenes, which reflects the degree of latent factors. In a classification scheme, the value of κ represents the number of clusters, and the goal of NMF is to find two nonnegative matrices V and H such that · clusters optimally characterize the intrinsic structure of samples in A. The NMF algorithm starts by initializing V and H to random values and iteratively updates their values to minimize the distance between A and $V H$. A number of the divergence functionals have been proposed to measure the distance, including Euclidian distance and Kullback-Leibler (KL) divergence [12, 13, 14]. The KL divergence functional is given by the Poisson likelihood of generating A from V and H,

$$KL(A \| VH) = \sum [A_{ij} \log \frac{A_{ij}}{(VH)_{ij}} - A_{ij} + (VH)_{ij}] \quad (4)$$

The divergence $KL(\cdot)$ is non-increasing under the following multiplicative update rules [15],

$$H_{a\mu} \leftarrow H_{a\mu} \frac{\sum_i V_{ia} A_{i\mu} / (VH)_{i\mu}}{\sum_k V_{ka}} \quad (5)$$

$$V_{ia} \leftarrow V_{ia} \frac{\sum_\mu H_{a\mu} A_{i\mu} / (VH)_{i\mu}}{\sum_\nu H_{a\nu}} \quad (6)$$

The updates of the two matrices are iteratively performed until the divergence of Equation (4) converges to a (local) minimum. Each sample is then considered to determine its membership to one of the κ clusters by the highest value of metagene expression pattern (column of H).

$$q = \sum_{j=1}^{M} \delta(j) \Big/ M, \quad (7)$$

where $\delta(j)$ is 1 if j-th sample is correctly classified and 0 otherwise.

Table 1: Number of 1-valued elements in $F^w(N,M)$ for the Leukemia data. F^i - F^j denotes the number of elements whose values take 1 both in class i and j. Also F^i - F^j - F^k denotes the number of elements that have 1 for all three classes i, j, and k.

F^1	F^2	F^3	F^1 - F^2	F^1 - F^3	F^2 - F^3	F^1 - F^2 - F^3
23,115	22,758	24,785	1,640	1,127	4,468	7

Table 2: Number of 1-valued elements in $F^w(N,M)$ for the Medulloblastoma data

F^1	F^2	F^1 - F^2
22,665	20,913	72

Table 3: Number of 1-valued elements in $F^w(N,M)$ for the Central nervous system tumors data. The value in the first column is the average of four F^w's.

Avg. of F^w's	F^1 - F^2 - F^3	F^1 - F^2 – F^4	F^1 – F^3 – F^4	F^2 – F^3 – F^4	F^1 - F^2 - F^3 - F^4
26,710	3,627	1,829	2,682	210	98