Identifying Genes with The Concept of Customization

Dong Hua
George Washington University
Washington, DC 20052
gwuhua@gwu.edu

Dechang Chen
USUHS
Bethesda, MD 20814
dchen@usuhs.mil

Abdou Youssef
George Washington University
Washington, DC 20052
ayoussef@gwu.edu

Abstract

Gene selection with microarray data is an important task towards the study of genomics. The goal is to identify the optimal subset of genes such that maximum discrimination power across samples (e.g., tumor types) while minimum redundancy among genes are achieved. Essentially, it is NP-complete. Approximation algorithms are usually solicited including individual ranking and sequential forward selection. Typically, from source input microarray data to output selected genes, multiple steps including preprocessing, discretization, discrimination modeling, redundancy modeling, optimization formularization, classification, and evaluation are involved in the presence of a number of options (techniques) for each of them. Putting them together, we form the concept of customization for gene selection in this paper, that is, configure the entire scenario such that various maybe trivial techniques can team work with superior performance rather than focus on certain technique within a single step (e.g., discrimination power modeling). One configuration following the principle of simplicity is constructed in this paper which identifies genes effectively shown by experiments.

1 Introduction

A significant step towards the current information revolution can be appreciated from the successfully applied new techniques and tools in molecular biology and genetics research. Such technologies make it possible to collect biological information rapidly at an unprecedented level of detail in large quantities. Among the most powerful technologies, microarrays provide the tool to extract biological significance such as the changes in expression profiling of genes under distinct types (e.g., normal vs cancer type), which shed the light on use of them in many fields including pharmacogenomics, medical diagnostics, drug target identification and underlying gene regulatory networks.

Microarrays, often interrogating thousands or tens of thousands of genes simultaneously, are capable of extracting huge amounts of biological information. It opens rich opportunities but also poses a great challenge on study of genomics. One critical step is called discriminant analysis, i.e., classifying samples according to gene expression profiling, for instance, distinguish cancer tissues from normal ones [2] or one subtype of cancer vs another [1]. Efforts have been made intensively to identify genes really contributing to the disease under study, which is often achieved via gene selection. This is necessary mainly due to the fact that, high dimension of features often result in more classification errors. Usually, when samples are limited while the number of features is very large beyond a certain point, classification accuracy will reduce. Instead of using all genes, one may look for a subset such that it can most discriminatively and compactly represent the expression patterns. In other words, genes with maximum discrimination power while minimum redundancy are preferred. It is NP-complete [6], i.e., feature selection in machine learning. We may not expect to find the optimal solution via brute-force. Rather, approximation algorithms are usually solicited including individual ranking (e.g., [9], rank the genes and choose the top) and sequential forward selection (e.g., [3], choose the best gene as the seed and add one more per iteration such that the obtained subset maximizes the given criterion function).

The entire procedure of gene selection includes multiple steps: preprocessing, discretization, discrimination modeling, redundancy modeling, optimization formularization, classification, and evaluation, in the presence of a number of options (techniques) for each of them. Instead of hard improving certain technique within a single step, we put them together and form the concept of customization for the entire scenario. By doing so, various maybe trivial techniques can team work with superior performance. In this paper, we configure the scenario following the principle of simplicity and experiments show its effectiveness. Specifically, we formularize the optimization issue by maximizing $U(\mathbf{f})$, discrimination power of genes in terms of Brown-Forsythe statistic, with the constraint on $V(\mathbf{f})$, i.e., redun-

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 IEEE

dancy, where Pearson correlation is employed. Advantages from both individual ranking and sequential forward selection are combined for the design of selection operation. The normalization technique is used for preprocessing. Ten-bins, naive Bayes, leave-one-out cross-validation are chosen in particular for discretization, classification and evaluation, respectively.

The rest of this paper is organized as follows. In Section 2, we present models and methods. Section 3 shows the experiments on real datasets. We conclude in Section 4.

2 Models and Methods

2.1 Mathematical Formularization

Consider a $k(\geq 2)$-class discriminant analysis with p genes, i.e., g_1, g_2, ..., g_p, and n microarray samples involved. Let X_{ij} be the value in terms of the measurement of g_i expression from the jth sample where $i = 1, \ldots, p$ and $j = 1, \ldots, n$. Typically, such microarray data can be written as a form of matrix, $\mathbf{M}$:

$$\mathbf{M} = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ X_{p1} & X_{p2} & \cdots & X_{pn} \end{pmatrix},$$

where the columns and rows correspond to samples and genes, respectively.

Given $\mathbf{M}$, to select m genes out of p genes for discriminant analysis can be viewed as the identification of representative rows (genes) to stand for the entire expression pattern across all the given samples instead of $\mathbf{M}$ itself. Effectiveness can be evaluated from two ways: (1) the combination of chosen rows can differentiate samples distinguishably; and (2) these rows contain redundancy as low as possible. In other words, selected genes should be discriminative and compact simultaneously. Let $\mathbf{f}$ be the selected genes, $U(\mathbf{f})$ the discriminative power of $\mathbf{f}$, and $V(\mathbf{f})$ redundancy in correspond. Generally, larger $U(\mathbf{f})$ implies higher discriminative power; while lower $V(\mathbf{f})$ implies less redundancy. As such, gene selection can be formularized by an optimization issue in the form of particular:

$$maximize: \ U(\mathbf{f}), \ subject\ to: \ V(\mathbf{f}) \leq T(U, V),$$

where T is a threshold function of U and V.

2.2 Functions Modeling: $U(\mathbf{f})$ and $V(\mathbf{f})$

To model $U(\mathbf{f})$, the key is to find some measurement such that the discrimination power of $\mathbf{f}$ is truly expressed. In this paper, we investigate this issue from the statistical point of view. It is assumed that the data $\mathbf{M}$ are normalized so that the genes have mean 0 and variance 1 across samples. Given a fixed gene, let Y_{ij} be the expression level from the jth sample of the ith class. Note that these Y_{ij} come from the corresponding row of $\mathbf{M}$. For example, for g_1, Y_{ij} are a rearrangement of the first row of $\mathbf{M}$. The following general model is considered for Y_{ij} in this paper:

$$Y_{ij} = \mu_i + \epsilon_{ij}, \quad \text{for} \quad i = 1, 2, \ldots, k; \ j = 1, 2, \ldots, n_i$$

where $n_1 + n_2 + \ldots + n_k = n$, μ_i is the mean expression level of the gene in class i, and ϵ_{ij} are the error terms, independent normal random variables with

$$E(\epsilon_{ij}) = 0, \ V(\epsilon_{ij}) = \sigma_i^2 < \infty,$$

for $i = 1, 2, \ldots, k; \ j = 1, 2, \ldots, n_i$.

An important task, associated with above model, is to detect whether or not there exists some difference among the means $\mu_1, \mu_2, \ldots, \mu_k$. It is often achieved by certain statistics, the well known ANOVA F test for instance, which is well suited for measuring the discriminative power of genes as thought in this paper. Specifically, given a test statistics $\mathcal{F}$, we define the *discrimination power* of a gene, $d(g_i)$, as the value of $\mathcal{F}$ evaluated over the samples. This definition is based on the fact that with larger $\mathcal{F}$ the null hypothesis $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ will be rejected more likely. Therefore, larger $\mathcal{F}$ implies higher discrimination power of the corresponding gene across classes of samples. We also note that discrimination power of genes could be determined equally well via p-values from $\mathcal{F}$. However, due to small sizes n_i, it is hard to justify the approximation of the known distribution to $\mathcal{F}$ and hence p-values may not reflect the real functionality of $\mathcal{F}$. Therefore, the value of $\mathcal{F}$ is preferred.

Usually, if the variances are equal, namely, $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$, then it is simply the commonly used one-way ANOVA model and hence the ANOVA F test is the optimal option [11, 13]. For microarray data, the existence of heterogeneity in variances is more realistic, since different σ_i may describe different variation of the gene expression across classes. It makes the above task challenging however, related to the well-known Behrens-Fisher problem [16]. When sample sizes of all classes are equal, i.e. $n_1 = n_2 = \cdots = n_k$, the presence of heterogeneous variances of the errors only slightly affects the F test. If sample sizes are *not* equal, the effect is serious [12]. The actual type I error is inflated when smaller sizes n_i are associated with larger variances σ_i^2. In contrast, the significance levels are smaller than anticipated when larger sizes n_i are associated with larger variances σ_i^2.

In this paper, the parametric Brown-Forsythe test statistic is chosen, which has been shown preferable over others [5]. It will be used as fitness or score function to measure the

discrimination power of selected genes given by [4]:

$$B = \frac{\sum n_i(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2}{\sum(1 - n_i/n)s_i^2}.$$

Under H_0, B is distributed approximately as $F_{k-1,\nu}$, where

$$\nu = \frac{[\sum(1 - n_i/n)s_i^2]^2}{\sum(1 - n_i/n)^2 s_i^4/(n_i - 1)}.$$

To model $V(\mathbf{f})$, we simply use Pearson correlation between genes. Given $\mathbf{M}$, the correlation of g_i and $g_{i\prime}$ is given by

$$\rho(g_i,\ g_{i\prime}) = \frac{\sum_j (X_{ij} - \bar{X}_i)(X_{i'j} - \bar{X}_{i'})}{\sqrt{\sum_j (X_{ij} - \bar{X}_i)^2 \sum_j (X_{i'j} - \bar{X}_{i'})^2}}\ ,$$

where $\bar{X}_i = \sum_j X_{ij}/n$ is the average level of g_i, based on the n samples in correspond.

Of particular, we simply formularize the optimization issue as follows:

$$maximize:\ U(\mathbf{f}) = \frac{1}{|\mathbf{f}|}\sum_{g_i \in \mathbf{f}} d(g_i)$$

subject to:

$$V(\mathbf{f}) = max\{\rho(g_i, g_j), \forall g_i, g_j \in \mathbf{f}\ and\ i \neq j\} \leq T.$$

For further simplicity, the threshold T is chosen as constant and adjusted dynamically according to certain requirement in reality.

The algorithm designed in this paper, approximating the above optimization issue, inherits advantages from both individual ranking and sequential forward selection.

Given a test statistic $\mathcal{F}$, rank all genes with $d(.)$ descending and choose the top as the seed which has the highest discrimination. Consider the rest whose correlation to the chosen gene is below T. Similarly, the top is chosen as the second. And then perform the next iteration. Note that we rank genes only once before the seed selection. As such, the kth informative gene is the one receiving the highest discrimination power from the set of all genes with correlation to each of the chosen $k-1$ genes below T. Above process will be repeated until the given number m of genes are obtained or all the genes have been scanned.

2.3 Classifier

As mentioned before, we follow the principle of simplicity to configure the entire selection procedure. Naive Bayes is employed. Additionally, input is restricted in categorical data. The discretization will be discussed in next subsection. Consider a k-class ($k \geq 2$) classification issue. Let

Algorithm 1 Gene selection algorithm

function Σ=**GeneSel**($\mathbf{M}, m, T$) ▷ $\mathbf{M}$ is the data matrix, Σ is the target feature set
  $\Sigma \leftarrow \phi$
  $\mathbf{M}' \leftarrow rank(\mathbf{M})$ ▷ Feature sorting
  $\Sigma \leftarrow \Sigma \cup \{first\ feature\ of\ \mathbf{M}'\}$ ▷ Choose the first one as the seed
  $\mathbf{M}' \leftarrow \mathbf{M}' \backslash \{first\ row\ of\ \mathbf{M}'\}$ ▷ Remove it
  while $\mathbf{M}' \neq \phi$ and $|\Sigma| < m$ **do** ▷ Loop for qualified features
    if $\max\{\rho(first\ feature\ of\ \mathbf{M}', \Sigma)\} \leq T$ **then** ▷ Check correlation criterion
      $\Sigma \leftarrow \Sigma \cup \{first\ feature\ of\ \mathbf{M}'\}$
      $\mathbf{M}' \leftarrow \mathbf{M}' \backslash \{first\ row\ of\ \mathbf{M}'\}$
    else
      $\mathbf{M}' \leftarrow \mathbf{M}' \backslash \{first\ row\ of\ \mathbf{M}'\}$
    end if
  end while
  return Σ
end function

$\mathbf{X} = (X_1, X_2, \cdots, X_p)^T$ be a p-dimensional feature vector, where T is the transpose operation. We use C to denote the class label of $\mathbf{X}$ with π_i referring to the prior probability, $P(C = i)$, for $i = 1, 2, \cdots, k$. Suppose that given $C = i$, the joint distribution of $\mathbf{X}$ is given by $P_i(\mathbf{X})$. If $\mathbf{x}$ is an observed value of $\mathbf{X}$, then it follows from the Bayes formula that the posterior probability of class i given $\mathbf{X} = \mathbf{x}$ is

$$P(i|\mathbf{x}) = \frac{\pi_i P_i(\mathbf{x})}{\sum_{i=1}^k \pi_i P_i(\mathbf{x})}\ . \tag{1}$$

With $0-1$ loss function, the Bayes rule states that we classify $\mathbf{x}$ to the most probable class according to posterior probabilities, that is, given $\mathbf{X} = \mathbf{x}$, the class label is chosen to be

$$C(\mathbf{x}) = \text{argmax}_{\ i}\ \{P(i|\mathbf{x})\}. \tag{2}$$

The naive Bayes model makes the additional assumption that given a class $C = i$, the features $X_1, X_2, \cdots, X_p$ are independent with each other. Under this assumption, we have, for each i,

$$P_i(\mathbf{x}) = \prod_{l=1}^{p} P_{il}(x_l), \tag{3}$$

where $P_{il}(x_l)$ is the class-conditional density of X_l for $l = 1, 2, \cdots, p$. Using (1), (2) and (3), it is easy to see that the naive Bayes classifier assigns $\mathbf{x}$ into the following class by taking the natural logarithm:

$$C(\mathbf{x}) = \text{argmax}_{\ i}\ \Big\{\ln \pi_i + \sum_{l=1}^{p} \ln P_{il}(x_l)\Big\}. \tag{4}$$

2.4 Discretizer

Input for naive Bayes classifier in this paper is restricted in categorical data. Continuous values need to be discretized. Equal width interval binning is used for simplicity. Actually, it is the simplest approach perhaps to achieve nominal values generation. Given a continuous value x in an array, then the corresponding discrete value is given by:

$$x^d = \lceil \frac{x - x_{min}}{x_{max} - x_{min}} \rceil \times L,$$

where L is customized number of bins. This approach is chosen, besides its simplicity, also because it is a *unsupervised* discretizer which makes no use of any sample class information. Moreover, it often achieves good performance for Naive Bayes classifier vs continuous values as compared by Dougherty et al. (1995) using 16 data sets [8]. In this paper, ten-bins are used.

3 Experimental Results

In this section, we perform the experiments with real datasets. Leave-one-out cross-validation (LOOCV) is used. Datasets include Ovarian [17], MLL Leukemia [15], Lung Cancer [10], and Colon [2].

3.1 Ovarian

It contains 36 samples: 5 normal tissues, 27 epithelial ovarian tumor samples, and 4 malignant epithelial ovarian cell lines. 7129 genes are employed. Table 1 summarizes the result, where the numbers of the first line (from .05 to 1) correspond to different threshold T, and the number (m) of the first column (e.g., 45) correspond to the number of genes required. The bottom line, i.e., $\hat{m}$ corresponds to the real maximum number of genes that can be chosen with the threshold T. The rest numbers represent the accuracy of LOOCV. The smaller is the value, the higher quality the chosen genes achieve. Usually, there are a number of options with the best performance. Which is better is based on certain requirement or metric in reality, for example, we can choose the least number of genes with smallest errors. In this paper, we take the best accuracy regardless of the number of genes within the limit 50. 100% accuracy is obtained for this dataset (0 LOOCV error).

3.2 MLL leukemia

This dataset contains both training data and test data. We combine them together for LOOCV evaluation. The training part summarizes 57 leukemia samples (20 ALL, 17 MLL and 20 AML), while the test part summarizes 4 ALL, 3 MLL and 8 AML samples. There are 3 classes in total. The number of genes is 12582.

Table 2 shows the LOOCV errors. The best result, 100% accuracy, i.e., 0 LOOCV error, is achieved.

3.3 Lung cancer

Classification is conducted between two classes, i.e., malignant pleural mesothelioma (MPM) and adenocarcinoma (ADCA) of the lung. 32 training samples and 149 test samples are combined together. In total, there are 31 MPM and 150 ADCA described by 12533 genes. The best result (100% accuracy) is achieved along the column with $T = .25$ (Table is omitted due to page limit).

3.4 Colon

There are 62 samples in this dataset collected from colon-cancer patients. Among them, 40 tumor biopsies are from tumors, labelled with 'negative' and 22 normal biopsies, labelled with 'positive', from healthy parts of the colons of the same patients. Two thousands out of 6500 genes were chosen for attributes serving these samples based on the confidence in the measured expression levels. The best result is 2 LOOCV errors, improving the previous best result 4 LOOCV errors (Table is omitted due to page limit).

From above experiments, we can see, instead of hard improving certain technique within a single step of the gene selection, the simple customization of the entire scenario can make various maybe trivial techniques to team work with superior performance. Specifically, the best result is achieved as before for Ovarian and MLL Leukemia datasets, and better than before for Lung cancer and Colon datasets. In particular, the Colon dataset involves high errors which is very difficult to reduce. The best previous result is 4 LOOCV errors, while above configuration gives only 2.

4 Conclusion

Gene selection is essentially NP-complete. It is an important procedure for identifying genes to the target disease as well as improving the classification accuracy. Typically, multiple steps are involved including preprocessing, discretization, discrimination modeling, redundancy modeling, optimization formularization, classification, and evaluation, with a number of options for each of them. We form the concept of customization in this paper for gene selection. By taking all these steps as an entire scenario, various maybe trivial techniques are allowed to team work with superior performance instead of hard improving certain individual techniques. Of particular, we configure the entire

m	T=.05	.1	.15	.2	.25	.3	.35	.4	.45	.5	.55	.6	.65	.7	.75	.8	.85	.9	.95	1
1	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
2	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	4	4	4
3	2	2	2	2	2	2	2	2	2	2	2	2	2	5	5	4	4	5	4	4
4	1	2	1	2	0	1	1	1	1	1	2	2	2	1	0	5	5	1	4	4
5	2	2	2	2	0	0	0	1	1	1	0	0	2	0	0	1	1	1	2	2
6	4	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	4	1	2	2
7	3	3	0	1	0	0	0	0	0	0	0	0	1	0	0	0	1	1	2	2
8	4	2	0	2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
9	4	2	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
10	4	2	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
11	4	4	2	1	1	0	1	0	1	0	0	0	0	0	0	0	0	0	1	1
12	4	2	2	4	1	0	1	0	0	1	0	0	0	0	0	0	0	0	1	1
13	4	2	2	4	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	1
14	4	2	3	3	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
15	4	2	2	4	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
16	4	2	4	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	4	2	3	6	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	4	2	4	2	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	4	2	4	3	2	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
20	4	2	4	4	3	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0
21	4	2	4	4	3	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
22	4	2	4	3	5	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
23	4	2	4	3	4	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
24	4	2	4	3	4	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
25	4	2	4	3	4	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0
26	4	2	4	6	5	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0
27	4	2	4	5	6	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
28	4	2	4	5	3	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
29	4	2	4	6	5	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
30	4	2	4	6	5	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0
31	4	2	4	7	6	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0
32	4	2	4	8	4	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0
33	4	2	4	8	4	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0
34	4	2	4	8	4	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0
35	4	2	4	8	4	1	0	0	1	0	0	0	1	0	0	0	0	0	0	0
36	4	2	4	8	3	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
37	4	2	4	8	3	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
38	4	2	4	8	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
39	4	2	4	8	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
40	4	2	4	8	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
41	4	2	4	8	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
42	4	2	4	8	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
43	4	2	4	8	4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
44	4	2	4	8	4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
45	4	2	4	8	5	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
46	4	2	4	8	5	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0
47	4	2	4	8	5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
48	4	2	4	8	5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
49	4	2	4	8	5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
50	4	2	4	8	5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
$\hat{m}$	9	12	18	32	47	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50

Table 1. LOOCV Errors for Ovarian Dataset

m	T=.05	.1	.15	.2	.25	.3	.35	.4	.45	.5	.55	.6	.65	.7	.75	.8	.85	.9	.95	1
1	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17
2	17	17	17	17	17	17	17	17	17	17	17	13	13	13	13	16	16	16	16	16
3	12	14	14	14	14	7	7	15	14	14	15	10	10	10	12	12	9	9	9	9
4	10	10	10	13	9	6	8	6	7	14	14	11	11	11	11	10	7	7	7	7
5	10	11	11	11	6	6	5	7	4	4	10	8	10	8	10	7	6	6	6	6
6	9	9	12	9	6	4	5	5	5	2	9	6	6	10	7	8	7	7	7	7
7	12	10	11	5	8	5	5	3	5	1	8	4	7	12	10	7	6	6	6	6
8	13	12	7	6	6	5	4	4	4	3	8	4	7	11	8	5	8	7	7	7
9	13	13	8	7	7	3	3	4	4	1	9	3	5	8	6	4	9	8	8	8
10	13	8	9	4	6	4	2	4	4	2	8	3	8	6	8	4	8	7	7	7
11	14	7	10	5	6	3	4	3	4	1	8	3	6	7	8	4	8	8	8	8
12	14	14	9	7	8	3	4	1	2	0	8	1	6	7	8	4	6	10	10	10
13	14	12	15	5	9	2	3	3	3	0	6	1	6	8	8	5	4	7	7	7
14	14	13	13	5	10	2	3	3	3	0	5	1	5	8	8	6	3	6	6	6
15	14	12	12	7	7	2	5	3	2	0	4	1	6	7	9	6	4	4	4	4
16	14	11	12	9	9	3	5	4	2	0	7	1	6	8	8	4	5	4	4	4
17	14	13	10	9	8	2	4	4	2	1	6	1	6	6	8	6	5	4	4	4
18	14	11	11	9	7	2	4	2	2	1	4	0	6	4	7	5	5	4	4	4
19	14	11	11	8	7	2	4	4	2	1	5	0	6	6	6	5	4	4	4	4
20	14	10	12	7	8	1	3	3	3	0	5	0	6	5	7	5	4	4	4	4
21	14	11	11	7	9	1	3	3	3	0	5	0	7	5	6	5	5	4	4	4
22	14	11	11	7	6	2	2	3	1	0	4	0	4	2	8	4	5	4	4	4
23	14	11	12	9	5	2	2	3	3	0	4	1	3	2	7	5	5	4	4	4
24	14	11	7	9	6	2	1	3	3	1	3	1	1	2	6	5	5	5	5	5
25	14	11	10	8	6	0	2	3	1	0	3	1	0	2	6	4	5	5	5	5
26	14	11	10	9	6	0	3	3	2	0	3	1	0	2	5	4	5	5	5	5
27	14	11	10	7	5	1	2	3	3	0	3	1	0	2	6	4	5	6	6	6
28	14	11	10	9	8	1	3	3	2	0	3	1	0	2	7	5	5	6	6	6
29	14	11	11	6	9	2	3	3	2	0	3	1	0	2	4	5	4	6	6	6
30	14	11	14	5	8	2	3	3	2	1	3	2	0	1	3	4	4	6	6	6
31	14	11	14	6	8	1	3	3	2	1	3	2	0	1	4	4	4	6	6	6
32	14	11	14	7	8	3	3	3	3	1	3	3	1	1	3	4	4	6	6	6
33	14	11	14	8	7	3	3	3	2	1	3	2	1	1	3	3	4	5	5	5
34	14	11	14	7	7	2	3	3	2	0	3	2	2	1	3	2	3	6	6	6
35	14	11	14	6	7	1	3	3	2	0	3	2	0	2	3	4	3	5	5	5
36	14	11	14	6	7	2	3	3	2	0	3	2	0	2	3	4	4	5	5	5
37	14	11	14	6	7	2	3	3	2	0	3	0	0	1	3	4	4	4	4	4
38	14	11	14	6	7	2	3	2	1	0	2	0	1	1	3	4	4	6	6	6
39	14	11	14	6	7	2	3	2	2	0	2	0	1	2	3	4	4	5	5	5
40	14	11	14	6	7	2	3	1	2	0	3	0	1	1	3	4	4	4	4	4
41	14	11	14	8	7	2	3	2	2	0	2	1	1	1	3	3	3	6	6	6
42	14	11	14	8	7	2	3	2	3	0	2	0	1	1	3	3	3	5	5	5
43	14	11	14	8	6	2	3	4	3	0	1	1	1	1	3	2	3	5	5	5
44	14	11	14	8	6	2	3	4	3	0	2	0	1	1	3	2	3	5	5	5
45	14	11	14	7	7	2	4	4	2	0	2	1	1	1	3	3	3	5	5	5
46	14	11	14	8	6	2	4	3	2	0	2	1	1	1	3	3	3	5	5	5
47	14	11	14	8	7	2	4	3	2	0	2	1	2	1	3	3	4	5	5	5
48	14	11	14	8	6	3	3	3	2	0	2	1	0	0	3	2	4	5	5	5
49	14	11	14	10	6	3	3	3	2	0	2	1	1	0	3	2	4	5	5	5
50	14	11	14	10	6	2	2	5	1	0	2	1	2	2	2	2	4	5	5	5
$\hat{m}$	11	21	31	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50	50

Table 2. LOOCV Errors for MLL Leukemia Dataset.

gene selection following the principle of simplicity. We formularize the optimization issue by maximizing the discrimination power of genes in terms of Brown-Forsythe statistic, with the constraint of redundancy modeled by Pearson correlation. The normalization, ten-bins, naive Bayes, and leave-one-out cross-validation are chosen for data preprocessing, discretization, classification and evaluation, respectively. Experiments show its effectiveness with best performance as before for Ovarian dataset (0 LOOCV error) and MLL Leukemia dataset (0 LOOCV error), and better than before for Lung Cancer dataset (0 LOOCV error) and Colon dataset (2 LOOCV errors vs the best previous result 4).

Acknowledgement

D. Chen was supported by the National Science Foundation grant CCR-0311252. Note the opinions expressed herein are those of the authors and do not necessarily represent those of the Uniformed Services University of the Health Sciences and the Department of Defense.

References

[1] Alizadeh, A., et. al. (2000) Identification of clinically distinct types of diffuse large B-cell lymphoma based on gene expression patterns. *Nature*, 403, 503-511.

[2] Alon, U., et. al. (1999) Broad patterns of gene expression revealed by clustering of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96, 6745-6750.

[3] Ambroise, C. and McLachlan, G. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA*, vol. 99 (10), 6562-6566.

[4] Brown, M. B. and Forsythe, A. B. (1974) The small sample behavior of some statistics which test the equality of means. *Technometrics*, 16, 129-132.

[5] Chen, D., Liu, Z., Ma, X. and Hua, D. Selecting Genes by Test Statistics. *Journal of Biomedicine and Biotechnology*, to appear.

[6] Cover, T. and Campenhout, J. (1977) On the possible orderings in the measurement selection problem. *IEEE Trans. Systems, Man, and Cybernetics*, SMC-7(9), 657-661.

[7] Ding, C. and Peng, H. (2003) Minimum Redundancy Feature Selection from Microarray Gene Expression Data. *CSB 2003*, 523-529.

[8] Dougherty, J., Kohavi, R. and Sahami, M. (1995) Supervised and unsupervised discretization of continuous features. In *Proc. 12th Int. Conf. on Machine Learning (ICML-95)* , 194-202.

[9] Dudoit, S., Fridlyand, J. and Speed, T. P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association,* 97, 77-87.

[10] Gavin J. Gordon, et. al. (2002) Translation of Microarray Data into Clinically Relevant Cancer Diagnostic Tests Using Gege Expression Ratios in Lung Cancer And Mesothelioma. *Cancer Research*, 62:4963-4967.

[11] Lehman, E. L. (1986) *Testing Statistical Hypotheses*. 2nd edition, New York: Wiley.

[12] Montgomery, D. C. (2001) *Design and Analysis of Experiments*. 5th edition, New York: Wiley.

[13] Neter, J., Kutner, M. H., Nachtsheim, C. J. and Wasserman, W. (1996) *Applied Linear Statistical Models*. 4th edition, McGraw-Hill.

[14] Ripley, B. D. (1996) *Pattern Recognition and Neural Networks*. Cambridge University Press.

[15] Scott A. Armstrong, et. al. (2002) MLL Translocations Specify A Distinct Gene Expression Profile that Distinguishes A Unique Leukemia. *Nature Genetics*, 30:41-47, January.

[16] Stuart, A., Ord, J. K. and Arnold, S. (1999) *Advanced Theory of Statistics, Volume 2A: Classical Inference and the Linear Model*. 6th edition, London: Oxford University Press.

[17] Welsh, J. B., et. al. (2001) Analysis of gene expression in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proc. Natl. Acad. Sci. U.S.A.*, 98:1176-1181.

[18] Xing, E., Jordan, M. and Karp, R. (2001) Feature selection for high-dimensional genomic microarray data. In *Proc. 12th Int. Conf. on Machine Learning (ICML-01)* , San Francisco.

[19] Xiong, M., Li, W., Zhao, J., Jin, L. and Boerwinkle, E. (2001) Feature (gene) selection in gene expression-based tumor classification. *Mol. Genet. Metab.* vol. 73, 239-247.