

# Evaluation of Gene Selection Using Support Vector Machine Recursive Feature Elimination

A report presented in fulfillment of internship  
requirements of the CBS PSM Degree

Committee:

Dr. Rosemary Renaut<sup>1</sup>

Professor Department of Mathematics and Statistics,  
Director Computational Biosciences PSM  
Arizona State University

Dr. Kenneth Hooper<sup>2</sup>

Professor Department of Plant Biology, Life Science  
Co-director Computational Biosciences PSM  
Arizona State University

Dr. Bradford Kirkman-Liff<sup>3</sup>

Professor School of Health Administration and Policy,  
College of Business  
Arizona State University

Advisor: Dr. Adrienne C. Scheck<sup>4</sup>

Senior Staff Scientist  
Neuro-Oncology Research  
Barrow Neurological Institute

Student: John Anh Huynh<sup>5</sup>

Computational Bioscience PSM  
Arizona State University

May 26, 2004

- <sup>1</sup>email: [renaut@asu.edu](mailto:renaut@asu.edu)  
<sup>2</sup>email: [khoober@asu.edu](mailto:khoober@asu.edu)  
<sup>3</sup>email: [bradford.kirkman.liff@asu.edu](mailto:bradford.kirkman.liff@asu.edu)  
<sup>4</sup>email: [AScheck@chw.edu](mailto:AScheck@chw.edu)  
<sup>5</sup>email: [jahuynh@asu.edu](mailto:jahuynh@asu.edu)

## Abstract

One advantage of the microarray technique is that it allows scientists to explore the expression of thousands of genes in a single expression microarray. This creates a set of high dimensional data with more than fifty thousand features. The price of a microarray is still high, thus we may not have many samples in a single microarray experiment. So this small data set with high dimensional feature space requires a special technique for its analysis. One common requirement is a small subset of genes so that we can build the most accurate classifier for predicting in-coming data.

Working in high dimensional space requires excessive computational cost. The main problem of small data set is overfitting. Feature ranking reduces the computational cost and overfitting. Guyon et al.[7] proved that support vector recursive feature elimination performs better in gene selection compared to other existing methods. Recursive feature elimination is a top-down wrapper method using weight vector as feature ranking criterion. In this experiment, we explore the interaction of the support vector machine correctness with the number of genes and the rate of feature elimination.

The result shows that there is no significance in the rate of feature elimination when the surviving subset is large, thus we can eliminate the genes quickly in feature selection. The smaller data sets still need more investigation.

# 1 Introduction

Meningiomas account for about 20% of all primary intracranial tumors. One tenth of meningiomas are located in spine. The peak incidence is 45-55 years old. It is common in women [12]. Most of them are benign (WHO grade I) and atypical (WHO grade II). The anaplastic (WHO grade III) meningiomas are more aggressive and can infiltrate bone, muscle, dura, or brain tissue. If the aggressive phenotypes of these meningiomas can be recognized early in clinical diagnosis, then physicians can apply appropriate therapies to improve the patient prognosis. Histopathology alone is not always sufficient to identify patients with a poorer prognosis. The molecular profile of these tumors may provide additional prognostic information. In order to provide information on the molecular basis of meningioma prognosis, one project in Dr. Adrienne C. Scheck's laboratory, Barrow Neurological Institute, is to correlate molecular genetics, biochemical profiles, and meningioma behavior using Affymetrix oligonucleotide microarray, FISH (fluorescence *in situ* hybridization), and NMR (nuclear magnetic resonance) spectroscopy.

Before we discuss further we would like to define some technical terminology used differently among scientists in different fields. A *data set* or *sample* is a set of data in which the rows are *examples*, and the columns are *attributes* while the raw data of microarray usually is reversed. In this experiment *attribute*, *feature*, and *gene* are equivalent. The term "sample" that is usually used in some biological documents is an "example" or a "microarray slide" in this experiment. A *small data set* is a set of data which has a small example size. The **bold** typeface indicates a vector.

# 2 Microarray Technique

Differentiation is the process whereby the complexity of the adult organism arises from the simplicity of the zygote. Every normal cell in the human body has a nucleus which contains the same amount of DNA. Every tissue has specific structures and functions due to the unique proteins of that tissue. The particular sequences of amino acids in each protein are decided by the gene sequence. There are different gene expression patterns for different tissues.

The genetic information flows from DNA to mRNA to protein. Transcription is the encoding process from DNA into mRNA in nucleus. The introns are sequences which do not encode for polypeptides and are removed in the process of transcription. The mRNA sequences are from the exons. Translation is the decoding process of genetic information from mRNA to protein by ribosomes and tRNA and takes place in the cytoplasm. There may be post-translation modifications of proteins before they become fully functional.

Microarray is a technique to explore the gene expression of tissue or cells. Messenger RNA (mRNA) is extracted from the target cells. The mRNA is used to synthesize to target cDNA (complementary deoxyribose nucleic acid) using reverse transcriptase reaction. The target cDNA is labeled by a dye. For example cy3 and cy5 are fluorescence dyes which light up under a specific light-band laser beam whenever the target binds to its complementary DNA strand. In cDNA microarray technique, the probes are single strand cDNAs which are built separately and then attached to the array surface. These probes vary in lengths (50-500) on

different platforms. In Affymetrix oligo-DNA microarray, the probes are 25 oligonucleotides which are specific for preselected genes. The Affymetrix microarray is manufactured using *in-situ* synthesis. The probes are attached to the array surface by the photolithographic fabrication. This technique produces high density arrays with a limited probe length. If the target is complementary to the probe, it binds with the corresponding probe during hybridization, and stays on the slide while the extra target cDNA is washed away [4].

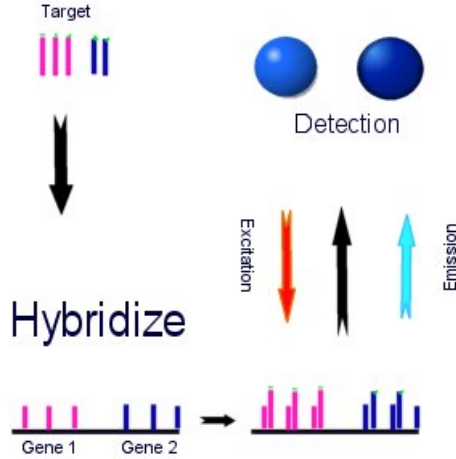


Figure 1: Microarray Technique Flow Chart. The mRNA is extracted from experimental cells. The mRNA is used to synthesize target cDNA using reverse transcriptase reaction. The target cDNA is labeled with an appropriate dye. The probes are single strand cDNAs or oligonucleotides of preselected genes that are attached to the surface of the microarray slide. During hybridization, the target binds to its complementary probe and stays on the slide while the extra target is washed away. The specific laser beams of the scanner are used to reveal the images of positive spots.

When the scanner beams a particular laser, the spots containing hybridized targets and probes light up. We say that the gene is present or expressed. Refer to Figure 1 for more details. Appropriate controls are needed for data analysis. In cDNA microarray technique, the intensity usually is the difference of foreground intensity median from the background intensity median. The Affymetrix HG-U133 Plus 2.0 is an oligonucleotide microarray, each feature is designed as a probe set of 11 probe pairs [8]. Each probe pair includes a perfect match (PM) probe and a mismatch probe (MM) which contains only one different nucleotide from the PM (Figure 2). The scanner software usually performs the spot preprocessing to give the *quantitative* (intensity values) and the *call*. The intensity  $I$  usually is the average of difference of PM and MM in the probe set [4],

$$I = \frac{\sum_{i=1}^n (PM_i - MM_i)}{n},$$

where  $i = 1, 2, \dots, n$ , and  $n$  is number of probes in a probe set which is 11 in this case [8]. The call can be A (absent), M (marginally present) or P (present). The details about the calculation of intensity and the call can be found in [9] and [4].

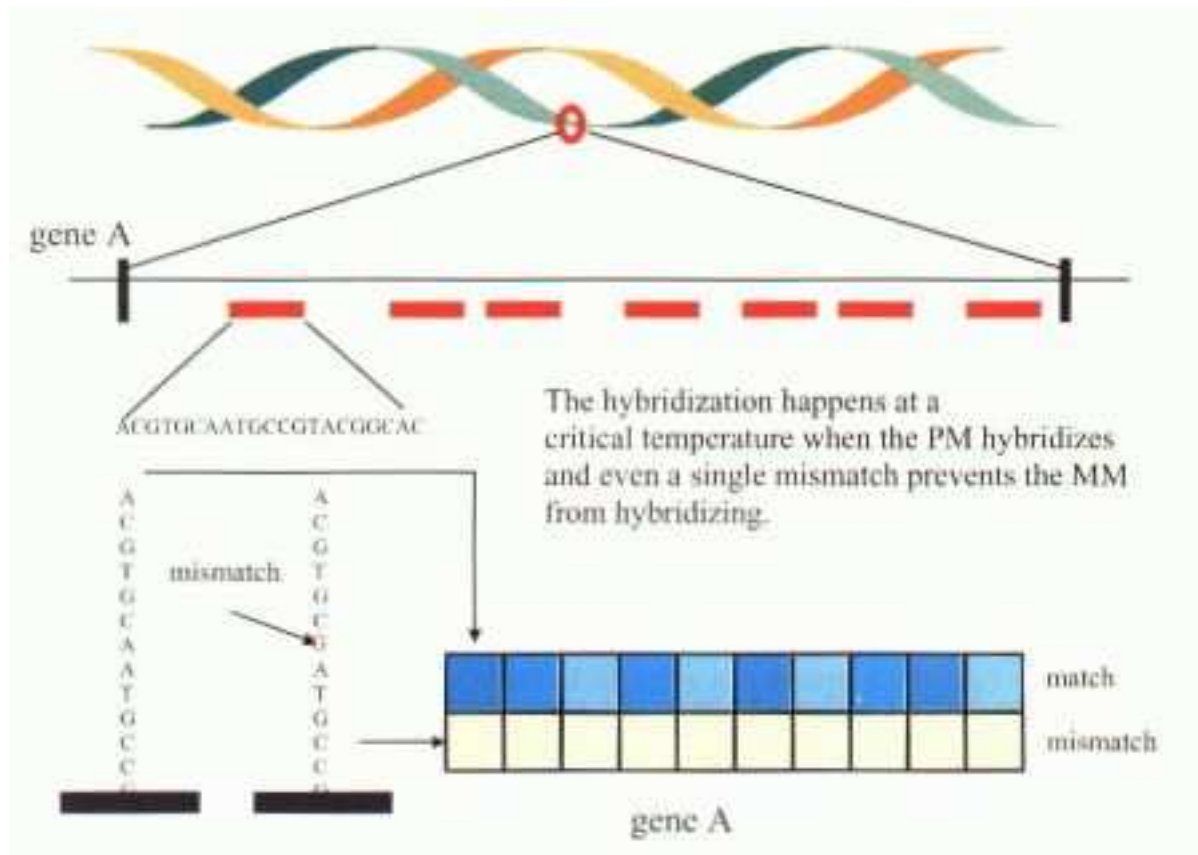


Figure 2: The Probe Set. There are 11 probe pairs in each feature (gene). Each probe pair includes 2 probes: perfect match (PM) and mismatch (MM). The picture is from [4].

One common use of expression microarray is to compare two cell lines: control and treatment. Thus the microarray data has binary class labels. The Affymetrix HG-U133 Plus 2.0 microarray includes 54,675 genes. Microarray technique is expensive, so it is difficult to have many examples in one experiment.

Microarray data owns two special characteristics: high dimensional feature space and small example size. A new technique is required to analyze this special database. In order to predict the classification of microarray data from a new patient, we need a low error rate classifier as a diagnostic tool. Working in high dimensional space requires excessive computational cost. The risk of over-fitting is high in training process with a small data set. Guyon et al. proved that support vector machine recursive feature elimination has a good performance in gene selection [7]. We would like to apply this method to build a support vector machine classifier to predict meningioma grade. Guyon et al. [7] also mentions the ability to remove a subset of genes in each loop of the recursive feature selection. Is there any difference in reducing one gene at a time and removing a group of genes at a time? What is the optimal gene subset size? In order to answer these questions, we would like to examine the effect of the proficiency of the support vector machine in the rate of feature elimination and the optimal subset size.

### 3 Inducer Problem

Our problem can be described mathematically in (1). Given data points  $\{\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_m\}$ ,  $\mathbf{x}_i \in \mathbb{R}^n$  with their target labels  $\{y_1, \dots, y_i, \dots, y_m\}$ ,  $y_i \in \{-1, 1\}$ , in which  $\mathbf{x}_i$  are row vectors, we would like to train a classifier  $f$ , a mapping function from the space of feature values to the set of class values so that

$$\begin{aligned} f : \mathbb{R}^n &\mapsto \{-1, 1\} \\ \mathbf{x} &\rightarrow y. \end{aligned} \tag{1}$$

### 4 Feature Selection

#### 4.1 Benefit of Feature Ranking

The main problem of small data training is the over-fitting. Guyon et al. [7] introduce a gene selection method as a feature selection. There are two benefits of feature reduction: overcoming the risk of *over-fitting* and reducing computational expense. Over-fitting happens when we have a huge dimensional feature space and a small number of examples in data set. In such situations, the classifier performs well on the training set and acts poorly on the test set [7]. Please refer to Section 4.3 for more details on the requirement of data set size.

#### 4.2 Optimal Feature Subset

The typical goal of classifier training is to find the most accurate classifier that performs well on both training set and test set. The feature selection is a search algorithm in feature space to find an optimal subset of features. Ron Kohavi et al. [10] defines the optimal feature subset as the following.

**Definition 4.1** *Given an inducer  $\mathcal{I}$ , and a data set  $\mathcal{D}$  with feature  $\{\mathcal{X}_1, \dots, \mathcal{X}_i, \dots, \mathcal{X}_n\}$ , from a distribution  $\mathcal{D}$  over the labelled instance space. An optimal feature subset,  $\mathcal{X}_{opt}$ , is a subset of the features such that the accuracy of the induced classifier  $\mathcal{C} = \mathcal{I}(\mathcal{D})$  is maximal.*

Thus the selected feature subset should be evaluated by the **correctness** of the evaluation inducer.

#### 4.3 Feature Characteristics

There are variant definitions of relevant feature that belongs to what we would like to compare. A relevant feature is either a weakly relevant or a strong relevant; otherwise it is irrelevant. In our inducer problem, we have  $n$  features and  $m$  examples. Each feature  $x_i$ ,  $i = 1, 2, \dots, n$  has a domain  $F_i$ . The values of the feature can be binary, nominal, or continuous. An example is a data point in the instance space  $F_1 \times F_2 \times \dots \times F_i \times \dots \times F_n$ . In order to learn well in a given training data set  $S$ , the examples must be diverse enough in the instance space; otherwise the learning algorithm  $f$  can not learn anything. This is called *over-fitting*. Avrim L. Blum [1] defines the “relevant feature to the target concept”:

**Definition 4.2 (*Relevant to the target*)** A feature  $x_i$  is relevant to a target concept  $c$  if there exists a pair of examples  $A$  and  $B$  in the instance space such that  $A$  and  $B$  differ only in their assignment to  $x_i$  and  $c(A) \neq c(B)$ .

In other words, the feature  $x_i$  is relevant if there are some examples in the instance space that helps the inducer  $f$  learn well in mapping from the feature space into the target space. Note that the training process scopes in the training set  $S$ ; there is a probability of the relevance of a feature. Avrim L. Blum [1] also defines the strongly relevant features and weakly relevant features as in the following.

**Definition 4.3 (*Strongly Relevant to the sample/distribution*)** A feature  $x_i$  is strongly relevant to sample  $S$  if there exist examples  $A$  and  $B$  in  $S$  that differ only in their assignment to  $x_i$  and have different labels (or have different distributions of labels if they appear in  $S$  multiple times). Similarly,  $x_i$  is strongly relevant to target  $c$  and distribution  $D$  if there exist examples  $A$  and  $B$  having non-zero probability over  $D$  that differ only in their assignment to  $x_i$  and satisfy  $c(A) \neq c(B)$ .

**Definition 4.4 (*Weakly Relevant to the sample/distribution*)** A feature  $x_i$  is weakly relevant to sample  $S$  (or to target  $c$  and distribution  $D$ ) if it is possible to remove a subset of the features so that  $x_i$  becomes strongly relevant.

The above definitions state that we can remove weakly relevant features and keep the strongly relevant features to improve the goodness of the inducer. The relevance of a feature is also a complexity measure [1].

**Definition 4.5 (*Relevance as a complexity measure*)** Given a sample of data  $S$  and a set of concepts  $C$ , let  $r(S, C)$  be the number of features relevant using Definition 4.2 to a concept in  $C$  that, out of all those whose error over  $S$  is least, has the fewest relevant features.

This definition requires the smallest relevant feature subset such that the inducer  $f$  still acts well in the training process to minimize the complexity. This definition explains the reason to eliminate the features for which the weights have least absolute values. Blum et al. [1] also mentions the possibility to incrementally add or remove features during the feature ranking process.

**Definition 4.6 (*Incremental usefulness*)** Given a sample of data  $S$ , a learning algorithm  $L$ , and a feature set  $\mathcal{A}$ , feature  $x_i$  is incrementally useful to  $L$  with respect to  $\mathcal{A}$  if the accuracy of the hypothesis that  $L$  produces using the feature set  $\{x_i\} \cup \mathcal{A}$  is better than the accuracy achieved using just the feature set  $\mathcal{A}$ .

The purpose of feature selection is to remove irrelevant and weakly relevant features so that data set becomes strongly relevant. For more details about feature selection, refer to [10].



Figure 3: Filter method uses the characteristics of data to select features in preprocessing phase.

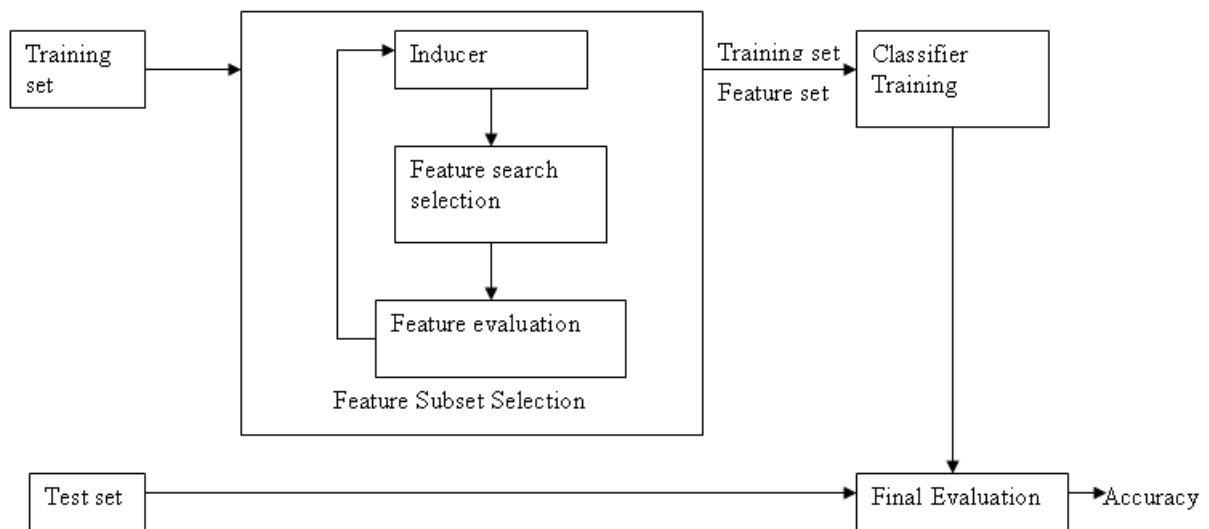


Figure 4: Wrapper method uses an inducer to select features.

#### 4.4 Feature Selection Methods

Generally feature selection methods can be classified into two categories: filter and wrapper. The filter model uses the characteristics of data to select features in a preprocessing phase (Figure 3). The simplest filter method is to evaluate each feature individually based on the correlation with the target function and then select the  $k$  features with the highest value. The selected feature subset is tested using holdout test set.

The wrapper model relies on an inducer as a “scale” to weigh the features and then select features (Figure 4) based on that cost. It finds features better suited to predetermined learning algorithm, but it is more computationally expensive than the filter model [11].

## 4.5 Feature Ranking Criteria

### 4.5.1 Correlation Coefficient

The feature ranking criteria can be a correlation coefficient method or a weight method. One well known measure is linear correlation coefficient  $\rho$  for a given pair of the variables  $(X, Y)$

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_i x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

where  $\bar{x}$  is the mean of  $X$ , and  $\bar{y}$  is the mean of  $Y$  [11].

Golub's coefficient is defined as

$$\gamma_i = \frac{\mu_i(+)-\mu_i(-)}{\sigma_i(+)+\sigma_i(-)},$$

where  $\mu_i$  and  $\sigma_i$  are the mean and standard deviation of the feature  $i$  for all examples of class (+) or class (-),  $i = 1, 2, \dots, n$  [5].

### 4.5.2 Weight

The weight method is derived from the Taylor series. Feature selection actually is an induction algorithm that performs in the feature space to classify the features into relevant and irrelevant ones. Machine learning is a problem of minimizing of the error function under certain constraints. The objective function  $E$  is the expected value of the error which is the cost function  $E$  in training process. To find the minimum of error we would like to find the change of the cost function  $dE$  and then set it to zero. The approximate  $dE$  can be calculated by the Taylor series, and dropping the first order term at the optimum point according to [3],

$$dE(i) = (1/2) \frac{\partial^2 E}{\partial w_i^2} (dw_i)^2.$$

The change in weight  $dw_i$  is  $w_i$  which corresponds to removing feature  $i$  [7]. Note that in this case we use the change of error function  $dE$  instead for the magnitude of the weights as a ranking criterion. In the case of discriminant functions, the cost function is a quadratic function of the vector of weights  $\mathbf{w}$ . In the case of the linear least square machines and neural network single layer with delta rule, the cost function is  $E = \sum \|\mathbf{x}\mathbf{w} - y\|^2$ . In the case of the linear support vector machine, the cost function is  $E = (1/2)\|\mathbf{w}\|^2$  under certain constraints. Therefore we can use  $w_i^2$  as a feature ranking criterion.

## 5 Recursive Feature Elimination

The recursive feature elimination algorithm is a feature selection algorithm. Since it is a top-down (backward) elimination algorithm, it starts to train the inducer with the whole feature space and then gradually removes the features until the stop condition is reached. The weight vector from the inducer is used as the feature ranking criterion. The recursive feature elimination removes the feature with its minimum weight (Figure 5). For more

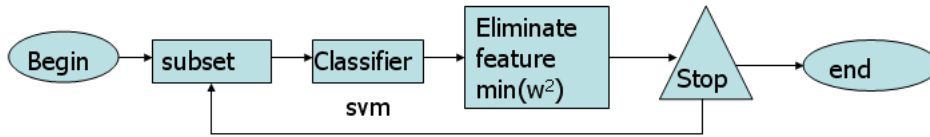


Figure 5: Recursive Feature Elimination Algorithm

details, refer to [7]. This algorithm eliminates one feature at every loop. However, Guyon et al. also mentioned that it can be modified to eliminate a subset every loop.

In this experiment we focus on number of features that can be eliminated in every loop, called “rate of elimination”, and the minimum number of strongly relevant features that can be chosen to build an optimal classifier, called “surviving subset.” When we perform feature selection, we choose the smaller subset nested inside the prior one. Every loop we eliminate a set of features. The final optimal subset is our destiny, “surviving subset” as in Figure 6. Our goal is to find the interaction between the rate of elimination, the subset size eliminated each step, and the surviving subset.

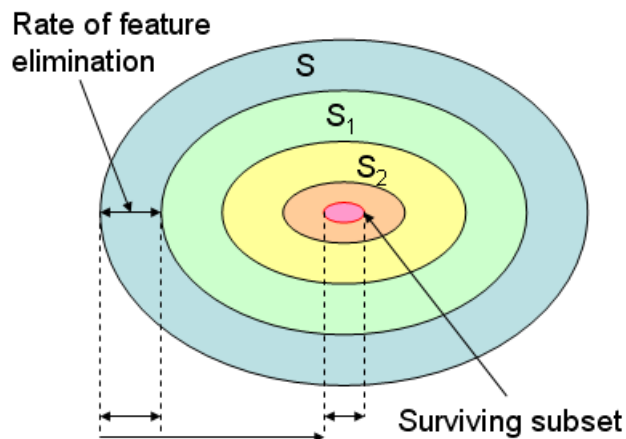


Figure 6: Nested subsets: rate of elimination and surviving subset

## 6 Support Vector Machine

Guyon et al. state that support vector machine is one of the best classifiers for feature selection of microarray data [7].

## Support Vector Machine: Separable Case and Non-Separable Case

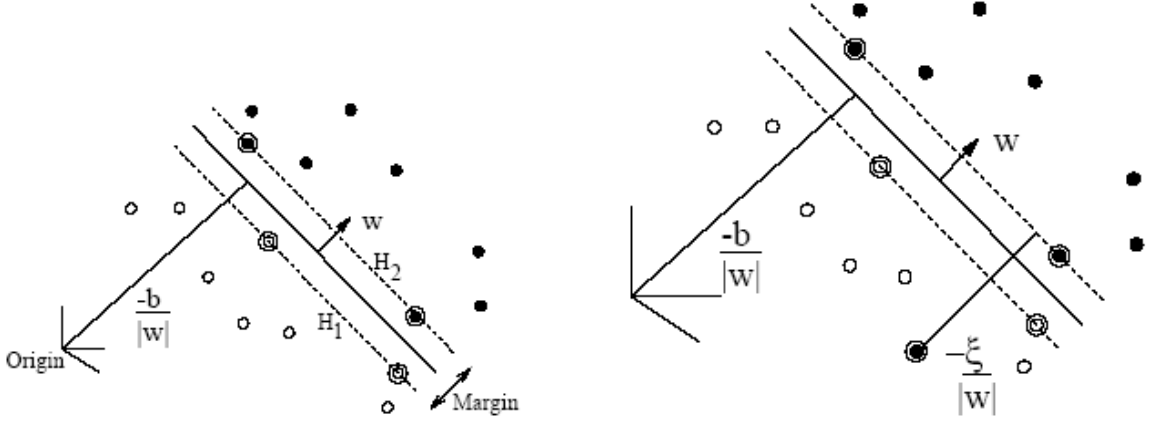


Figure 7: Separable Case SVM. Maximizing the margin is the goal of SVM. The optimal decision boundary is the middle hyperplane of  $H_1$  and  $H_2$ . The points on the marginal hyperplane  $H_1$  and  $H_2$  are support vectors and play an important role in defining the decision boundary [2].

Figure 8: Non-Separable Case SVM. Slacks  $\xi_i$  are introduced to adjust the choosing of support vectors when needed [2].

### 6.1 Linear Separable Case

Now we modify our inducer problem to understand support vector machine learning.

$$\begin{cases} \mathbf{x}_i \mathbf{w} + b \geq +1 & \text{for } y_i = +1, \\ \mathbf{x}_i \mathbf{w} + b \leq -1 & \text{for } y_i = -1, \end{cases}$$

where  $\mathbf{w}$  is the weight vector, and  $b$  is a bias value as in Figure 7.

In this case, the classifier is trained on the separable data. Suppose we have some “separating hyperplanes” that separate data points into class  $-1$  and class  $+1$ . Denote by  $H_1$  the hyperplane for which  $f(\mathbf{x}) = \mathbf{x} \mathbf{w} + b = +1$  and  $H_2$  by the hyperplane  $f(\mathbf{x}) = \mathbf{x} \mathbf{w} + b = -1$ . The shortest distance from  $H_1$  to  $H_2$  is the *margin*. The ideal of the support vector machine is to look for the **maximum margin**. The decision boundary is the middle hyperplane of the margin. Note that the weight vector is perpendicular to the decision boundary. Thus the distance from the origin to the decision boundary is  $|b|/||\mathbf{w}||$ , in which  $||\mathbf{w}||$  is the Euclidean norm of  $\mathbf{w}$ . The distance from the origin to  $H_1$  is  $|1 - b|/||\mathbf{w}||$ ; and the distance from the origin to  $H_2$  is  $|-1 - b|/||\mathbf{w}||$ . It is easy to see the margin which is  $2/||\mathbf{w}||$ .  $H_1$  and  $H_2$  are both perpendicular to  $\mathbf{w}$ , so they are parallel. There is no data point between them. Thus we can maximize the margin by minimizing  $||\mathbf{w}||^2$ . The inducer problem becomes the optimization problem with respect to weight vector.

$$\begin{cases} \text{minimize} & ||\mathbf{w}||^2 \\ \text{subject to} & y_i(\mathbf{x}_i \mathbf{w} + b) - 1 \geq 0 \quad \forall i = 1, 2, \dots, m. \end{cases}$$

This is quadratic with respect to weight vector  $\mathbf{w}$ . So we introduce Lagrangian multipliers, and then apply the Karush-Kuhn-Tucker (KKT) conditions, please refer to [2] for more details.

Note that by maximizing the margin, the support vector machine maximizes the ability of separating data points in the input space. In other words, the decision boundary of support vector machine is an *optimal* “separating hyperplane”. The data points on  $H_1$  and  $H_2$  are *support vectors*. They are critical in training the decision boundary of the support vector machine.

## 6.2 Linear Non-Separable Case

When we train the support vector machine on non-separable data, the above algorithm does not always find the feasible solution especially when the dual Lagrangian objective function is very large. To overcome this problem, *slack* variables  $\xi_i$  are introduced to adjust the choice of support vectors when necessary as in Figure 8. The inducer problem becomes

$$\begin{cases} \mathbf{x}_i \mathbf{w} + b & \geq +1 - \xi_i & \text{for } y_i = +1, \\ \mathbf{x}_i \mathbf{w} + b & \leq -1 + \xi_i & \text{for } y_i = -1, \\ \xi_i & \geq C & \forall i = 1, 2, \dots, m. \end{cases}$$

Even though the optimization problem is still the same, the only difference is that the Lagrangian multipliers now have a constraint  $C$ , which is chosen by the user [2]. For microarray data, Guyon et al. set  $C = 100$  [7].

## 6.3 Non-Linear Case

The decision boundary of support vector machine can be a non-linear function. We map our data from  $\mathbb{R}^n$  to some Euclidean space  $\mathcal{H}$  using mapping function  $\Phi$ ,

$$\Phi : \mathbb{R}^n \mapsto \mathcal{H}.$$

Some authors strictly require an infinite dimensional space in which the Euclidean space is a special case. Let  $\mathcal{L}$  be the data space  $\mathcal{L} = \mathbb{R}^n$ . From now on, the training and testing processes work on our data through dot products. In training algorithm we use a *kernel function*  $K$ ,

$$K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j).$$

Since our data and weight vector are in hyperspace  $\mathcal{H}$ , the training and testing processes must take place in space  $\mathcal{H}$ . This means that we just simply replace  $\mathbf{x}_i \cdot \mathbf{x}_j$  every where in the training and testing algorithm with  $K(\mathbf{x}_i, \mathbf{x}_j)$ . The dimension of space  $\mathcal{H}$  is usually much higher than the dimension of space  $\mathcal{L}$ . Assuming the nonlinear case, there is no vector in  $\mathcal{L}$  that maps to  $\mathbf{w}$  in space  $\mathcal{H}$  using mapping function  $\Phi$ .

There are two important characteristics of support vector machine: maximal margin and low computational cost. Support vector machine is a *supervised* training algorithm which maps data in  $\mathbb{R}^n = \mathcal{L}$  onto a high dimensional Euclidean space  $\mathcal{H}$  using a kernel function.

The training and testing processes occur in high dimensional space  $\mathcal{H}$  in the non-linear case. The decision boundary of the classifier is the middle hyperplane of the maximal margin.

The weight vector is a function of a small subset of data points on the marginal hyperplanes, called “support vectors”. It shows good performance on the high dimensional data due to its low computational complexity. The kernel function requires  $\mathcal{O}(n)$  in dot product case. In the training process, the worst complexity is  $\mathcal{O}(n_{sv}^3 + n_{sv}^2 m + n_{sv} n m)$  and the best case is  $\mathcal{O}(n m^2)$ . In the test phase, the complexity is  $\mathcal{O}(n)$  in dot product case [2]. In which  $m$  is number of examples,  $n_{sv}$  is number of support vectors, and  $n$  is feature dimension of data. For more details on support vector machine, see [2]. In this experiment we limit to the linear support vector machine with scalar dot product kernel. The microarray has small example number  $m$ , and large feature dimension  $n$ . The number of support vector  $n_{sv}$  is usually small, since support vectors are a subset of example set  $m$ . The complexity is linear to the feature dimension  $n$ , which can be 50,000 in microarray data. The low complexity of support vector machine proves a strong characteristic to apply it to microarray data.

## 7 Experimental Design

### 7.1 Evaluation Parameter

It is possible to use support vector machine to perform feature selection then use the feature subset result to build a different kind of classifier; however, Guyon et al.’s research shows that the trained classifier has better performance if it is the same kind that is used in feature selection. We use support vector machine in the feature selection module and then use the resulting feature subset to build a new support vector machine for evaluation purposes. The **success rate** of the test set which is returned by this evaluation SVM is the evaluation cost of the resulting feature subset, Definition 4.1.

### 7.2 Evaluation Methods

#### 7.2.1 Independent Test

To avoid bias in the training set and test set, we use stratified independent test. In the stratified procedure, we randomly sample the training set and test set. This ensures that both training set and test set represent the classes. Usually the training set accounts for two-third of data and the test set is one-third. The independent test is available when our data is large enough. In microarray data, we may not have large sample size; thus, cross-validation is a good method for microarray data.

#### 7.2.2 Cross-Validation Test

The holdout methods provide numerous methods for evaluation. In  $n_t$ -cross-validation method, the data is randomly divided into  $n_t$  equal buckets. One bucket is used for testing purpose while the rest,  $n_t - 1$  buckets, are for training. The next loop, the test bucket will be swapped to the next bucket in the row of  $n_t$  buckets. The process repeats  $n_t$  times. Every

loop returns an error rate. The error rate in  $n_t$  times can be considered as  $n_t$  replicates for a statistical test. An alternate way is to return the mean and standard deviation of error rate.

In  $k$ - $n_t$ -cross validation, the inner loop does exactly the same as in  $n_t$ -cross-validation. The outer loop is done  $k$  times to shuffle data. The result of this test can be used as  $k \times n_t$  replicates.

The number  $n_t$  of cross validation test commonly is 3, 5, 10, and 20. In our experiment, we apply  $k = 3$ , and  $n_t = 3$ ; therefore, we have 9 replicates for z-test and ANOVA.

### 7.2.3 Leave-One-Out

Leave-One-Out method is one special  $n_t$ -cross-validation where  $n_t$  is the number of data points. The test set size is 1 while the training set has  $n_t - 1$  points. The test repeats  $n_t$  times by alternating the test set. It returns a value of 0 or 1 every loop and the result will be the average of these values over the  $n_t$ . There is no replicate for a statistical test.

The test set should be never seen in both feature selection and training processes; otherwise the success rate of the test is unexpectedly high [10].

## 7.3 Optimal Subset Size Criteria

The criteria to choose the optimal subset size are:

- the smallest subset,
- the maximum accuracy of the evaluation test,
- independent elimination rate among the groups of surviving subsets from all feature subset to the optimal subset.

## 7.4 Full Two Factorial Experimental Design

This is a full two factorial experimental design. One factor is the rate of elimination and another is the surviving subset. The effective model of our experiment design is

$$T_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}.$$

In which the  $T_{ijk}$  measure the goodness of SVM,  $\mu$  is the overall mean effect,  $\tau_i$  are the effect of the  $i$ th level of the row factor which is the rate of feature elimination;  $\beta_j$  are the effect of the  $j$ th level of column factor which is the subset size;  $\epsilon_{ijk}$  are the error. Denote by  $a$  the number of rate of feature elimination,  $b$  by the number of subset sizes, and  $c$  by the number of replicates in each cell, then  $i = 1, 2, \dots, a$ ,  $j = 1, 2, \dots, b$ , and  $k = 1, 2, \dots, c$ .

Data is randomly divided into training set and test set. The training set is used for feature selection algorithm and the independent test with selected feature subset. The test set is used for the independent test and 3-3-cross-validation test. We implement the algorithm using Matlab 6.5R13 as in Figure 9.

The rates of feature elimination are determined by dropping a subset of size from 10 to 1000, step 10,  $r = 10 : 10 : 1000$ . For each rate of feature elimination, the support vector machine is trained using the training set with all features at the beginning. In each loop

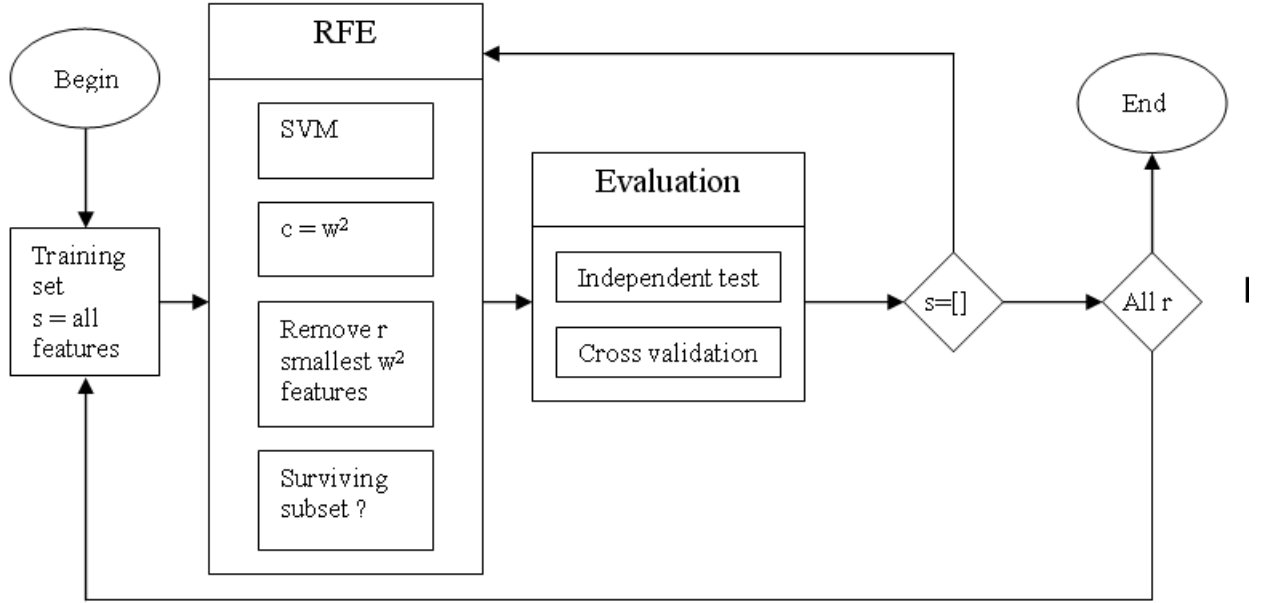


Figure 9: Flow Chart of Recursive Feature Selection and Evaluation.

of recursive feature elimination (RFE), we reduce a certain number of genes. Every time the subset size reaches the observed subset size, we evaluate the feature subset using an independent test and 3-3-cross-validation test. The first surviving subset is all features, and the next is  $2^z$  where  $z = \text{floor}(\log_2(n))$ . As a sequence, the observed subsets are

$$s = \begin{cases} [n \ 2^{[z:-1:0]}] & \text{for } 2^z < n, \\ 2^{[z:-1:0]} & \text{for } 2^z = n. \end{cases}$$

In the evaluation module, the independent test and 3-3-cross validation results are recorded. Refer to Figure 9 for more details. ANOVA II is used to estimate the variances between the elimination rate factor and surviving subset factor. The optimal surviving subset is the most accurate and smallest subset that is not affected by the elimination rate.

## 8 Data

In microarray raw data, each attribute includes two columns. One is a real number which is the intensity. Another is the call. The call can be P (positive), M (marginal), and A (absent). In our experiment we do not use the call.

Gene expression intensity has negative and positive values. The range of the gene expression could be different from slide to slide, so in the preprocessing process we simply apply linear normalization to slide vectors and gene vectors and apply logarithm transformation.

## Scatter Plots

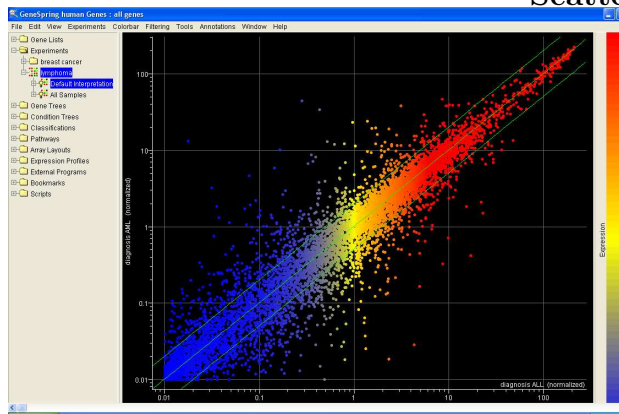


Figure 10: Lymphoma Data Scatter Plot. The horizontal axis is the logarithm base 2 of the average intensity of ALL. The vertical axis is the logarithm base 2 of the average intensity of AML. The points that are above the oblique line represent the ALL expression levels are higher than the AML expression levels. The two outer oblique lines represent the two-fold change between AML and ALL. The data distributes around the oblique line. The variance is high; however, there are no outliers. This plot was created using GeneSpring 6.2 [6].

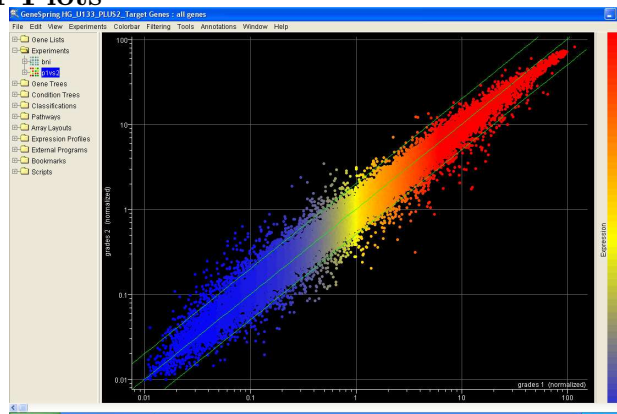


Figure 11: BNI Data Scatter Plot. The horizontal axis is the logarithm base 2 of the meningioma grade I average intensity. The vertical axis is the logarithm base 2 of the meningioma grade II average intensity. The points that are above the oblique line  $y = x$  show the expression levels of meningioma Grade I are higher than the expression levels of meningioma Grade II. The two outer oblique lines represent the two-fold change between Grade I and Grade II. The data distributes neatly around the oblique line. There are no outliers. This plot was created using GeneSpring 6.2 [6].

### 8.1 Lymphoma Evaluation Data

We evaluate the approach with data that has been used in literature [5] and [7], Amersham cDNA with 7,129 genes (Table 1). The test samples are blood and bone marrow.

The scatter plot is the plot of the ALL (acute lymphoblastic leukemia) intensities (horizontal axis) over the AML (acute myeloid leukemia) intensities both in logarithm base 2 scale. In the scatter plot (Figure 10), the data distributes around the oblique line. The variance is especially high; however, there are no outliers. The hierarchical tree is a clustering method in which the similar expression levels are clustered in the same node. By comparing the pattern of expression level, the conclusion about the quality of lymphoma microarrays can be made. The patterns of all slides are the same in the hierarchical tree (Figure 12). They are good.

### 8.2 BNI Meningioma Data

This data is Affymetrix HG-U133 Plus 2.0 oligo-DNA microarray with 54,675 genes. We randomly divide it into training set and test set as in Table 2.

The quality of this data is much better than the lymphoma data as in the scatter plot

## Hierarchical Trees

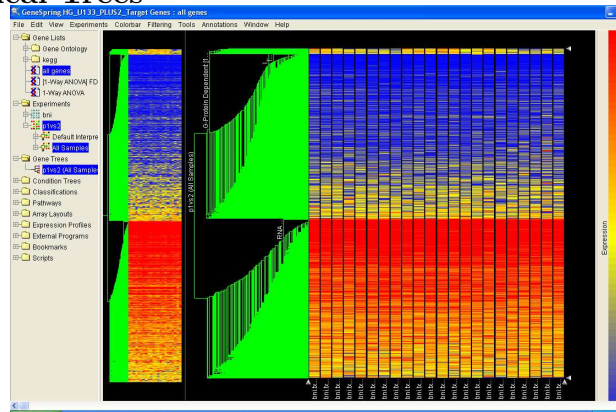
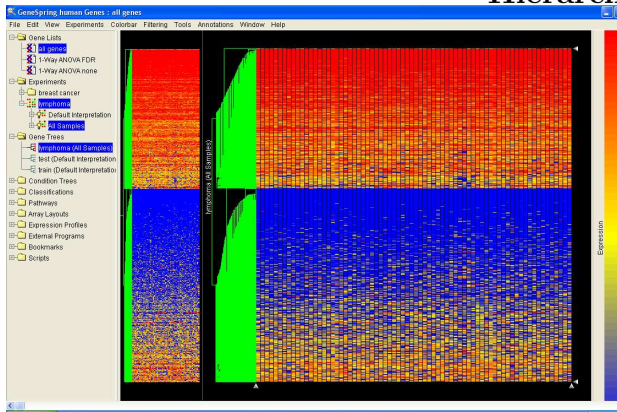


Figure 12: Lymphoma Data Hierarchical Trees. The horizontal axis shows lymphoma microarray slides. The same intensity genes among the slides are clustered in to one group. The patterns of the slides can be used to consider the quality of the lymphoma microarrays. The top half is the high intensity cluster. The bottom half is the low intensity cluster. The patterns of all examples are the same. These are good slides. This plot was obtained using GeneSpring 6.2 [6].

Figure 13: BNI Data Hierarchical Trees. The horizontal axis shows BNI microarray slides. The same intensity genes among the slides are clustered in to one group. The patterns of the slides can be used to consider the quality of the BNI meningioma microarrays. The top half is the high intensity cluster. The bottom half is the low intensity cluster. The patterns of all examples are the same. This plot was obtained using GeneSpring 6.2 [6].

(Figure 11) and the hierarchical tree plot (Figure 13). This data set is small. We run with 2 strategies:

- Consider as a large data set and do SVMRFE with exactly the same as lymphoma data set.
- Combine the training and test set to a large data set which we call master data. We provide the master data to SVMRFE and cross validation. In this case, the accuracy of both test will be unexpectedly high due to the affect of over-fitting. However, the cross-validation test is preferred, because it produces replicates for ANOVA.

## 9 Result

### 9.1 Lymphoma Data

In the plots,  $T_{suc}$  is the success rate of independent test.

$$T_{suc} = \frac{c_{test}}{m_{test}},$$

	Train	Test	Total
ALL	27	20	47
AML	11	14	25
Total	38	34	72

Table 1: Lymphoma Data. ALL is acute lymphoblastic leukemia, and AML is acute myeloid leukemia

	Train	Test	Total
Grade 1	11	5	16
Grade 2	4	2	6
Total	15	7	22

Table 2: BNI Meningioma Data.

where  $c_{test}$  is the correct number of the test result and  $m_{test}$  is the size of test set.  $Vsuc$  is the average of success rates of 3-3-cross-validation test.

$$Vsuc = \frac{\sum_{i=1}^k \sum_{j=1}^{n_t} c_{test}}{k n_t},$$

where  $k = 3$  and  $n_t = 3$  are the  $k$  and  $n_t$  times of  $k$ - $n_t$ -cross-validation test,  $i = 1, 2, \dots, k$ ,  $j = 1, 2, \dots, n_t$ ,  $c_{test}$  is the correct number of the test result, and  $m_{test}$  is the size of the test set. The 3-D plot of  $Tsuc$  over the elimination rate and the subset size is in Figure 14. This plot shows a smooth surface spreading the elimination rates from 10 to 1,000 and the subset size from 7,129 to less than 1,000. This flat surface suggests that there is no interaction between the elimination rate and the surviving subset size. It also shows that the success rate does not relate to the rate of elimination. When the surviving subset size is less than 1,000, the success rate starts to increase artificially due to reducing over-fitting, SVM acts better in the test set. When the surviving subset is very small, the correctness of the test decreases and the interaction with the rate of elimination appears. The 3-D plot of  $Vsuc$  over the rate of elimination and the subset size is in the Figure 15. Although it is not smooth like in the  $Tsuc$  plot, it also suggests that there is no interaction between the rate of elimination and the surviving subset.

We use *anova2.m* in Matlab Statistics Toolbox to assembly ANOVA II. ANOVA II test provides a box plot and an ANOVA table. The plot in Figure 16 shows box and whisker plots of the distribution of the success rate  $Tsuc$  (vertical axis) in the groups of surviving subset sizes (horizontal axis). The box shows the 50% of data from the lower quartile to the upper quartile. The whiskers show the extent from the box to the rest of data. A red dot is on the bottom whisker. The red lines (-) are the medians. The red plus signs (+) are the outliers. The medians of success rate  $Tsuc$  start and stay stable at about 0.97 when the surviving subset sizes are from 7,129 to 512. When the subset sizes are from 512 to 64, the medians of  $Tsuc$  increase to 1. When the subset size is smaller than 64, the medians of  $Tsuc$  decrease. It suggests the optimal surviving subset is 64.

### 3-D Plots of Lymphoma Data

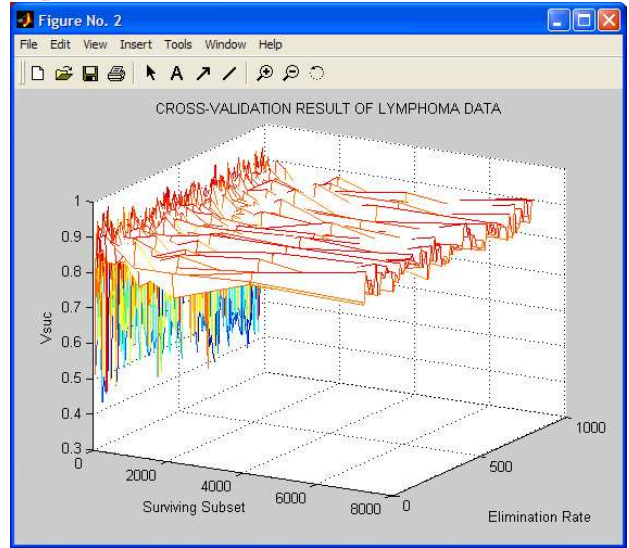
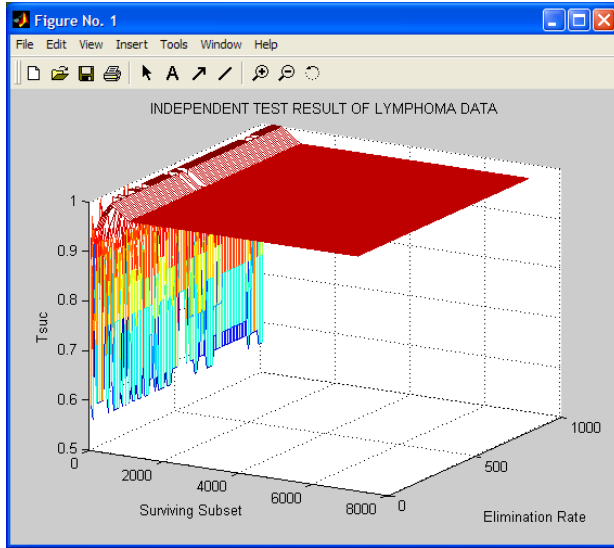


Figure 14: Lymphoma Data Independent Test Result Plot. The flat surface shows that there is no relation of the success rate to the elimination rate for a period of surviving subset sizes from 7,129 to less than 1,000. It starts to increase at the subset size about 1,000, but there is no relation to the elimination rate. The success rate decreases when the surviving subsets are small and the interaction with the elimination rate appears.

Figure 15: Lymphoma Data 3-3-Cross-validation Result Plot. The “surface” of the success rate  $V_{suc}$  (vertical axis) in this plot is not smooth as the  $T_{suc}$  plot 14. The success rate is lower than the one in the plot 14, but it still shows the trends are the same.

The independent test ANOVA II table is in Figure 18. The columns are the surviving subset size groups, and the rows are the elimination rate groups. In the “Columns” factor, the null hypothesis of F-test is “there is no surviving subset effect.” The F-test p-value is 0 which is smaller than 0.05. We reject the null hypothesis at the significance level of 95%. This means that there are variances among the groups of the surviving subset sizes. For the “Rows” factor, the null hypothesis of F-test is “There is no effect of elimination rate.”. The F-test p-value is 0.13 which is greater than 0.05. We can not reject the null hypothesis at 95% significance. This confirms our conclusion about the optimal surviving subset which is 64. The optimal subset size 64 satisfies all three criteria in Section 7.3.

The box plot of the cross-validation test of lymphoma data is in Figure 17. This plot shows a box and whisker of the distribution of the success rate  $V_{suc}$  (vertical axis) in the groups of surviving subset sizes (horizontal axis). There are outliers in this plot. The medians of success rate  $V_{suc}$  vary from 0.8 to 0.9 from when the surviving subset sizes are from 7, 129 to 64. The maximum median of the correctness is at the surviving subset 32. It suggests the optimal surviving subset is 32. The ANOVA table of  $V_{suc}$  (Figure 19) also confirms this.

## Box Plots of Lymphoma Data

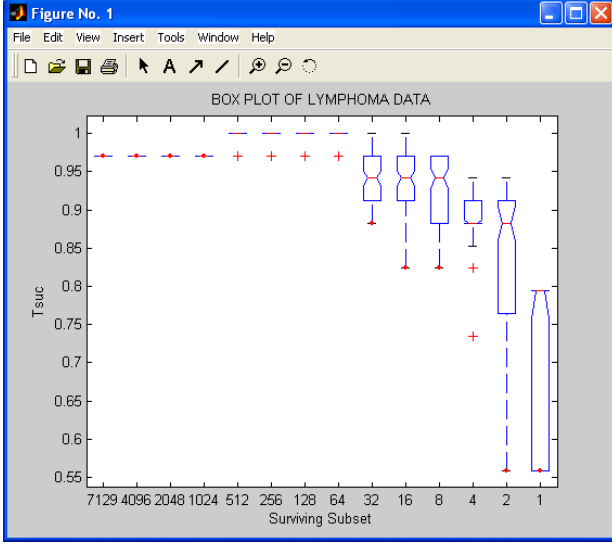


Figure 16: Independent Test Box Plot. This plot shows a box and whisker of the distribution of the success rate  $Tsuc$  (vertical axis) in the groups of surviving subset sizes (horizontal axis). The red lines (-) are the medians. The red plus signs (+) are the outliers. The medians of success rate  $Tsuc$  are stable at about 0.97 when the surviving subset sizes are from 7,129 to 512. When the subset sizes are from 512 to 64, the medians of  $Tsuc$  increase to 1. When the subset size decreases lower than 64, the medians of  $Tsuc$  decrease. It suggests the optimal surviving subset is 64. We obtain this plot using Matlab Statistic Toolbox.

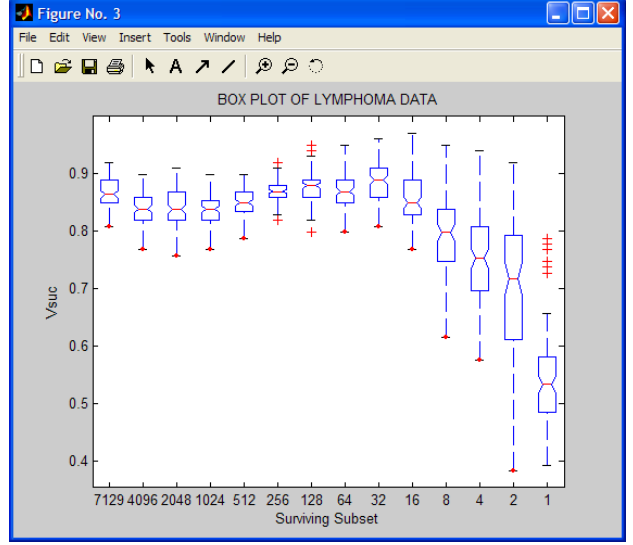


Figure 17: Cross-Validation Test Box Plot. This plot shows a box and whisker of the distribution of the success rate  $Vsuc$  (vertical axis) in the groups of surviving subset sizes (horizontal axis). The red lines (-) are the medians. There are outliers in this plot. The medians of success rate  $Vsuc$  vary from 0.8 to 0.9 from when the surviving subset sizes are from 7,129 to 64. The maximum median of the correctness is at the surviving subset 32. It suggests the optimal surviving subset is 32. We obtain this plot using Matlab Statistic Toolbox.

## 9.2 BNI Meningioma Data

### 9.2.1 Consider BNI Meningioma Data As A Large Data Set

When we treat the data as a large data set, the test set is never seen in the feature selection and the training process. The test result will be more reliable. Can SVM act well on small training set like this one? Treating the BNI data as a large data set, the results are reported in Figures 20-25, for the same measures as Figures 14-19 for lymphoma data.

The 3-D plot of the independent test (Figure 20) shows that the overall accuracy is lower than in the lymphoma data (Figure 14). The flat surface of the success rate  $Tsuc$  spreads equally through all groups of the elimination rate and the surviving subsets from about 50,000 to less than 10,000. The trend of this surface is the same as in the case of the lymphoma data. The 3-D plot of the cross-validation test is in Figure 21. It is not a smooth surface like in the independent test results (Figure 20). There is a variance of

## The ANOVA Tables of Lymphoma Data

ANOVA Table					
Source	SS	df	MS	F	Prob>F
Columns	0.14868	7	0.02124	735.81	0
Rows	0.00335	99	0.00003	1.17	0.1339
Error	0.02	693	0.00003		
Total	0.17204	799			

Figure 18: Independent Test ANOVA II Table of Lymphoma Data. The column factor groups are the surviving subsets from 64 to 7,129. The row factor groups are the elimination rates from 10 to 1,000. The F-test p-values show that the surviving subset sizes are dependent and the elimination rates are independent. We obtain this plot using Matlab Statistic Toolbox.

ANOVA Table					
Source	SS	df	MS	F	Prob>F
Columns	0.27285	8	0.03411	43.7	0
Rows	0.08938	99	0.0009	1.16	0.153
Error	0.61806	792	0.00078		
Total	0.98029	899			

Figure 19: 3-3-Cross-validation Test ANOVA II Table. The column factor is the surviving subset from 32 to 7,129. The row factor is the elimination rate from 10 to 1,000. The F-test p-values show that the surviving subset sizes are dependent; the elimination rates are independent and there is no interaction between the surviving subset factor and the elimination rate factor. We obtain this plot using Matlab Statistic Toolbox.

the accuracy in this plot; however, we still see the trend. The accuracy of cross-validation increases when the feature subset becomes smaller. When the subset is about 10,000, the accuracy amplitudes are varied and the independence of the elimination rate appears. The *Vsuc* 3-D plot of lymphoma data appears more stable than this plot. The independence of the elimination rate still can be revealed by these 3-D plots.

The box plot of independent test result is in Figure 22. The medians of success rate *Tsuc* stay stable at about 0.72 when the surviving subset sizes are from 54,675 to 4,096. They start to increase when the subset sizes is 2,048. They reach the maximum at the subset size 64, then they start to decrease after the surviving subset 64. This suggests the optimal surviving subset is 64.

The ANOVA II table of independent test is in Figure 24. The p-value of the F-test in the row factor, the elimination rate, is 0.14 which is greater than 0.05. Thus we can not reject the null hypothesis “There is no elimination rate interaction” at the 95% significance. This confirms the choice of the optimal feature subset 64.

The box plot of the cross-validation test (Figure 23) shows boxes and whiskers of the distribution of the success rate *Vsuc* (vertical axis) in the groups of surviving subset sizes (horizontal axis). The medians of success rate *Vsuc* is maximum at 8. It suggests the optimal surviving subset is 8. The ANOVA table of cross-validation test is in Figure 25. The F-test p-value of the null hypothesis “There is no elimination rate interaction” is 0.76 which is greater than 0.05. Thus we can not reject the null hypothesis at the 95% significance. This confirms the choice of the optimal subset 8.

We can state that the optimal feature subset is 64 based on the independent test or 8 based on the cross-validation test.

### 9.2.2 Consider BNI Meningioma Data As A Small Data Set

When we treat the data as a small data set, the accuracy of the tests are higher than when we treat the data as a large data set, especially with independent tests as mentioned in [10]. The 3-D plot of independent test result is in Figure 26. This plot shows a flat surface which indicates there is no interaction between the elimination rate and the surviving subset factors. It also tell us that there is no interaction of the success rate  $T_{suc}$  and the elimination rate. The correctness increases to 1 at the surviving subset 32,768 and stays at that level until the surviving subset is 1. This is unexpectedly high correctness. The 3-D plot of the cross-validation test is in Figure 27. In this plot, there are variances of the correctness  $V_{suc}$  among the elimination rate and the surviving subset. The plot is not a smooth surface, but it still shows the overall trend of the increasing correctness over the surviving subset. The correctness increases to 1 at the end, surviving subset 1.

The independent test box plot in Figure 28 shows the high medians when the surviving subset decreases from 32,768 to 1. This suggests that the optimal subset is 1. It is too good to be true. We do not trust the independent test when the data is treated as a small set. The accuracy of this test is affected by over-fitting. The cross-validation test box plot is in Figure 29. The means of the correctness increase to 1 when the surviving subset decreases to 128. They stay at 1 until the surviving subset is 16. When the surviving subsets progress smaller than 16, the means of correctness decrease. The maximum value of correctness mean is 1 at the surviving subset 16. Thus 16 is the optimal feature subset which satisfies 2 of the top criteria of the optimal subset in Section 7.3.

The cross-validation test provides better information for the optimal feature subset selection. The ANOVA II table of the cross-validation test is in Figure 30. The “Rows” are the groups of the elimination factor. The null hypothesis of F-test of this factor is “There is no elimination rate factor effect.” The p-value of the F-test is 0.85 which is greater than 0.05. Thus we can not reject the null hypothesis at the significant level of 95%. In other words, the elimination rate factor is independent in the interval of the surviving subset from 16 to 54,675. This satisfies the third criterion of the optimal feature subset in Section 7.3.

We believe that this method provides a large enough training data set for the feature selection and the evaluation SVM. So the optimal subset is 16 according to cross-validation test.

## 10 Conclusion

Guyon et al. proved that the support vector machine serves as an effective inducer for the recursive feature elimination algorithm in feature selection [7]. In our experiment, we conclude that the rate of elimination does not affect the accuracy of feature selection. We also recognize that there is no interaction between the rate of elimination and the surviving subset size. In the scope of our experiment, we can eliminate a thousand features in every loop of the support vector machine recursive feature elimination and it does not affect the quality of the optimal surviving subset. Note that in our experiment we limit ourselves in support vector machine recursive feature elimination to select features for support vector machine classifier training only.

### 3-D Plots of BNI Data As A Large Data Set

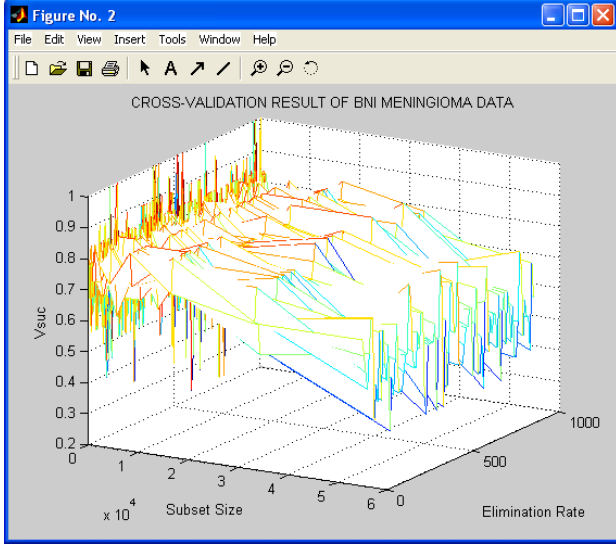
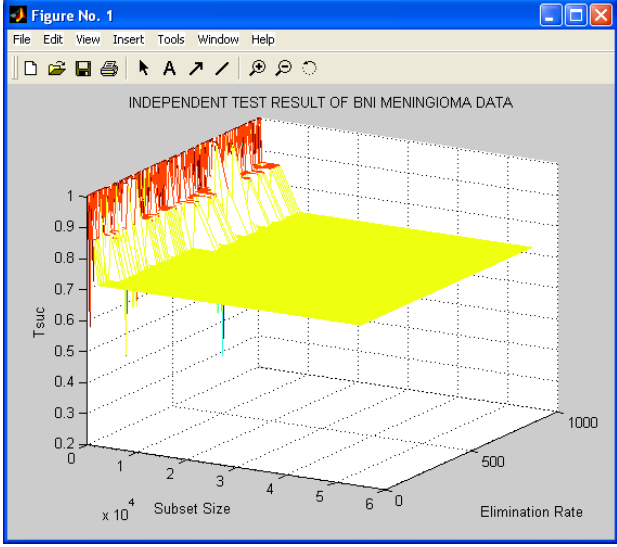


Figure 20: BNI Data Independent Test Result Plot. The flat surface shows that there is no relation of the success rate to the elimination rate for a period of surviving subset sizes from 50,000 to less than 10,000. It starts to increase at the subset size about 1,000, but there is no relation to the elimination rate. The success rate decreases when the surviving subsets are small and the interaction with the elimination rate appears.

Figure 21: BNI Data 3-3-Cross-Validation Result Plot. The “surface” of the success rate  $V_{suc}$  (vertical axis) in this plot is not smooth as the  $T_{suc}$  plot 20. The success rate is lower than the one in plot 14, but it still shows the trend.

In the case of the small data set, the result is difficult to interpret. The evaluation feature subset algorithm plays a central role in the small data set case. If we divide the data into the training set and the test set, then the training set is too small for both feature selection and evaluation. If we “feed” the whole data set into the feature selection, then the test result is too high to believe, especially, in the independent test. The test set in the evaluation has been seen in the feature selection. The independent test set is not available; we must rely only on the cross-validation test. To choose the optimal subset we use the cross validation test when we treat data as small. It suggests the optimal subset is 16 genes in BNI meningioma data.

## 11 Future Work

1. Small data sets need further investigation in order to answer the question “What is the minimal data set size which is needed for SVMRFE to act well?” We need to consider the small subset effect on different aspects rather than only the correctness of the evaluation test.

## Box Plots of BNI Data As A Large Data Set

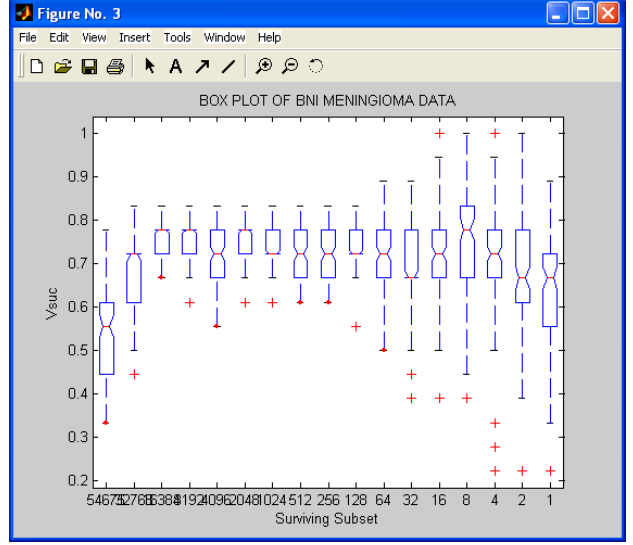
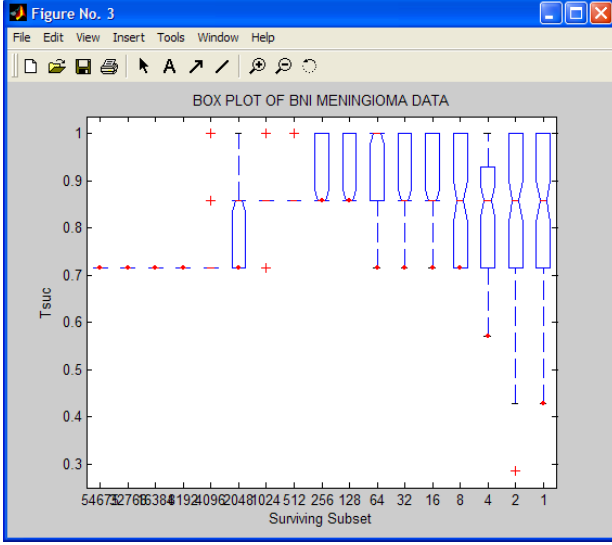


Figure 22: Independent Test Box Plot of BNI Data As Large Data Set. This plot shows boxes and whiskers of the distribution of the success rate  $T_{suc}$  (vertical axis) in the groups of surviving subset sizes (horizontal axis). The medians of success rate  $T_{suc}$  stay stable at about 0.72 when the surviving subset sizes are from 54,675 to 4,096. They start to increase when the subset sizes is 2,048. It reaches the maximum at the subset size 64, then it starts to decrease after the surviving subset 64. It suggests the optimal surviving subset is 64. We obtain this plot using Matlab Statistic Toolbox.

Figure 23: 3-3-Cross-Validation Test Box Plot of BNI Meningioma Data. This plot shows boxes and whiskers of the distribution of the success rate  $V_{suc}$  (vertical axis) in the groups of surviving subset sizes (horizontal axis). The medians of success rate  $V_{suc}$  vary when the surviving subsets are from 54,675 to 8. The maximum median is at 8. It suggests the optimal surviving subset is 8. We obtain this plot using Matlab Statistic Toolbox.

2. We need a better evaluation method for a small data set. We suggest that the leave-one-out cross validation is appropriate. The training set needs to feed into the SVMRFE and evaluation module while the test set is hold out to obey the “never-seen rule” for the test set. The test set size is one in this case. We need to develop a new method to choose and evaluate the surviving subset.
3. The effect of over-fitting to the cross-validation test algorithm needs more investigation in the small data set case.
4. In our current experiment, we investigate only a small number of data sets. We will examine more data sets with different example sizes in order to confirm our conclusion.
5. This experiment scopes on linear support vector machine. We need to test the performance of non-linear support vector machine.

## The ANOVA Tables of of BNI Data As A Large Data Set

ANOVA Table					
Source	SS	df	MS	F	Prob>F
Columns	8.1842	10	0.81842	294.94	0
Rows	0.3453	99	0.00349	1.26	0.0524
Error	2.7472	990	0.00277		
Total	11.2766	1099			

Figure 24: Independent Test ANOVA II Table of BNI Data As A Large Data Set. The column factor groups are the surviving subsets from 64 to 54, 675. The row factor groups are the elimination rates from 10 to 1,000. The F-test p-values show that the surviving subset sizes are dependent and the elimination rates are independent. We obtain this plot using Matlab Statistic Toolbox.

ANOVA Table					
Source	SS	df	MS	F	Prob>F
Columns	3.9878	13	0.30675	48.56	0
Rows	0.5576	99	0.00563	0.89	0.7648
Error	8.1298	1287	0.00632		
Total	12.6751	1399			

Figure 25: Cross-Validation Test ANOVA II Table of BNI Data As A Large Data Set. The column factor groups are the surviving subsets from 8 to 54, 675. The row factor groups are the elimination rates from 10 to 1,000. The F-test p-values show that the surviving subset sizes are dependent and the elimination rates are independent. We obtain this plot using Matlab Statistic Toolbox.

6. We do not yet have any effective biological evaluation of our selected features.

## References

- [1] Avrim L. Blum and Pat Langley, *Selection of Relevant Features and Examples in Machine Learning*, Web-print: <http://citeseer.ist.psu.edu/blum97selection.html>.
- [2] Christopher J.C. Burges, *A Tutorial on Support Vector Machines for Pattern Recognition*, (1998), Web-print: <http://citeseer.ist.psu.edu/397919.html>.
- [3] Yann Le Cun, John S. Denker, and Sara A. Solla, *Optimum Brain Damage*, (1990), Web-print: <http://citeseer.ist.psu.edu/lecun90optimal.html>.
- [4] Soroin Drăghici, *Data Analysis Tools for DNA Microarrays*, Chapman and Hall/CRC, 2003.
- [5] Golub et al, *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*, *Science* **286** (1999), 531–7, [http://www.broad.mit.edu/mpr/publications/projects/Leukemia/Golub\\_et\\_al.1999.pdf](http://www.broad.mit.edu/mpr/publications/projects/Leukemia/Golub_et_al.1999.pdf).
- [6] Silicon Genetics, *Genespring 6.2*, software, <http://www.silicongenetics.com/cgi/SiG.cgi/index.smf>.
- [7] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik, *Gene Selection for Cancer Classification using Support Vector Machines*, *Machine Learning* **46** (2002), no. 1-3, 389–422, Web-print: <http://citeseer.ist.psu.edu/guyon00gene.html>.
- [8] Affymetrix Inc., *Data Sheet GeneChip Human Genome Arrays*, Tech. report, [http://www.affymetrix.com/support/technical/datasheets/human\\_datasheet.pdf](http://www.affymetrix.com/support/technical/datasheets/human_datasheet.pdf).

### 3-D Plot of BNI Data As A Small Data Set

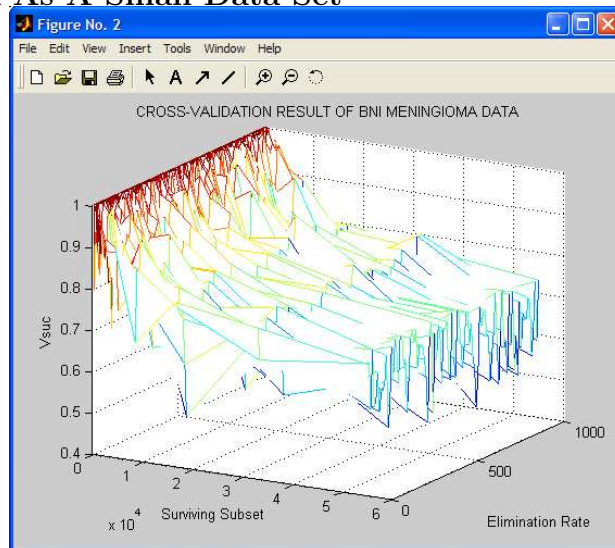
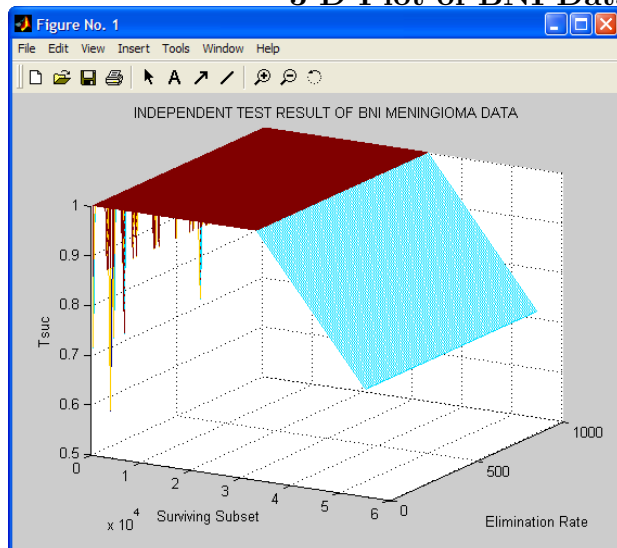


Figure 26: Independent Test Result 3-D Plot of BNI Meningioma Data As A Small Data Set. The overall accuracy increases quickly to the maximum value 1 at about 30,000 surviving subset size. It stays high until the surviving subset is 1. The flat surface suggests that there is no elimination rate effect. This also shows that there is interaction of the elimination rate effect and the surviving subset effect.

Figure 27: 3-3-Cross-Validation Result 3-D Plot of BNI Meningioma Data As A Small Data Set. There are variance of the cross-validation correctness  $V_{suc}$  over the elimination rates; however, it still shows the trend of increasing  $V_{suc}$  when the surviving subset decreases. The accuracy is high at the end, but the shape is not very different from when we treat data as a large set.

- [9] Affymetrix Inc, *Statistical Algorithms Description Document*, Tech. report, Affymetrix Inc., <http://www.affymetrix.com/support/technical/whitepapers/sadd.whitepaper.pdf>, 2002.
- [10] Ron Kohavi and George H. John, *Wrappers for Feature Subset Selection*, *Artificial Intelligence* **97** (97), 273–324.
- [11] Huan Liu and Lei Yu, *Feature Selection for High-Dimensional Data: A Fast Correlation-based Filter Solution*, (2003), Web-print: <http://www.hpl.hp.com/conferences/icml2003/papers/144.pdf>.
- [12] Kar-Ming Fung M.D. Ph.D.1 Gregory N. Fuller M.D. Ph.D Walter F. Bierbaum, M.D.1, *45 year old Woman with an Enhancing Brain Mass*, Tech. report, Department of Pathology, University of Oklahoma Health Science Center, Oklahoma City, Oklahoma, 2 Department of Pathology, University of Texas M.D. Anderson Cancer Center, Houston, Texas, June 2003, <http://moon.ouhsc.edu/kfung/JTY1/Com/Com306-1-Diss.htm>.

## Box Plots of BNI Data As A Small Data Set

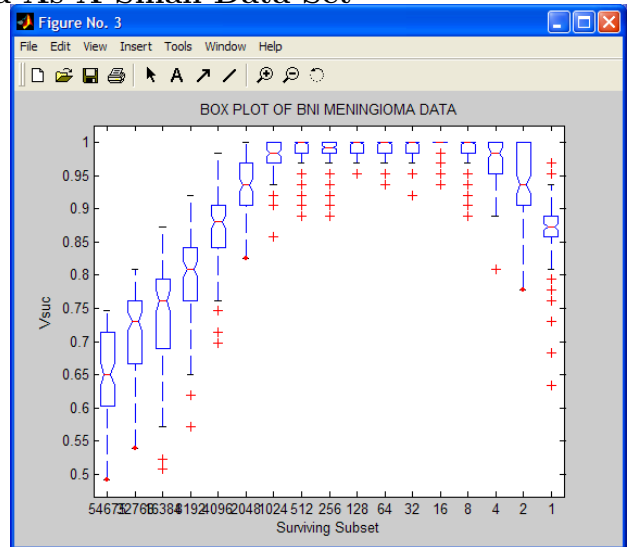
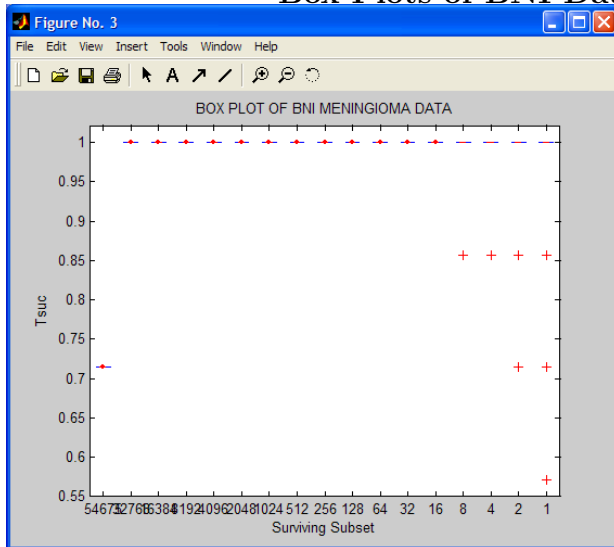


Figure 28: Independent Test Box Plot of BNI Data As A Small Data Set. This plot shows a box and whisker of the distribution of the success rate  $T_{suc}$  (vertical axis) in the groups of surviving subset sizes (horizontal axis). The red lines (-) are the medians. The medians of the success rate  $T_{suc}$  increase to the maximum 1 at 32,768. They are stable until the surviving subset is 1. We obtain this plot using Matlab Statistic Toolbox.

Figure 29: 3-3-Cross-Validation Test Box Plot of BNI Data As A Small Data Set. When the surviving subset decreases, the mean of accuracy increases. It reaches the maximum 1 at the surviving subset 128. The smallest subset that has the maximum accuracy is 16. We obtain this plot using Matlab Statistic Toolbox.

ANOVA Table					
Source	SS	df	MS	F	Prob>F
Columns	18.9164	12	1.57637	760.15	0
Rows	0.1747	99	0.00176	0.85	0.8475
Error	2.4636	1188	0.00207		
Total	21.5547	1299			

Figure 30: Cross-Validation Test ANOVA II Table of BNI Data As A Small Data Set. The column factor groups are the surviving subsets from 16 to 54,675. The row factor groups are the elimination rates from 10 to 1,000. The F-test p-values show that the surviving subset sizes are dependent and the elimination rates are independent. We obtain this plot using Matlab Statistic Toolbox.