Feature Selection for Gene Expression using Model-based Entropy

Shenghuo Zhu, Dingding Wang, Kai Yu, Tao Li and Yihong Gong

Abstract—**Gene expression data usually contain a large number of genes, but a small number of samples. Feature selection for gene expression data aims at finding a set of genes that best discriminate biological samples of different types. Using machine learning techniques, traditional gene selection based on empirical mutual information suffers the data sparseness issue due to the small number of samples. To overcome the sparseness issue, we propose a model-based approach to estimate the entropy of class variables on the model, instead of on the data themselves. Here, we use multivariate normal distributions to fit the data, because multivariate normal distributions have maximum entropy among all real-valued distributions with specified mean and standard deviation, and are widely used to approximate various distributions. Given that the data follow a multivariate normal distribution, since the conditional distribution of class variables given the selected features is normal distribution, its entropy can be computed with the log-determinant of its covariance matrix. Because of the large number of genes, the computation of all possible log-determinants is not efficient. We propose several algorithms to largely reduce the computational cost. The experiments on seven gene datasets and the comparison with other five approaches show the accuracy of the multivariate Gaussian generative model for feature selection, and the efficiency of our algorithms.**

Keywords— **feature selection, multivariate Gaussian generative model, entropy**

I. Introduction

GENE expression refers to the level of production of protein molecules defined by a gene. Monitoring of gene expression is one of the most fundamental approach in genetics and molecular biology. The standard technique for measuring gene expression is to measure the mRNA instead of proteins, because mRNA sequences hybridize with their complementary RNA or DNA sequences while this property lacks in proteins. The DNA arrays, pioneered in [Chee et al.1996], [Fodor et al.1991], are novel technologies that are designed to measure gene expression of tens of thousands of genes in a single experiment. The ability of measuring gene expression for a very large number of genes, covering the entire genome for some small organisms, raises the issue of characterizing cells in terms of gene expression, that is, using gene expression to determine the fate and functions of the cells. The most fundamental of the characterization problem is that of identifying a set of genes and its expression patterns that either characterize a certain cell state or predict a certain cell state in the future [Li et al.2004].

When the expression dataset contains multiple classes, the problem of classifying samples according to their gene expression becomes much more challenging, especially when the number of classes exceeds five [Ooi and Tan2003]. Moreover, the special characteristics of expression data adds more challenge to the classification problem. Expression data usually contains a large number of genes (in thousands) and a small number of experiments (in dozens). In machine learning terminology, these datasets are usually of very high dimensions with undersized samples. In microarray data analysis, many gene selection methods have been proposed to reduce the data dimensionality [Su et al.2003].

Gene selection aims to find a set of genes that best discriminate biological samples of different types. The selected genes are "biomarkers", and they form "marker panel" for analysis. Most gene selection schemes are based on binary discrimination using rank-based schemes [Dudoit et al.2002], such as information gain, which reduces the entropy of the class variables given the selected features. One critical issue in these rank-based methods is data sparseness. For example, the estimation of the traditional information gain is an empirical estimation directly on the data. Suppose we select the eleventh gene for a dataset. The ten selected genes split the training data into $1024 = 2^{10}$ groups (assuming each gene does a binary split). Since we have very few samples in most groups, the estimations of mutual information between the eleventh gene and the target in each group are not accurate. Thus the information gain, which is the sum of the mutual information over all groups, is not accurate.

To overcome the issue of data sparseness, we propose a model-based approach to estimate the entropy on the model, instead of on the data themselves. Here, we use multivariate Gaussian generative models, which model the data with multivariate normal distributions. Multivariate normal distributions are widely used in various areas, including gene expression data [Yeung et al.2001], because of their *generality* and *simplicity*. The means of variables (expression data of genes and class labels) and the covariances between them are two basic measures of variables themselves and the interaction between them. To predict the classes of data, we have to model the interaction between genes and class labels. Given the mean and covariance, multivariate Gaussian is the distribution with the maximum entropy, which implies its *generality* according to the principle of maximum entropy [Jaynes1957]. Usually we can explicitly and efficiently estimate parameters of multivariate Gaussian models via a few matrix operations, which implies the *simplicity* of the models. Though the class variables are binary or categorical, we relax them as numerical values, which bring us the simplicity. Our experiments suggest that this approximation in the feature selection does not affect the classification accuracy.

A nice property of multivariate Gaussian distributions is that the conditional distribution of a subset of variables given another subset of variables is still multivariate Gaussian distribution. We can explicitly compute the entropy of class variables given the selected features with the log-determinant of the covariance matrix of the conditional distribution. Therefore, the objective of minimizing the entropy becomes to find a set of

S. Zhu, K. Yu, and Y. Gong are with NEC Lab. America, Inc., Cupertino, CA 95014.

D. Wang and T. Li are with School of Computer Science, Florida International University, Miami, FL 33199.

features to minimize the log-determinant of the conditional covariance matrix. Because of the large number of genes, the computation of all log-determinants of the conditional covariance matrix is not time consuming. We propose several algorithms to largely reduce the computational cost.

In summary, our contributions are: (1) We propose a model-based approach to estimate the entropy based on multivariate normal distributions. The model-based approach addresses the data sparseness problem in gene selection; (2) We propose several algorithms to efficiently compute all log-determinants of the conditional covariance matrix and largely reduce the computational cost. The assumption of multivariate Gaussian generative models leads to simple, robust and effective computation methods for gene selection; and (3) We perform extensive experimental study on seven gene datasets and compare our algorithms with other five approaches. The rest of the paper is organized as the follows. A brief note on the related work is given in Section II. The notation used in this paper is listed in Section III. Our algorithms are presented in Section IV and the comparison methodologies are described in Section V. We show the experimental results in Section VI and discuss the idea of experimental designs and its connection with gene selection in Section VII. Finally Section VIII concludes.

II. Related work

Generally two types of feature selection methods have been studied in the literature: filter methods [Langley1994] and wrapper methods [Kohavi and John1997]. Filter-type methods are essentially data pre-processing or data filtering methods. Features are selected based on the intrinsic characteristics which determine their relevance or discriminative powers with regard to the target classes. In wrapper-type methods, feature selection is "wrapped" around a learning method: the usefulness of a feature is directly judged by the estimated accuracy of the learning method. Wrapper methods typically require extensive computation to search for the best features. As pointed out in [Xing et al.2001], the essential differences between the two methods are:

(1) that a wrapper method makes use of the algorithm that will be used to build the final classifier while a filter method does not, and

(2) that a wrapper method uses cross validation to compare the performance of the final classifier and searches for an optimal subset while a filter method uses simple statistics computed from the empirical distribution to select attribute subset.

Wrapper methods could perform better but would require much more computational costs than filter methods. Most gene selection schemes are based on binary discrimination using rank-based filter methods [Dudoit et al.2002], such as t-statistics and information gain [Su et al.2003]. One critical issue in these rank-based methods is data sparseness, and most of the rank-based methods do not take redundancy into consideration. In order to remove the redundancy among features, a Min-Redundancy and Max-Relevance (mRMR) framework is proposed in [Peng et al.2005]. Yu and Liu [Yu et al.2004] also explored the the relationship between feature relevance and redundancy and proposed a method that can effectively remove redundant genes. In addition, ReliefF, a widely used feature subset selection method, has also been applied to gene selection [Marko and Igor2003]. In Section V, we will describe several gene selection methods used in our experimental comparisons in detail.

In this paper, we propose a model-based approach to estimate the information gain, instead of on the data itself. This would overcome the limitations of data sparseness. In addition, the model parameters can be explicitly and efficiently estimated via a few matrix operations.

III. Notation

A summary of the notation we use in this paper is shown in Table I.

$\mathbf{K}$	a matrix
$\mathbf{K}_{SR}$	the sub-matrix of $\mathbf{K}$, with row indices S and column indices R
$\mathbf{K}_{sR}$	the sub row vector of $\mathbf{K}$, with row index s and column indices R
$\mathbf{\Sigma}^{(S)}$	the matrix $\mathbf{\Sigma}$ when a feature set S is selected.
$\mathbf{I}_p$	an identity matrix of size $p \times p$
$\mathbf{x}$	a column vector
$\mathbf{x}^\top$	the transposition of vector $\mathbf{x}$
$\mathbf{1}$	a vector whose elements are all ones
$\|\mathbf{x}\|^2$	the square of the norm of $\mathbf{x}$, i.e., $\mathbf{x}^\top\mathbf{x}$
λ	a scalar. regularization parameter.
$\|D\|$	the cardinality of set D.
D	the full index set in $\mathbf{Z}$. $\|D\| = d$.
F	the index set in $\mathbf{Z}$ corresponding to the features, $\mathbf{X}$. $\|F\| = f$.
T	the index set in $\mathbf{Z}$ corresponding to the targets, $\mathbf{Y}$. $\|T\| = t$.
S	the index set of selected features.

TABLE I
A SUMMARY OF NOTATION.

IV. Model-based Feature Selection

A. Feature Selection using Entropy Measure

Suppose we have f feature variables of the underlying data, denoted by $\{X_i | i \in F\}$, where F is the full feature index set, having $|F| = f$. We have the class variable, Y, represented by multiple class indicator variables. For example, in a three-class classification problem, the class variables are represented by vectors $(1, -1, -1)$, $(-1, 1, -1)$ and $(-1, -1, 1)$. The problem of feature selection is to select a subset features, $S \subset F$, to accurately predict the target Y, given that the cardinality of S is m ($m < f$). Let us denote $\{X_i | i \in S\}$ by X_S, for any set S.

The prediction capability of Y given X_S can be measured by the entropy of Y given X_S, which is defined as

$$H(Y|X_S) \stackrel{\text{def}}{=} -\mathbb{E}_{p(Y,X_S)}(\ln p(Y|X_S)), \quad (1)$$

where $\mathbb{E}_p(\cdot)$ is the expectation given the distribution p, and p stands for the underlying data distribution, i.e. the joint distribution $p(Y, X_s)$. The feature selection problem using the mutual information criterion is

$$\arg\min_S H(Y|X_S). \tag{2}$$

Selecting an optimal subset of features is a combinatorial optimization problem, which is an NP problem. For the effective practice is to take a greedy approach, i.e. sequentially selecting features to achieve a sub-optimal solution. Given a selected feature set, S, the one-step goal of feature selection is to select one feature to minimize the entropy. The one-step objective, named as information gain, is to find i to maximize

$$\mathrm{IG(i;S)} \stackrel{\text{def}}{=} \mathrm{H(Y|X_S)} - \mathrm{H(Y;X_{S\cup\{i\}})}.$$

Then the greedy procedure of feature selection based on mutual information is shown in Algorithm 1.

Algorithm 1 Feature selection by information gain

1: Let $S = \emptyset$;
2: **repeat**
3: $i = \arg\max_{i\in F} \mathrm{IG(i;S)}$;
4: $S \leftarrow S \cup \{i\}$;
5: **until** $|S| = m$.

The distribution $p(Y, X_S)$ can be estimated by the empirical distribution, i.e. measuring the proportion of Y and X_S values in the given data. The empirical distribution faces data sparseness problem. Thus, we discuss the estimation based on a multivariate Gaussian generative model in the next section.

B. Multivariate Gaussian Model

We assume that the joint distribution of $\{X_i\}$ and Y is a multivariate normal (Gaussian) distribution,

$$\mathbf{z} = [X_F Y] \sim \mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{3}$$

where $\boldsymbol{\mu}$ is the mean vector, $\boldsymbol{\Sigma}$ is the covariance matrix. Let F be the index set of X in $\mathbf{z}$, and T be the index set of Y in $\mathbf{z}$. The reason is that Gaussian assumption results a linear model, which is simple and scalable.

We denote the feature values in the training data by $\widetilde{\mathbf{X}}$, where each row represents a sample, and each column represents a feature (a gene). We consider multiple target variables. For training data, we denote the target values as $\widetilde{\mathbf{Y}}$, where each row represents target variables of a sample, and each column represents a target variable.

Given the training data, we can estimate the parameters of Eq. (3) by

$$\widehat{\boldsymbol{\mu}} \stackrel{\text{def}}{=} \frac{1}{n}\mathbf{1}^\top[\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}], \qquad \widehat{\boldsymbol{\Sigma}} \stackrel{\text{def}}{=} \frac{1}{n}\mathbf{Z}^\top\mathbf{Z}, \tag{4}$$

where n is the number of rows of matrix $\widetilde{\mathbf{X}}$, $\mathbf{1}$ is a column vector of size n, whose elements are all ones, and

$$\mathbf{Z} \stackrel{\text{def}}{=} [\mathbf{X}, \mathbf{Y}] \stackrel{\text{def}}{=} [\widetilde{\mathbf{X}}, \widetilde{\mathbf{Y}}] - \mathbf{1}\widehat{\boldsymbol{\mu}}^\top. \tag{5}$$

Though the estimation of Eq. (4) is an unbiased estimation of the covariance matrix, such estimation may suffer ill-posed problems. By adding a regularization term, a robust estimation can be obtained:

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\left(\mathbf{Z}^\top\mathbf{Z} + \lambda\mathbf{I}_d\right). \tag{6}$$

This is first proposed in [Stein1975]. Since Eq. (4) is a special case of Eq. (6), we use Eq. (6) as the estimation of $\boldsymbol{\Sigma}$.

After the parameters of the model being estimated, we do not differentiate the parameters and the estimated parameters. For simplicity, we write $\widehat{\boldsymbol{\Sigma}}$ by $\boldsymbol{\Sigma}$. Since the computation of the entropy does not involve $\boldsymbol{\mu}$, we let $\boldsymbol{\mu} = \mathbf{0}$ without loss of generality. Let $\mathbf{z}$ be a d dimensional vector, following the multivariate Gaussian distribution $\mathcal{N}(\mathbf{z}; \mathbf{0}, \boldsymbol{\Sigma})$. There are some properties of multivariate Gaussian distribution.

Property 1: Let $\mathbf{z}_S$ and $\mathbf{z}_T$ be two sub-vectors of $\mathbf{z}$, where S and T are the index sets. We denote the sub-vector of $\boldsymbol{\mu}$ corresponding to an index set S by $\boldsymbol{\mu}_S$, and the sub-matrix of $\boldsymbol{\Sigma}$ corresponding to index sets S and T by $\boldsymbol{\Sigma}_{ST}$. We have

$$\Pr(\mathbf{z}_T|\mathbf{z}_S) \stackrel{\text{def}}{=} \mathcal{N}(\mathbf{z}_T; \boldsymbol{\mu}_{T|S}, \boldsymbol{\Sigma}_{T|S}),$$

where

$$\boldsymbol{\mu}_{T|S} \stackrel{\text{def}}{=} \boldsymbol{\mu}_T + \boldsymbol{\Sigma}_{TS}\boldsymbol{\Sigma}_{SS}^{-1}(\mathbf{z}_S - \boldsymbol{\mu}_S), \tag{7}$$

$$\boldsymbol{\Sigma}_{T|S} \stackrel{\text{def}}{=} \boldsymbol{\Sigma}_{TT} - \boldsymbol{\Sigma}_{TS}\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{ST}. \tag{8}$$

This is the property of conditional distribution [Petersen and Pedersen2006].

Then, the property of the incremental updates on

Property 2: Let D be the full index set of $\mathbf{z}$, $S \subset F$, $i \in F - S$. We have

$$\boldsymbol{\Sigma}_{T|S\cup\{i\}} = \boldsymbol{\Sigma}_{TT}^{(S)} - \frac{1}{\boldsymbol{\Sigma}_{ii}^{(S)}}\boldsymbol{\Sigma}_{Ti}^{(S)}\boldsymbol{\Sigma}_{iT}^{(S)}, \tag{9}$$

where

$$\boldsymbol{\Sigma}^{(S)} \stackrel{\text{def}}{=} \boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{DS}\boldsymbol{\Sigma}_{SS}^{-1}\boldsymbol{\Sigma}_{SD}. \tag{10}$$

Proof: Let $T' = T \cup \{i\}$. By Property 1, $\Pr(\mathbf{z}_{T'}|\mathbf{z}_S)$ follows Gaussian distribution with covariance $\boldsymbol{\Sigma}_{T'|S}$ (see the definition in Eq. (8)). By the definition of $\boldsymbol{\Sigma}^{(S)}$, $\boldsymbol{\Sigma}_{T'|S}$ is the submatrix of $\boldsymbol{\Sigma}^{(S)}$, whose column and row indices are T'. Since $\Pr(\mathbf{z}_T|\mathbf{z}_{S\cup\{i\}}) = \Pr(\mathbf{z}_T|\mathbf{z}_{\{i\}}, \mathbf{z}_S)$, applying Property 1 again, we obtain

$$\boldsymbol{\Sigma}_{T|S\cup\{i\}} = \boldsymbol{\Sigma}_{TT}^{(S)} - \boldsymbol{\Sigma}_{Ti}^{(S)}[\boldsymbol{\Sigma}_{ii}^{(S)}]^{-1}\boldsymbol{\Sigma}_{iT}^{(S)}.$$

As $\boldsymbol{\Sigma}_{ii}^{(S)}$ is a scalar, we have the Eq. (9). ■

The differential entropy of the multivariate normal distribution can be computed by the following Property.

Property 3:

$$H(\mathbf{z}) = -\int \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \ln \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) d\mathbf{z}$$

$$= \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \frac{d}{2}\ln(2\pi e).$$

This can be found in [Petersen and Pedersen2006].

Property 4:

$$H(Y|X_S) = \frac{1}{2}\ln\left|\mathbf{\Sigma}_{T|S}\right| + \frac{d}{2}\ln(2\pi e), \tag{11}$$

where T be its set of indices for Y in $\mathbf{z}$.
By Property 1 and Property 3, we have the property of the conditional multivariate normal distribution.

C. Feature Selection Algorithms

Now we propose two sets of feature selection algorithms based on multivariate Gaussian model and entropy measure.

C.1 D-optimality Feature Selection

In the multivariate Gaussian model, the problem of feature selection Eq. (2) becomes

$$\arg\min_{S} \quad \ln\left|\mathbf{\Sigma}_{T|S}\right|. \tag{12}$$

As the determinant of the covariance matrix is known as *generalized variance*. This criterion is to minimize the generalized variance of the joint distribution of targets. We name the feature selection based on the determinant criterion (Eq. (12)) as *D-optimality Feature Selection* after determinant. We borrow the name of D-optimality from experimental designs[Fedorov1972].

As solving Eq. (12) is still an NP-problem, we use the greedy approach as Algorithm 1. Let $\mathbf{\Sigma}^{(S)} = \mathbf{K} + \lambda\mathbf{I}_d$. By Eq. (9), we have

$$\begin{aligned}\ln\left|\mathbf{\Sigma}_{T|S\cup\{i\}}\right| &= \ln\left|\mathbf{\Sigma}^{(S)}_{TT} - \frac{1}{\mathbf{\Sigma}^{(S)}_{ii}}(\mathbf{\Sigma}^{(S)}_{Ti}\mathbf{\Sigma}^{(S)}_{iT})\right| \\ &= \ln\left|\mathbf{K}_{TT} + \lambda\mathbf{I}_t - \frac{1}{\mathbf{K}_{ii}+\lambda}(\mathbf{K}_{Ti}\mathbf{K}_{iT})\right| \\ &= \ln|\mathbf{K}_{TT} + \lambda\mathbf{I}_t| + \ln(1 - \frac{\mathbf{K}_{iT}(\mathbf{K}_{TT}+\lambda\mathbf{I}_t)^{-1}\mathbf{K}_{Ti}}{\mathbf{K}_{ii}+\lambda}),\end{aligned} \tag{13}$$

where $t = |T|$. Therefore

$$\begin{aligned}&\arg\min_i \ln\left|\mathbf{\Sigma}_{T|S\cup\{i\}}\right| \\ &\quad = \arg\max_i \frac{\mathbf{K}_{iT}(\mathbf{K}_{TT}+\lambda\mathbf{I}_t)^{-1}\mathbf{K}_{Ti}}{\mathbf{K}_{ii}+\lambda}\end{aligned} \tag{14}$$

We can compute $\mathbf{\Sigma}^{(S\cup\{i\})}$ from $\mathbf{\Sigma}^{(S)}$ by Eq. (10).

$$\begin{aligned}\mathbf{\Sigma}^{(S\cup\{i\})} &= \mathbf{\Sigma}^{(S)} - \frac{1}{\mathbf{\Sigma}^{(S)}_{ii}}(\mathbf{\Sigma}^{(S)}_{Di}\mathbf{\Sigma}^{(S)}_{iD}) \\ &= \mathbf{K} + \lambda\mathbf{I}_d - \frac{1}{\mathbf{K}_{ii}+\lambda}(\mathbf{K}_{Di}\mathbf{K}_{iD} + \lambda\boldsymbol{\delta}_i\mathbf{K}_{iD} \\ &\quad + \lambda\mathbf{K}_{Di}\boldsymbol{\delta}_i^\top + \lambda^2\boldsymbol{\delta}_i\boldsymbol{\delta}_i^\top),\end{aligned} \tag{15}$$

where $\boldsymbol{\delta}_i$ is a column vector whose elements are zeros except that the i-th element is one. Since we shall not select a feature twice, we shall not concern the values in the i-th row or column in $\mathbf{K}$ any more. Therefore we can update $\mathbf{K}$ by $\mathbf{K} - \frac{1}{\mathbf{K}_{ii}+\lambda}(\mathbf{K}_{Di}\mathbf{K}_{iD})$. By sequentially updating $\mathbf{K}$, we have Algorithm 2. Note that since we only compare the value for features, we can drop the scale factor $\frac{1}{n}$ in Eq. (6) for simplicity. The complexity of Algorithm 2 is $O(m(d^2 + dt^2))$, where $d = |D|$ and $t = |T|$.

Since the complexity of the algorithm contains md^2, the algorithm is very inefficient when d is large. Especially, the memory complexity is $O(d^2)$. When the sample size n is much smaller than d, we can take advantage of it to speed up the algorithm. Assume that $\mathbf{K}$ has the form of $\mathbf{Z}^\top\mathbf{\Phi}\mathbf{Z}$. Note that $\mathbf{\Phi}$ is symmetric since $\mathbf{K}$ is symmetric. Initially, $\mathbf{\Phi} = \mathbf{I}_n$ in Step 2 of Algorithm 2.

Algorithm 2 D-optimality Feature Selection-I

1: $S = \emptyset$;
2: $\mathbf{K} = \mathbf{Z}^\top\mathbf{Z}$;
3: **repeat**
4: Let $\mathbf{U} = (\mathbf{K}_{TT} + \lambda\mathbf{I}_t)^{-1}$;
5: $i = \arg\max_{i\in F-S} \frac{\mathbf{K}_{iT}\mathbf{U}\mathbf{K}_{Ti}}{\mathbf{K}_{ii}+\lambda}$;
6: $\mathbf{K} \leftarrow \mathbf{K} - \frac{1}{\mathbf{K}_{ii}+\lambda}(\mathbf{K}_{Di}\mathbf{K}_{iD})$;
7: $S \leftarrow S \cup \{i\}$;
8: **until** $|S| = m$.

Let us denote the i-th column vector of $\mathbf{X}$ as $\mathbf{x}_i$. The update of $\mathbf{K}$ can be written as

$$\begin{aligned}&\mathbf{K} - \frac{1}{\mathbf{K}_{ii}+\lambda}(\mathbf{K}_{Di}\mathbf{K}_{iD}) \\ &\quad = \mathbf{Z}^\top\mathbf{\Phi}\mathbf{Z} - \frac{1}{\mathbf{x}_i^\top\mathbf{\Phi}\mathbf{x}_i+\lambda}(\mathbf{Z}^\top\mathbf{\Phi}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{\Phi}\mathbf{Z}).\end{aligned} \tag{16}$$

Therefore, we derive the update for $\mathbf{\Phi}$ in Algorithm 3. The complexity of the algorithm is $O(m(dn^2 + nt^2 + t^3))$.

Algorithm 3 D-optimality Feature Selection-II

1: $S = \emptyset$;
2: $\mathbf{\Phi} = \mathbf{I}_n$;
3: **repeat**
4: Let $\mathbf{\Omega} = \mathbf{\Phi}\mathbf{Y}(\mathbf{Y}^\top\mathbf{\Phi}\mathbf{Y} + \lambda\mathbf{I}_t)^{-1}\mathbf{Y}^\top\mathbf{\Phi}$;
5: $i = \arg\max_{i\in F-S} \frac{\mathbf{x}_i^\top\mathbf{\Omega}\mathbf{x}_i}{\mathbf{x}_i^\top\mathbf{\Phi}\mathbf{x}_i+\lambda}$;
6: $\mathbf{\Phi} \leftarrow \mathbf{\Phi} - \frac{1}{\mathbf{x}_i^\top\mathbf{\Phi}\mathbf{x}_i+\lambda}(\mathbf{\Phi}\mathbf{x}_i\mathbf{x}_i^\top\mathbf{\Phi})$;
7: $S \leftarrow S \cup \{i\}$;
8: **until** $|S| = m$.

N.B. $\mathbf{x}_i$ is the i-th column of centered feature matrix $\mathbf{X}$.

When $t < n$, we can reduce the complexity by sequentially compute $\mathbf{\Phi}\mathbf{X} \stackrel{\text{def}}{=} \mathbf{P}$. We have Algorithm 4, whose complexity is $O(mdnt)$. Note that Phi and P in Algorithm 3 and Algorithm 4 are used to save the intermediate results in the matrix computation to reduce the computation complexity.

C.2 A-optimality Feature Selection

Because of the complexity in computing determinants and non-convexity of log-determinants, we can replace the log-

Algorithm 4 D-optimality Feature Selection-III

$S = \emptyset$;
$\mathbf{\Phi} = \mathbf{I}_n$;
$\mathbf{P} = \mathbf{X}$;
repeat
Let $\mathbf{R} = \mathbf{Y}((\mathbf{Y}^\top \mathbf{\Phi} \mathbf{Y} + \lambda \mathbf{I}_t)^{-1} \mathbf{Y}^\top \mathbf{P})$;
$i = \arg\max_{i \in F - S} \frac{\mathbf{r}_i^\top \mathbf{p}_i}{\mathbf{x}_i^\top \mathbf{p}_i + \lambda}$;
$\mathbf{\Phi} \leftarrow \mathbf{\Phi} - \frac{1}{\mathbf{x}_i^\top \mathbf{p}_i + \lambda}(\mathbf{p}_i \mathbf{p}_i^\top)$;
$\mathbf{P} \leftarrow \mathbf{P} - \frac{1}{\mathbf{x}_i^\top \mathbf{p}_i + \lambda}(\mathbf{p}_i (\mathbf{p}_i^\top \mathbf{X}))$;
$S \leftarrow S \cup \{i\}$;
until $|S| = m$.

N.B. $\mathbf{x}_i$ is the i-th column of centered feature matrix $\mathbf{X}$. $\mathbf{p}_i$ is the i-th column of $\mathbf{P}$, $\mathbf{r}_i$ the i-th column of $\mathbf{R}$.

determinant of the covariance matrix with the trace of the covariance matrix, which is the upper-bound of the log-determinant of the covariance matrix.

Lemma 1: If $\mathbf{X}$ is a $p \times p$ positive definite matrices, it holds that $\ln |\mathbf{X}| \leq \mathrm{tr}(\mathbf{X}) - \mathrm{p}$. The equality holds when $\mathbf{X}$ is an orthonormal matrix.

Proof: Let $\{\lambda_1, \cdots, \lambda_p\}$ be the eigenvalues of $\mathbf{X}$. We have $\ln |\mathbf{X}| = \sum_i \ln \lambda_i$ and $\mathrm{tr}(\mathbf{X}) = \sum_{\mathrm{i}} \lambda_{\mathrm{i}}$. Since $\ln \lambda_i \leq \lambda_i - 1$, we have the inequality. The equality holds when $\lambda_i = 1$. Therefore, when $\mathbf{X}$ is an orthonormal matrix (especially $\mathbf{X} = \mathbf{I}_p$), the equality holds. ■

As $\ln \left|\mathbf{\Sigma}_{T|S}\right| \leq \mathrm{tr}\left(\mathbf{\Sigma}_{\mathrm{T|S}}\right) - \mathrm{t}$, the problem feature selection (12) can be approximated by

$$\arg\min_{S} \quad \mathrm{tr}\left(\mathbf{\Sigma}_{\mathrm{T|S}}\right) = \arg\max_{\mathrm{S}} \quad \mathrm{tr}\left(\mathbf{\Sigma}_{\mathrm{TS}} \mathbf{\Sigma}_{\mathrm{SS}}^{-1} \mathbf{\Sigma}_{\mathrm{ST}}\right). \quad (17)$$

Since the trace of the covariance divided by the number of variables is the *average* covariance, Eq. (17) is called *A-optimality feature selection*. We also borrow the name of A-optimality from experimental designs, which is an alternative of the D-optimality criterion[Fedorov1972].

To sequentially solve Eq. (17), we have Algorithm 5. The algorithm is similar to the sequential algorithm in [Yu et al.2006], which is to solve transductive active learning problems. The complexity of Algorithm 5 is $O(md^2)$.

Algorithm 5 A-optimality Feature Selection-I

$S = \emptyset$;
$\mathbf{K} = \mathbf{Z}^\top \mathbf{Z} + \lambda \mathbf{I}_d$;
repeat
$i = \arg\max_{i \in F - S} \frac{\mathbf{K}_{iT} \mathbf{K}_{Ti}}{\mathbf{K}_{ii} + \lambda}$;
$\mathbf{K} \leftarrow \mathbf{K} - \frac{1}{\mathbf{K}_{ii} + \lambda}(\mathbf{\Sigma}_{Di} \mathbf{K}_{iD})$;
$S \leftarrow S \cup \{i\}$;
until $|S| = m$.

When $n \ll d$, we can use a similar method as Algorithm 3 to speed up Algorithm 5. We can obtain Algorithm 5 by letting $\mathbf{U} = \mathbf{I}_t$ in Algorithm 2. Then we can use

$$\mathbf{\Omega} = \mathbf{\Phi} \mathbf{Y} \mathbf{Y}^\top \mathbf{\Phi}$$

in Step 4 of Algorithm 3 to obtain Algorithm 6. The complexity of Algorithm 6 is $O(mdn^2)$.

Algorithm 6 A-optimality Feature Selection-II

$S = \emptyset$;
$\mathbf{\Phi} = \mathbf{I}_n$;
repeat
Let $\mathbf{\Omega} = \mathbf{\Phi} \mathbf{Y} \mathbf{Y}^\top \mathbf{\Phi}$;
$i = \arg\max_{i \in F - S} \frac{\mathbf{x}_i^\top \mathbf{\Omega} \mathbf{x}_i}{\mathbf{x}_i^\top \mathbf{\Phi} \mathbf{x}_i + \lambda}$;
$\mathbf{\Phi} \leftarrow \mathbf{\Phi} - \frac{1}{\mathbf{x}_i^\top \mathbf{\Phi} \mathbf{x}_i + \lambda}(\mathbf{\Phi} \mathbf{x}_i \mathbf{x}_i^\top \mathbf{\Phi})$;
$S \leftarrow S \cup \{i\}$;
until $|S| = m$.

N.B. $\mathbf{x}_i$ is the i-th column of centered feature matrix $\mathbf{X}$.

We can also sequentially compute $\mathbf{\Phi X} \stackrel{\text{def}}{=} \mathbf{P}$ as shown in Algorithm 7, whose time complexity is $O(mdn + dnt)$.

Algorithm 7 A-optimality Feature Selection-III

$S = \emptyset$;
$\mathbf{P} = \mathbf{X}$;
$\mathbf{R} = \mathbf{Y}(\mathbf{Y}^\top \mathbf{P})$;
repeat
$i = \arg\max_{i \in F - S} \frac{\mathbf{r}_i^\top \mathbf{p}_i}{\mathbf{x}_i^\top \mathbf{p}_i + \lambda}$;
$\mathbf{R} \leftarrow \mathbf{R} - \frac{1}{\mathbf{x}_i^\top \mathbf{p}_i + \lambda}(\mathbf{r}_i (\mathbf{p}_i^\top \mathbf{X}))$;
$\mathbf{P} \leftarrow \mathbf{P} - \frac{1}{\mathbf{x}_i^\top \mathbf{p}_i + \lambda}(\mathbf{p}_i (\mathbf{p}_i^\top \mathbf{X}))$;
$S \leftarrow S \cup \{i\}$;
until $|S| = m$.

N.B. $\mathbf{x}_i$ is the i-th column of centered feature matrix $\mathbf{X}$. $\mathbf{p}_i$ is the i-th column of $\mathbf{P}$, $\mathbf{r}_i$ the i-th column of $\mathbf{R}$.

Table II shows the summary of the complexity of the above algorithms. In the gene expression data, which contain a large number of genes, but a small number of samples, the D-opt III and A-opt III are the good choice for computational efficiency.

Algorithm	Time	Space
D-opt I	$O(m(d^2 + dt^2))$	$O(d^2)$
D-opt II	$O(m(dn^2 + nt^2 + t^3))$	$O(n^2 + d)$
D-opt III	$O(mdnt)$	$O(n^2 + nd)$
A-opt I	$O(m(d^2 + dt^2))$	$O(d^2)$
A-opt II	$O(mdn^2)$	$O(n^2 + d)$
A-opt III	$O(mdn + dnt)$	$O(nd)$

TABLE II
THE COMPLEXITY OF ALGORITHMS. THE SPACE COMPLEXITY MEASURES THE EXTRA REQUIRED SPACE BESIDES DATA.

V. Methods Used for Comparison

In this section, we describe several feature selection methods used in our experimental comparisons.

A. Rankgene

We use the following feature selection methods provided in the program Rankgene [Su et al.2003]: *information gain*, *twoing rule*, and *sum minority*. These methods have been widely used either in machine learning (information gain) or in statistical learning theory (twoing rule and sum minority). All these methods measure the effectiveness of a feature by evaluating the strength of class prediction when the prediction is made by splitting it into two regions, the high region and the low region, by considering all possible split points.

B. Max-Relevance

The Max-Relevance method selects a set of genes with the highest relevance to the target class [Peng et al.2005]. Given g_i which represents the gene i, and the class label c, their mutual information is defined in terms of their frequencies of appearances $p(g_i)$, $p(c)$, and $p(g_i, c)$ as follows.

$$I(g_i, c) = \iint p(g_i, c) \ln \frac{p(g_i, c)}{p(g_i)p(c)} dg_i dc \qquad (18)$$

The Max-Relevance method selects the top m genes in the descent order of $I(g_i, c)$, i.e. the best m individual features correlated to the class labels.

C. mRMR

Although we can choose the top individual genes using Max-Relevance algorithm, it has been recognized that "the m best features are not the best m features" since the correlations among those top features may also be high [Cover1974]. In order to remove the redundancy among features, a Min-Redundancy and Max-Relevance (mRMR) framework is proposed in [Peng et al.2005]. In mRMR, the mutual information between each pair of genes is taken into consideration. Suppose set G represents the set of genes and we already have S_{m-1}, the feature set with m-1 genes, then the task is to select the m-th feature from the set $\{G - S_{m-1}\}$. In the following formula, we see that minimizing the redundancy and maximizing the relevance can be achieved concordantly [Peng et al.2005]. Methods proposed in [Yu et al.2004] shares the similar idea with mRMR.

$$\max_{g_j \in G - S_{m-1}} [I(g_i; c) - \frac{1}{m-1} \sum_{g_i \in G_{m-1}} I(g_j; g_i)] \qquad (19)$$

D. ReliefF

ReliefF is a simple yet efficient procedure to estimate the quality of attributes in problems with strong dependencies between attributes [Marko and Igor2003]. In practice, ReliefF is usually applied as a feature subset selection method.

The key idea of the ReliefF is to estimate the quality of genes according to how well their values distinguish between instances that are near to each other. Given a randomly selected instance Ins_m from class L, ReliefF searches for K of its nearest neighbors from the same class called nearest hits H, and also K nearest neighbors from each of the different classes, called nearest misses M. It then updates the quality estimation W_i for gene i based on their values for Ins_m, H, M. If instance Ins_m and those in H have different values on gene i, then the quality estimation W_i is decreased. On the other hand, if instance Ins_m and those in M have different values on the the gene i, then W_i is increased. The whole process is repeated n times which is set by users. The equation below can be used to update W_i.

$$W_i = W_i - \frac{\sum_{k=1}^{K} D_H}{n \cdot K} + \sum_{c=1}^{C-1} P_c \cdot \frac{\sum_{k=1}^{K} D_{M_c}}{n \cdot K} \qquad (20)$$

where n_c is the number of instances in class c, D_H (or D_{M_c}) is the sum of distance between the selected instance and each H (or M_c), P_c is the prior probability of class c. Detailed discussions on ReliefF can be found i [Marko and Igor2003].

E. D-opt and A-opt Methods

Consider the large number of features (genes) and the relative small number of samples, we use Algorithm 4 for the D-optimality feature selection (denoted by D-opt) and Algorithm 7 for the A-optimality feature selection (denoted by A-opt). The mean of each data set is removed as Eq. (5). Note that the standardization is a way of increasing the degree of normality for the gene expression data [Yeung et al.2001]. The regularization parameter, λ, is set to 0.5 in our experiments.

VI. Experiments

We conduct three sets of experiments using seven datasets as described in Section VI-A. In the first set of experiments, we compare the classification accuracy of data with gene selection and without gene selection using Support Vector Machine (SVM) classifiers implemented in the LIBSVM package [Chang and Lin2001]. The second set of experiments provides a comprehensive study on the performance of different gene selection methods under different conditions. In the third set of experiments, we discuss the number of selected genes.

A. Datasets Description

The datasets and their characteristics are summarized in Table III.

The ALL dataset [Yeoh et al.2002] is a dataset that covers six subtypes of acute lymphoblastic leukemia: BCR (15), E2A (27), Hyperdip (64), MLL (20), T (43), and TEL (79). Here the numbers in the parentheses are the numbers of samples. The dataset is available at [ALL]. The GCM dataset [Ramaswamy et al.2001] consists of 198 human tumor samples of fifteen types. The HBC dataset consists of 22 hereditary breast cancer samples and was first studied in [Hedenfalk et al.2001]. The dataset has three classes and can be downloaded at [HBC]. The Lymphoma dataset is a dataset of the three most prevalent adult lymphoid malignancies and available at [LYM] and it was first studied in [Alizadeh et al.2000]. The MLL-leukemia dataset consists of three classes and can be downloaded at [MLL]. The

NCI60 dataset was first studied in [Ross et al.2000]. cDNA microarrays were used to examine the variation in gene expression among the 60 cell lines from the National Center Institute's anticancer drug screen. The dataset spans nine classes and can be downloaded at [NCI60]. SRBCT [Khan et al.2001] is the dataset of small, round blue cell tumors of childhood and can be downloaded at [SRBCT].

Dataset	# Samples	# Genes	# Classes
ALL	248	12558	6
GCM	198	16063	14
HBC	22	3226	3
Lymphoma	62	4026	3
MLL	72	12582	3
NCI60	60	1123	9
SRBCT	83	2308	4

TABLE III
THE DATASET DESCRIPTION.

B. Effectiveness of Gene Selection

Table IV presents the accuracy values of applying SVM on the top 30 genes selected by different methods and also on all the genes without selection. The accuracy values are obtained via 10-fold cross validation. The table shows that gene selection improves classification performance, at least the accuracy of SVM on genes selected by both the D-opt and A-opt methods outperform that without feature selection. We will discuss the number of selected genes in Section VI-D.

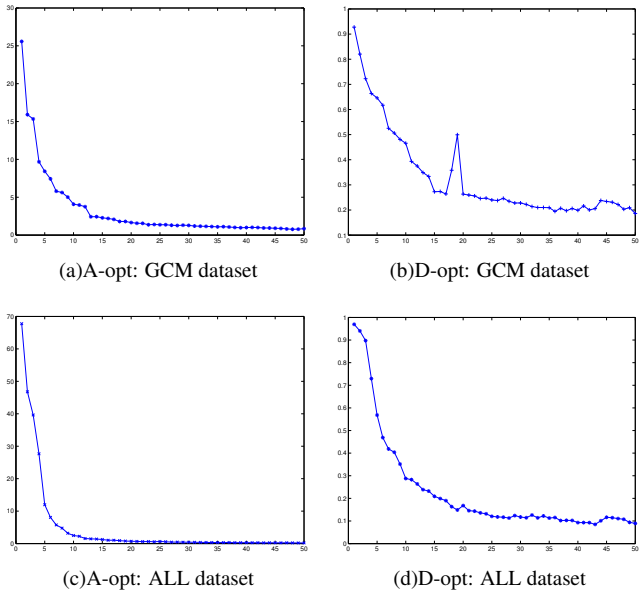
C. Performance of Different Gene Selection Methods

In this section, we present a comparative study of various gene selection methods using SVM and Naive Bayes algorithms on the seven datasets. Both SVM and Naive Bayes have been widely used in previous studies(e.g., [Li et al.2004], [Peng et al.2005]). Figure 1 and 2 show the classification accuracy results as a function of the number of selected genes on the seven datasets, respectively. From the comparative study, we observe that:

- Gene selection by experimental design (D-opt and A-opt) outperforms other gene selection methods such as information gain, etc. It largely owes to the generality of the multivariate Gaussian generative model. In addition, our methods estimate the information gain based on models, instead of on the data itself. This overcomes the limitations of data sparseness and provides more robust and accurate estimations.
- The results of the A-opt method are similar to those of the D-opt method. Besides the simplicity of A-opt, the A-opt method outperforms the D-opt method in most cases. There are some discussion of comparing A-optimality and D-optimality in the literature experimental designs [Fedorov1972].
- Gene selection by D-opt and A-opt implicitly selects the features with the minimum redundancy. In Step 6 of Algorithm 2 and Step 5 of Algorithm 5, the covariance matrices are updated, which remove the second-order redundancy. We can find similar actions in other algorithms as well.

D. Number of Selected Genes

From the above experiment, it can be observed that when the number of selected genes is greater than 30, the variation of the performance is small. In Step 6 of Algorithm 4, we select genes to reduce the generalized variance. In Step 5 of Algorithm 7, we select genes to reduce the total variance. Figure 3, 4 and 5 show the variance reduction as the function of the number of genes on the seven datasets, respectively. The number of selected genes is varied from 1 to 50, and the results show the change of classification accuracy.



(a)A-opt: GCM dataset (b)D-opt: GCM dataset

(c)A-opt: ALL dataset (d)D-opt: ALL dataset

Fig. 3. Variance Reduction on datasets (I).

The experiment results demonstrate that only a small number of genes are needed for classification purpose. In our experiments, we observe that when the number of selected genes is greater than 30, the variation of the classification performance is small. We find that the cumulative reduction in generalized variance or total variance converges after 30 steps.

E. Other Discussion

This set of experiments aims to study the choice of the regularization parameter λ in our proposed A-opt and D-opt methods. We set the number of selected gene to be 30, and change λ from 0.1 to 0.9. Figure 6 shows that the accuracy is not sensitive to the regularization parameter. Note that on LYM and HBC datasets, the accuracies of both methods are 100% under different regularization parameters. In our experiments, we choose 0.5 as λ.

VII. DISCUSSIONS

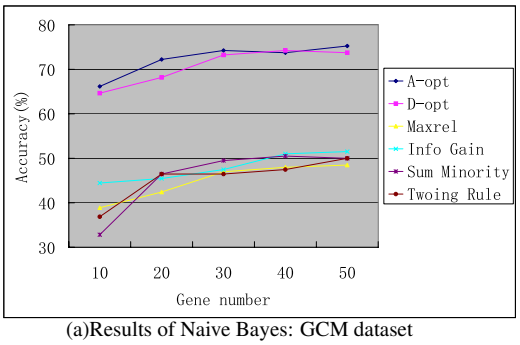
Though we are studying the feature selection problem in this paper, the idea largely owes to that of experimental designs.

In the statistics literature, the *experimental designs* can be backtracked to the ideas presented in [Kiefer1959]. The goal of experimental designs is usually to extract the maximum amount of information from as few observations as possible. For experimental designs, several criteria can be used, such as D-

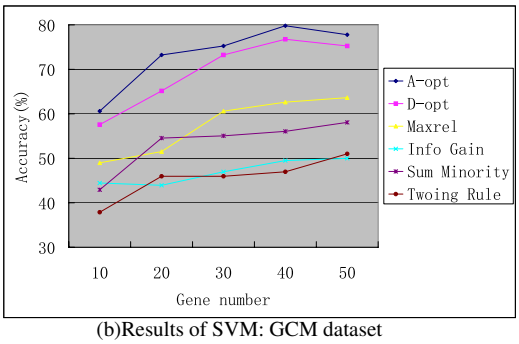
	SRBCT	NCI60	Lymphoma	GCM	HBC	ALL	MLL
No Gene Selection	85.22%	63.33%	95.16%	51.52%	77.27%	91.94%	97.22%
mRMR	81.82%	53.33%	**100.0%**	N/A	95.45%	N/A	N/A
Maxrel	84.09%	51.67%	**100.0%**	60.61%	72.73%	89.11%	77.78%
ReliefF	89.77%	58.33%	**100.0%**	55.25%	95.45%	96.37%	94.44%
Information Gain	89.77%	61.67%	98.39%	46.97%	**100.0%**	97.58%	98.67%
Sum Minority	78.41%	65.00%	98.39%	55.05%	95.45%	93.95%	90.28%
Twoing Rule	84.09%	61.67%	98.39%	45.96%	90.91%	96.77%	97.22%
A-opt (Alg. 7)	**94.32%**	**88.33%**	**100.0%**	**75.25%**	**100.0%**	99.19%	**100.0%**
D-opt (Alg. 4)	90.91%	80.00%	**100.0%**	73.23%	**100.0%**	**100.0%**	**100.0%**

TABLE IV

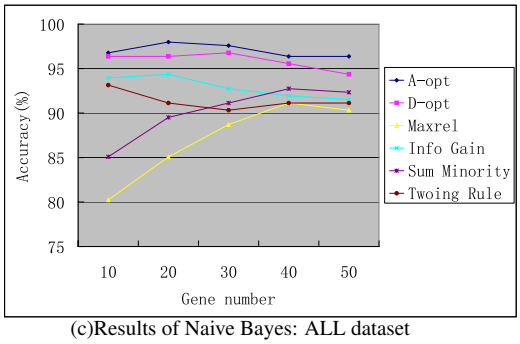
COMPARATIVE ACCURACY OF DIFFERENT SELECTION METHODS ON 7 DATA SETS (GENE NUMBER = 30). BECAUSE OF LIMITATION OF MEMORY, MRMR CAN NOT RUN ON GCM, ALL AND MLL.



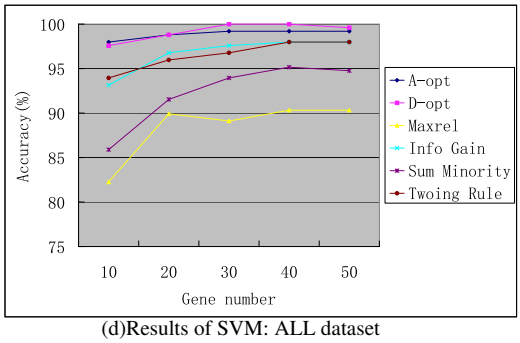
(a)Results of Naive Bayes: GCM dataset



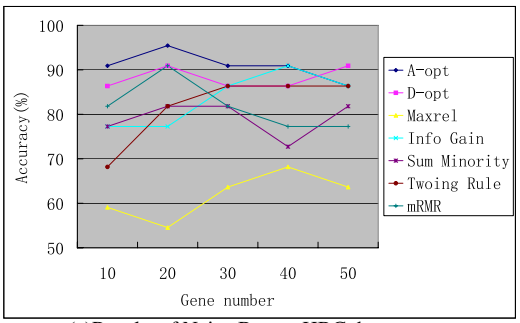
(b)Results of SVM: GCM dataset



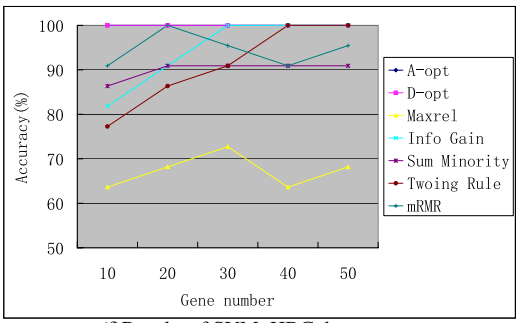
(c)Results of Naive Bayes: ALL dataset



(d)Results of SVM: ALL dataset

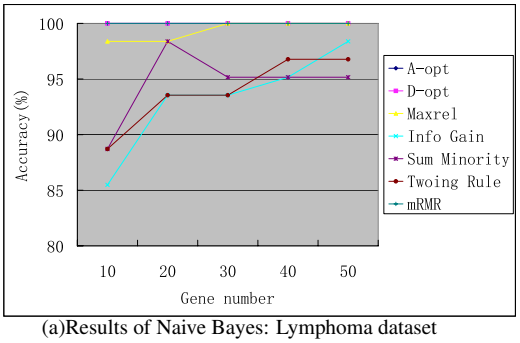


(e)Results of Naive Bayes: HBC dataset

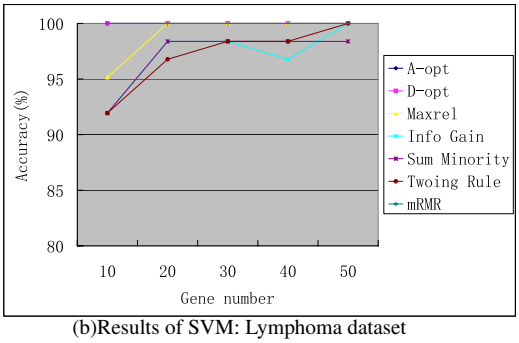


(f)Results of SVM: HBC dataset

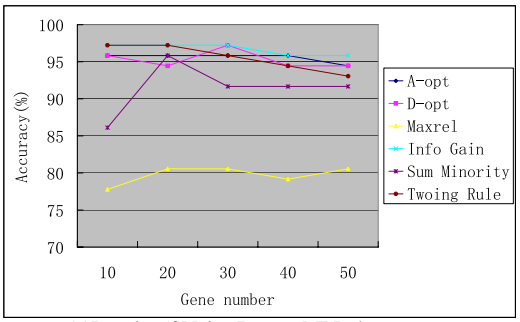
Fig. 1. Comparison of Various Gene Selection Methods (I).



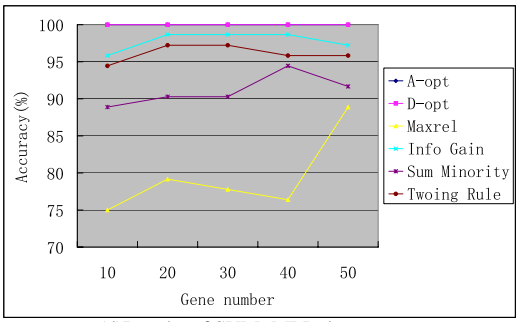
(a)Results of Naive Bayes: Lymphoma dataset



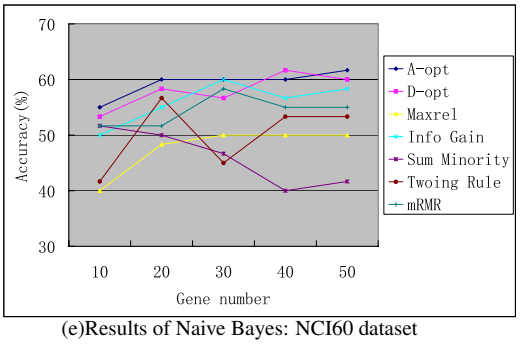
(b)Results of SVM: Lymphoma dataset



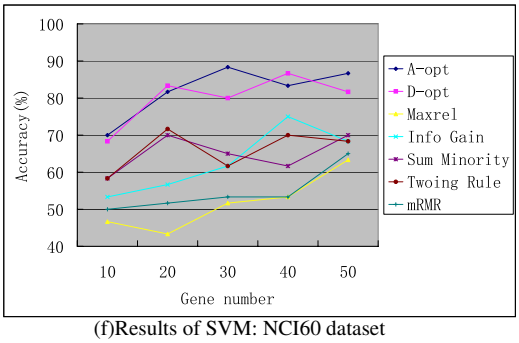
(c)Results of Naive Bayes: MLL dataset



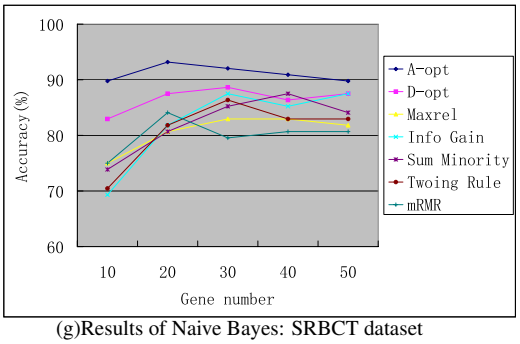
(d)Results of SVM: MLL dataset



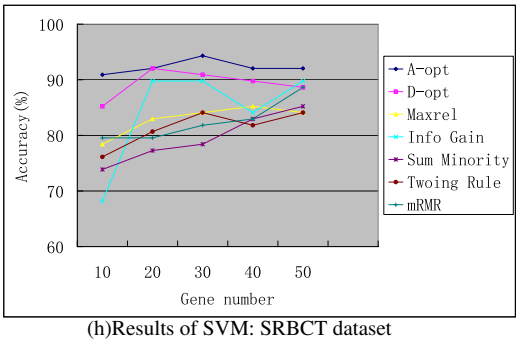
(e)Results of Naive Bayes: NCI60 dataset



(f)Results of SVM: NCI60 dataset



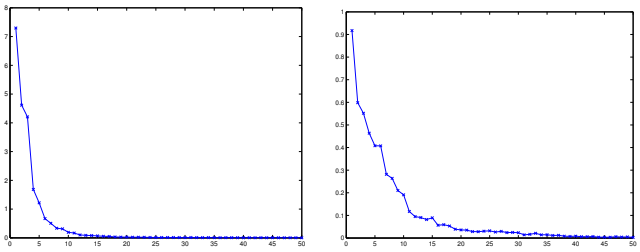
(g)Results of Naive Bayes: SRBCT dataset



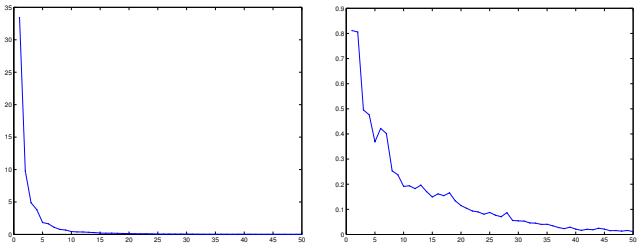
(h)Results of SVM: SRBCT dataset

Fig. 2. Comparison of Various Gene Selection Methods (II).

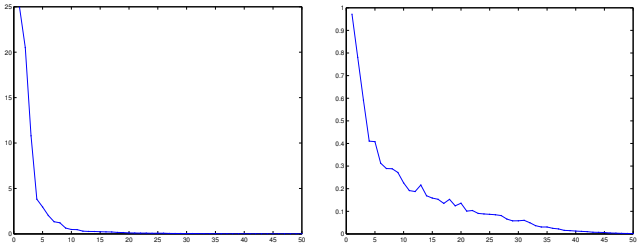
optimality and A-optimality [Fedorov1972]. They all concern about reducing the uncertainties of estimated parameters. The criterion of the D-optimality is minimizing the generalized variance of joint distribution of parameters, i.e. the *determinant* of



(a)A-opt: HBC dataset (b)D-opt: HBC dataset

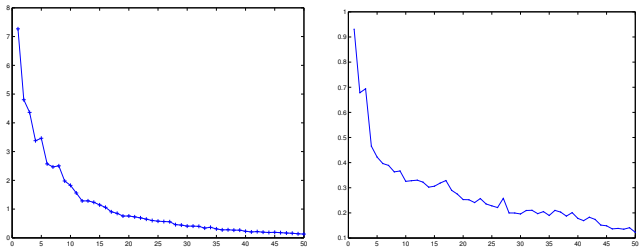


(c)A-opt: Lymphoma Dataset (d)D-opt: Lymphoma Dataset

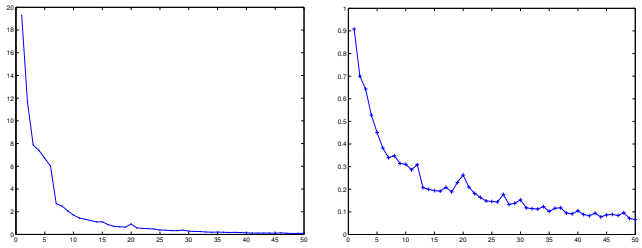


(e)A-opt: MLL dataset (f)D-opt: MLL dataset

Fig. 4. Variance Reduction on datasets (II).



(a)A-opt: NCI60 dataset (b)D-opt: NCI60 dataset



(c)A-opt: SRBCT dataset (d)D-opt: SRBCT dataset

Fig. 5. Variance Reduction on datasets (III).

the multivariate variance, which gives its name. The criterion of the A-optimality is minimizing the *average* variance of all parameters.

As concentrating on the predictive variance of a target set of data, Yu et al [Yu et al.2006] propose *transductive experimental designs* for least-squares linear (or kernel) regression. The idea is to add samples to training set in order to improve the numerical stability of predictions on the target test data, measured by the inversion of the Fisher information matrix. It has been shown that the predictive stability only depends on the *locations* of the selected training data, while does not depend on their *label values*, which leads to a very simple active learning approach[Yu et al.2006].

Though there is a big difference between experimental designs and feature selection at the first glance, we find a *duality property* between them.

Let us consider the problem of predicting target Y given a row of feature random vector X. We assume that the model is a linear model,

$$Y = X^\top \mathbf{w} + \epsilon, \tag{21}$$

where $\mathbf{w}$ is the weight vector and ϵ is the error. The reason of using linear models is because linear models are simple and scalable.

Given the training data $\mathbf{y}$ and $\mathbf{X}$, where $\mathbf{y}$ is the column target vector and $\mathbf{X}$ is the feature matrix, each row of $\mathbf{X}$ is a feature vector. We can write Eq. (21) in matrix format as

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon}$ is the error vector.

We further assume the loss function is a square loss, therefore we want to minimize $\frac{1}{2}\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon}$. Meanwhile, we prefer a robust estimation of $\mathbf{w}$, i.e., a regularization term, $\frac{\lambda}{2}\mathbf{w}^\top\mathbf{w}$. Combining them, the estimation problem becomes

$$\arg\min_{\mathbf{w}} \frac{1}{2}(\mathbf{y}-\mathbf{X}\mathbf{w})^\top(\mathbf{y}-\mathbf{X}\mathbf{w}) + \frac{\lambda}{\mathbf{2}}\mathbf{w}^\top\mathbf{w}. \tag{22}$$

The problem (22) can be explicitly solved as

$$\widehat{\mathbf{w}} = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}. \tag{23}$$

This is also known as ridge regression.

Given a feature vector $\mathbf{x}$, the estimation of y is

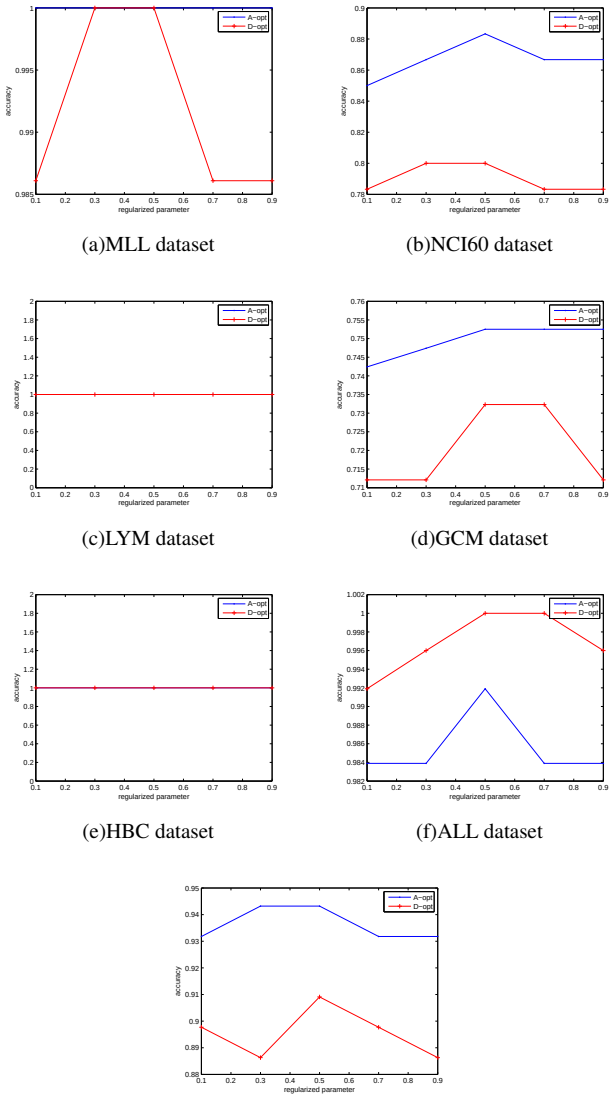
$$\widehat{y} = \mathbf{x}^\top(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y}. \tag{24}$$

On the other hand, we can estimate y by the multivariate Gaussian model. For simplicity, we assume that $\boldsymbol{\mu} = \mathbf{0}$. By Eq. (7), we know

$$\begin{aligned}\hat{y} &= \boldsymbol{\mu}_{T|F} = \boldsymbol{\Sigma}_{TF}(\boldsymbol{\Sigma}_{FF})^{-1}\mathbf{x} \\ &= \mathbf{y}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{x},\end{aligned}$$

which is equal to Eq. (24) as long as we have the same λ.

This shows the duality between the target label and the feature. This property motivate us to treat the feature selection as a dual problem of selecting data samples to label in active learning or experimental designs. Then we can apply experimental design approaches, more precisely *transductive experimental design* [Yu et al.2006], onto the feature selection problem.



(a)MLL dataset (b)NCI60 dataset

(c)LYM dataset (d)GCM dataset

(e)HBC dataset (f)ALL dataset

(g)SRBCT dataset

Fig. 6. Different regularization parameters on 7 datasets.

VIII. Conclusions

In this paper, we suggest multivariate Gaussian generative models for feature (gene) selection because multivariate normal (Gaussian) distributions are maximum entropy probability distribution. Using the model based entropy estimation, we avoid the data sparseness problem which commonly happens in the empirical information gain approach.

Using the properties of multivariate normal distributions, we derive the feature selection methods based on the D-optimality criterion and its approximation, A-optimality criterion.

To efficiently select genes from gene expression data, where the numbers of features are large and the numbers of samples are relatively small, we propose several simple algorithms (a few lines of code). Among them, Algorithm 4 and Algorithm 7 are most suitable for gene expression data. The time complexity of the proposed algorithms is linear to the product of the number of genes and the number of samples for each iteration of selection.

The experiments on seven gene datasets and the comparison with other five approaches show the accuracy and efficiency of our approach.

Acknowledgment

Tao Li would like to thank NIH/NIGMS S06 GM008205 and NSF IIS-0546280 for partially support this work.

References

[Alizadeh et al.2000] Alizadeh, A. A., Eisen, M. B., David, R. E., Ma, C., Lossos, I. S., osenwald, A. R., Boldrick, H. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Martu, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, G. P., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botsten, D., Brown, P. O., and Staudt, L. M. 2000. Distinct types of diffuse large B-cell lymphoma identified by gene expression pr ofiling. *Nature 403*, 503–511.

[ALL] ALL. `http://www.stjuderesearch.org/data/ALL1/`.

[Chang and Lin2001] Chang, C.-C. and Lin, C.-J. 2001. *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[Chee et al.1996] Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X., Stern, D., Winkler, J., Lockhart, D., Morris, M., and Fodor, S. 1996. Accessing genetic information with high density DNA arrays. *Science 274*, 610–614.

[Cover1974] Cover, T. 1974. The best two independent measurements are not the two best. *IEEE Trans. Systems, and Cybernetics 4*, 116–117.

[Dudoit et al.2002] 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,* 457, 77–87.

[Fedorov1972] Fedorov, V. V. 1972. *Theory of Optimal Experiments*. Academic Press, New York.

[Fodor et al.1991] Fodor, S., Read, J., Pirrung, M., Stryer, L., Lu, A., and Solas, D. 1991. Light-directed, spatioally addressable parallel chemical synthesis. *Science 251*, 767–783.

[HBC] HBC. `http://www.columbia.edu/~xy56/project.htm`.

[Hedenfalk et al.2001] Hedenfalk, I., Duggan, D., Y, Y. C., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Kallioniemi, O. P., B, B. W., Borg, A., and Trent, J. 2001. Gene-expression profiles in hereditary breast cancer. *The New England Journal of Medcine 344,* 8, 539–548.

[Jaynes1957] Jaynes, E. T. 1957. Information theory and statistical mechanics. *Physical Review 106,* 4 (May), 620–630.

[Khan et al.2001] Khan, J., Wei, J., Ringner, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. 2001. Classificatiion and diagnostic prediction of cancers using expression profiling and artificial neural networks. *Nature Medicine 7,* 6, 673–679.

[Kiefer1959] Kiefer, J. 1959. Optimum experimental designs. *J. R. Statist. Sco. B 21*, 272–319.

[Kohavi and John1997] Kohavi, R. and John, G. H. 1997. Wrappers for feature subset selection. *Artificial Intelligence 97,* 1-2, 273–324.

[Langley1994] Langley, P. 1994. Selection of relevant features in machine learning. In *AAAI Fall Symposium on Relevance*. 140–144.

[Li et al.2004] Li, T., Zhang, C., and Ogihara, M. 2004. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics 20,* 15, 2429–2437.

[LYM] LYM. `http://genome-www.stanford.edu/lymphoma`.

[MLL] MLL. `http://research.dfci.harvard.edu/korsmeyer/MLL.htm`.

[Marko and Igor2003] Marko, R. and Igor, K. 2003. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning Journal*, 23–69.

[NCI60] NCI60. `http://genome-www.stanford.edu/nci60/`.

[Ooi and Tan2003] Ooi, C. and Tan, P. 2003. Genetic algorithms applied to multi-class prediction for the a nalysis of gene expression data. *Bioinformatics 19,*, 37–44.

[Peng et al.2005] Peng, H., Long, F., and Ding, C. 2005. Feature selection based on mutual information: Criteria of max-depe ndency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Analysis and Machine Intelligence 27*.

[Petersen and Pedersen2006] Petersen, K. B. and Pedersen, M. S. 2006. The matrix cookbook. Version 20051003.

[Ramaswamy et al.2001] RAMASWAMY, S., TAMAYO, P., RIFKIN, R., MUKHERJEE, S., YEANG, C.-H., ANGELO, M., LADD, C., REICH, M., LATULIPPE, E., MESIROV, J. P., POGGIO, T., GERALD, W., LODA, M., LANDER, E. S., AND R.GOLUB, T. 2001. Multiclass cancer diagnosis using tumor gene expression signatures. *98,* 26, 15149–15154.

[Ross et al.2000] ROSS, D. T., SCHERF, U., EISEN, M. B., PEROU, C. M., REES, C., SPELLMAND, P., IYER, V., JEFFREY, S. S., VAN DE RIJN, M., WALTHAM, M., PERGAMENSCHIKOV, A., LEE, J. C. F., LASHKARI, D., SHALON, D., MYERS, T. G., WEINSTEIN, J. N., BOTSTEIN, D., AND BROWN, M. P. O. 2000. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics 24*, 227–235.

[SRBCT] SRBCT. `http://research.nhgri.nih.gov/microarray/Supplement/`.

[Stein1975] STEIN, C. 1975. Estimation of a covariance matrix. In *Rietz Lecture, 39th IMS Annual Meeting*. Atlanta, Georgia.

[Su et al.2003] SU, Y., MURALI, T. M., PAVLOVIC, V., AND KASIF, S. 2003. Rankgene: Identification of diagnostic genes based on expression data. *Bioinformatics*. The program can be downloaded from `http://genomics10.bu.edu/yangsu/rankgene/`.

[Xing et al.2001] XING, E. P., JORDAN, M. I., AND KARP, R. M. 2001. Feature selection for high-dimensional genomic microarray data. In *Proc. 18th International Conf. on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 601–608.

[Yeoh et al.2002] YEOH, E.-J., ROSS, M. E., SHURTLEFF, S. A., WILLIAMS, W. K., PATEL, D., MAHROUZ, R., BEHM, F. G., RAIMONDI, S. C., RELLING, M. V., PATEL, A., CHENG, C., CAMPANA, D., WILKINS, D., ZHOU, X., LI, J. ., LIU, H., PUI, C.-H., EVANS, W. E., NAEVE, C., WONG, L., AND DOWNING, J. R. 2002. Classification, subtype diswcovery, and prediction of outcome in ped iatric lymphoblastic leukemia by gene expression profiling. *Cancer Cell 1,* 2, 133–143.

[Yeung et al.2001] YEUNG, K. Y., FRALEY, C., MURUA, A., RAFTERY, A. E., AND RUZZO, W. L. 2001. Model-based clustering and data transformations for gene expression data. *Bioinformatics 17,* 10, 977–987.

[Yu et al.2006] YU, K., BI, J., AND TRESP, V. 2006. Active learning via transductive experimental design. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*. ACM Press, New York, NY, USA, 1081–1088.

[Yu et al.2004] YU, L., LIU, H., AND TRESP, V. 2004. Redundancy based feature selection for microarray data. In *KDD '04: Proceedings of the 10th international conference on Knowledge discovery and data mining*. ACM Press, Seattle, WA, USA.