7-29-2010

Enhancing Gene Expression Signatures in Cancer Prediction Models: Understanding and Managing Classification Complexity

Vidya P. Kamath
University of South Florida

Enhancing Gene Expression Signatures in Cancer Prediction Models:

Understanding and Managing Classification Complexity

by

Vidya P. Kamath

A dissertation submitted in partial fulfillment
of the requirement for the degree of
Doctor of Philosophy
Department of Chemical and Biomedical Engineering
College of Engineering
University of South Florida

Co-Major Professor: Steven A. Eschrich, Ph.D.
Co-Major Professor: Dmitry Goldgof, Ph.D.
John Heine, Ph.D.
Rangachar Kasturi, Ph.D.
Ji-Hyun Lee, Dr.PH.
Timothy Yeatman, M.D.

Date of Approval
July 29, 2010

Keywords: quantization, survival analysis, random subspaces, cost-sensitive analysis, biological covariates

DEDICATION

I dedicate this work to everyone in my life that has encouraged me and helped me enjoy life's little challenges, and to a never-ending quest to understand the fascinating world we live in.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Dr. Steven A. Eschrich for his skillful guidance, support and encouragement during my graduate work. His careful and critical review of my work is greatly appreciated.

I am grateful for the financial support provided by Dr. Timothy Yeatman, Dr. Javier Torres-Roca and Dr. Rangachar Kasturi during the course of my study.

I thank my committee members, Dr. Dmitry Goldgof, Dr. John Heine, Dr. Rangachar Kasturi, Dr. Ji-Hyun Lee and Dr. Timothy Yeatman for the insightful discussions that helped shape this work.

I am also grateful for the constant support and timely help provided by Karen Bray at the Department of Chemical and Biomedical Engineering, USF and Paula Price at the Department of Biomedical Informatics at H. Lee Moffitt Cancer Center.

The gene expression analysis of colorectal adenocarcinoma was partially supported by the Department of Defense, National Functional Genomics Center project, under award DAMD17-02-2-0051. Views and opinions of, and endorsements by, the author do not reflect those of the US Army or the Department of Defense.

The gene expression analysis for prediction of radiosensitivity was supported in part by the State of Florida Department of Health, Bankhead-Coley Cancer Research Program, 09BB-22.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

MRC-CRC:	Moffitt Colorectal Adenocarcinoma dataset
NSCLC:	Non-small cell lung cancer dataset
NCI60	Cell line dataset obtained from the NCI60 panel of cell lines
SVM:	Support vector machines
NN:	Neural networks
C4.5 DT:	C4.5 decision trees
K-M:	Kaplan-Meier survivor estimates
L-R:	Log-rank test
CoxPH:	Cox proportional hazards models
n-fold CV:	n-fold cross validation
SAM:	Significance analysis of microarrays
MFS-RS:	Multivariable feature selection using random subspaces
MFS-RSc:	MFS-RS with cost-sensitive analysis
ΔR^2:	Change in Adj-R^2 value between two groups of samples

ABSTRACT

Cancer can develop through a series of genetic events in combination with external influential factors that alter the progression of the disease. Gene expression studies are designed to provide an enhanced understanding of the progression of cancer and to develop clinically relevant biomarkers of disease, prognosis and response to treatment. One of the main aims of microarray gene expression analyses is to develop signatures that are highly predictive of specific biological states, such as the molecular stage of cancer. This dissertation analyzes the classification complexity inherent in gene expression studies, proposing both techniques for measuring complexity and algorithms for reducing this complexity.

Classifier algorithms that generate predictive signatures of cancer models must generalize to independent datasets for successful translation to clinical practice. The predictive performance of classifier models is shown to be dependent on the inherent complexity of the gene expression data. Three specific quantitative measures of classification complexity are proposed and one measure (ϕ) is shown to correlate highly (R^2=0.82) with classifier accuracy in experimental data.

Three quantization methods are proposed to enhance contrast in gene expression data and reduce classification complexity. The accuracy for cancer prognosis prediction is shown to improve using quantization in two datasets studied: from 67% to 90% in lung

cancer and from 56% to 68% in colorectal cancer. A corresponding reduction in classification complexity is also observed.

A random subspace based multivariable feature selection approach using cost-sensitive analysis is proposed to model the underlying heterogeneous cancer biology and address complexity due to multiple molecular pathways and unbalanced distribution of samples into classes. The technique is shown to be more accurate than the univariate t-test method. The classifier accuracy improves from 56% to 68% for colorectal cancer prognosis prediction.

A published gene expression signature to predict radiosensitivity of tumor cells is augmented with clinical indicators to enhance modeling of the data and represent the underlying biology more closely. Statistical tests and experiments indicate that the improvement in the model fit is a result of modeling the underlying biology rather than statistical over-fitting of the data, thereby accommodating classification complexity through the use of additional variables.

CHAPTER 1 INTRODUCTION

1.1 Introduction

Cancer is the second leading cause of death in the United States [1]. Studies of the molecular basis of cancer [2-5] have shown that progression of cancer is influenced by a series of genetic events in combination with external factors such as age, dietary conditions or smoking history [6-7]. Gene expression studies probe genetic activity to develop clinically-relevant biomarkers of disease, prognosis and response to treatment [8-11]. Microarray gene expression data (see Section 2.3) has been used for discovery of genes involved in one or more specific biological functions of tumor cells [2-4, 12-28]. One of the main aims of microarray gene expression analyses is to extract signatures that are highly predictive of specific biological states, such as the molecular stage of cancer [5, 29]. This has opened up the possibility of translational science that moves basic biological findings from laboratory discoveries to clinical tests [30]. Molecular signatures may offer the opportunity to develop a personalized medicine approach to disease management, in which therapy is tailored to the individual [31].

1.2 Contribution and Organization

Microarray gene expression data has been used to generate models to understand the development and progression of cancer. However, models predictive in the dataset in which they were developed may not generalize to independent samples accurately [3].

This issue is examined and used to understand some of the fundamental issues in building gene expression models and to develop a framework to extract more accurate and generalized signatures.

Chapter 2 provides a brief overview of the problem domain (gene expression studies in cancer), along with the datasets and the data analysis tools used for various modeling techniques described in the subsequent chapters.

Chapter 3 examines the sources of data complexity and develops three specific measures of complexity. These measures are used to quantitatively compare the complexity of different gene expression datasets to establish a relationship between the proposed complexity measures and classification accuracy.

Chapter 4 proposes simplification of high-resolution microarray data by the use of quantization techniques. Classifier experiments are performed to provide a quantitative assessment of the quantized datasets in developing predictive models of survival.

Chapter 5 discusses a cost-sensitive random subspace based approach that is proposed for extracting subsets of genes that are, in combination, associated with the outcome. Data complexity due to multiple gene pathways is used as a motivation for the multivariable feature selection approach for improving classification accuracy.

Chapter 6 explores the effect of inclusion of specific biological indicators into gene expression models. The fit of the enhanced model to the underlying data is explored and verified.

Finally, conclusions from the analysis of the proposed methods are summarized in Chapter 7 and recommendations for future work are provided.

CHAPTER 2 BACKGROUND

2.1 Introduction

Classifier models and statistical tools are used on microarray gene expression datasets to extract patterns of expression differences between samples [32-34]. These patterns, called signatures in this body of work, are used to develop clinically-relevant biomarkers of disease, prognosis and response to treatment [5, 19, 29, 35-38]. This chapter presents basic background information regarding the data and techniques used in the following chapters.

An overview of cancer is provided in Section 2.2 followed by a description of gene expression, the basic setup of microarray technology and gene expression analysis in Section 2.3. Section 2.4 describes the specific datasets used for analysis and the basic data analysis models used in the following chapters are presented in Section 2.5 Model validation techniques and performance measures are presented in Section 2.6.

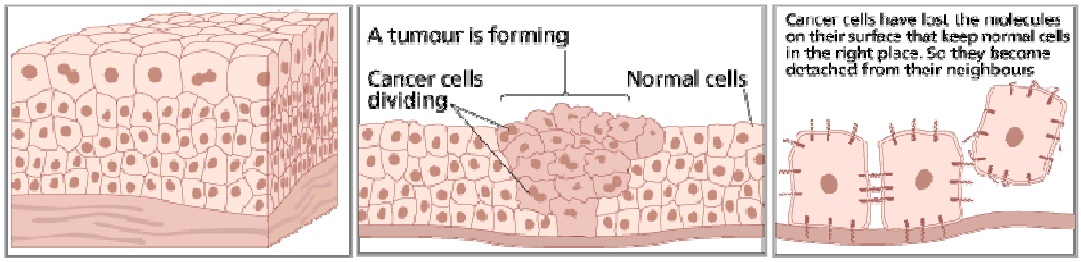
2.2 Cancer

Cancer is the second leading cause of death in the United States [1]. More people succumb to lung cancer per year than any other type of cancer. Almost 80% of the 196,454 people diagnosed with lung cancer in 2006 died from the disease. Colorectal cancer (cancer of the colon and rectum) is the second leading cause of death due to cancer. Close to 37% of the 139,127 people diagnosed with colorectal cancer in 2006 died due to the disease. In women, breast cancer is one of the leading causes of cancer

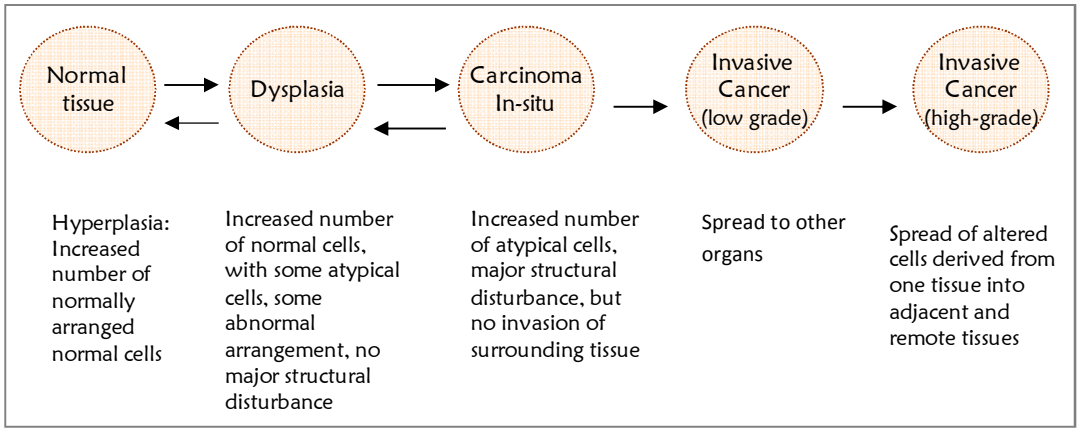
related deaths, second only to skin cancer. 191,410 women were diagnosed with breast cancer in 2006, and 40,820 died from the disease. Development of biomarkers for early detection of these cancers and treatment planning in advanced stages of the disease can greatly aid in reducing the fatality due to the disease [13, 14, 35].

Genetic mutations in normal cells, and environmental stimuli in combination with external factors such as age, smoking history or diet, can cause a normal cell to multiply uncontrollably, leading to cancer (Figure 2-1a) [7,39]. It is hypothesized that the specific sequence of molecular events that initiated the deviation from normalcy can be an indicator for the progression of the disease [7]. A molecular pathway is defined as a series of actions between molecules at the cellular level that leads to a certain cell function [8]. The progression of the tumor from normal to invasive depends on the molecular pathways that are active in the cells as indicated in Figure 2-1b.

Physical and molecular characteristics of a tumor sample can aid the physician in planning treatment for a specific patient. General information such as patient age and gender, and more specific clinical information including biological indicators for molecular pathways such as the RAS status [7, 8, 40] and p53 status [41] may also be used for treatment planning. Models of disease progression use retrospective data such as patient survival information in addition to the physical and molecular data.



(a)



(b)

Figure 2-1: Description of cancer development. (a) Behavior of cells at different stages of cancer [42] (b) Pathways for progression from normal cell to invasive

2.3 Gene Expression

2.3.1 Measuring Gene Expression Using Microarrays

The human body is estimated to have approximately 30,000 genes [8, 42, 43]. Although each cell of the body contains an exact copy of these genes for the individual, only certain genes are active in any given cell. The specific genes that are active and the level of their activity, or the expression level, in the cell govern how the cell functions in its environment. Measuring the level of activity of genes in a cell can provide information on the function it performs, and the influence it exerts on its environment [7, 8]. Figure 2-2 provides an example of differences in expression levels of certain genes in normal and cancerous cells. Identification of genes that are expressed differently under different

biological conditions, such as normal or tumor, or different stages of tumor can aid in understanding the disease process and progression.

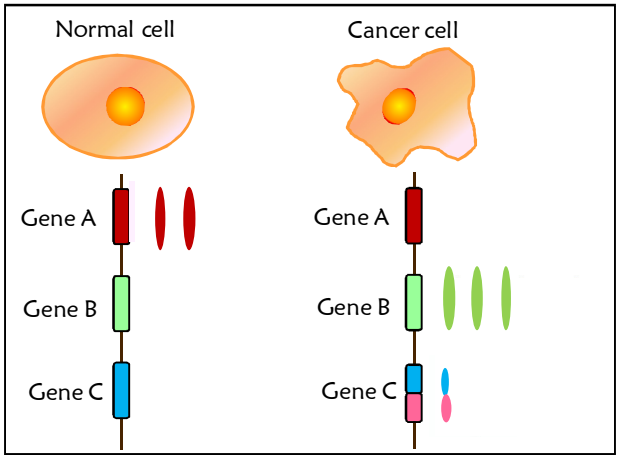
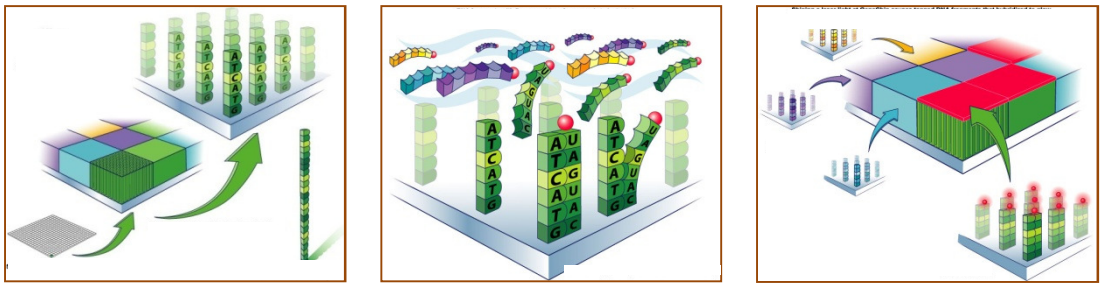


Figure 2-2: Differences in expression levels of genes can be used to distinguish normal from cancerous cells

The activity of thousands of genes in a target tissue may be measured simultaneously using mRNA microarray chips [44, 45]. A microarray chip consists of specific sequences of nucleic acids, called probesets, that are designed to identify and hybridize [45] with specific target sequences representing genes of interest. A probeset is typically designed to probe only a small section of the target gene, and a single gene may be probed by multiple probesets. Detailed information regarding the design and functioning of different types of microarrays is provided in [44, 46]. For a microarray experiment, RNA from the sample is extracted from the tissue and fluorescence-tagged. This tagged RNA is then washed over the microarray chip under specific experimental conditions. If the sample RNA contains sequences of interest, these sequences will hybridize with the corresponding probe sequences, resulting in fluorescence in specific areas of the chip. Detection of fluorescence is used as an indication of expression of the

target gene while lack of fluorescence indicates an absence of expression. The level of fluorescence for each probe is detected and quantified by acquiring a high-resolution digital image of the microarray chip (Figure 2-3). The image data is converted to a numerical quantity that is used as the expression level for the probeset of the specific gene.



(a) (b) (c)

Figure 2-3: Microarray experiment to detect expression of target genes on Affymetrix GeneChip®. (a) RNA probes attached to the microarray chip designed to identify specific target sequences (b) Hybridization of sample RNA onto the microarray chip (c) Detected fluorescence indicates expression level of the target gene (Images courtesy of Affymetrix®)

2.3.2 Building Gene Expression Models

The main steps involved in building gene expression models are shown in Figure 2-4. The model building process begins with the framing of a specific biological question followed by an experimental design that selects samples such that at least a few examples are included for each aspect of the disease intended for study. For example, in an experiment designed to study the molecular differences between different stages of colorectal cancer, at least a few samples must be included for each of the four stages of disease.

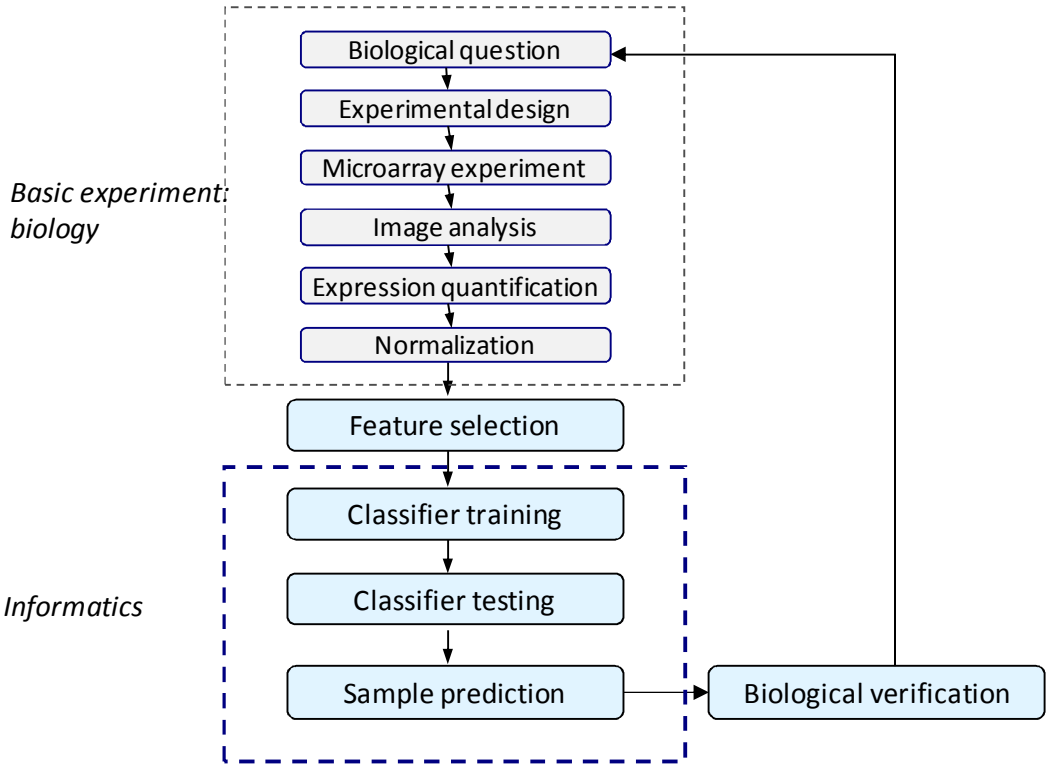


Figure 2-4: A general approach for gene expression analysis to build models of cancer

Microarray experiments include image analysis and normalization of the data to yield the gene expression dataset for analysis [9, 10, 47]. Since microarray data typically consists of a few thousand probesets or features, the first step of the analysis is to select a small set of features that are strongly correlated to the outcome. These features are used to train a classifier model which is then tested on a set of independent test samples. Performance measures such as accuracy, specificity and sensitivity (see Section 2.6) may be used to determine if the classifier model of gene expression is predictive.

2.3.3 Gene Expression Signatures

Microarray gene expression analysis involves studying patterns of expression for genes e.g. across tissues of various types [4], diseased vs. normal tissue [26], or tissue under varying environmental conditions, such as tumor cells treated with radiation therapy [27]. Other examples of microarray analysis include analyzing expression across different stages of development of cancer [13] or for different types of patient outcome [14].

Feature selection, or elimination of noisy probesets, is popularly used as a first step in gene expression analysis [34]. Feature selection may be achieved in an unsupervised or supervised manner. Some simple methods for unsupervised analysis include computation of a few basic statistics from the dataset, such as variance [48], ranking of features [49], linear dependency [50] and others [50, 51] to provide information on the existence (or lack thereof) of some structure or order within the data. Many methods have been proposed for supervised feature selection, including information gain-based methods [18, 52-55]. Supervised methods of feature selection are useful when a specific signature is being investigated, for example, a signature to describe the drug effect for cancer patients.

Golub et al [4] analyzed two types of acute leukemia, (ALL: acute lymphoblastic leukemia and AML: acute myeloid leukemia), to develop a general strategy for discovering and predicting types of cancer. Neighborhood analysis was used to identify a set of informative genes that could predict the class of an unknown sample of leukemia. Each informative gene was used to cast a weighted vote on the class of the sample, and

the summation of the votes predicted the class of the sample. Self-organizing maps (SOM) were used to cluster tumors by gene expression to discover new tumor types.

van ‘t Veer et al [14] utilized a hierarchical clustering algorithm to identify a gene expression signature that could predict the prognosis of breast cancer. Two subgroups were created using the clustering technique, with genes that were highly correlated with the prognosis of cancer. The number of genes in each cluster was then optimized by sequentially adding subsets of 5 genes and evaluating the power of prediction in a leave-one-out cross-validation scheme. Expression profiles of tumors with correlation coefficients above the optimized sensitivity threshold were classified as good prognosis, and the rest as poor prognosis.

Alon et al [26] distinguished between normal and tumor samples of colon cancer using a deterministic annealing algorithm. Genes were clustered into separate groups sequentially to build a binary “gene tree”, and tissues were clustered to create a “tissue tree”. Genes that showed strong correlation were found closer to each other on the “gene tree”, and tissues with strong similarities were found close together on the “tissue tree”. A two-way ordering of genes and tissues was used to identify families of genes and tissue based on the gene expressions in the dataset.

Glinsky et al [56] identified an 11-gene signature that was shown to be a powerful predictor of a short interval to distant metastasis and poor survival after therapy in breast and lung cancer patients, when diagnosed with an early-stage disease. The method clustered genes exhibiting concordant changes of transcript abundance. The degree of resemblance of the transcript abundance rank order within a gene cluster between a test sample and a reference standard was measured by the Pearson correlation coefficient.

Eschrich at al [13] showed that molecular staging of cancer, using the gene expression profile of the tumor at diagnosis, can predict the long-term survival outcome more accurately than clinical staging of the tumor. A feed-forward-back-propagation neural network used 43 genes to predict the molecular stage of a tumor sample.

Fan et al, [57] address the issue of disagreement of gene expression models for the same tumor type, in terms of the genes used for the models. The models described in the article were developed to analyze characteristics of breast cancer samples. A 70-gene model was used to predict "good-versus-poor" prognosis of patients and a recurrence model predicted a high or low recurrence score for the samples. A wound response model predicted samples with poor or good response. An intrinsic sub-type model classified the samples as "luminal A", "luminal B", "basal-like", "HER2-positive", "estrogen-receptor-negative" (HER2+ and ER-) and "normal breast-like". The fifth model was a two-gene ratio model developed to predict outcome for ER+ samples receiving tamoxifen. Clearly, each of these 5 models addressed a different clinical characteristic than the others, or more explicitly, as the authors allude to, these models address clinically different biological phenotypes.

Each of these models studied by Fan et al [57] selected a few genes to create a model for prediction. Although these models were claimed to perform well on breast cancer samples, a comparison of the five lists of selected genes showed that very few genes were actually common.

Ho and Basu [58] explored several popular methods of defining classification complexity, including overlap in individual features, separability of classes, and the topology of the problem space. They demonstrated through a large set of problems (both

real and artificial) that many of the measures identify complex (or even random) problems from simpler problems.

The literature review presented here demonstrates that several models have been developed to address various biological questions and generate predictive signatures using microarray gene expression data.

2.3.4 Problem Definition for Data Analysis

The choice of feature selection and classifier methods for gene expression data analysis is largely dependent on the problem definition. For many problems, the samples are split into discrete groups or classes, with samples in each class having some common characteristic/s and differing from samples in the other classes. For example, a study based on the stages of colorectal adenocarcinoma will split the sample set into four classes with one class for each stage of the tumor. Other problems require data analysis techniques that model a continuous variable as the outcome. For example radiation sensitivity of a patient may be modeled as a continuous outcome.

Outcomes such as patient survival time can be modeled both as a discrete or continuous variable. The data can be split into two distinct groups of patients with specific survival characteristics, for example, patients who have survived longer than 60 months may be categorized as “Good prognosis” patients and those who survived a lesser time in the “Bad prognosis” group.

When modeled as a continuous variable, the actual patient survival times (number of months survived after surgery) are used as the outcome. This data includes a censoring variable when complete information regarding patient survival is unavailable.

Other two-class outcomes are also defined for some of the datasets described in this chapter. These outcomes include gender (Male, Female), clinical stage (I, III), and tissue type (Colon, Rectum).

2.4 Gene Expression Datasets

The microarray gene expression datasets used for describing the techniques proposed in the subsequent chapters are presented in the following sections. All datasets except the MRC-CRC dataset (Section 2.4.2) are publicly available.

2.4.1 Lung Adenocarcinoma (NSCLC)

This dataset was arrayed on the Affymetrix HuGeneFL GeneChips® (n = 62) with 7,129 features and stored in the MAS5.0 data format [47] with a 6-digit precision and range of 10.0-6000.0. A study previously published on this dataset identified a signature between patients with higher and lower risk of death from the cancer [26].

For the two-class survival analysis models developed in Chapters 3-4, the risk of death was transformed using a cut-off for survival time of 30 months (median survival time). Patients who died within 30 months were considered poor prognosis (n = 20), else they had a good prognosis (n = 42). Two additional classification problems using this data are: predicting cancer stages I (n = 49) and III (n = 13); and predicting gender (Male: n=25, Female: n=37) of the patients. The overall survival time for the patients was a continuous variable with information on the vital status of the patient at the end of the study.

2.4.2 Colorectal Adenocarcinoma (MRC-CRC)

Colorectal cancer patient samples (MRC-CRC), collected at the H. Lee Moffitt Cancer Center and Research Institute, were arrayed on the Affymetrix U133Plus2.0 GeneChip®[45], consisting of 54,675 features. The data (n = 121) was processed using RMA normalization [47] and represented as a continuous value from 0.0 to 15.0 with a 6 digit precision. An outcome study was previously published using a subset of this data [13].

For the models developed in Chapters 3-5, the survival times were stratified into high risk (less than 36 months of survival, n = 37) and low risk (greater than 36 months of survival, n = 84). Additional classification problems include determining the gender of the patients (Male: n=59, Female: n=62), as well as the differentiation between colon (n = 85) and rectal (n = 36) cancer. The overall survival time for the patients was available as a continuous variable along with information on the vital status of the patient at the end of the study.

2.4.3 Cell Line Data (NCI60)

Gene expression profiles were obtained from a previously published study [59] using Affymetrix HU6800® chips consisting of 7,129 features [45]. The data was normalized using the Affymetrix MAS 4.0 algorithm in average difference units [47], with the gene expression represented as a continuous value from 0.0-122000.0. Radiation sensitivity data, defined by survival fraction after 2 Gy (SF2), were obtained from literature and used as a continuous-valued outcome for developing a radiation sensitivity model (Chapter 6).

2.5 Data Modeling Techniques

Several data modeling techniques including data mining classifiers and statistical modeling techniques are described in the following sections. These techniques may be used at the feature selection stage and/or as the classifier for building gene expression signatures.

2.5.1 C4.5 Decision Trees

Decision trees are learning algorithms that employ a "divide and conquer" strategy [60] to create nodes at various levels of the tree. C4.5 [61] is a variant of the basic decision tree that uses information-gain as a measure of purity at each node. Information gain can be described as the effective decrease in entropy resulting from making a choice as to which feature to use and at what level. The entropy is computed as:

$$\text{entropy}(p_i) = -p_i \log(p_i)$$

where p_i = (# samples at node i)/(total samples at parent node)

The entropy of subsets created by splitting the samples on a feature value is compared to the entropy of the system prior to the split. The feature that yields the maximum information gain by splitting the dataset is chosen as the best split attribute. Thus a tree can be built up of decisions that allow navigation from the root of the tree to a leaf node by continually examining these split attributes (Figure 2-5). The USF implementation of C4.5 decision trees (C4.5 DT) is used for the models described here.

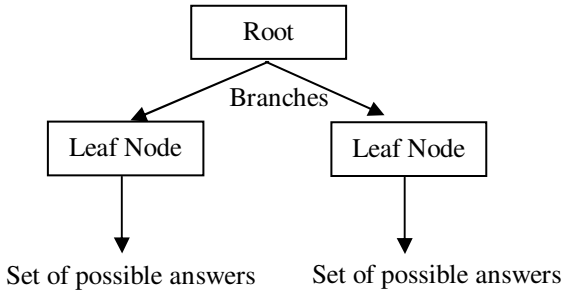


Figure 2-5: Structure of a decision tree

2.5.2 Feed-Forward-Back-Propagation Neural Network

A feed-forward-back-propagation neural network (NN) [62, 63] typically consists of an input layer followed by one or more layers of hidden units or computational nodes (Figure 2-6), ending in a layer of output nodes. The learning algorithm employs a forward and a backward pass of signals through the different layers of the network. The forward pass involves propagation of an input vector through the layers of the network, producing a response at the output layer of the network. An error signal is computed by subtracting the actual response of the network from a desired or target response and propagated backward through the network to make the actual response of the network closer to the desired response.

Quickprop, a fast implementation of the feed-forward-back-propagation network, is used for the models described in the subsequent chapters. The network is designed with 10 hidden units and 2 output nodes. The training of the classifier is designed to stop either when the training set error rate dropped to 0, or 500 epochs.

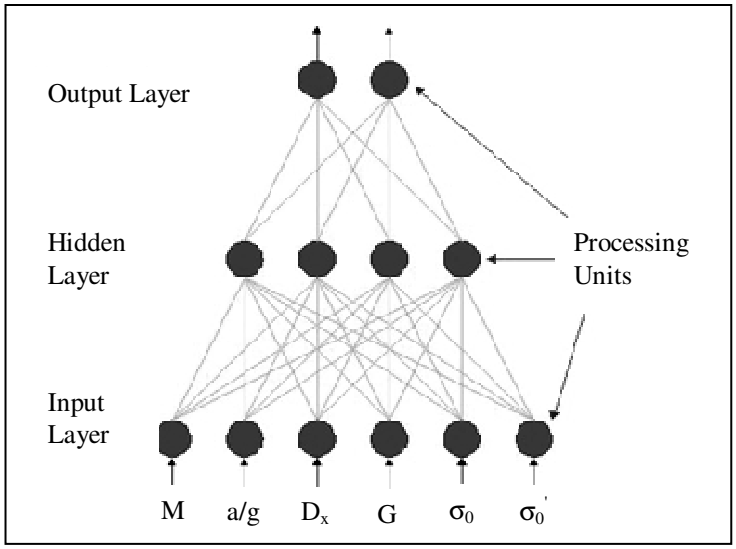


Figure 2-6: Architecture of a feed-forward-back-propagation neural network

2.5.3 Support Vector Machines

Support vector machines (SVM) use linear models to represent non-linear boundaries between classes [32, 64]. Input feature vectors are transformed into a higher dimensional space using a non-linear mapping. Hyper-planes are defined in this high dimensional space so that data from any two classes can be separated (Figure 2-7). The hyper-plane that achieves the highest separation of the classes, known as the maximum margin hyper-plane, generalizes the solution of the classifier and is completely defined by specifying the vectors closest to it, called the support vectors. The support vector machine implementation in WEKA [32] is used for the models described in the following chapters. A linear kernel is used with standard normalization.

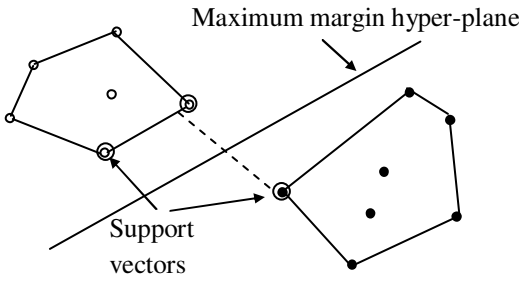


Figure 2-7: A maximum margin hyper-plane [69]

2.5.4 Linear Regression Analysis

Linear regression analysis [65] is a statistical technique used to model the relationship between a scalar outcome variable *y* and one or more covariates *x*. The model depends linearly on unknown parameters that are determined via regression techniques. The linear model is represented as:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}\,; \qquad \text{for } i=1, \ldots, N$$

where *N* is the total number of samples in the data, with *y* as the outcome variable that is linearly dependent on the covariates *x* via the model parameters β. Thus, the linear effect of each covariate x_i on the outcome is governed by the regression parameter β_i.

2.5.5 Student's T-Test

The Student's t-test is a popular technique used to test the difference in means of two groups of data [65]. The computation of the t-statistic, as shown in Figure 2-8, indicates the ease of distinguishing between two groups in presence of variability. The null hypothesis states that there is no difference in the means of the two groups. A p value

of less than the α-level (typically set at 0.05 or lower) indicates that the difference between the two groups is statistically significant thereby rejecting the null hypothesis.

$$t - statistic = \frac{signal}{noise} = \frac{difference\ between\ group\ means}{variability\ of\ groups}$$

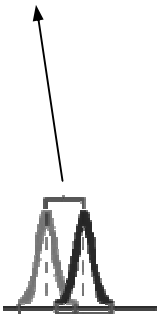


Figure 2-8: Formulation of the t-statistic for Student's t-test

2.5.6 Kaplan-Meier Survivor Estimates and the Log-Rank Test

The Kaplan-Meier product limit estimate (K-M) is used to compute the survival probabilities and the survivor curves for a cohort of patients [66, 67]. The product limit estimate computes the survival probability as:

$$S(t) = \prod_{j=1}^{t} \left[\frac{(n-j)}{(n-j+1)} \right]^{\delta_{(j)}} \quad ; n\text{: total number of cases,}$$

$$\delta(j) = \begin{cases} 1; & \text{if } j^{\text{th}} \text{ case is uncensored} \\ 0; & \text{if } j^{\text{th}} \text{ case is censored} \end{cases}$$

These estimates are used to draw survivor curves for each group of patients in the dataset. A log-rank (L-R) test [66, 67] is used to compare these curves for differences in survival. This test statistic is approximately chi-square distributed with one degree of freedom under the null hypothesis that the two K-M curves are statistically equivalent.

$$Log-rank-statistic = \frac{(O_i - E_i)^2}{Var(O_i - E_i)}$$

where O_i: observed score for the group *i* and E_i: expected score for the group *i*

The p value obtained at an α (e.g. 0.05) confidence level from the chi-square distribution tables is used to determine if the null hypothesis is rejected. A rejection of the null hypothesis indicates that the two curves are statistically different.

2.5.7 Cox Proportional Hazards Model

The Cox proportional hazards model [66, 67] (CoxPH) is a semi-parametric survival regression analysis technique used to model the effect of secondary variables or covariates on survival. The strength of the technique lies in its ability to model and test many inferences about survival without making any specific assumptions about the parametric form of the hazard or survival functions. For any two individuals with covariate vectors x_1 and x_2, the hazard ratio is specified by a constant of proportionality:

$$\rho(x) = \frac{\lambda(t, x_1)}{\lambda(t, x_2)} = \frac{\lambda_0(t)\exp(\beta' x_1)}{\lambda_0(t)\exp(\beta' x_2)} = \exp(\beta'(x_1 - x_2))$$

where $\lambda(t, x_i)$ is the hazard function for individual *i* with covariates x_i at time *t*. The hazard is interpreted as:

$\rho(x)=1$: $S(t,x) = S_o(t)$: no difference in survival between groups

$\rho(x)<1$: $S(t,x) > S_o(t)$: better survival than baseline

$\rho(x)>1$: $S(t,x) < S_o(t)$: worse survival than baseline

2.6 Validation Techniques

Validation techniques [68] are used to test the predictive performance of classifier models. A simple validation method used in this work is a hold-out procedure known as *n*-Fold-Cross-Validation (n-fold CV) that involves dividing the dataset into a fixed number of partitions (*n*). All but one partition are used to train the classifier and the left-out partition is used for testing. The training-and-testing procedure is repeated enough number of times (called folds) so that each partition is used as a test set exactly once. The 10-fold cross-validation setup shown in Figure 2-9 is used for validating classifier models in the subsequent chapters.

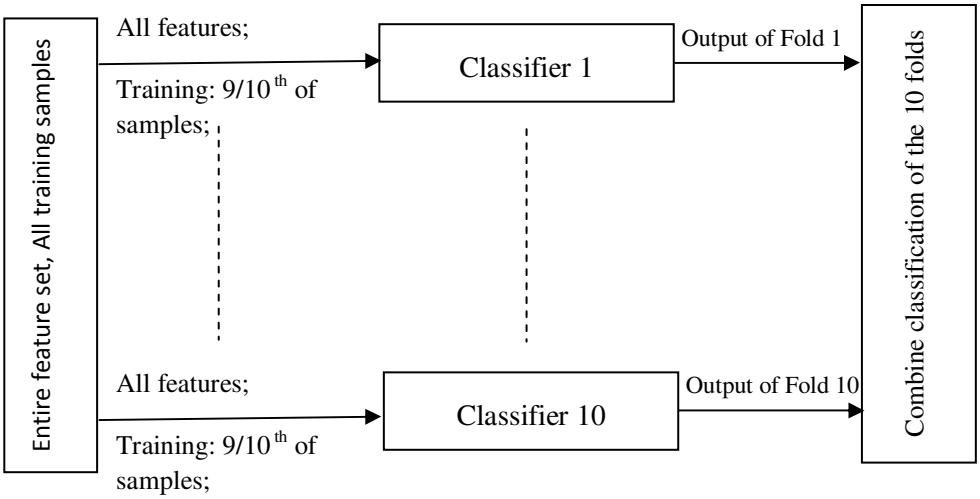


Figure 2-9: 10-fold cross validation setup

For two-class problems, the samples in each partition represents a proportional selection of samples from all the classes under consideration to ensure that the classifier learns all the classes equally well, and is not over-trained on any one class.

2.6.1 Performance Measures

Performance measures [68] are used to determine the expected prediction accuracy of a classifier model on independent test samples. The choice of the measures used depends on the basic construct of the problem definition (see Section 2.3.4).

The predictive performance of a classifier for a two-class problem can be analyzed by means of a confusion matrix (Table 2-1), and the performance measures listed in (Table 2-2). Although the total accuracy of prediction is commonly used as a preliminary measure of performance, computation of a weighted accuracy is useful in datasets with unbalanced class distributions. Measures of sensitivity and specificity [69] are popularly used to gauge performance when dealing with clinical data. Sensitivity is defined as the true positive rate and specificity as the true negative rate for the classifier. Since the weighted accuracy reports the average of these rates it may be used as a convenient measure to evaluate the performance of the classifier.

Table 2-1: Confusion matrix

Classified As / **True condition**	**Class A (positive)**	**Class B (negative)**
Class A (positive)	True Positive (*a*)	False Negative (*b*)
Class B (negative)	False Positive (*c*)	True Negative (*d*)

Table 2-2: Performance measures for two-class problems

Performance measure	Formula used
Total accuracy	$\frac{a+b}{a+b+c+d}$
Weighted accuracy	$\left(\frac{a}{a+b}+\frac{d}{c+d}\right)/2$
Sensitivity	$\frac{d}{c+d}$
Specificity	$\frac{a}{a+b}$

Continuous-valued predictions are obtained for statistical techniques that model continuous-valued problems. Such outcomes can be assessed for predictive ability in two ways. In the first approach, the continuous-valued predictions are split into two classes based on some threshold. Statistical tests such as the Student's t-test or K-M curves and L-R test may be used to determine if these groups are significantly different. For example, the survival estimates obtained from the CoxPH model may be split on the median value to obtain two groups. K-M survivor estimates and L-R test can be used to test the difference of survival between the two groups.

The second method used in the following chapters to determine model performance is a goodness-of-fit test for a specific statistical technique. This is measured by a model R^2 value. The high values of R^2 indicate good correlation of the modeled covariates with the outcome.

2.7 Summary

A basic overview of the process for developing molecular signatures was provided, from a description of cancer to the data analysis and statistical tools used in the

subsequent chapters. Two different ways to state a problem definition for experimental design were described. These included splitting the data into discrete number of classes, or modeling the outcome as a continuous variable. Three different microarray gene expression datasets were described and multiple outcomes for study were outlined for two of these datasets. Several data analysis models were described that may be used for feature selection as well as classifier design. Finally, validation techniques for assessing the predictive performance of both types of problem definitions were discussed.

CHAPTER 3 MEASURING THE CLASSIFICATION COMPLEXITY OF GENE EXPRESSION DATASETS

3.1 Introduction

Microarray gene expression signatures that are predictive in the training datasets in which they were developed can perform poorly when tested on samples from independent sources [57]. Further, classifier models that generate predictive signatures in certain datasets can fail to be as predictive on other datasets [58]. Although methodological mistakes can lead to poor estimates of signature accuracy estimates [17], even correctly developed classifiers suffer from this difficulty. This chapter shows that the inherent complexity of gene expression data can limit the ability of classification schemes to generate accurate signatures.

A case study is presented in Section 3.2 to illustrate the behavior of classifiers on complex datasets. Data complexity is explored in detail in Section 3.3 and three specific quantitative measures of complexity are proposed in Section 3.4. The need for internal controls in a dataset, and a method to establish the control is outlined Section 3.5. The proposed measures of complexity are applied to datasets in Section 3.6 and discussed in Section 3.7. Finally, a methodology is outlined in Section 3.8 to assess the complexity of a dataset given a problem definition and a classifier model.

3.2 Case Study: Survival Analysis of MRC-CRC and NSCLC

The MRC-CRC dataset (n = 121) was analyzed as a two-class problem to generate a predictive signature for patient survival (MRC-CRC/Survival). The survival times were stratified into high risk (less than 36 months of survival, n = 37) and low risk (greater than 36 months of survival, n = 84). Features with smallest Student's t-test p values were selected to train the three classifiers (C4.5 DT, SVM and NN) and classification accuracy was estimated using 10-fold CV. Figure 3-1 shows that the best weighted accuracy of the classifiers was found to be 56%, indicating survival prediction was only slightly better than chance. Since the main aim was to build accurate predictive signatures, a further refinement of the models was required. However, refining a classifier model significantly to obtain high prediction accuracies even within a 10-fold CV can lead to over-training of the classifiers [58, 68]. Such over-trained classifiers rarely perform with the expected prediction accuracy on independent datasets.

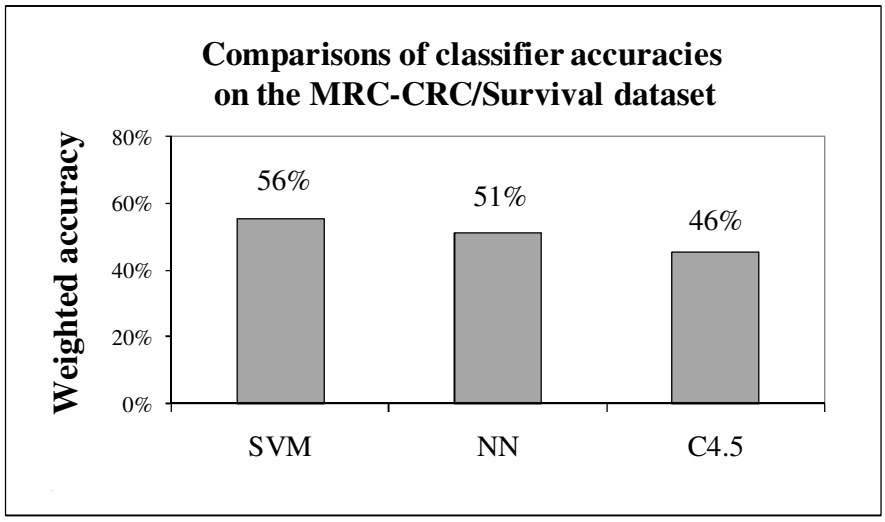


Figure 3-1: Classifier accuracies for MRC-CRC/Survival dataset

To explain poor classification accuracy, it was hypothesized that the classifiers chosen were ineffective in modeling the underlying characteristics of survival in the data. The survival outcome is inherently a difficult problem to model due to the lack of complete follow-up information as well as confounding factors such as age, existing medical conditions and other physiological parameters. To determine the ability of the three classifiers (C4.5 DT, SVM and NN) to model a survival outcome, the same techniques were used to predict survival for a different dataset (NSCLC). A previously published K-M survival model was shown to be highly predictive for this dataset [26]. In this work, the dataset (n=62) was transformed into a two-class problem using a cut-off for survival time of 30 months (median survival time). Patients who died within 30 months were considered poor prognosis (n = 20), otherwise they had a good prognosis (n = 42) (NSCLC/Survival).

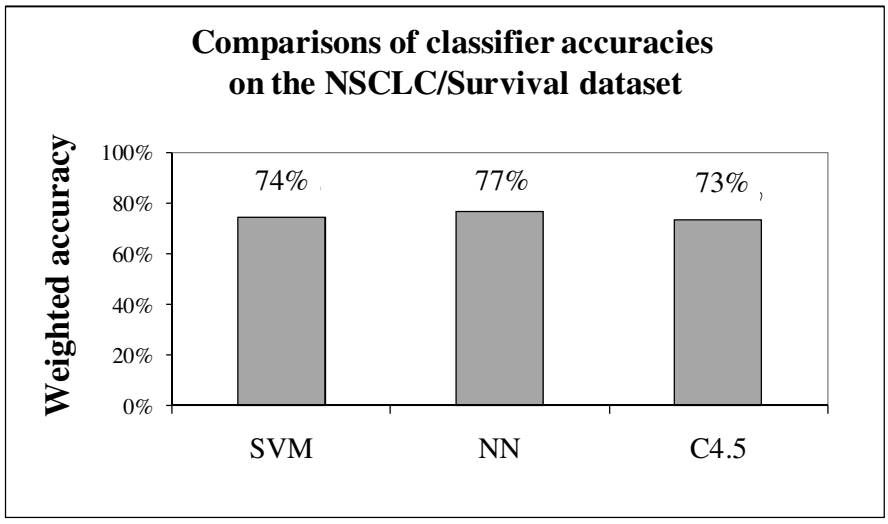


Figure 3-2: Classifier accuracies for NSCLC/Survival dataset

Figure 3-2 shows that the classifiers were able to predict the survival outcome with an accuracy of 77%, suggesting that the classifier models can be reasonably

effective in modeling two-class survival data. Thus, it was hypothesized that the poor accuracy in the MRC-CRC/Survival dataset must result from some intrinsic property of the dataset.

3.3 Data Complexity

A key issue in gene expression studies is to determine if a predictive signature can be developed for a dataset given an appropriate classification method. The case study illustrates two difficulties in choosing classifier models for gene expression analysis. The first difficulty is that a classifier method that is shown to work well in one dataset may not yield satisfactory results in a different dataset. The second difficulty is that for a given classification problem one type of classifier may outperform other types of classifiers, as shown in Figure 3-1 and Figure 3-2. This was also shown in the No Free Lunch Theorem [70]. In some cases, this difference is explained by the decision boundary created by the classifier to separate the defined classes. For example, a decision tree creates only axis-parallel decision boundaries while neural networks can create arbitrary boundaries [62, 63]. Another possible cause for these issues is that the data itself imposes a limit on the classification accuracy that can be obtained from any classifier [58]. This may be due to several reasons including noisy data, omission of informative variables, incorrect assignment of the examples into specific classes or perhaps an incomplete understanding of the ground truth. Unfortunately, each of these problems is prevalent in gene expression studies, particularly when the tissue originates within humans. This work proposes and develops quantitative measures of data complexity to estimate an upper limit on classification accuracy for a dataset given a classification method.

3.3.1 Example: Intrinsic Heterogeneity in Datasets

Figure 3-3 illustrates the problem of classification complexity in a heterogeneous dataset with a simple example. The dataset consists of two variables: *Color* (*Black* or *White*) and *Pattern* (*Solid* or *Stripes*). *Size* (*Big* and *Small* shapes) is used as the outcome.

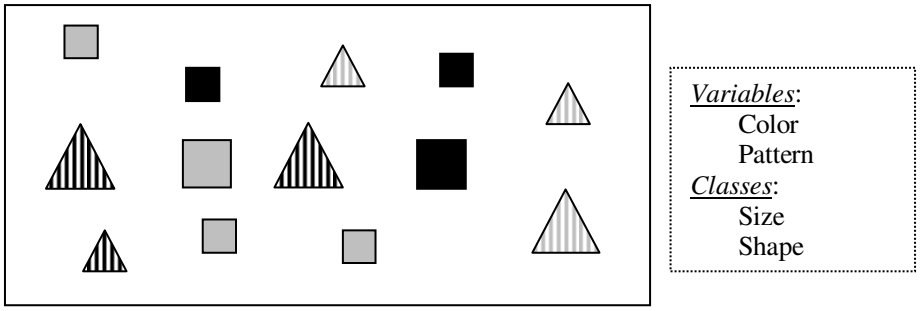


Figure 3-3: Example of a heterogeneous dataset

Since there is no direct correlation between *Color* and *Size*, or *Pattern* and *Size*, classifier models using these variables will be inaccurate (see Figure 3-4). However, defining a second outcome, type of *Shape* (*Squares* or *Triangles*) can lead to a more trivial grouping of the samples, and accurate classifiers can be created using *Pattern* as a variable (Figure 3-4). Here, the samples are perfectly split into the two classes: all striped shapes are *Triangles* and all solid shapes are *Squares*. Thus, in a heterogeneous dataset, the problem definition can have an impact on the complexity of a classifier model. The definitions that are easy to model tend to yield simple classifiers, while other questions can lead to complex classifier boundaries.

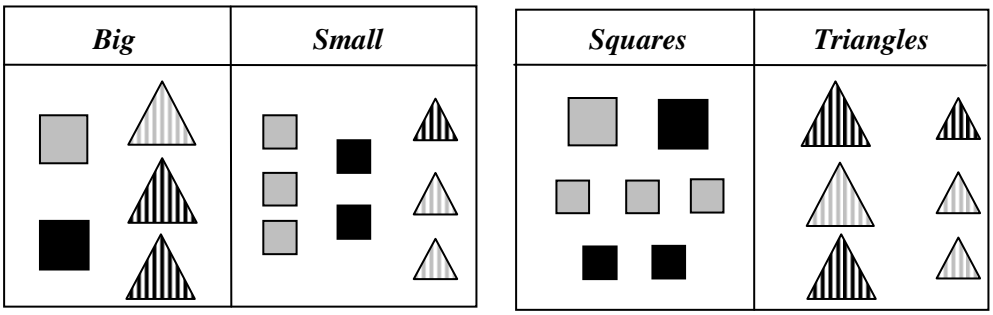
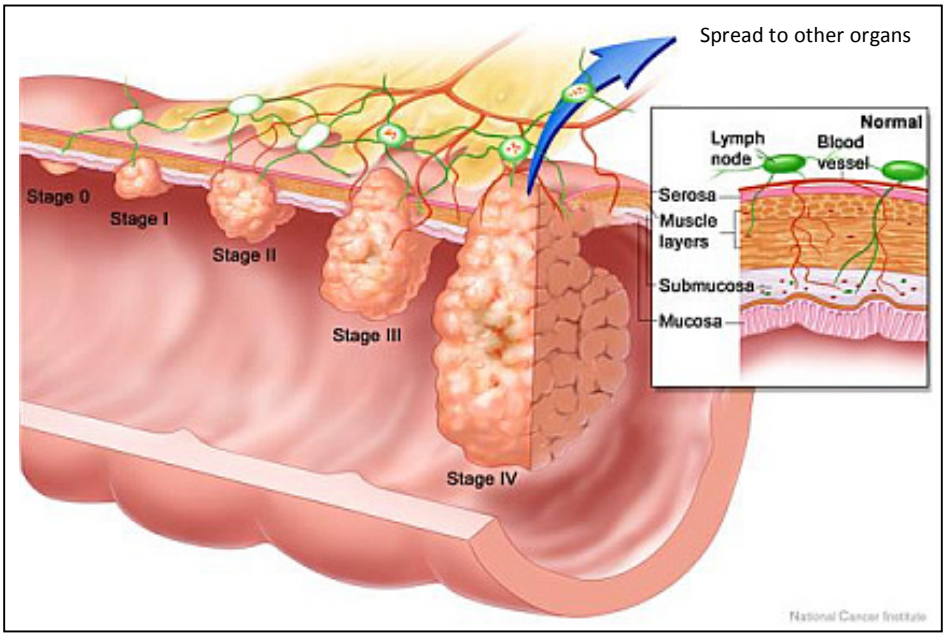


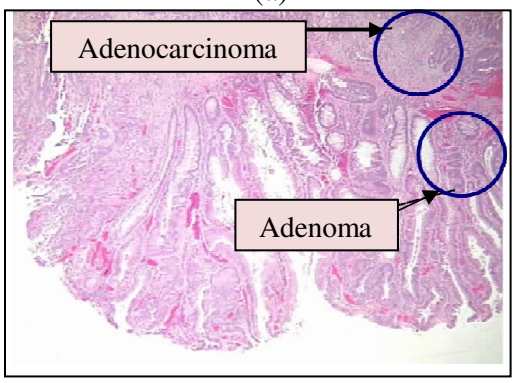
Figure 3-4: Examples of two possible classifications. (a) Dividing samples by *Size* (b) Dividing samples by type of *Shape*

3.3.2 Example: Heterogeneity from Sampling Process

Inherent biological characteristics of tissue samples can introduce heterogeneity within gene expression signatures of cancer. An example may be observed in solid tumors such as colorectal cancer. Figure 3-5a shows that the colon tissue is composed of several different layers of epithelial cells surrounded by connective and muscle tissue, and the specific composition of a tissue sample changes based on the location of the tumor in the colon or rectum. Inconsistent extraction of tissue across samples can lead to microarray datasets with a mixture of cell types with varying proportions [71] and introduces signatures that are inherently different. The task of classifying samples for a specific biological question has to then overcome the distinction between the basic tissue types to find more subtle differences. In the worst case, the sample may consist of non-malignant cell types. This sample may be erroneously labeled as tumor, along with the clinical factors that are attributed to the patient, such as age, stage of tumor, surgery and overall survival time. A classifier model using this as a training sample could generate an inaccurate model.



(a)



(b)

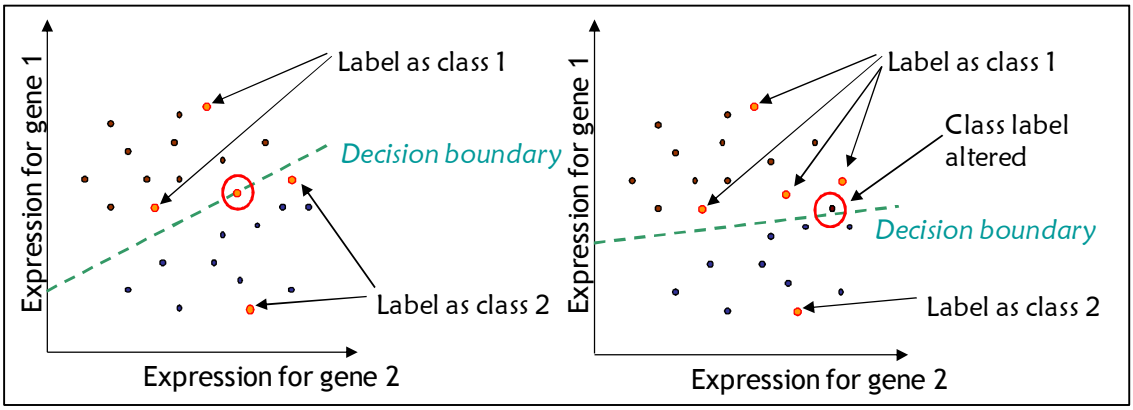
Figure 3-5 :Cross-section of colorectal tumor. (a) At different stages of development (Image courtesy of http://www.cancersociety.com/cancer_information/colon.html) (b) Surrounded by adenoma and normal tissue

3.3.3 Other Examples of Heterogeneity

Other less obvious differences in samples such as age of the patient, gender, smoking history or ethnic background may introduce further heterogeneity in the data that may not be modeled by the classifiers and may confound the analysis. Errors in documenting these factors can exacerbate the problem. The microarray experiment itself could introduce some heterogeneity in the final data due to differences in processing

conditions or book-keeping errors [71]. Image analysis and normalization of the data in the final processing steps can add to the noise or introduce undesirable signals into the dataset [72].

Figure 3-6 depicts the impact of sample mislabeling on a classifier decision boundary. A test sample is assigned to a specific class depending on the side of the decision boundary it lies on. Changing the class label of a single training sample from Figure 3-6a to Figure 3-6b (circled in red in Figure 3-6b), alters the decision boundary such that four of the five test samples are labeled as Class 1 in Figure 3-6b when only two of the test samples were labeled as Class 1 in Figure 3-6a.



(a) (b)

Figure 3-6: Understanding the impact of sample mislabeling on classifier decision boundaries. Training samples for the two classes are represented as filled (Class 1) or hollow (Class 2) samples. Test samples are depicted as orange circles in the graphs

If the actual class labels for the test samples were known in advance, it is relatively easy to determine which of these two decision boundaries is more accurate. However, since in a practical situation there is no way of knowing the actual class of test samples in advance, there is little guiding information on whether a chosen classifier

model is performing poorly on the data because it is an ineffective choice for the data, or if the data itself is imposing an upper limit on the accuracy that can be obtained. Further, if the source of the mislabeling is unknown, there may be little chance of rectifying the error.

3.4 Measures of Classification Complexity

As discussed in the preceding sections, intrinsic complexity of gene expression datasets can affect classifier performance. Three specific measures are formulated here to quantitatively evaluate the complexity of these datasets and provide insight into the expected classifier performance for a given problem definition.

A biological basis for measuring the complexity of a classification problem is that large numbers of gene expression changes occur in significantly different tissue types. For instance, the differences in gene expression between epithelial tissue of the colon and surrounding connective tissue are dramatically high; correspondingly, tissue type-based signatures have been demonstrated to be accurate and robust (e.g. [12]). This rationale suggests consideration of the number of genes differentially expressed across two classes as a measure. Separately, large changes in individual gene expressions (e. g. 5 fold differences) between classes can be indicative of functional or morphological differences within cells; therefore the maximum univariate gene discrimination can also be considered. The following sections describe the proposed measures of classification complexity.

3.4.1 Complexity Measure I: Student's T-Test: τ

The first measure of complexity (τ) is defined as the proportion of genes identified as significant when considered using a Student's t-test. P values (p_i) are calculated for the difference in gene expression between two groups.

The t-statistic represents the ease of distinguishing between two groups in the presence of inherent variability in the data or noise in measurement. A p value of less than the α-level (typically set at 0.05 or lower) indicates that the difference between the two groups is statistically significant thereby rejecting the null hypothesis. The assumption for this measure is that if a dataset has a large number of features that are significant, an accurate classifier model may be designed. The larger the number of significant features, the less complex the classification problem. Thus τ can be formally defined as shown below, where s represents the number of features tested.

$$\tau = \frac{|p_i < 0.05|}{s}; \qquad i = 1, \ldots, s$$

3.4.2 Complexity Measure II: Fisher's Discriminant Ratio: ϕ

A second complexity measure (ϕ) is based on Fisher's discriminant ratio that was used in [58]. The ratio measures the separation of two classes, adjusted by the spread of the samples in each class. The method is primarily used to find an axis in the feature space along which the separation of the two classes is a maximum [73, 74]. The samples are then projected onto this axis for classification.

The Fisher's discriminant ratio is computed using each feature univariately to determine the separabilty of each feature. The ratio is defined as given below, where μ_{1i};

μ_{2i}; σ_{1i} and σ_{2i}, are the means and variances of the two classes for feature *i* and *s* is the number of features tested. The proposed summary complexity measure, ϕ, is the maximum ratio obtained from the dataset.

$$F_i = \frac{(\mu_{1i}-\mu_{2i})^2}{(\sigma_{1i}^2+\sigma_{2i}^2)^2}; \qquad i = 1 ,..., s$$

$$\phi = max(F_i)$$

Higher ratios indicate better separation between the classes for the selected feature. Since the first step in gene expression analysis selects features with good discrimination between classes, considering the maximum separation of a single gene provides an upper bound on classification.

3.4.3 Complexity Measure III: SAM π_0

A third measure of complexity is π_0, an estimate of the number of unchanged (true null) features in a series of statistical tests. This measure is used by the SAM (Significance Analysis for Microarrays) [15, 75] algorithm. The samples are repeatedly shuffled around by permuting their class labels (or response states) and the statistic is computed for each permutation. SAM identifies genes as significant when they change stably and significantly with a minimum pre-specified change in expression level across the repeated measurements.

The overall error rate is summarized by a measure π_0 that estimates the probability of erroneously rejecting the null hypothesis. π_0 is specified for a rejection region, (e.g. α

= 0.05 or lower) and is computed as the proportion of features with p values that fall in this rejection region, normalized by the range of the region.

$$\pi_0 = \frac{|p_i > \gamma|}{(1-\gamma)s}; \qquad i = 1, \ldots, s$$

Thus π_0 indicates the proportion of features whose values do not change between the classes. As stated earlier, a dataset with classes that have strong differences is expected to have a small proportion of features that are not associated with the outcome. Thus an increase in π_0 values from one outcome to another on a specific dataset can be used as an indication of increasing complexity of classification.

3.5 Internal Controls

While a universal measure of classification complexity is desirable, individual datasets may have different baseline complexities. One approach to alleviate this concern is through the use of internal controls within each dataset. Identifying different outcomes (e.g. gender, staging, and patient survival) that are believed to be more or less complex within the same set of samples provides an internal control on the measurement of complexity. This approach also allows for a normalization factor in the form of a similar outcome across datasets.

Identification of gender from gene expression data is an example of a low-complexity classification problem, in particular when the dataset includes gender-related features. Gender of an individual is indicated in the chromosomal composition of the cells. Males have one X and one Y chromosome, while females have two X

chromosomes. Since the chromosomes define the genetic composition of the cells, male and female samples are likely to have strong differences in the expression of gender related genes. Further, secondary effects of gender, such as differences in hormonal levels, can also be measured at the genetic level. If the features measured in the dataset include genes associated with gender, then distinguishing between males and females becomes a straightforward classification problem.

3.6 Assessing the Complexity of MRC-CRC and NSCLC Datasets

The MRC-CRC and NSCLC datasets were used to measure complexity and these measures were compared with prediction accuracies of survival models. Gender was used as the internal control for each dataset: (MRC-CRC/Gender: Male: n=59, Female: n=62 and NSCLC/Gender: Male: n=25, Female: n=37). An additional outcome was specified for each dataset to provide a further data point for assessing the complexity measures. Tissue type was defined for MRC-CRC dataset (MRC-CRC/Site: Colon: n=85, Rectal: n=36). Stage was defined for the NSCLC dataset. (NSCLC/Stage: I: n==49, III: n=13).

The three measures of classification complexity were computed for each dataset and compared against the best weighted accuracy classifier for each problem, regardless of classifier type. These results were published in [76]. Table 3-1 details the complexity measures and accuracies for each problem. As expected, for both MRC-CRC and NSCLC the gender classification achieved the highest accuracy (98% and 95% respectively) compared to the other problems specified for each dataset. However, both τ and π_0 measures indicate relatively few features that are significant in these problems. In retrospect this result is not surprising, since the genes associated with gender are often

very distinct (e.g. absent or present) but may not be numerous. Although the number of significant genes may be a sufficient measure of complexity, it is not a complete measure. The complexity measure ϕ estimates the best univariate separation in the data, and hence is a more reasonable measure of expected classification accuracy.

Table 3-1: Classification complexity and classifier accuracy for the MRC-CRC and NSCLC datasets

Dataset	τ	ϕ	π_0	Best weighted classifier accuracy (%)
MRC-CRC/Gender	7.8	9.98	0.93	98.4
MRC-CRC/Site	12.3	0.64	0.76	77.7
MRC-CRC/Survival	11.8	0.33	0.75	55.5
NSCLC/Gender	4.9	2.47	0.98	95.3
NSCLC/Stage	13.2	1.10	0.88	87.4
NSCLC/Survival	8.8	0.75	0.96	76.8

Table 3-1 also reports the complexity for the two additional problems (MRC-CRC/Site and MRC-CRC/Survival; NSCLC/Stage and NSCLC/Survival) for the two datasets, along with the corresponding maximum classifier accuracy. Again two measures of significant genes (τ and π_0) do not reflect the differences in accuracy that are observed. For instance, in the MRC-CRC/Site and MRC-CRC/Survival datasets, the differences are small in τ and π_0 however there is almost a 20% difference in best accuracies between the two problems. For example, in the MRC-CRC dataset, τ is 12.3% for Site and 11.8% for Survival and π_0 is 0.757 for Site and 0.750 for Survival; however the classifier accuracies are very different (78% for Site and 56% for Survival). Note that despite the differences for these two outcomes, the trend in accuracy vs. complexity is maintained: accuracy drops as fewer features are found to be significant.

The complexity measure ϕ captures the classification complexity better than the remaining two measures in this data. However, it can be seen from the results for gender and survival outcomes that the measure of complexity is not directly comparable across datasets. Thus, the internal control for each dataset is required to provide information on the maximum attainable classifier accuracy.

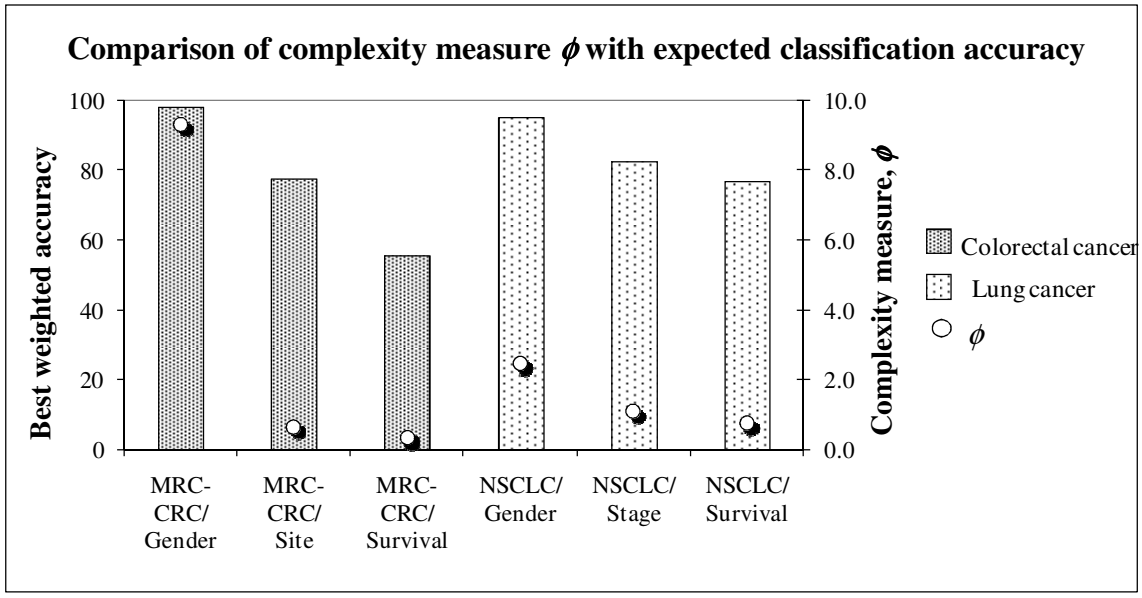


Figure 3-7: Classification Accuracy vs. Complexity measure ϕ

Of the three measures tested, ϕ correlated highly with maximum classification accuracy (R^2 for correlation is 0.82 for the MRC-CRC dataset, Table 3-2). Figure 3-7 provides a detailed view of the classifier accuracy (bar) and complexity measure (point) for each of these datasets. The equation for correlation can be used to estimate an upper bound on the expected classifier accuracy using the specified classifier models for any new outcome on the dataset.

Table 3-2: Correlation of complexity measures with classifier accuracies

Dataset	Correlation coefficient for comparison of complexity measures with classifier accuracy (R^2)		
	τ	ϕ	π_0
MRC-CRC	0.58	0.82	0.69
NSCLC	0.53	0.99	0.16

3.7 Discussion

Genes with large univariate differences can aid in achieving high classifier accuracies. Examples of such differences are gender related genes that are present or absent in each class. If a large number of these genes are available, the classifier accuracy can be expected to be very high. However, with such large distinctions, even a small set of genes is sufficient to create an accurate classifier. In such a case, the number of distinct genes may not provide much information on the expected performance. However information on the largest separation between the classes can provide insight into the quality of a classifier decision boundary. In datasets where such large distinctions are not available, the best univariate separation between the classes can provide an indication of the classifier performance.

Figure 3-8 depicts complex datasets with multiple probesets. In case 1 with a single probeset, all the complexity measures provide the same information. When more genes are added to the model, the measures provide slightly different types of information. In case 2, where Gene 1 has a reasonable separation between the samples and Gene 2 has very poor separation, values of τ and π_0 indicate that a decision boundary can be found (here, $\tau = 50\%$).

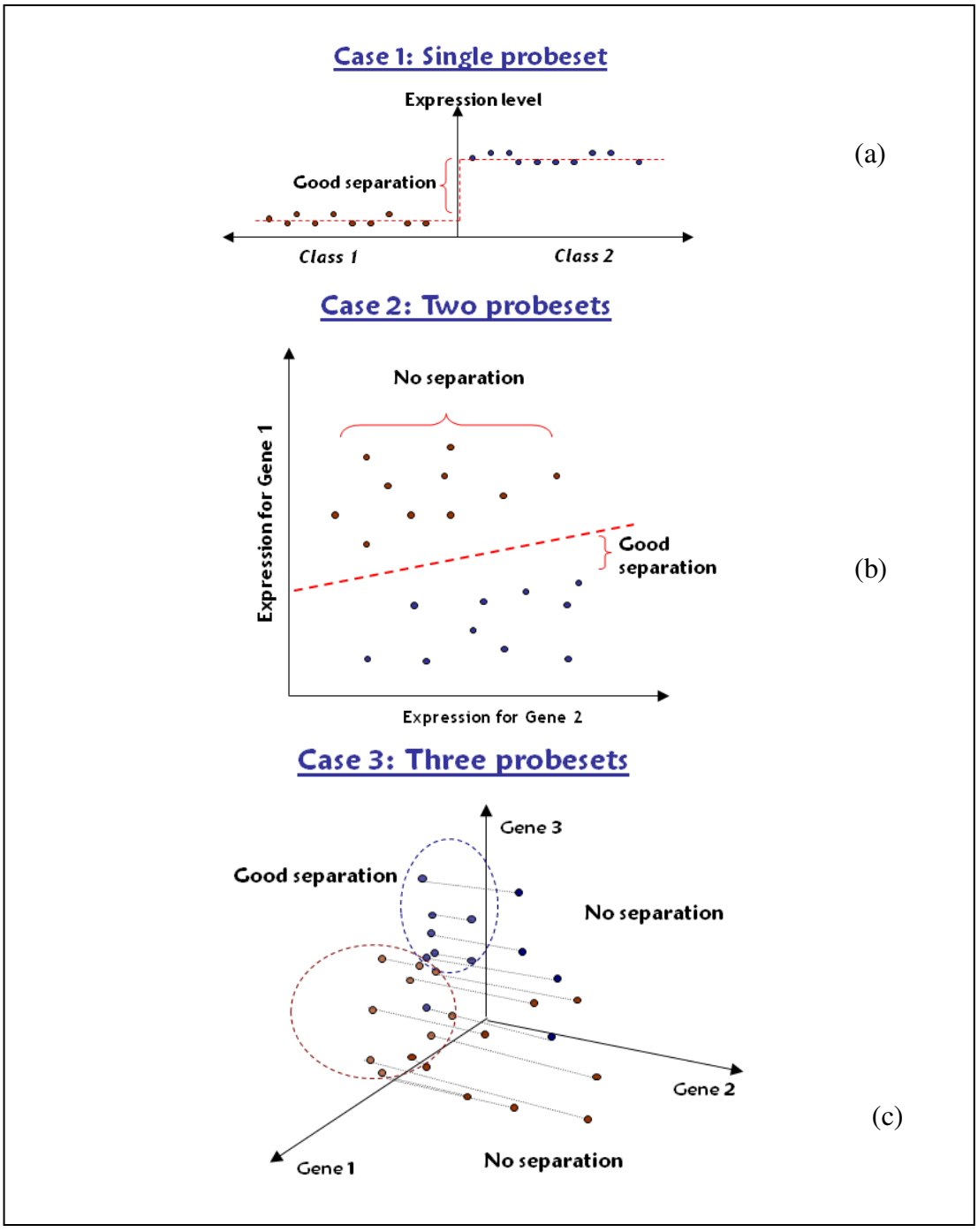


Figure 3-8: An example to demonstrate the applicability of the complexity measures. # significant features is a useful measure, but ϕ provides a measure of the maximum univariate separability

The value of ϕ is unaffected by the addition of the variable Gene 2 in the model, and will provide the same measure of separation as case 1. As indicated by the complexity measures, the separation provided by Gene 1 allows for a decision boundary between the samples. Similarly, Gene 1 and Gene 2 in case 3 do not have a good separation but Gene 3 does. Here, τ and π_0 will be lower ($\tau = 33.3\%$), but ϕ still measures the separation provided by Gene 3, indicating that a predictive classifier may be created. This supports the conclusion that complexity measure ϕ is a good indicator of the complexity of a dataset, and provides an estimate of expected classifier performance.

3.8 A Method to Assess the Classification Complexity of a Microarray Gene Expression Dataset

The case study and experimental results indicate that when using the t-test for feature selection in the classifier model, the complexity measure (ϕ) provides a reasonable estimate of the complexity across different outcomes. The table below outlines the proposed steps for evaluating classification complexity.

Step 1: Establish an internal control for the dataset using covariate information to define the easiest classification problem. Ensure that the gene expression dataset contains at least a few relevant probesets for this problem. For example, gender will not work as an easy problem if the microarray chip contains no probesets for the primary or secondary aspects of gender.

Step 2: Define one or more additional problems for the dataset, if possible. For example, tissue site was used as an additional outcome for the MRC-CRC

dataset. As before, each outcome defined here must have relevant probesets in the data.

Step 3: Compute the complexity measure on all the outcomes proposed.

Step 4: Compute the correlation to describe the change in complexity across the different outcomes as a function of classifier accuracy.

Step 5: Given the maximum classifier accuracy for the defined outcomes, the maximum expected classifier accuracy for a new outcome can be estimated from the correlation.

Step 6: If the classifier performs much worse than the predicted accuracy, a further refinement of the method is warranted. Else, the classifier model is shown to perform as well as it possible can on the dataset. In this case, the method recommends investigation of the intrinsic properties of the data before refining the model further.

Consider the case study presented in Section 3.2. The survival model for the MRC-CRC/Survival dataset had very poor accuracy. To investigate the reason for this lowered accuracy, two outcomes were defined using the same classifier method. MRC-CRC/Gender was used as the internal control and MRC-CRC/Site was used as a problem with medium level of difficulty to provide more information on the complexity of the data. ϕ was very high (9.98) in the gender outcome, with a correspondingly high classifier accuracy (98.35%). The low classifier outcome for survival (56%) was found to correspond to the low value of ϕ (0.33). This result indicates that the low accuracy in the survival outcome is a result of the inherent complexity in the data and further refinement

of the classifier models will not aid in improving accuracy without a severe loss of generality. Here, the data may be investigated further to reduce the inherent complexity by some means. Alternately, other classifier models may be tested that can deal with the complex feature space in a more efficient manner. An example of this will be provided in Chapter 5.

3.9 Summary

Data complexity was proposed as a means to explain the classifier performance on two gene expression datasets (the MRC-CRC and NSCLC datasets). Three methods of quantitatively measuring classification complexity in gene expression data were proposed. The sources of data complexity were explored and used to propose three measures of complexity (τ, ϕ and π_0). Experimental results were used to compare the complexity of microarray gene expression datasets with maximum achieved classification accuracy. Correlation of these measures with classifier performance was used to determine the usefulness of the measures. In this study, outcomes with larger π_0 or lower τ and ϕ values tended to have lower overall classification accuracy except in the case of gender classification where a strong signal exists in a small number of genes. The complexity measure ϕ was shown to have a clear relationship with classifier accuracy. A methodology was proposed to assess the complexity of a dataset given a problem definition and a classifier model.

CHAPTER 4 REDUCTION OF DATA COMPLEXITY FOR GENE EXPRESSION MODELS USING QUANTIZATION

4.1 Introduction

As discussed in Chapter 3, the intrinsic heterogeneity of samples in a microarray gene expression dataset can lead to complex classifier models with low predictive accuracy. The large number of features available in a typical microarray experiment, along with the high resolution of the data can lead to complex decision boundaries. For instance, the MRC-CRC data contains 54,675 probesets, each taking a value of 0.0 to 15.0 with a 6-digit precision. Methods to reduce these sources of data complexity aim at creating simpler classifier models and extracting predictive signatures. This chapter explores the use of quantization of gene expression datasets for data reduction.

Section 4.2 describes the MRC-CRC dataset as a case study of a complex dataset. Reduction of data complexity by quantization is discussed in Section 4.3. Three different methods of data quantization are presented in Section 4.4, followed by results of their application to the MRC-CRC dataset in Section 4.5. Finally, the modified complexity of the quantized datasets is examined in Section 4.6.

4.2 Case Study: Survival Analysis of MRC-CRC Dataset

Chapter 3 described some of the sources of heterogeneity in gene expression datasets and demonstrated the impact of complexity on the predictive accuracy of classifier models on MRC-CRC and NSCLC datasets. The MRC-CRC dataset was

analyzed as a two-class survival analysis problem to generate a predictive signature for patient survival (MRC-CRC/Survival) and the best weighted classifier accuracy was found to be 56%.

The complexity measures (τ, ϕ and π_0) proposed in Chapter 3 indicated that the MRC-CRC dataset contained significant information, as demonstrated by the high predictive accuracies on the MRC-CRC/Gender problem. However, re-organizing the samples to setup the survival problem resulted in a higher complexity dataset. The measures indicated that higher accuracies were probably not attainable on the dataset in the original form for the classifier models considered. Two steps were recommended to generate predictive signatures from the data: first, the data had to be refined in some way to reduce the complexity and second, a better classifier model could be designed to address the characteristics of the underlying class information.

4.3 Reduction of Data Complexity

It was shown in Chapter 3 that when studying gene expression datasets with a relatively heterogeneous cohort of samples, small differences in gene expression could be lost in the experimental and biological level of variability (or noise) in the data. For example, when studying the effects of a drug on a cohort of cancer patients, gene expression differences due to secondary aspects of the study such as gender, age or race may in fact be more prominent than the primary effect of the drug. The high resolution of the data, relative to the expected effect size, can further add to the complexity of the data. To efficiently extract the drug's effect in this example, the relevant differences in gene expression between samples of different classes need to be magnified relative to the

minor differences between samples of the same class. In noisy datasets, this magnification of the signal may reduce the complexity of analysis. Feature selection is a popular approach to achieve this magnification of important gene expression differences [48, 50, 51, 77]. Retaining only a small set of informative features aids in simplifying classifier boundaries (see Section 2.3.3). A complex algorithm is more prone to being fine-tuned to the specific dataset and often fails when applied to newer samples [58, 68]. Thus, most algorithms incorporate some mechanism of limiting the noise in a dataset to improve the accuracy of analysis and to build robust models for prediction [13, 14, 25, 34, 77].

4.3.1 Quantization to Reduce Data Complexity

One approach to reducing data complexity is to enhance the contrast within the dataset by altering the individual expression levels either at the probeset level or for the data as a whole [78]. In general, this approach aims at magnifying the differences between distinct groups of samples. Small differences in expression consistent with the sample grouping are magnified along with the larger and more pronounced differences and hence can contribute to the analysis more effectively [79].

One way to achieve this contrast enhancement of the data is quantization of the continuous gene expression data into a distinct number of levels [80]. For example, when working with a single gene to separate samples into two classes, having more than two levels for expression could potentially render the analysis complex. This issue is illustrated in Figure 4-1, where the variable *Color*, with three distinct levels (one level for each shade of gray) is used to separate the data into the two classes of *Shape*

(*Squares* and *Triangles*). While using the three levels of *Color* yields a complex classifier, quantizing the variable to two levels (*Light* - white and light gray and *Dark* - black), as shown in Figure 4-2, can yield a simpler classifier for prediction.

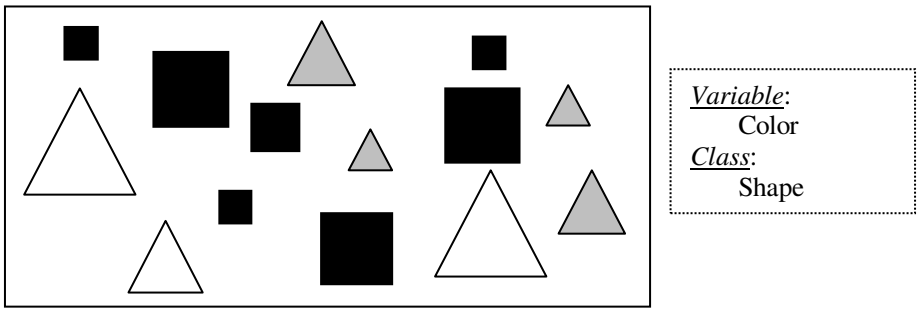


Figure 4-1: Example of a two-class dataset with multiple levels for a feature

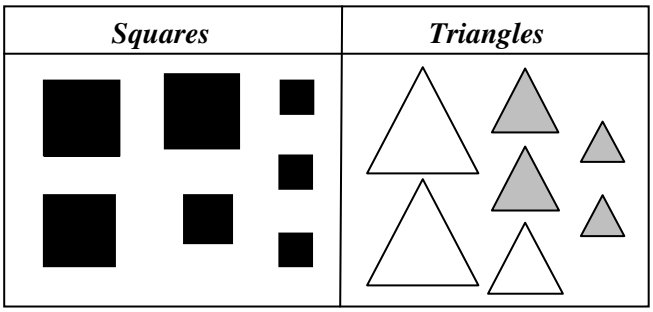


Figure 4-2: An example of quantizing a feature from three levels to two levels to represent a two-class problem

4.4 Quantization Techniques for Microarray Data

Several techniques exist in literature for quantizing gene expression data into meaningful levels to aid in improving the accuracy of subsequent analysis. In [79], the authors proposed the use of clustering techniques to find a natural grouping of expression levels in the data. The probesets were reassigned values based on their group membership. Parametric analysis of the data has also been used in a similar manner to

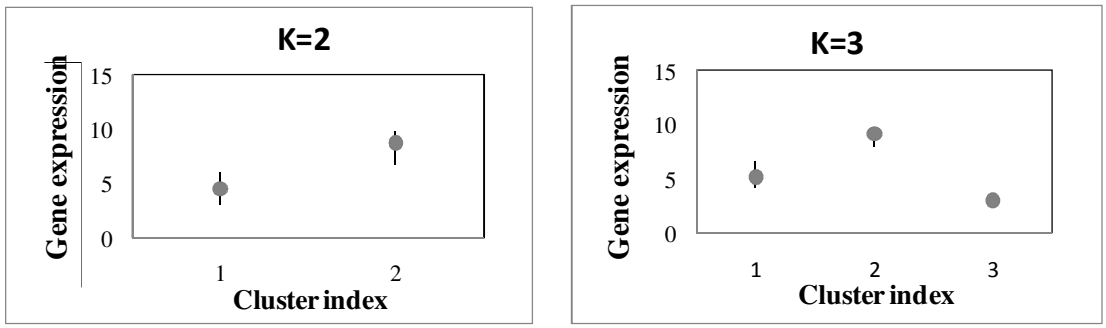
find overlapping Gaussian distributions that described the spread of the data [78]. Each of these methods were described in the context of specific analysis such as finding genes that were turned "on" or "off" in different tissue types. However, these techniques have not been applied to understand if resulting classification accuracy is altered as a result. Modifications of some of these techniques are proposed in the following sections to be more suitable for developing cancer-related signatures.

4.4.1 K-Means Clustering

K-means clustering estimates the number of groups that exist within a given dataset [81]. When the number of groups the data is known in advance, the method is straightforward to use. However, the technique also proves to be useful in exploring the types and numbers of sub-groups within a cohort of samples. Here, each probeset is analyzed separately using varying values of *K* to indicate the number of possibly distinct groups of expression values that exist in the samples for the selected probeset. The value of *K* that yields the tightest clusters (lowest within-cluster variation) as well as the largest between-cluster variation is chosen to represent the number of “levels” for that probeset [79].

After clusters or levels are determined, a typical application of this method re-labels the gene expression of the probeset for individual samples by the level or cluster that it belongs to. This re-labeling technique works quite well when just information regarding group membership is required. However, when expression values are compared across genes or used to build classifiers, the method can fail to maintain the ordering of the samples in the expression space, and hence the distinction between the groups. Figure

4-3 uses the gene expression value for 10 samples and a single probeset to demonstrate a loss of information on relative expression differences when the expression values are reassigned based on cluster labels, and highlight the advantage of using cluster centroids to retain information on the ordering of samples with the data. Thus, instead of re-labeling the samples by the group or cluster index, it is proposed that the expression value for an individual sample is replaced by the centroid of the cluster to which it belongs.



(a) (b)

Figure 4-3: Example of K-means clustering. (a) Use of cluster index as sample label does not affect analysis (b) Use of cluster index results in loss of information on relative differences in expression levels

A practical drawback of this method is the limitation in the maximum value of *K* that can be explored. The K-means algorithm is designed to find *K* distinct groups in the dataset, and in the worst case scenario, each sample in the dataset forms an individual cluster. Thus, *K* can take on a maximum value equal to the number of samples in the dataset. When the range of expression levels is on a compressed scale, such as in the case of RMA normalized data [47] represented in the log-2 scale (3.0-14.0), a few hundred samples can adequately represent the spread of the data samples. In cases where the range of data is very large, for example a range from 1.0 to 6000.0, as in a MAS5.0 normalized

dataset [47], several thousand samples may be required to adequately represent the entire data range. However, in practical situations, the gene expression datasets are limited to only a few hundred samples. Thus, the method works when the values consist of very tight clusters around a few expression levels and can fail when the clusters span a large range of expression values. The effect of quantization on the dataset due to clustering must be carefully examined before proceeding with gene expression analyses to ensure that the quantized data contains meaningful information.

4.4.2 Noise Removal

A method is described in [80] to reduce the noise in a gene expression dataset by re-labeling the numerical levels in the data. The actual number L of distinct levels α_l in the gene expression matrix $[A_{m\ x\ n}]$ is used to re-organize the data. The gene expression values are rank-ordered by magnitude, and each level is first redefined as:

$$\alpha_l = \frac{b_l + b_{l-1}}{2}$$

$$\text{where:} \quad b_l = b_0 + le\ ;\ e = \frac{b_L - b_0}{L}\ ;\quad b_0 = \min([a_{mn}])\ ;\quad b_L = \max([a_{mn}])$$

The interval $[b_0\ b_L]$ is divided into equal sub-intervals. The new data matrix is created by analyzing each expression level – if the value a_{nm} falls in the sub-interval $[b_{l-1}\ b_l]$ then, it is quantized to the centroid of that sub-interval.

High resolution gene expression datasets are expected to have a very large number of distinct levels. To reduce the number of distinct levels, a slight modification is

proposed. The number of labels L in the final dataset is pre-specified rather than computed from the data. The data is organized into bins that represent the range between the ordered levels. Each gene expression level is analyzed to determine which bin it belongs to. The probeset is then assigned a new expression value equal to the median or mean of that bin. This method aims at reducing the noise in the dataset by eliminating unnecessary levels, regardless of the number of levels or groups within each probeset. Figure 4-4 demonstrates the working of the method using simulated gene expression data for 10 samples and a single gene.

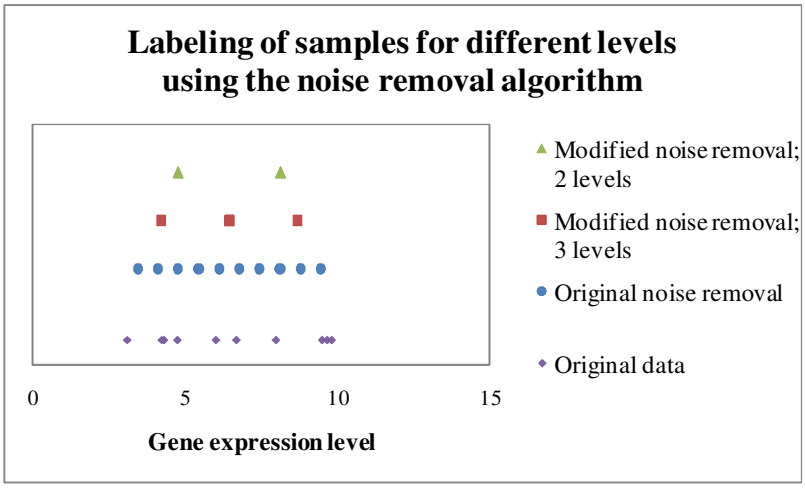


Figure 4-4: An example to demonstrate the noise removal algorithm for quantization of gene expression data

As with K-means clustering, the effectiveness of the method depends on the characteristics of the data being analyzed. The method can maintain the integrity of the data at the probeset level when working with a small range of expression levels. Data represented on a log-2 scale inherently has a lower range than the original data, and can be represented more easily with a relatively small L. As the value of L is increased, the

new expression values begin to converge around the original values. Thus, a small L would be adequate to represent the data without losing significant information. However, since the method converts the data into a set of L uniform intervals, when the range of expression values is very large, an adequate number of levels L have to be used in order to maintain the relative differences between probesets. For example use of L=10 in a MAS5.0 dataset with a range from 1.0 to 6000.0 can severely distort the contrast between low and high expressing samples. Use of a large L on the other hand, can maintain the contrast between the extreme values as well as limit the noise in the data. The selection of the quantization parameter L is thus dependent on the characteristics of the numerical data.

4.4.3 Simple Rounding

A simple method for reducing the resolution of data is to limit the numerical precision of the data [82]. Practically, many generalized gene expression analysis algorithms ignore the higher significant digits. The use of all the significant digits to create a numerical or mathematical model of the biological problem tends to generate models that are very specific to the given sample set. Such highly specific models rarely work well in predicting the class of new samples. Slight perturbations in gene expression values, either due to experimental variation or genetic differences, can lead to significantly different models that cannot be validated on independent samples. Thus, a straightforward way to reduce the resolution of a gene expression dataset is to reduce the number of significant digits in the numerical representation. The number of significant digits that are retained can have an impact on the outcome of analysis and the accuracy of

prediction models. Hence, it is necessary to experiment with the level of quantization, and choose an optimal tradeoff between resolution of the data and loss of accuracy.

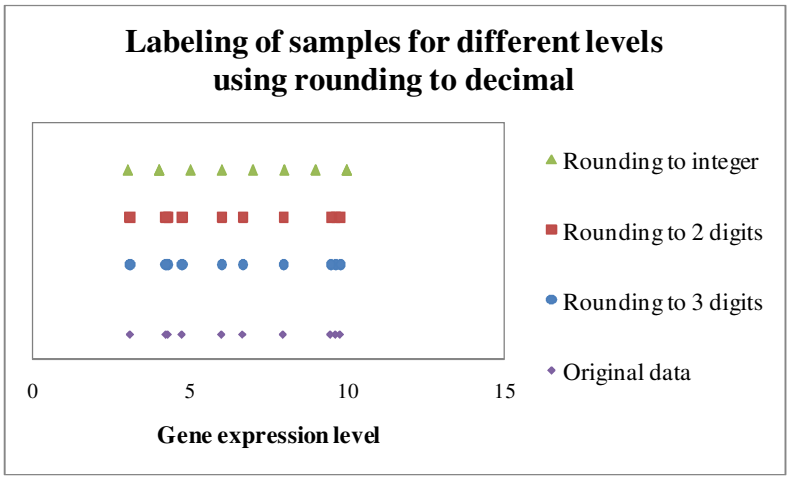


Figure 4-5: Example of the effect of rounding to decimal on a gene expression dataset

Figure 4-5 shows the effect of rounding on the distribution of expression levels for 10 samples using 4 significant digits. The figure shows the change in spread of the samples as the number of significant digits is reduced. The rounding technique used here analyzes the dataset by examining each individual expression value in the dataset. Thus, the relative expression levels in the data as well as the ranking of the probesets or samples within the dataset remain unaltered.

4.5 Experiments Using Quantization

The goal of the case study presented in Section 4.2 was to generate a predictive signature for patient survival in the MRC-CRC dataset. The survival times were stratified into high risk (less than 36 months of survival: n = 37) and low risk (greater than 36

months of survival: n = 84). The data (54675 features) was processed using RMA normalization and represented as a continuous value from 0.0 to 15.0 with 6 digit precision. Three classifiers (C4.5 DT, SVM and NN) were used to build models in a 10-fold CV setup and the best weighted accuracy of the classifiers was found to be 56%.

The motivation for quantization of gene expression data is largely dependent on the problem definition and the type of analysis to be performed on the data. Here, the usefulness of the quantization algorithms and the selection of quantization parameters are analyzed in the context of survival analysis of gene expression data, a highly complex classification problem. The NSCLC dataset is also studied here to compare the effect of quantization on classifier performance when working with datasets of different complexity. As before, the NSCLC dataset (n=62) was transformed into a two-class problem using a cut-off for survival time of 30 months (median survival time). Patients who died within 30 months were considered poor prognosis (n = 20), otherwise they had a good prognosis (n = 42). The data consisted of 7129 features stored in MAS5.0 data format with a 6-digit precision and range of 10.0-6000.0.

4.5.1 Experimental Setup to Test the Effectiveness of Quantization Algorithms

Table 4-1 provides the range of values for the quantization parameters used for each of the datasets by the three methods. The K-means algorithm uses *K*, the number of clusters as a parameter. The noise removal method considers *L*, the number of discrete levels and the rounding algorithm uses *R*, the number of significant digits to retain after rounding, as parameters.

Table 4-1: Quantitative description of quantization parameters

MRC-CRC dataset			
Quantization method	**Parameter used**	**Min value**	**Max value**
K-means clustering	K	2	100
Noise-removal	L	10	100
Rounding	R	0	6
NSCLC dataset			
Quantization method	**Parameter used**	**Min value**	**Max value**
K-means clustering	K	2	60
Noise-removal	L	10	2000
Rounding	R	0	6

Chapter 3 showed that the MRC-CRC/Survival as well as the NSCLC/Survival datasets contained several probesets that were significantly correlated with the survival outcome (Section 3.6). The quantization methods aim at improving the contrast in probesets that have small signals while also retaining the effect of large signals. Thus, the number of probesets that are significantly associated with survival outcome in a quantized dataset can be used as one of the indicators of the effectiveness of quantization.

The quantization algorithms may be categorized based on whether the algorithm operates at the probeset level, or at a global level (using the entire dataset) to alter the data. Both methods of quantization retain the integrity of the probeset level data, such as the relative ranking of the samples and the number of distinct groups of samples. Thus, univariate gene expression analysis can be used as an initial screening test of effectiveness for both types of quantization schemes. Multivariable models are useful in practical situations to understand the collective effect of a set of genes on the survival outcome and used to test the selection of quantization parameters for model building.

Survival analysis may be performed on continuous survival data, or on dichotomized data, as described in Chapter 2. The CoxPH method works with continuous

survival data and requires a few levels in the data for effective modeling. Other methods such as the Student's t-test and the K-M survivor estimates work exclusively on two groups of data. The two groups of data are created in slightly different ways for each of these tests.

The Student's t-test is used to determine if the two groups of data have a significantly different expression profile for a selected probeset [33, 65]. Hence, the two groups are formed by choosing an appropriate cut-off for patient survival time. For example, patients with survival time less than 36 months are grouped in a "Bad prognosis" group, and the rest of the samples are grouped in the "Good prognosis" group.

On the other hand, K-M curves are used to determine if the two groups have significantly different survival characteristics [66, 67]. In this case, the two groups are formed by defining an appropriate cut-off for expression values. Often the median expression level is used as a threshold to form two groups of patients. K-M curves are estimated for each of these groups. A log-rank test is used to determine if the two survivor curves are significantly different.

Each of these methods provides different means to understand the data, and uses information in the gene expression dataset in slightly different ways. However, each method aims at answering a single question – can a mathematical model be generated from the dataset to distinguish the groups of survival? If such a model can be created, it would then be used to suggest the survival group, or expected survival time for a new patient. The effect of the three quantization algorithms was tested on the Student's t-test, K-M and CoxPH in a univariate manner. Since all the three methods aim at analyzing the same aspect of the data, the outcomes of the analyses are expected to concur. Only those

parameter settings that yield a reasonable number of significant probesets and the most consistent results across the analysis methods are retained for further inspection (see Table 4-1).

4.5.2 Effect of Quantization on Survival Analysis of MRC-CRC/Survival and NSCLC/Survival Datasets

The number of probesets found to be significant using univariate tests in a quantized dataset is compared with the original full resolution data to assess the effectiveness of the quantization method. This information is shown in Figure 4-6 and Figure 4-7 for the two datasets. The number of significant probesets is expected to stabilize as the resolution of the data is altered from very coarse to very fine resolution (e.g. original resolution). For the MRC-CRC/Survival dataset, the number of significant probesets for each test remains stable except at the coarsest resolutions (for example at R=0; L=40 and below; and K=2).

A similar effect is seen with the noise removal method for the NSCLC/Survival dataset, with lower number of significant probesets for resolution L=200 and below. However, the opposite trend is seen for rounding and K-means quantization. One reason for this effect due to rounding could be the scale of the data (0-6000.0). When working with such a large range of values, the small effect of the significant digits may add more noise than information for the univariate tests. Reducing this data may enhance the contrast between the classes to improve the significance of correlation with survival. K-means clustering assigns cluster centroids as the expression value. Given the large range of the data, at quantization levels of K=2 or K=10, an extreme contrast is introduced at

the probeset level. Such an effect is not observed in the MRC-CRC/Survival dataset that is represented on a log-2 scale with a range of 0-15.0. This result supports the hypothesis that the selection of the quantization method and its parameters should take into account the range of the data and the inherent data complexity.

It was shown in Chapter 3 that predictive classifiers may be built on datasets with small numbers of significant features, if these features have a high contrast between the classes. Thus, it is important to assess the quality of each quantized dataset individually. The reduction in the number of significant probesets may result from information being lost with the resolution. Alternately, these settings may be improving the probesets with high level of contrast, while all the probesets with lower levels of contrast are suppressed.

Although the variation in the total number of significant probesets is lower at the higher resolution settings, these datasets may include probesets that are noisy and hence not highly correlated to the outcome. As the resolution is lowered, the noise is expected to diminish since the contrast is expected to be enhanced in the highly correlated probesets as well as probesets with low levels of correlation. The lower variation in the number of significant probesets for the medium-resolution settings suggests that a low complexity dataset can be used in place of the high resolution original data while still retaining all the relevant features of the dataset.

If the dataset created by a specific quantization method is consistent with the original dataset and maintains the integrity of the expression values at the probeset level, the result of the three analyses should be consistent for the probeset. Figure 4-8 (MRC-CRC/Survival) and Figure 4-9 (NSCLC/Survival) summarize this consistency in results. Figure 4-8b a shows the number of probesets in each dataset for which all three tests had

significant p values. Note that in the baseline MRC-CRC/Survival dataset (Figure 4-6), the three tests individually found several thousand features to be significant (Student's t-test: 6000; CoxPH: 11000; K-M: 8500). However, only about 1100 probesets had significant p values for all three tests. The trend for the change in number of significant probesets that was observed in Figure 4-6 is still maintained. Figure 4-8b shows the total number of probesets that had concordant results across the tests. As before, the higher resolution datasets have very little variation in the total number of concordant probesets, indicating that the information content is consistent with the original dataset. Similar trends are observed in the NSCLC/Survival dataset. (Figure 4-9). These results suggest that at least a few significant probesets are being retained in the dataset by each quantization method and the different parameter settings. Figure 4-8b also shows that at coarser settings, a large number of probesets have consistent p values across the tests. However, only a small subset is significantly related to survival.

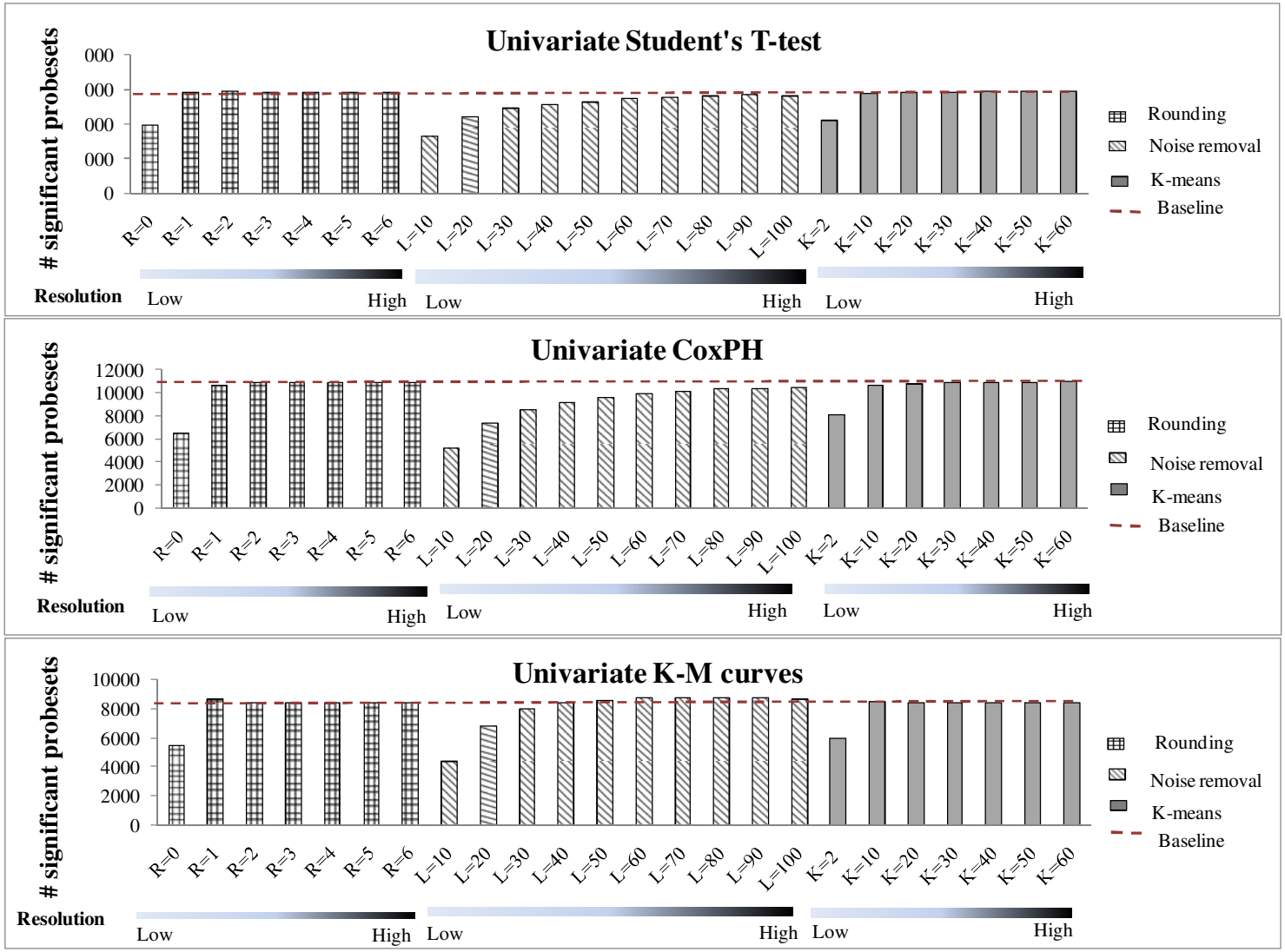


Figure 4-6: # Significant probesets in the MRC-CRC/Survival datasets for the quantized datasets

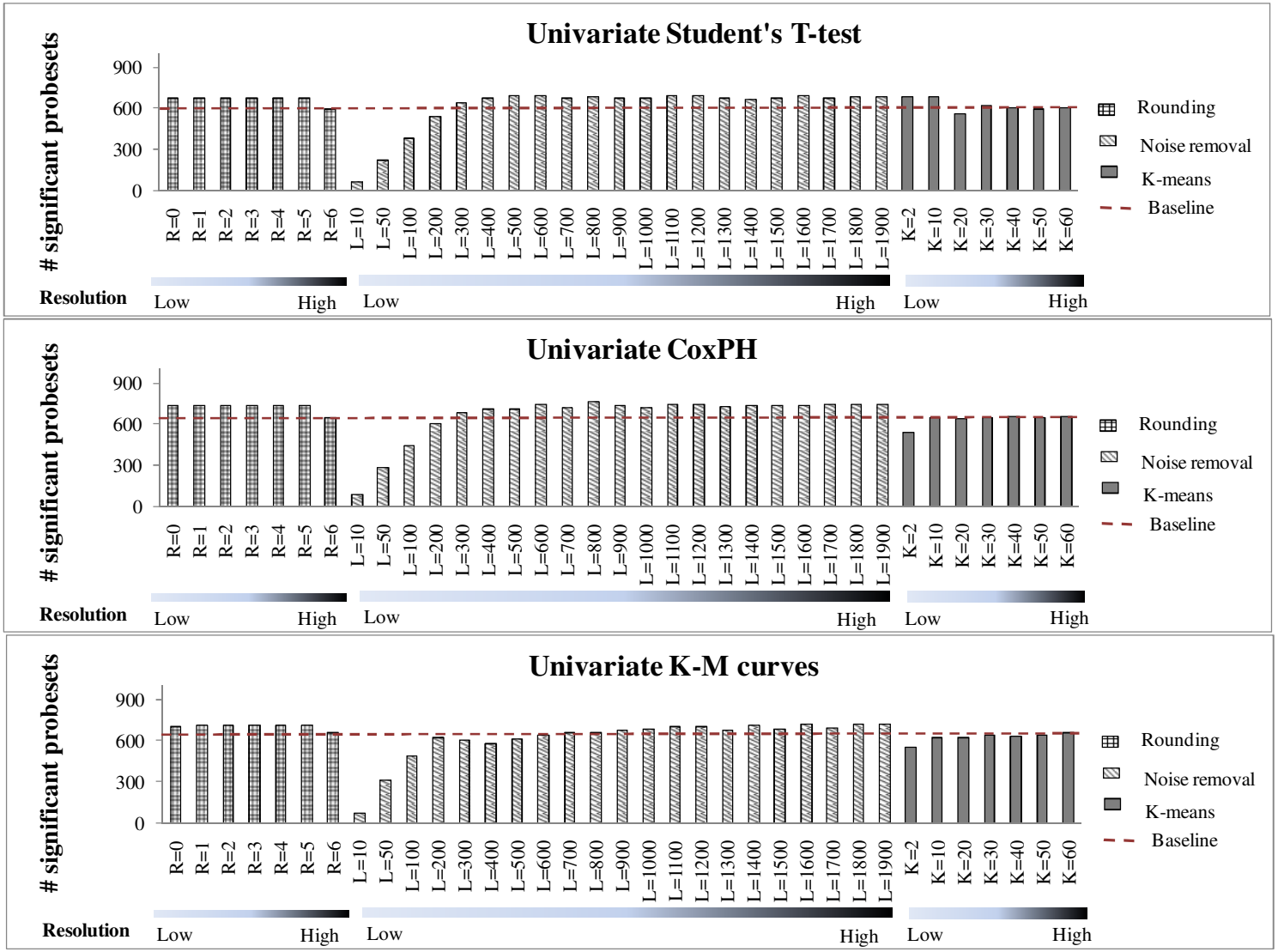
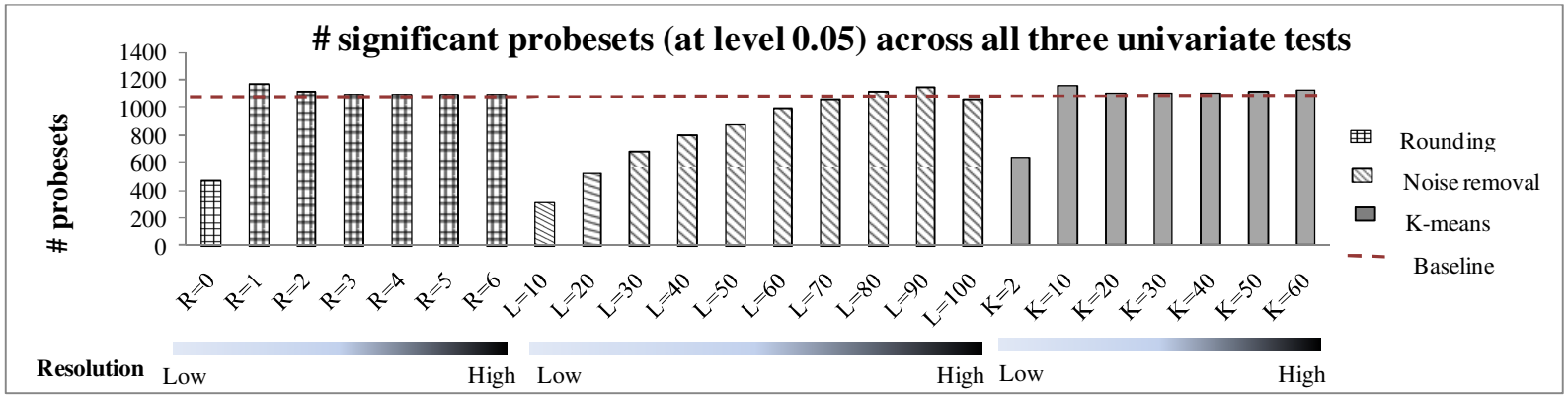
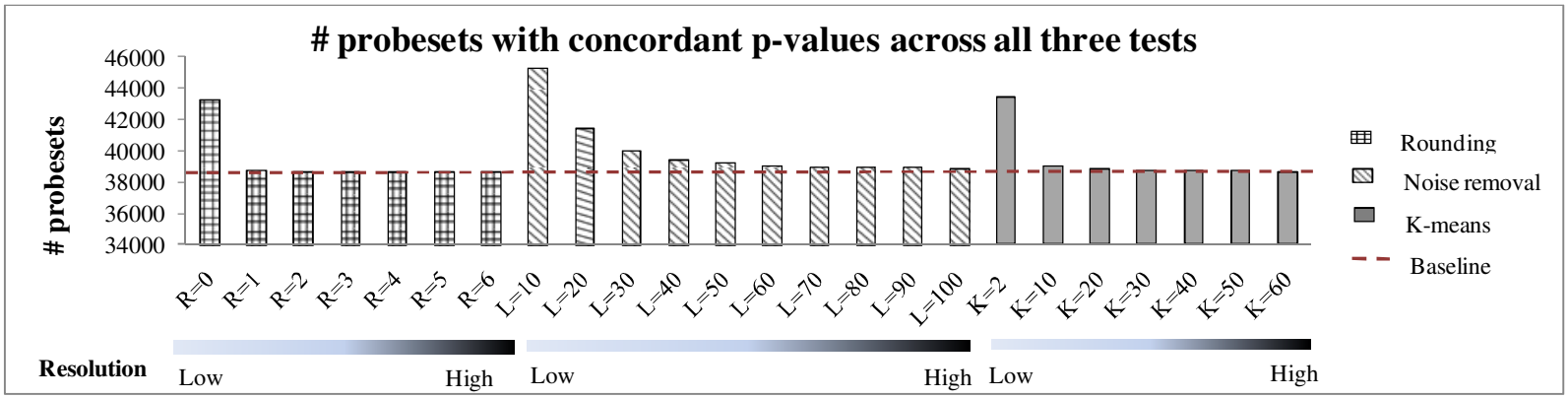


Figure 4-7: # Significant probesets in the NSCLC/Survival dataset for the quantized datasets

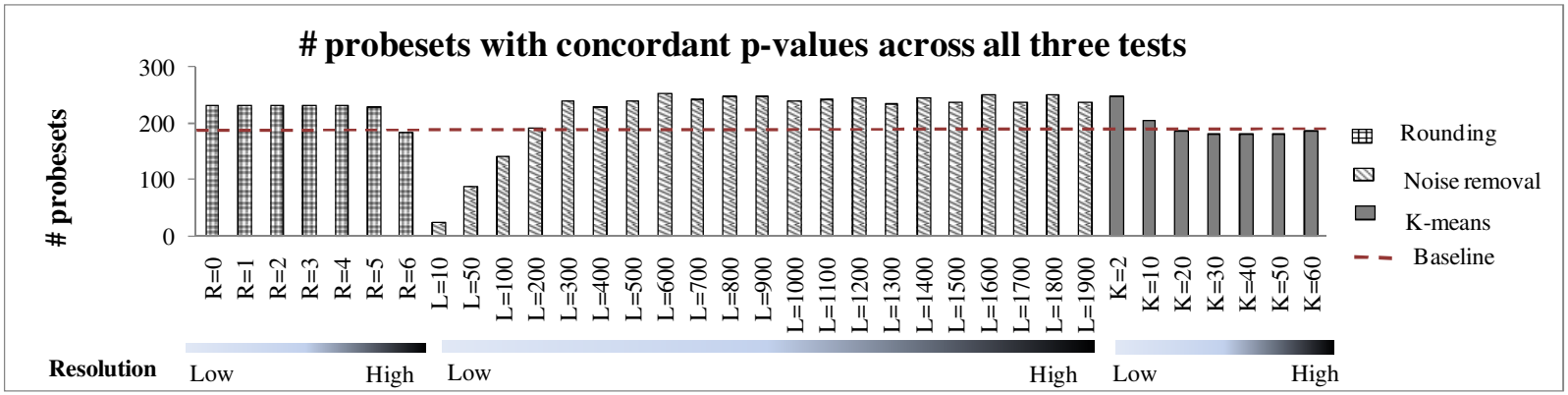


(a)

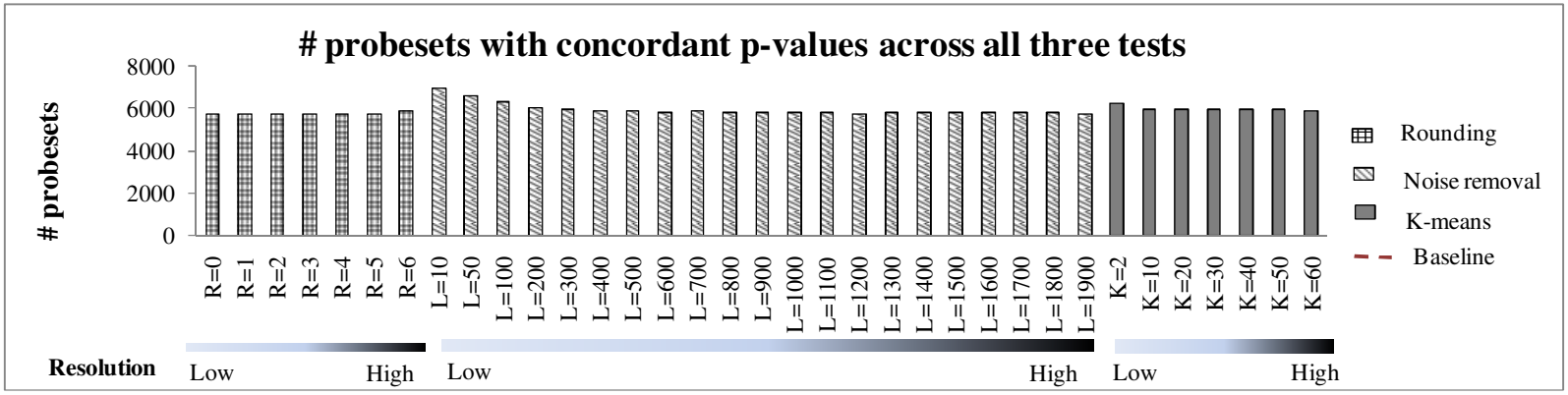


(b)

Figure 4-8: Number of probesets with concordant p values across all three univariate tests on the MRC-CRC/Survival dataset



(a)



(b)

Figure 4-9: Number of probesets with concordant p values across all three univariate tests on the NSCLC /Survival dataset

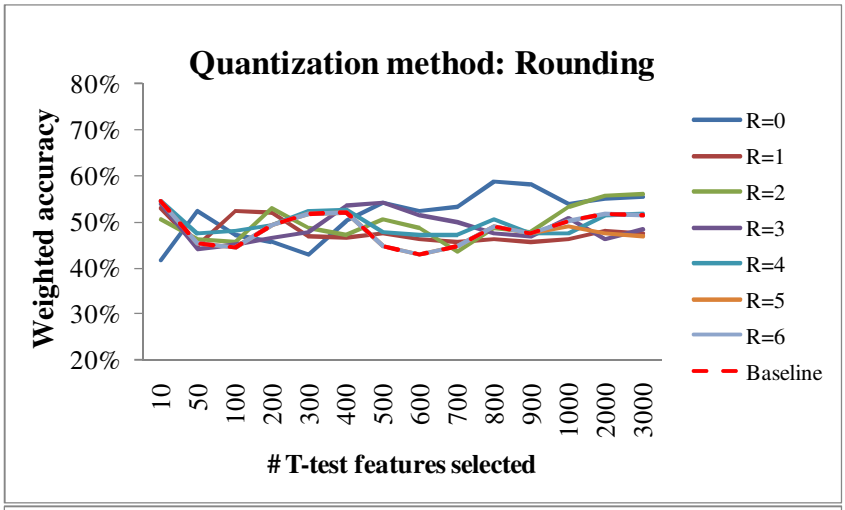
4.5.3 Multivariable Analysis

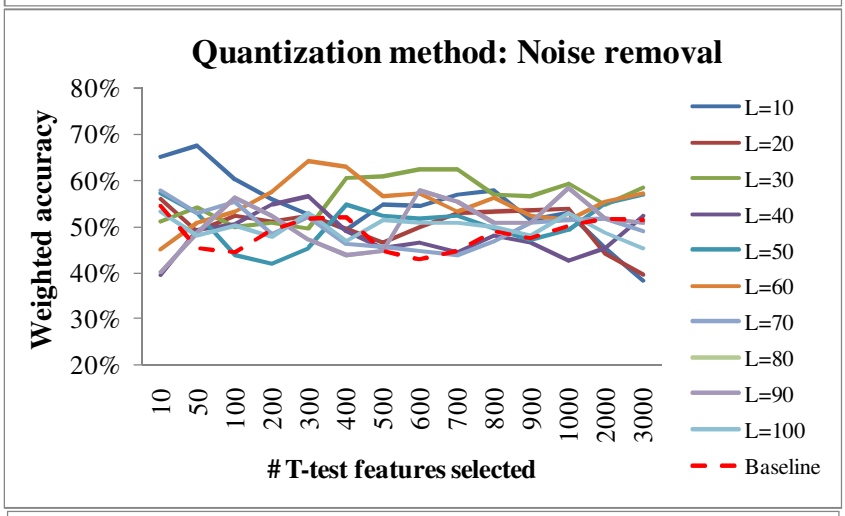
Quantization of the gene expression data was used to reduce the complexity of the data to aid in the use of simpler classifiers as well as creating simpler classifier boundaries. Chapter 3 used three classifiers (C4.5 DT, SVM and NN) in a 10-fold CV setup to generate models of survival. The results indicated that a better model needed to be designed for the MRC-CRC/Survival dataset. The same classifier experiments were repeated on the quantized datasets to determine if the modified data was better suited for use with the described classifier setup. Student's t-test was used as the initial feature selection step to compare the effect of feature selection with the effect of quantization as a data reduction technique.

Figure 4-10 and Figure 4-11 show the weighted accuracy of the classifiers for the quantized datasets (MRC-CRC/Survival) using C4.5 DT and NN for a varying number of features. Figure 4-12 and Figure 4-13 show the same for the NSCLC/Survival dataset for NN and SVM. SVM for MRC-CRC/Survival and C4.5 DT for NSCLC/Survival did not perform better than the other classifiers presented here, and thus are not represented in the graphs. For the MRC-CRC/Survival dataset (Figure 4-10 and Figure 4-11), it is seen that the behavior is complex, however, in general, the classifiers tend to perform with better accuracy with the quantized datasets than with the original full resolution data that is used as the baseline for comparison.

The results indicate that accuracy is impacted by interaction between the number of features and quantization parameters. For example, L=10 performs better than L=100 when 50 features are used (L-10 : 68% and L-100 : 48%), but worse when the number of features is 3000 (L-10 : 38% and L-100 : 45%). The graphs do not show clear trends that

can be used to determine the best combination of parameters for quantization and feature selection. This suggests that it is important to explore the settings for both types of data reduction to obtain the best accuracy for prediction. However, the results also indicate that quantization may lead to improved accuracy. Figure 4-14 and Figure 4-15 compares the best performing quantization parameters with the baseline accuracies for the MRC-CRC/Survival dataset. Figure 4-16 and Figure 4-17 show the same for the NSCLC/Survival dataset. The graphs show that each quantization method performs differently. For example, maximum accuracy for rounding is obtained when higher numbers of features are selected, however, for noise removal the maximum accuracy occurs with fewer features selected. The data also suggests that relatively coarse data produces the highest accuracy. Note here that the range of accuracies for the NSCLC/Survival dataset is lower than the accuracy obtained in Chapter 3 due to different numbers of t-test features selected for classification.





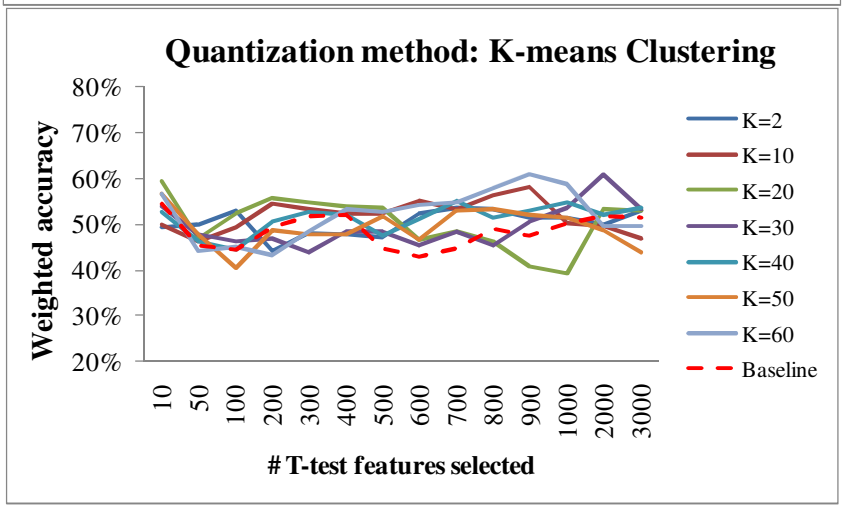


Figure 4-10: Performance of C4.5 DT on the quantized MRC-CRC/Survival datasets. Each classifier result is compared to the performance on the baseline dataset (shown in red dashed lines)

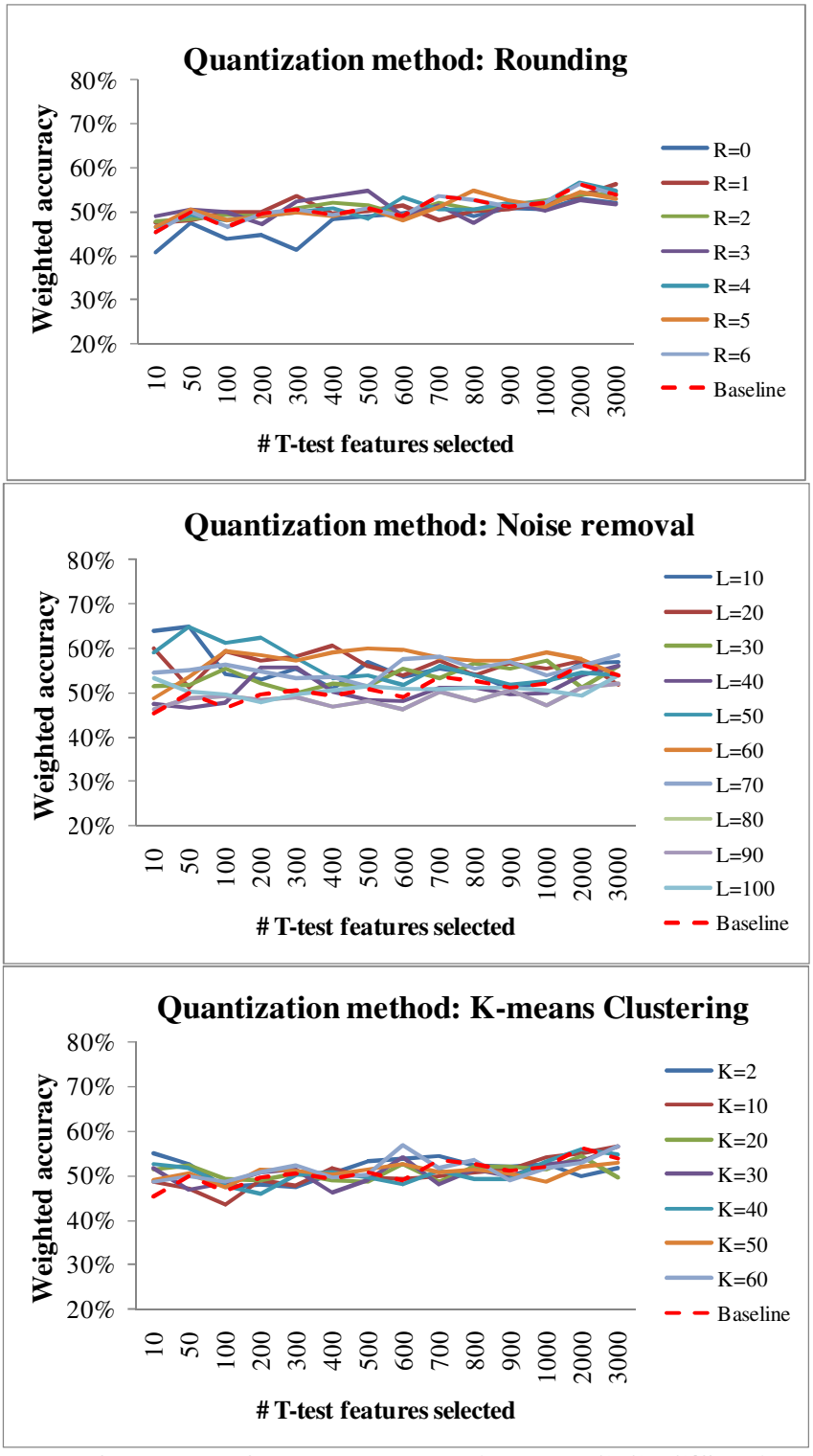


Figure 4-11: Performance of NN on the quantized MRC-CRC/Survival datasets. Each classifier result is compared to the performance on the baseline dataset (shown in red dashed line)

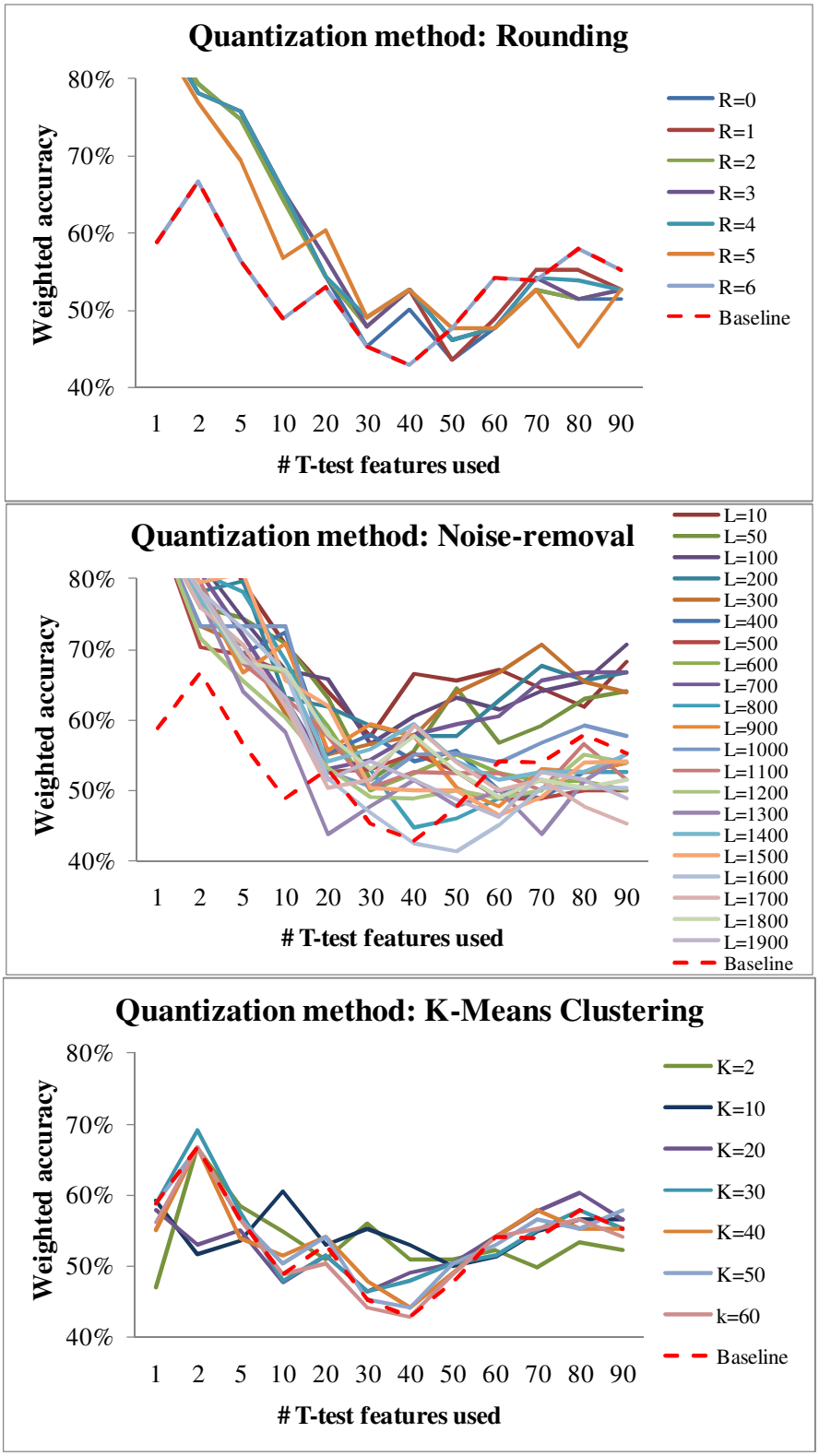
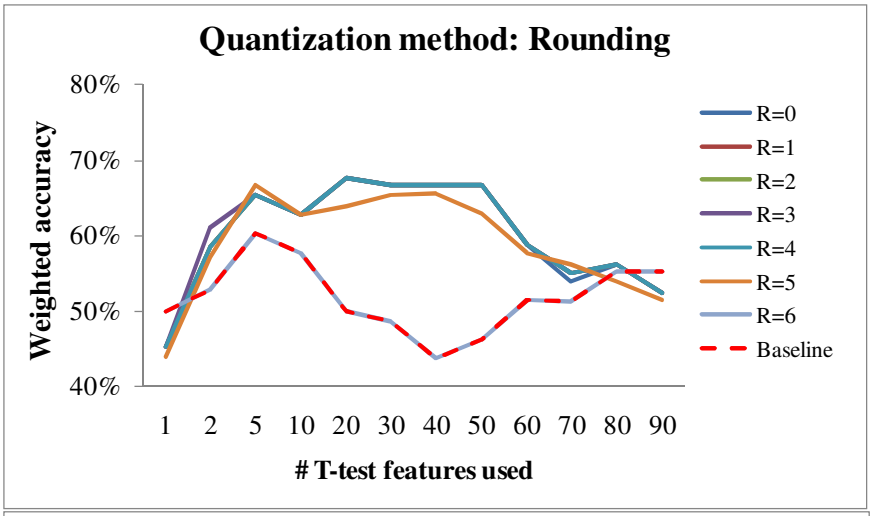
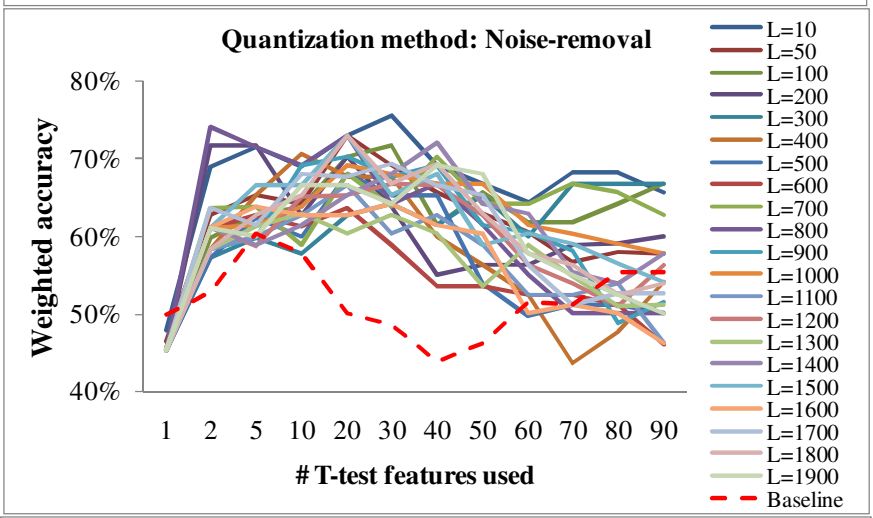


Figure 4-12: Performance of NN on the quantized NSCLC/Survival datasets. Each classifier result is compared to the performance on the baseline dataset (shown in red dashed line)





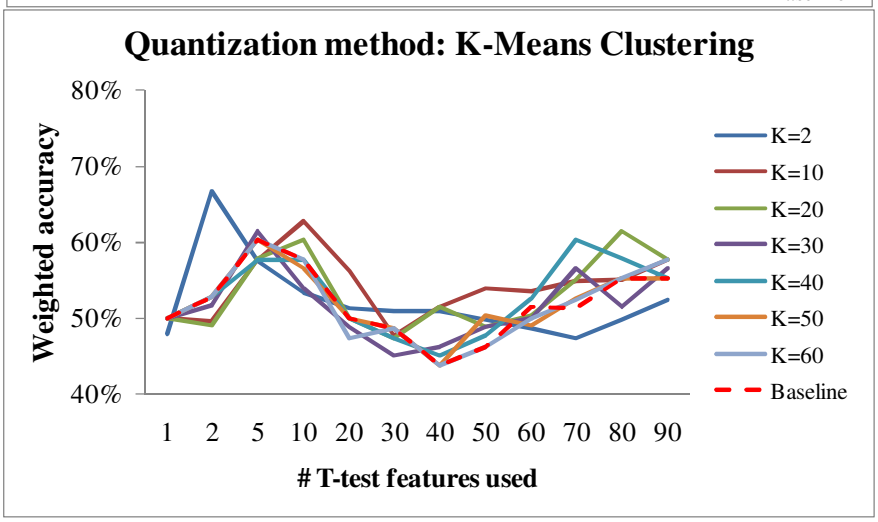


Figure 4-13: Performance of SVM on the quantized NSCLC/Survival datasets. Each classifier result is compared to the performance on the baseline dataset (shown in red dashed line)

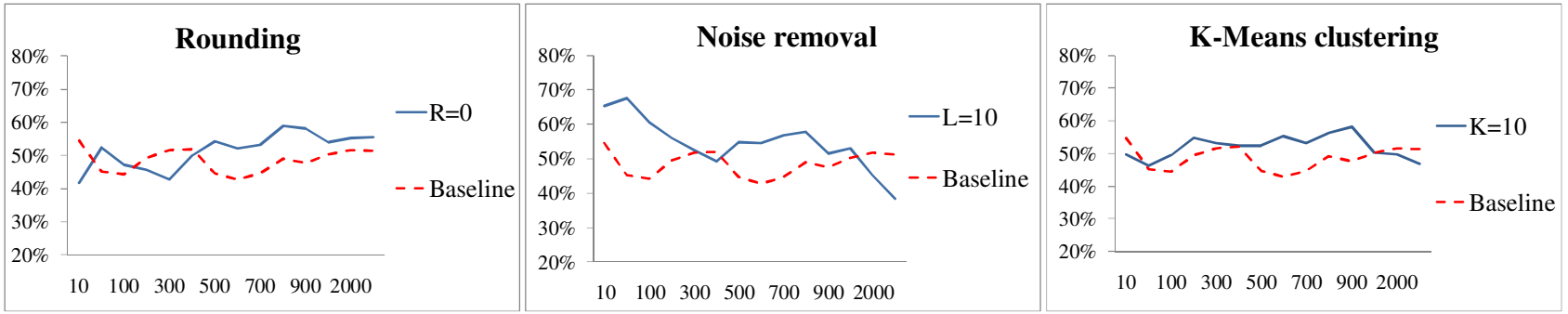


Figure 4-14: Comparison of the weighted accuracies for C4.5 DT using the best parameter setting for quantization on the MRC-CRC/Survival dataset. Graph legends: x-axis: T-test features used, y-axis: Weighted accuracy for classifier

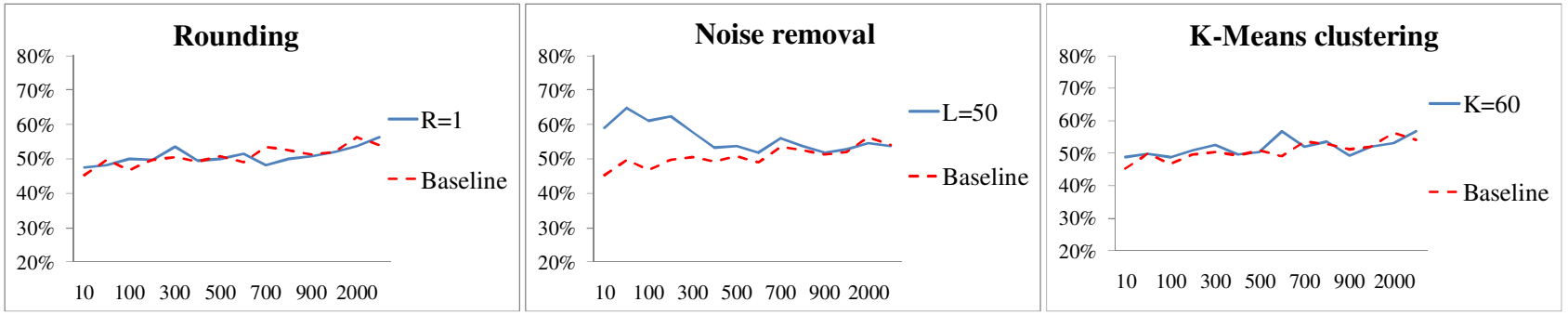


Figure 4-15: Comparison of the weighted accuracies for NN using the best parameter setting for quantization on the MRC-CRC/Survival dataset. Graph legends: x-axis: T-test features used, y-axis: Weighted accuracy for classifier

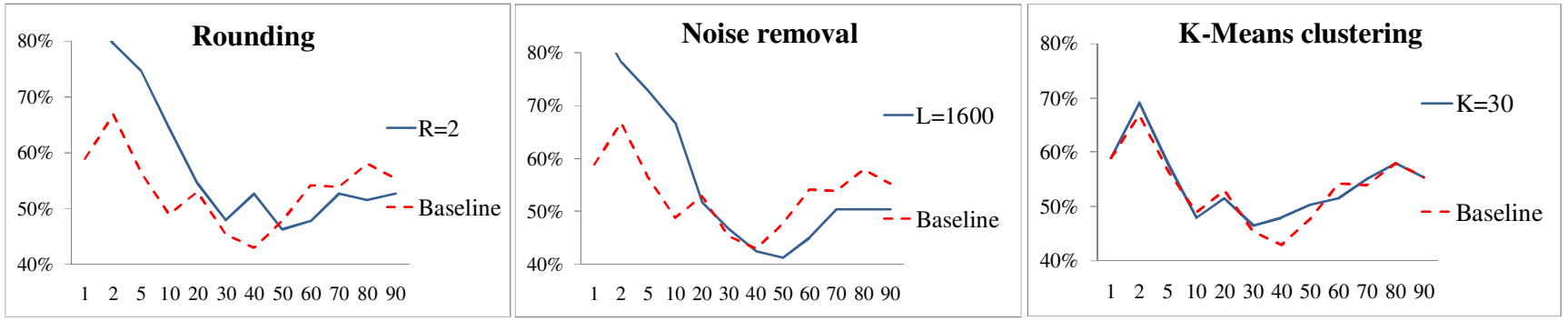


Figure 4-16: Comparison of the weighted accuracies for NN using the best parameter setting for quantization on the NSCLC /Survival dataset. Graph legends: x-axis: T-test features used, y-axis: Weighted accuracy for classifier

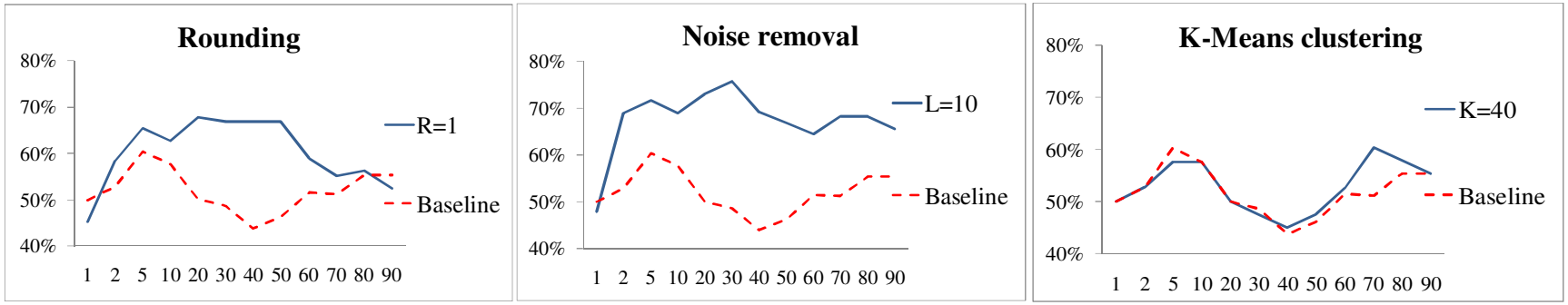
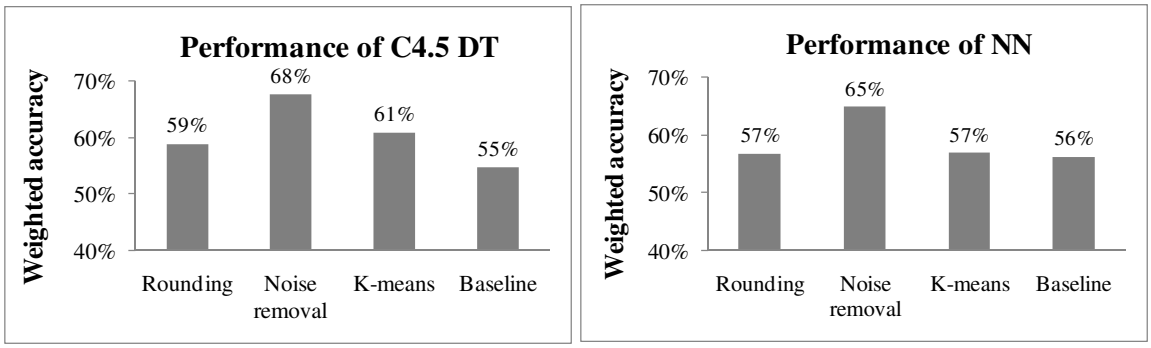


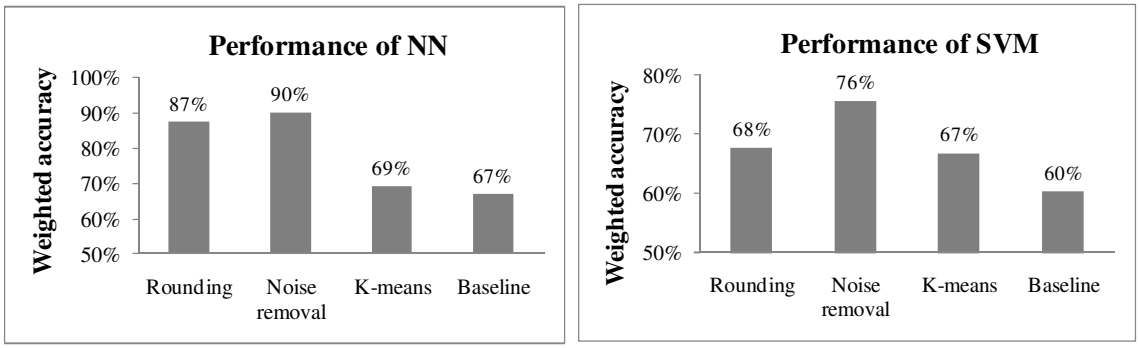
Figure 4-17: Comparison of the weighted accuracies for SVM using the best parameter setting for quantization on the NSCLC /Survival dataset. Graph legends: x-axis: T-test features used, y-axis: Weighted accuracy for classifier

Figure 4-18 (MRC-CRC/Survival) and Figure 4-19 (NSCLC/Survival) compare the best weighted accuracies for each method of quantization independent of the features used for modeling. The figures clearly indicate that the predictive accuracy of the classifier models improves with quantization of the data. The noise removal algorithm tends to do the best in improving accuracy while reducing the resolution of the dataset. The weighted accuracy for the MRC-CRC/Survival dataset was improved from 56% to 68% and the accuracy for the NSCLC/Survival dataset improved from 67% to 90% for the feature selection settings presented here. Statistical tests were used to determine if the improvement per fold of the 10-fold CV was significant. For each dataset, this change was found to be significant at a level of 0.05 (MRC-CRC/Survival: p value=0.035; NSCLC/Survival: p value=0.008).



(a) (b)

Figure 4-18: Comparison of the best weighted accuracies using the three methods of quantization for the MRC-CRC/Survival dataset



(a) (b)

Figure 4-19: Comparison of the best weighted average accuracies using the three methods of quantization for the NSCLC/Survival dataset

4.6 Classification Complexity Using Quantization

Figure 4-20 shows the complexity of a quantized dataset with the measure ϕ as defined in Chapter 3. ϕ is shown for the best quantized dataset (L=10) for the MRC-CRC/Survival dataset. It can be seen that the complexity of the MRC-CRC/Survival dataset has been reduced, and the classifier accuracy has increased. The complexity of this dataset is now equivalent to the MRC-CRC/Site dataset (see Chapter 2). As predicted by ϕ, the corresponding classifier accuracies are similar. It can be noted that the accuracy of the MRC-CRC/Site dataset has decreased. This suggests that the quantized dataset at L=10 can be used to generate a predictive model for the survival problem. A better quantization parameter setting has to be determined to obtain better accuracies with the MRC-CRC/Site dataset. The complexity and the classifier accuracy of the MRC-CRC/Gender dataset remain the same as before suggesting that the information regarding gender is maintained in the quantized dataset.

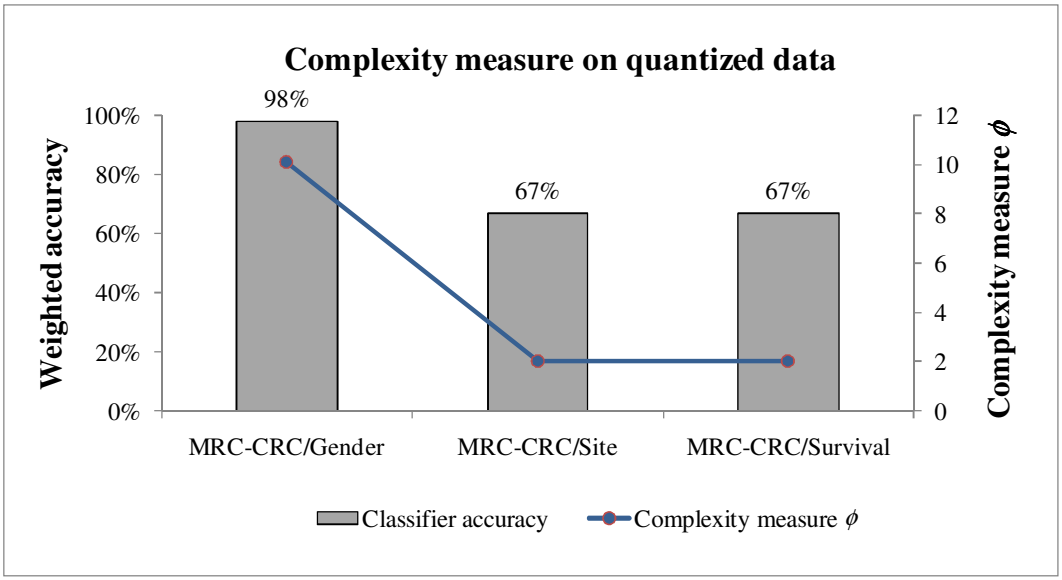


Figure 4-20: Measure of complexity on the best quantized MRC-CRC/Survival dataset

This result further emphasizes that the quantization methods retain useful information in the data at all the parameter settings and lower resolution datasets can be used instead of the original high resolution datasets to create better predictive classifier models. Such models created on simpler datasets are expected to have simpler decision boundaries and hence be able to generalize to independent samples for prediction.

4.7 Summary

Data quantization was explored to limit the resolution of gene expression data to yield low complexity data for analysis. Three methods of quantization were proposed and tested on the MRC-CRC/Survival and NSCLC/Survival. Concordance in the results of univariate analyses indicated that the three methods altered the data in a consistent manner. Experiments with classifier models indicated that the quantization techniques aided in improving classification accuracy by creating simpler models for analysis. Thus,

quantization of gene expression data creates datasets with low complexity and provides the ability to build robust and reliable prediction models.

CHAPTER 5 A COST-SENSITIVE MULTIVARIABLE FEATURE SELECTION FOR GENE EXPRESSION ANALYSIS USING RANDOM SUBSPACES

5.1 Introduction

As stated in earlier chapters, one aim for building gene expression models is to identify signatures that provide accurate clinically-relevant biomarkers of disease. Chapter 3 analyzed the impact of data complexity on classifier accuracy, while Chapter 4 used quantization to reduce data complexity and improve classifier accuracy. This chapter explores the use of feature selection to improve classifier accuracy.

Section 5.2 illustrates the need for a multivariable approach in modeling biological processes using the example of a molecular pathway involving the Ras family of proteins. The random subspace approach is described in Section 5.3. A previously developed random-subspace based approach for multivariable feature selection (MFS-RS) is described in Section 5.4 and extended to incorporate a cost sensitive aspect (MFS-RSc) in Section 5.5. Results are summarized in this section and demonstrate the improved classification accuracy from using the extended method. Future work is summarized in Section 5.6.

5.2 Multivariable Models

One of the first stages of gene expression analysis is to reduce the dimensionality of the data by selecting a small set of features that are reliably expressed at different

levels across different classes of samples. Many techniques for feature selection analyze the data in a univariate fashion to determine if a gene is significantly associated with the outcome of interest. The classifier models presented thus far utilize univariate tests (Student's t-test) as the basis to select relevant features. However, many biological processes are governed by multiple genes acting along pathways [8, 83]. This domain knowledge suggests that a mathematical process that incorporates multiple variables in the feature selection process is likely to capture the biological process better than univariate feature selection and thus improve the predictive ability of the classifier.

5.2.1 Molecular Pathways - An Example

A genetic pathway is defined by the interactions between groups of genes with individual functions [8]. These genes are typically dependent on specific interactions for the cell to function normally. Mutation in a gene active within a pathway can disrupt the functioning of the pathway. For example, consider a molecular pathway for the Ras family of proteins [40, 83, 84]. These proteins deliver signals from cell surface receptors via several protein-to-protein signals to ultimately affect cell growth, differentiation and cell survival [8]. Ras communicates signals from the cell surface to the nucleus and mutation of the Ras gene can disrupt these sequences of protein signaling and cause transmission of the signaling even in the absence of an extracellular stimulus. This can ultimately lead to the development of cancer [40, 83].

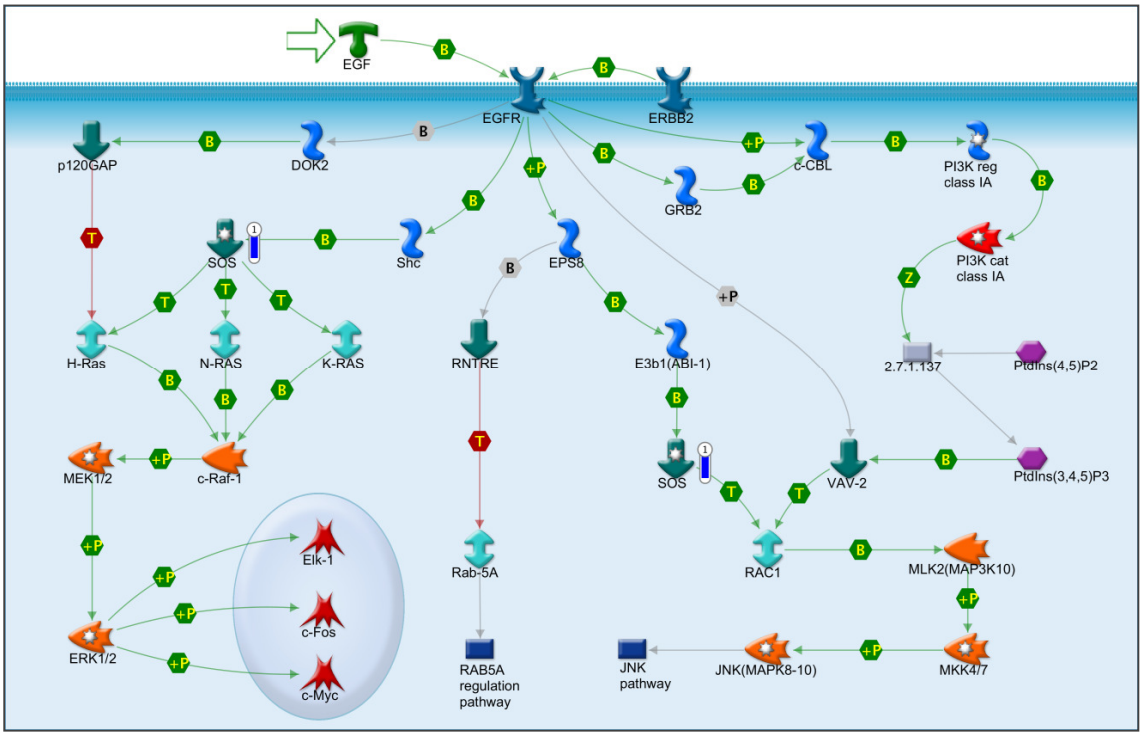


Figure 5-1: Example of a molecular pathway involving Ras. (Image generated from GeneGO, St. Louis, MO)

Figure 5-1 provides an example of a molecular pathway involving Ras. Disrupting the expression of Ras can cause changes in transcription downstream from Ras in the pathway [8, 40]. Thus, information on expression of genes in a pathway can provide a hint regarding where a disruption in signaling may have occurred. Further, it can be seen that the change in expression of a single gene can rarely provide an indication of the complete picture of the tissue function. In addition to this, a gene may be a part of multiple pathways that could drive the biological state of a cell. The supporting genes in a selected pathway can aid in determining which of these pathways is active in the cell. Hence, multivariate selection of gene probesets in a gene expression dataset would be expected to provide more information regarding the state of the cell than univariate analysis.

5.2.2 Existing Multivariable Gene Expression Techniques

Multivariable gene expression models have been developed using different types of analysis methods. Continuous models such as CoxPH have been used to select subsets of features that are correlated with survival to develop clinical relevant signatures for breast cancer [23].

A supervised principal component approach was published in [85] that modified the unsupervised technique of principal component analysis by selecting subsets of components that were shown to be related to a specified outcome. The genes used for the component analysis were univariately selected in a supervised manner using CoxPH. This method allowed selection of subsets of genes that were related to outcome via a specific combination, defined by the principal component. The contribution of each gene to the signature was altered by the component scores rather than the traditional CoxPH coefficients. The technique was constructed so as to allow inclusion of covariates in the model to improve predictive ability. The models were shown to extract biologically relevant gene expression signatures for various models of disease.

The principal component approach was used in combination with a maximum entropy linear discriminant analysis (MLDA) [86] to discriminate normal from tumor prostatic tissue. In this case, the MLDA weights assigned to each feature was modified by the principal components and the features ranked in decreasing order of the weights. The method was found to identify clinical known biomarkers of the disease.

A SVM-based feature selection method is described in [87] that steps through the gene expression dataset to generate a feature set that best fits the problem. At each step of the feature selection process, a feature that maximizes the correlation of the feature set to

outcome is included into the model. The resultant feature set is expected to be representative of the multi-gene pathways of the underlying biology.

An entropy based multivariate feature selection method is described in [88]. The method estimates the entropy of the class variables on the model rather than on the data. Multivariate normal distributions are used to model the sparse data. The method was tested on several datasets and shown to perform with high predictive accuracy.

5.3 Random Subspace Approach

The random subspace technique has been used to create ensembles of classifiers to achieve accuracies higher than those obtained from a single classifier [60]. Considering a problem in which many features are present ($p >> n$), selection of the best features for distinguishing the samples of the two classes can create a projection on the feature space that can greatly aid in creating simpler classifier boundaries. The greater the separation between the classes for each feature, the better the ability to design a simple decision boundary.

However, in complex datasets, it may be difficult to find many features that can individually separate the samples into the two classes. Statistical techniques have the advantage of a significant theoretical basis and allow partially overlapping distributions between two classes. However, when features with overlap are used in a classifier model, a complex decision boundary may be created, as illustrated in Figure 5-2.

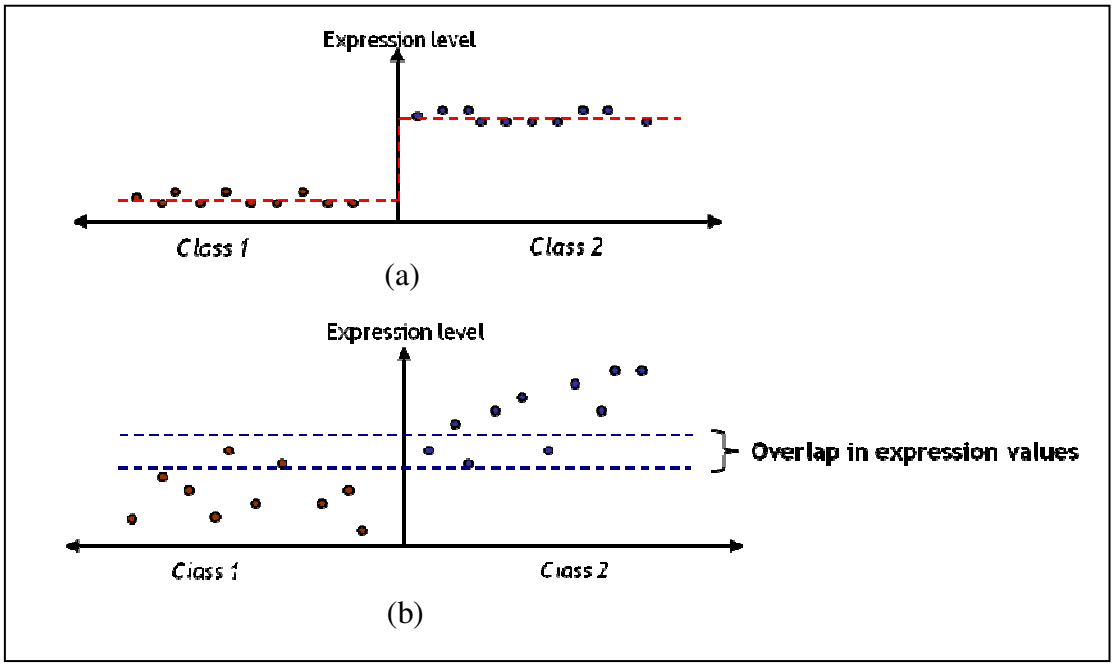


Figure 5-2: Illustration of distributions of features or probesets

Figure 5-3 illustrates the advantage of using random subspaces for multivariable feature selection in such complex cases. In a dataset with multiple features, where the individual features have poor separation between the classes, use of the entire feature set for classification may lead to poor accuracy. However, a projection of the data into a subset of the feature space could provide a better separation of the samples. In Figure 5-3, projection of the data onto the plane created by Gene 1 and Gene 3 provides a separation of the samples, while projection onto the other two planes yields poor separation. A classifier that uses this projection space for modeling a gene expression signature may be expected to perform better than a classifier that used the entire feature space, or the features univariately.

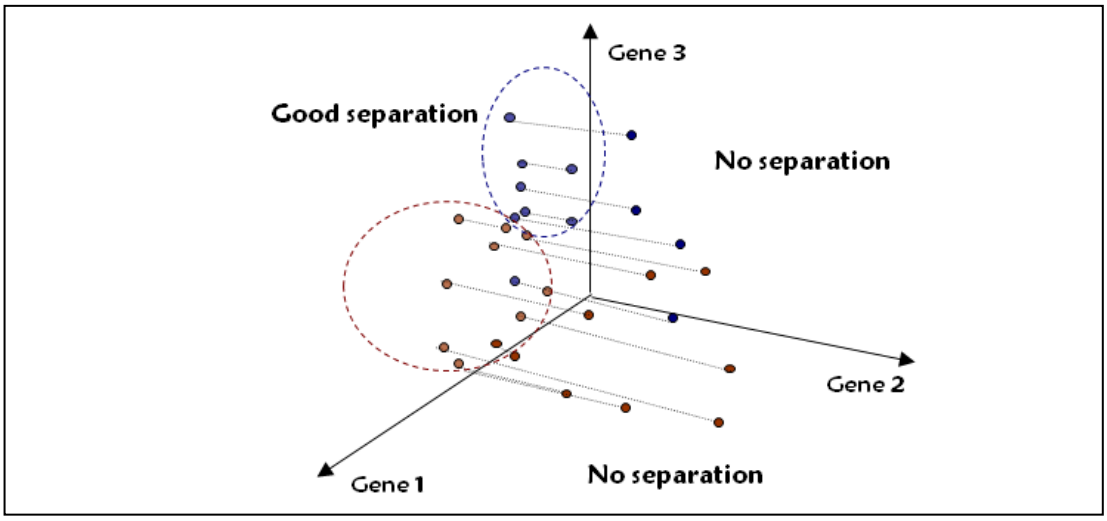


Figure 5-3: A random projection of the data provides better separation of samples

The random subspace technique uses this concept to create ensembles of classifiers [60] as illustrated by Figure 5-4. A subset of features is randomly sampled from the entire set of features. This is a random subspace or a random projection of the feature space. A classifier is constructed from this random projection on the feature space. The process is repeated many times, each time selecting another random subset of features. If enough such random subspaces are created then several subspaces may be obtained that optimally represent all the important features in the samples. Further, if the random subspaces cover all the important features effectively, then each classifier would potentially be tuned to learn a few characteristics of the population. This process inherently identifies subsets of features that are important for describing the underlying samples in a multivariate sense. The combination of feature subsets can provide a better understanding of the underlying data than using a single feature set or creating a single classifier.

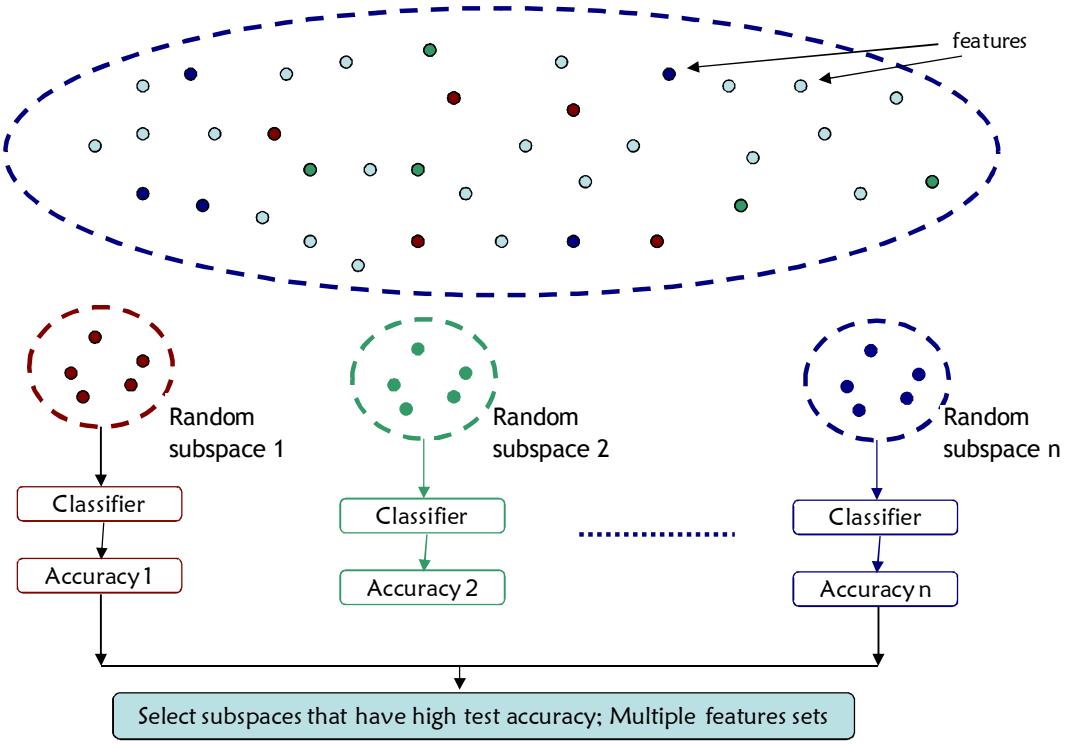


Figure 5-4: Random subspace approach for feature selection

A typical application of the original random subspace classifier model described in [60] uses each of these classifiers to predict the class of an unseen test example, and the resulting predictions are combined by a majority vote. The majority vote assigns the test sample to the class that was predicted by a majority of the random subspace classifiers. Since each classifier is tuned to learn slightly different characteristics of the population, the class assigned by the majority vote will indicate that the test sample displayed characteristics of that class for a majority of the random subspace classifiers. Further, since the sample characteristics are analyzed from multiple points of view, the majority vote is expected to perform better at learning the classes of samples than any single classifier.

5.4 Multivariable Feature Selection Using Random Subspaces (MFS-RS)

The random-subspace approach was used for multivariable feature selection in [46] by identifying the features used by these random subspace classifiers rather than use the ensemble of classifiers as the model for the gene expression signature. Thus, instead of using the random subspace classifiers to assign the class of a test sample, this new technique, termed MFS-RS, extracted the features used by the subspace classifiers. C4.5 DT [61] were used to build the subspace classifiers. Since these classifiers select important features from accurate random subspaces, they inherently select important features from an input set while creating the decision boundary. Standard C4.5 pruning is used to avoid over fitting that may occur by randomly selecting features from such a large space.

With large gene expression datasets that consist of several thousand features, a large percentage of the features could be unrelated to the underlying classes, as evidenced by signatures with gene sets that consist of hundreds of features [2, 4, 14]. These features do not provide any useful information for a classifier and may reduce the accuracy of prediction. Selecting random subspaces that yield extremely inaccurate classifiers aids in quickly removing uninformative features. Thus, classifiers that perform very poorly on either the training or test samples indicate that the specific combination of features used by the classifier may be safely dropped from the analysis. Conversely, classifiers that can create an efficient decision boundary between the training samples of the classes can be used as an indicator that at least some of the features are useful for describing the classes. A simple way to select the highest accuracy classifiers is to examine the prediction accuracy of each subspace classifier on a test set of samples. Use of these subsets of

features increases the likelihood that many of the important genes are included in the analysis as opposed to a univariate selection of features.

The MFS-RS method was applied to the MRC-CRC/Survival problem (see Section 2.4.2). Chapter 3 showed that the dataset was relatively complex and in general, the prediction accuracies for survival are expected to be low. Figure 5-5 indicates that a large percentage of the subspaces created on this dataset were extremely poor in predictive accuracy, and thus not very general. Less than 0.05% of the subspaces were found to have predictive accuracies better than 80%. Hence, the approach taken was to create as large a number of random subspaces as possible and sift through these subspaces to identify features used most often in subspaces that yield high-accuracy classifiers.

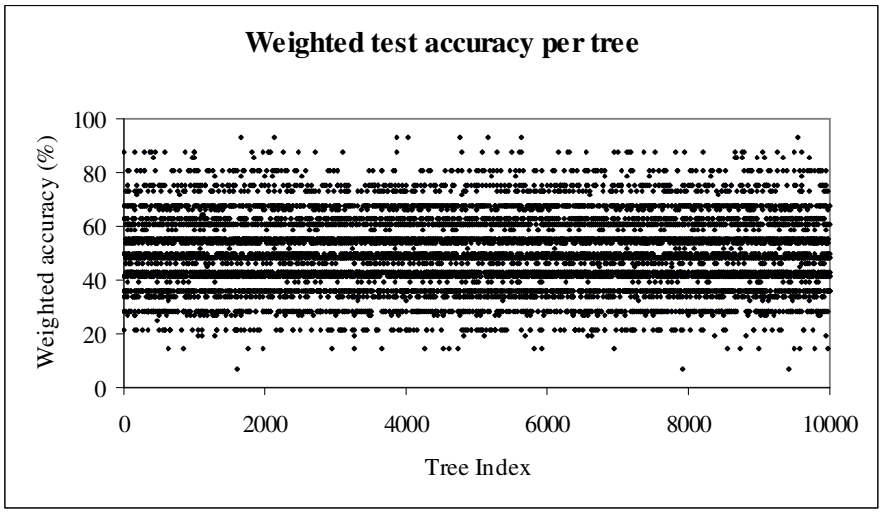


Figure 5-5: Weighted test accuracies of 10000 trees on MRC-CRC/Survival dataset

The random subspace classifiers were created in a 10-fold CV setup. Two thousand subspaces were generated of size 200 each and a C4.5 decision tree was constructed from each random subspace. Subspace classifiers within each fold were tested

for prediction accuracies and a single subspace that attained the highest train and test accuracy was selected as the final classification model. Thus, 10 subspaces were extracted from the dataset (representing the best subspace of each fold) and combined into a single feature set.

This procedure of extracting the best single feature set from the dataset was conducted within a 10-fold CV to provide independent samples for validation, thus yielding 10 feature sets for use. Survival models were created for these feature sets using three classifiers (C4.5 DT, NN and SVM). The average weighted accuracy of these classifiers on the 10 feature sets was used as a measure of the performance of the technique.

The Student's t-test was used as a univariate feature selection method for comparison on the same dataset. The *n* most significant features, ranked according to Student's t-test p values, were used for building feature sets for the same classifier methods in a 10-fold CV. The MFS-RS technique used an average of 81-96 features per feature set. To compare the performance of the two feature selection techniques, *n* in the Student's t-test approach was set to 100 and the average weighted accuracies of the 10-fold CV were compared.

Figure 5-6 shows that MFS-RS performed with better prediction accuracies than univariate feature selection [46]. Since the features were selected in a multivariate fashion, it is expected to mimic the underlying biology of the samples in a closer manner than the univariately chosen features.

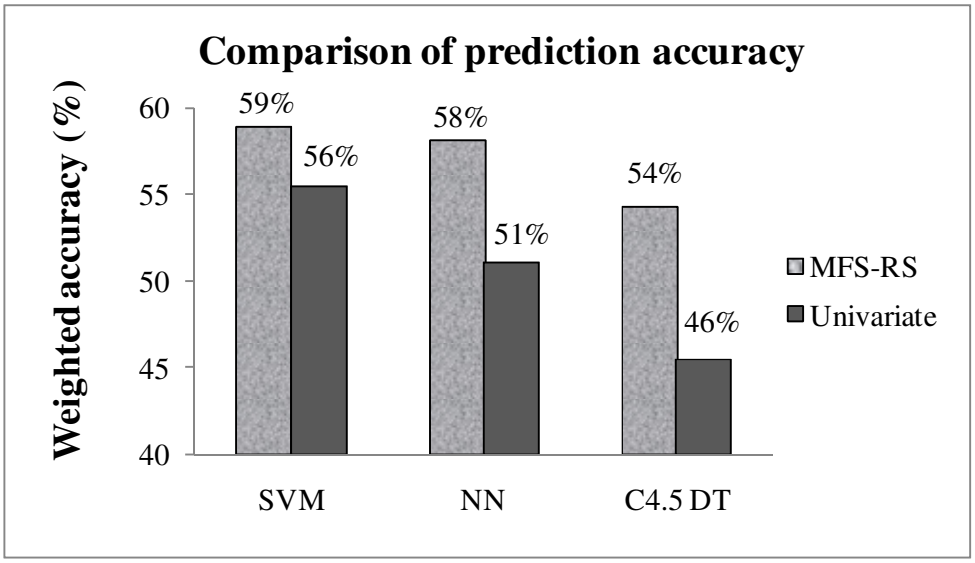


Figure 5-6: Comparison of prediction accuracies using MFS-RS and univariate feature selection methods for the MRC-CRC/Survival dataset

5.5 Cost-Sensitive Multivariable Feature Selection (MFS-RSc)

Many practical gene expression datasets including MRC-CRC/Survival consist of unbalanced classes, i.e., the number of samples in different classes may be unequal. Larger numbers of samples in one class could bias the classifier towards predicting the larger class a majority of the time. In doing so, the classifier may gain in accuracy of prediction but the sensitivity and specificity (see Section 2.6) of such classifiers may be drastically altered. When analyzing biomedical questions, it is often desirable to have a high sensitivity as well as a high specificity of prediction.

Classifier performance in such imbalanced datasets can be evaluated by using cost-sensitive learning tools [89]. Evaluation of cost-sensitive classifiers using cost curves was proposed in [90] to visualize classifier performance in imbalanced datasets. A wrapper approach was used in [91, 92] to address this issue and to improve minority class accuracy. The wrapper was used for optimization of a composite f-value to reduce the

average cost per test example for the datasets considered. The true positive rate of the minority class increased significantly without causing a significant change in the f-value.

A modification of the MFS-RS technique is proposed here to factor in this imbalance in the class distributions. The method, termed MFS-RSc, chooses the best random subspaces to maximize both the sensitivity and the specificity of a random subspace classifier based on a pre-determined threshold. The thresholds for prediction accuracy, specificity and sensitivity are set to a value lower than 100% to avoid selecting classifiers that are over-trained on the samples. In this modification, multiple feature sets may be selected to maximize the representation of good features in the dataset.

As before, 2000 subspaces were created of size 200 features each. A multivariable feature set was created by combining the features in subspaces that simultaneously yielded 80% or greater specificity and sensitivity. All the features in the selected random subspaces were pooled together to form a new feature space. Hence, each individual random subspace classifier may be re-created from this space. Further, since the features are now pooled together, more features could be included in the creation of a classifier than in the original subspace. A new classifier created on this feature space has the choice of several of the individually important subsets of features, as well the ability to combine these features into a single decision boundary. The decision boundary created by this classifier is used as the final predictor for all new and unseen samples, evaluated in a 10-fold CV setup.

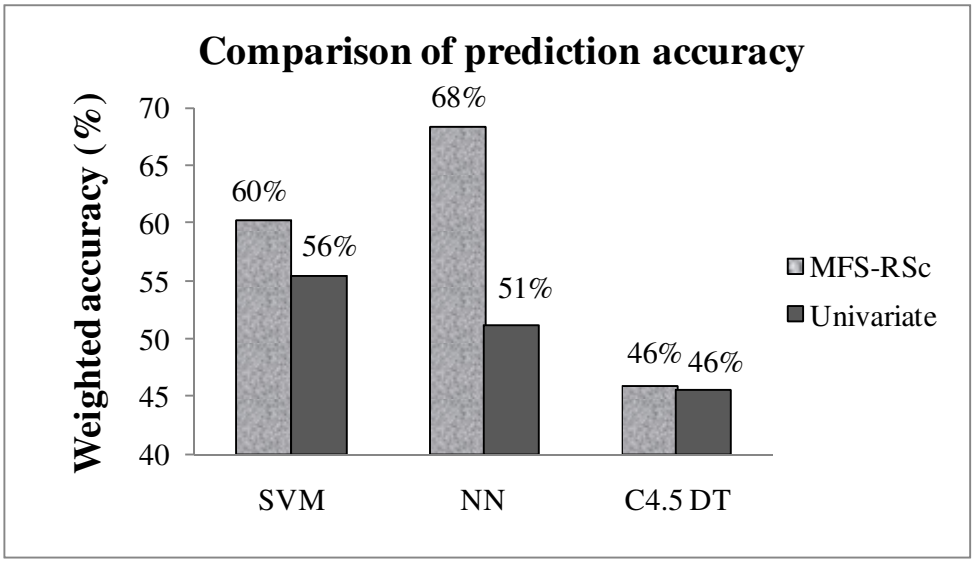


Figure 5-7: Comparison of prediction accuracies of classifiers using the proposed MFS-RSc technique and univariate feature selection on the MRC-CRC/Survival dataset

Figure 5-7 shows the improvement in prediction accuracy of the classifiers with use of the proposed multivariable feature selection. SVM and NN were found to perform with much better prediction accuracies with the random subspace features. However, the performance of C4.5 decision trees was found to be similar with both feature selection methods. The difference in performance of the three classifiers is most likely due to the difference in the use of features by the classifier methods. SVM and NN train on all the features used as an input and develop a decision boundary based on these. On the other hand, C4.5 DT select only a small set of the input features to represent the classifier boundary. In doing so, some of the features may be removed from consideration. In such cases, C4.5 DT are useful for the initial stages of multivariable feature selection, and the classifier models such as SVM and NN are useful in creating the final prediction model.

Figure 5-8 shows the sensitivity and specificity of the classifiers with multivariable and univariate feature selection. Here, the sensitivity of the classifier is the

accuracy of the majority class ("Good" prognosis) and has similar values using either feature selection method. However, the specificity, or the accuracy of the "Poor" prognosis class, indicates that when using the univariate feature selection method, the classifiers are focused on the majority class at the expense of the minority class. With the multivariable feature selection however, the accuracies of this class are significantly increased. MRF-RSc is able to boost the accuracy of the minority class without sacrificing the accuracy of the majority class, thereby increasing the overall accuracy of the classifier. These results were published in [93].

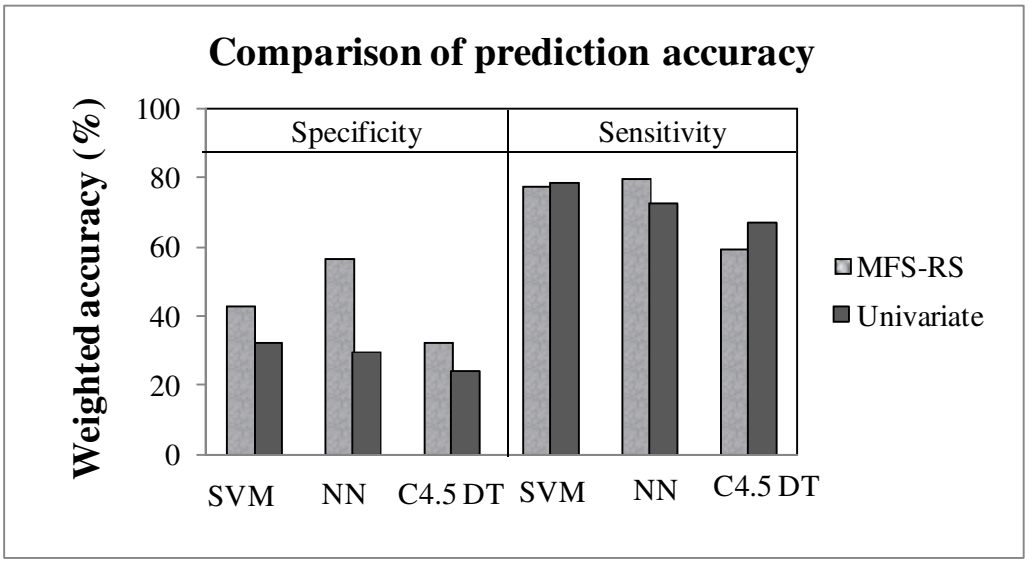


Figure 5-8: Comparison of the specificity and sensitivity of prediction using MFS-RS and univariate feature selection on the MRC-CRC/Survival dataset

Figure 5-9 shows that in general, the prediction accuracies improve with the proposed modification of feature selection using random subspaces. As before, SVM and NN perform with better accuracies than univariate analysis while C4.5 DT does not result in significant increase in accuracy.

The specificity and sensitivity of the best random-subspace based classifiers using the proposed modification is compared to the best subspace classifier created using the original method (Figure 5-10). MFS-RS selected features based on the weighted accuracy of each subspace classifier. This would potentially bias the classifier towards the majority class due to the unequal distribution of the classes. MFS-RSc was proposed to retain features that were equally representative of both classes of survival. It can be seen that the specificity as well as sensitivity of the classifier model are improved with the use of MFS-RSc.

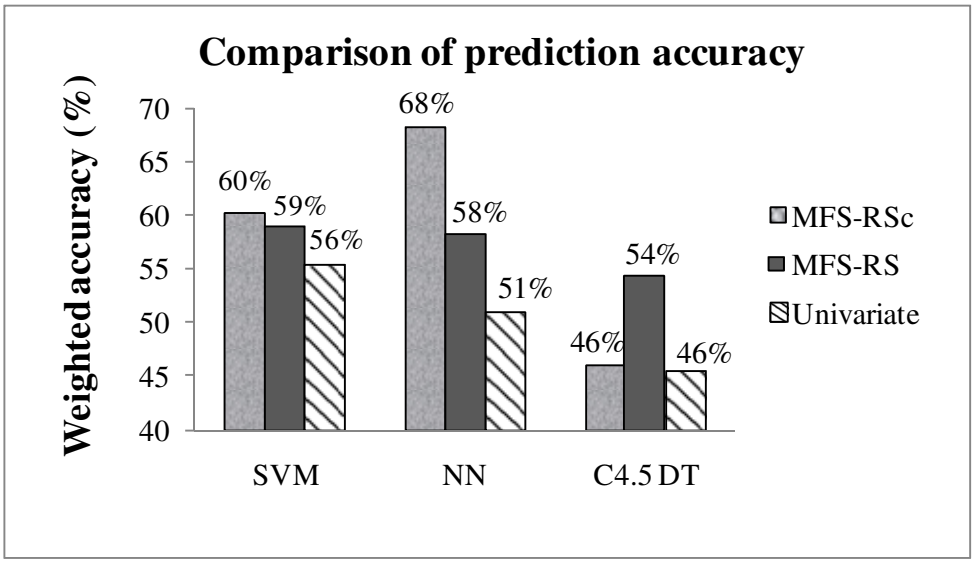


Figure 5-9: Comparison of classifier prediction accuracies for MFS-RS, MFS-RSc and univariate feature selection on the MRC-CRC/Survival dataset

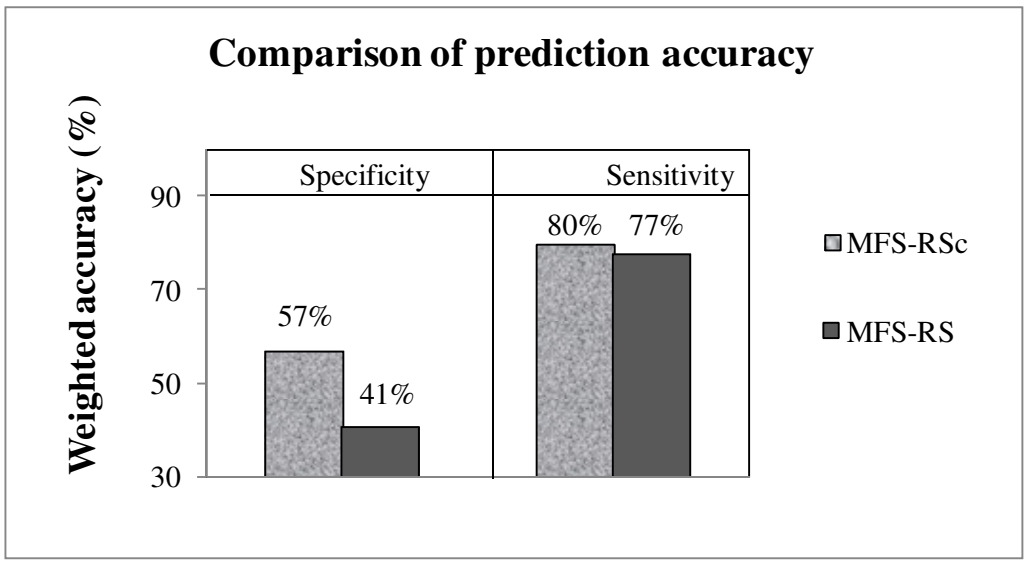


Figure 5-10: Comparison of the best classifier sensitivity and specificity using MFS-RS and MFS-RSc methods on the MRC-CRC/Survival dataset

5.6 Future Work

The random subspace technique was shown to work as a multivariable feature selection tool with C4.5 DT. Since MFS-RSc is used in conjunction with a classifier to pick the best features, it seems reasonable to try other classification schemes for the same purpose. Classifiers such as NN and SVM have been shown to create accurate models for classification, and can be considered as candidates for random subspace based feature selection. Since these methods use all the input features to create a decision boundary, it may difficult to select features from a subspace, resulting in the use of the entire subspace. Algorithms such as SVM-RFE [94] provide a means to evaluate the weight of each support vector. These weights could be used to select features that are important for classification within a subspace.

Other techniques such as regression models may be used instead of the C4.5 DT to select features. Regression models such as CoxPH or linear regression models can be

designed to select a subset of the input features to retain only those features that are correlated to the outcome and provide a good fit to the data [67]. The features can be selected in a forward or backward selection model [67]. Good random subspaces are selected as before, and the selected features pooled to form input features for the final classifier model.

Further work can be done in understanding the effect of features on classification accuracy. Features selected as important in subspaces that performed poorly could be assigned low weights while features in accurate subspaces could be assigned higher weights. These weights could be used to influence the classifier models to use better performing features to produce more predictive models.

The described models can also be extended to incorporate the quantization techniques described in Chapter 4 to further enhance the classification accuracy.

5.7 Summary

The complexity of heterogeneous cancer with multiple molecular pathways was used as a motivation for extending a feature selection method using random subspaces. The predictor built using the original formulation of this feature selection method (MFS-RS) was shown to perform with better accuracy than univariately selected features on the MRC-CRC/Survival dataset. A modification of the original formulation was proposed to account for the difference in sensitivity and specificity of predictors when working with imbalanced datasets (MFS-RSc). This new method of feature selection used cost-sensitive analysis and was shown to improve the overall weighted accuracy of prediction.

It was also shown to boost the prediction accuracy in the minority class while retaining the high accuracies in the majority class.

CHAPTER 6 INTEGRATING BIOLOGICAL COVARIATES IN GENE EXPRESSION MODELS

6.1 Introduction

In preceding chapters, classification complexity was measured and methods were proposed for managing or reducing this complexity. The use of quantization to reduce data resolution and a multivariable feature selection method to reduce data dimensionality were shown to improve the predictive accuracy of classifiers in complex datasets. However, additional options exist for managing complexity by accounting for the biological heterogeneity of tumor samples when creating gene expression models. One such approach was used in [95] for the prediction of radiation sensitivity. This chapter experiments with the methods employed in that paper and provides a more complete approach to the integration of selected biological indicators of cancer into a gene expression model.

Biological indicators of cancer models are introduced in Section 6.2. A brief description of the radiosensitivity dataset and multi-linear regression model developed for prediction of radiosensitivity is presented in Section 6.3. A systematic method to include biological variables in the linear regression model is discussed in Section 6.4 followed by results in Section 6.5. Verification of the proposed model is presented in Section 6.6.

6.2 Biological Indicators for Cancer Models

An important aspect in the management of cancer treatment is understanding how a patient will respond to a specific treatment such as radiation therapy. Customizing radiation therapy to maximize cancer cell death is beneficial, and predicting such a response of the cells to radiation therapy is important for effective patient management.

Genes such as Ras [40, 84] and p53 [41] influence the response of tumor cells to radiation treatment. For example, the presence of a mutant Ras can indicate a higher likelihood of non-response to radiation, while the wild type Ras gene does not predict response to radiation treatment. Similarly, presence of a mutant p53 gene is used as an indicator for uncontrolled proliferation of cells, while a wild type p53 gene is known to be a tumor suppressor. The effect of these genes, both wild type and mutant, has been studied with respect to radiation sensitivity of cancer patients [83]. Radiation sensitivity has been measured by applying a specific amount of radiation (2 Gy) and measuring survival fraction of the target cells. The measured survival fraction is referred to as SF2 and used to predict the sensitivity of patients to radiation treatment [96].

Combining clinical indicators, such as the influence of these genes, with gene expression models of tumor characteristics has the potential to provide meaningful insight into the tumor biology. Models have been developed that incorporate clinical indicators such as tumor grade and angio-invasion [97] to build predictive models for prognosis of breast cancer. Here, the radiation sensitivity of the NCI60 panel of cell lines is investigated by incorporating biological indicators into a gene expression model.

6.3 Multivariable Linear Regression for Prediction of Radiosensitivity

A multivariable linear regression model was developed using gene expression data to predict the radiosensitivity of tumor cell lines [95]. The model was built using a subset of 35 epithelial-based human tumor cell lines from the NCI60 panel of cancer cell lines representing significant biological diversity with respect to the tissue of origin (TO). Radiation sensitivity data, defined by survival fraction after 2 Gy (SF2), was available for each cell line. The method used the Significance Analysis of Microarrays (SAM) [75] to select probesets with a false discovery rate of 5%. The model was shown to achieve a statistically significant ($p=0.0002$) predictive accuracy of 62% for predicting radiosensitivity. The genes selected by the model were shown to be mechanistically involved in radiation sensitivity through wet-lab experiments, thus establishing the biological validity of the mathematical algorithm.

Further work on enhancing the model showed that although the gene expression–based predictor was found to be accurate, the classifier model was not accurate as the cell line population was increased to 48 (compared to 35) cell lines. The best linear regression-based classifier using the 48 cell lines correctly classified 28/48 samples (58%) compared to 25/35 (71%) for the best classifier in the 35 cell line dataset. This result suggested that the linear regression model based only on gene expression data may not be able to capture the complexity of the problem in detail. To address this issue, clinical indicators including tissue of origin (TO), Ras mutational status (wt/mut) (RAS) and p53 mutational status (wt/mut), that are known to be implicated in the biological regulation of radiosensitivity [83, 96], were included in the gene expression model for prediction of radiosensitivity. RAS and P53 status indicators are binary variables that

indicate wild type (wt) or mutational (mut) status of the gene for a cell line. The indicator for tissue of origin (TO) has 9 levels, one for each type of the tissue of origin for the tumor cell line [59].

6.4 Inclusion of Biological Covariates in Model Development

In the published model [95], gene expression was used to predict radiation sensitivity using the following mathematical equation.

$$\textit{Gene Expression Model}: \quad SF2_j = k_0 + k_1(y_{ij})$$

where k_i (i=0, 1) represents a model coefficient, computed during the training process; y_{ij} represents the gene expression value for probeset i in cell line j of the n predictive probesets selected by the classifier and $SF2_j$ is the predicted radiosensitivity for cell line j.

A drop in predictive performance of this basic model was observed when tested on newer samples. An attempt was made to include clinical indicators in the predictive model to capture the underlying biology more closely. The complexity of the data due to the large number of features increases the difficulty of integrating gene expression and biological (or clinical) parameters into a single model. The large number of gene expression measurements makes it much more likely that the most significant correlations to radiation sensitivity are gene expression probesets rather than biological variables.

Expanded linear mathematical models outlined by the following equations allow inclusion of biological variables to construct individual probeset models for explicitly integrating the biological parameters at the feature selection step.

$$\textit{Additive Model}: \quad SF2_j = k_0 + k_1(y_{ij}) + k_2(TO) + k_3(RAS) + k_4(p53)$$

Interactive Model:

$$SF2_j = k_0 + k_1(y_{ij}) + k_2(TO) + k_3(RAS) + k_4(p53)$$
$$+ k_5(y_{ij})(TO) + k_6(y_{ij})(RAS) + k_7(TO)(RAS) + k_8(y_{ij})(p53) + k_9(y_{ij})(TO)(RAS) + \ldots$$

where y_{ij} represents the gene expression value for probeset i in cell line j of the n predictive probesets selected by the classifier and $SF2_j$ is the predicted radiosensitivity for cell line j.

The goodness of fit, represented by an R^2 value, is used to estimate the fit of linear regression models to the underlying data. Higher R^2 values are a better fit for the data. Here, an adjusted R^2 value (Adj-R^2) was used instead of R^2 to adjust for the addition of regressors in the equations. While R^2 tends to increase with an increasing number of regressors, the Adj-R^2 value will penalize the statistic for inclusion of regressors that are not correlated with the outcome.

Thus, the usefulness of a modified linear model for the prediction of radiosensitivity in comparison to the existing model is assessed in terms of the least squares model fit parameter Adj-R^2. Once a set of probesets is selected, a full model combining selected probesets/genes and biological parameters can then be considered.

This effectively reduces the impact of a large number of gene expressions by forcing the biological parameters into the equation.

6.5 Analysis of Fit for the Linear Models

The original gene expression-only model, as well as the additive and the interactive linear regression models were used on 48 of the NCI60 panel of human cancer cell lines. The analysis was performed for each probeset and the model fit parameter (Adj-R^2) was used to determine if the model improved by inclusion of the covariates. Figure 6-1 shows a box plot of the Adj-R^2 values from modeling each probeset individually when correlated with radiation response in the 48-cell line database for each of the three linear models. In the original gene expression-only model, relatively fewer probesets could achieve a model fit better than Adj-R^2=0.2 (< 30 of the 7129 probesets), with the best fit being just above Adj-R^2=0.3. The least squares fit for the additive model as well as the interactive models improve considerably. The average fit for the additive model was found to be Adj-R^2=0.28 with a maximum fit of Adj-R^2=0.48. With the interactive model, the average fit improved to Adj-R^2=0.6 with a maximum value of Adj-R^2=0.84.

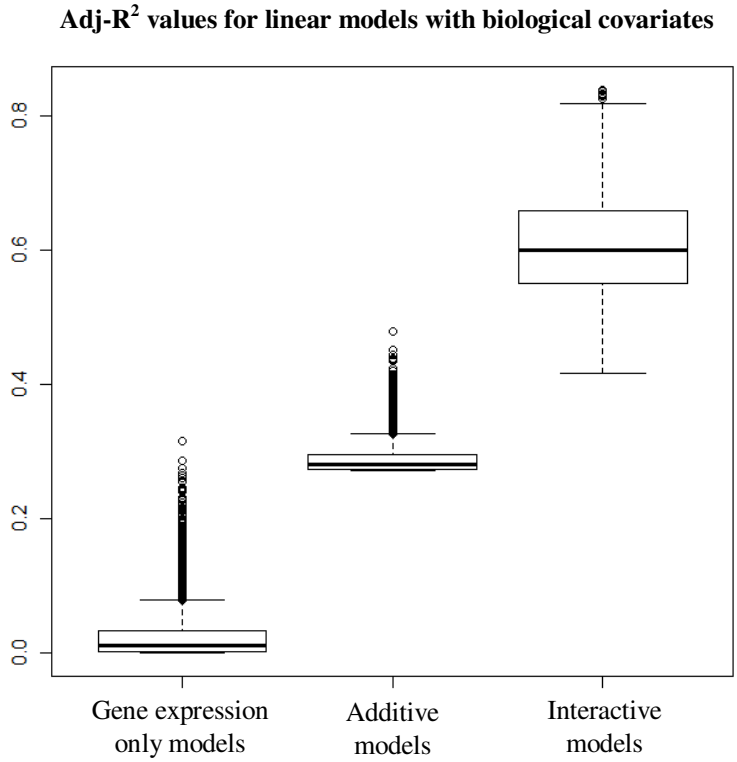


Figure 6-1: Adj-R^2 values for linear equations fitting SF2 on 48 cell lines

6.6 Verification of Model Fit

Figure 6-1 indicates that the inclusion of biological variables significantly improved the ability of most genes to describe the relationship between gene expression and radiosensitivity in a linear regression model. However, the inclusion of additional parameters and their interactions within the same equation almost certainly leads to over fitting. Since all genes are considered separately with the addition of these parameters, the use of this strategy for ranking genes does not require that over-fitting not occur. However, it is hypothesized that for some probesets and biological variables, the interaction will, in some instances, provide significantly better model fit.

An observed improvement in Adj-R^2 value of expanded linear models using biological variables such as RAS, p53 and TO could be from the addition of covariates that tends to improve the overall fit of the model regardless of information content [65]. However, this also suggests that an improvement in Adj-R^2 value of the linear fit can be similarly obtained when adding a randomly generated variable into the model instead of a variable that carries biological significance. It is hypothesized that the improvement in the model fit due to inclusion of the biological covariates is due to relevant biological information contained in the covariates. Random variables that do not have any meaningful information and are uncorrelated to the outcome are expected to produce models with lower Adj-R^2 values.

The random variables created for exploring the effect of RAS and p53 were created and uniformly distributed into two states (one each for the mutated and wild-type status). The frequencies of these states were similar to the true distributions in the data. Similarly, a random variable was defined for TO, with each sample being assigned a tissue type at random. This new dataset with randomly assigned biological parameters was used to develop the basic and expanded linear models as described earlier.

Table 6-1 documents the change in the model fit (ΔR^2: difference in Adj-R^2 values of two models) when terms are added to a linear model. The table documents the average difference in Adj-R^2 values observed across all the probesets. The variances of the measured ΔR^2 values were very small for each tabulated result (<0.006) and hence are not shown in the table. Both the change in fit from clinical indicators and randomly generated variables are recorded. For example, consider the linear model that includes gene expression values only. The Adj-R^2 of the model is expected to change when a

covariate (e.g. TO) is added into the linear model (e.g. additive). This change in the Adj-R^2 value obtained by including the additional covariate (ΔR^2=0.254) can be compared to the change obtained by including a randomly generated covariate (ΔR^2=0.256), as seen in Table 6-1. In this example, the finding suggests that the inclusion of TO provided no more information than would be expected by chance. The biological covariates are expected to carry meaningful information and hence expected to increase the model fit when included in a linear model. This result might suggest that the biological variables may not be adding much information to the model, and the improvement in model fit may be due to over-fitting of the data.

However, when the random variables are included in the expanded interaction models, the true impact of the biological variables becomes more apparent. For example, the additive model considered earlier included TO and gene expression. When RAS is included in this model in an additive manner, the model-fit improves by ΔR^2=0.256. The same improvement is observed when the random variable is added to the GeneEx: TO model (ΔR^2 = 0.257). When RAS is included into the interaction model, the correlation improves as before by 0.272. However, the interaction of the random variables for TO and RAS does not provide any meaningful information for modeling, and the correlation drops by 0.213 (ΔR^2 = -0.213).

A similar behavior can be observed for each biological covariate (RAS, p53 and TO), where inclusion of the variable in an additive manner improves the model fit just as well as including randomly generated variables. Inclusion of the variables in an interaction model improves the fit considerably, as opposed to the random variables which always causes a poorer fit to the data. This behavior is observed even when

including all three terms in the interaction models, where the Adj-R^2 improves by 0.317 but the random variables cause a drop in correlation (ΔR^2 = -0.103).

Table 6-1: Change in Adj-R^2 value (ΔR^2) obtained by adding terms and complexity to the linear model. Results obtained with clinical indicators TO, RAS and p53 are compared to Adj-R^2 values obtained using random variable for each indicator.

Model terms	Model Comparison	Mean ΔR^2 value	
		Clinical indicators	Random Variables
GeneEx: TO	GeneEx only vs. Additive	0.254	0.256
	Additive vs. Interaction	0.134	0.146
GeneEx: RAS	GeneEx only vs. Additive	0.060	0.004
	Additive vs. Interaction	0.030	0.031
GeneEx: p53	GeneEx only vs. Additive	0.026	0.0007
	Additive vs. Interaction	0.016	0.031
GeneEx: TO: RAS	Basic vs. Additive	0.256	0.257
	Additive vs. Interaction	0.272	-0.213
GeneEx: TO: p53	Basic vs. Additive	0.262	0.257
	Additive vs. Interaction	0.198	-0.211
GeneEx: RAS: TO	Basic vs. Additive	0.256	0.257
	Additive vs. Interaction	0.272	-0.214
GeneEx: RAS: p53	Basic vs. Additive	0.062	0.022
	Additive vs. Interaction	0.042	0.024
GeneEx: p53: TO	Basic vs. Additive	0.262	0.257
	Additive vs. Interaction	0.198	-0.212
GeneEx: p53: RAS	Basic vs. Additive	0.062	0.022
	Additive vs. Interaction	0.042	0.024
GeneEx: TO: RAS: p53	Basic vs. Additive	0.265	0.258
	Additive vs. Interaction	0.317	-0.103
GeneEx: RAS: TO: p53	Basic vs. Additive	0.265	0.258
	Additive vs. Interaction	0.317	-0.103
GeneEx: p53: TO: RAS	Basic vs. Additive	0.265	0.258
	Additive vs. Interaction	0.317	-0.103

Since inclusion of the biological variables in the additive models may not provide more information than inclusion of random variables in the model, there is a risk of over-

fitting to the data by including biological covariates. However, the behavior of the random variables in the interaction models clearly indicate that the biological variables do provide meaningful information, and rather than cause over-fitting of the model to the data, the biological covariates can be used to create a better model for predicting radiosensitivity of the tumor cells.

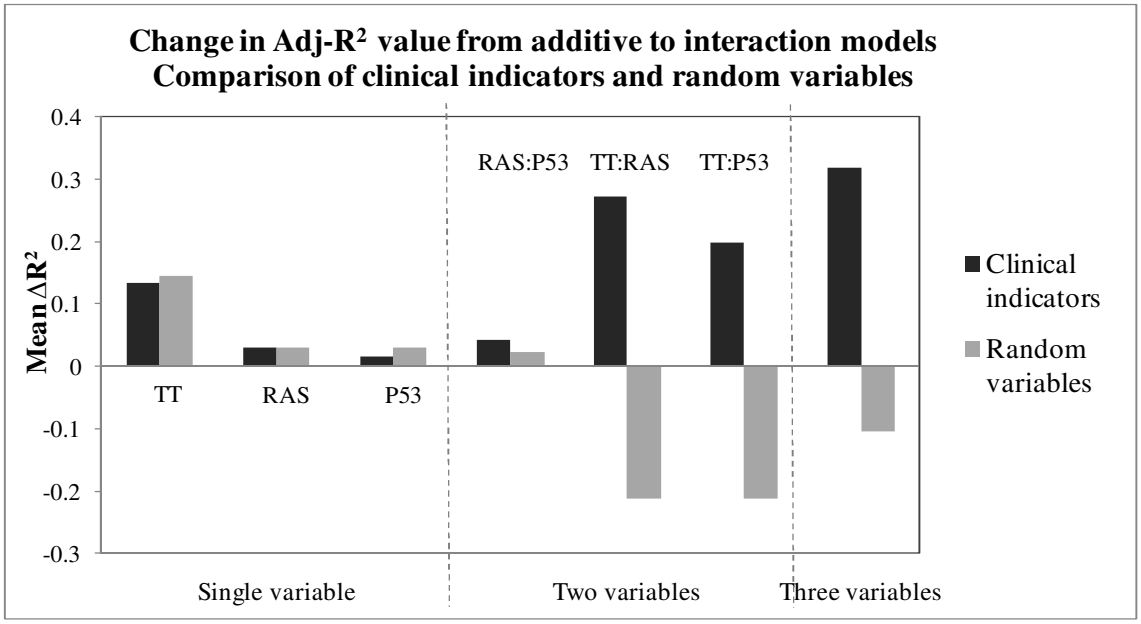


Figure 6-2: Change in Adj-R^2 values obtained by including interaction terms in the linear model

Figure 6-2 shows the change in the Adj-R^2 value when a term is added to a model. The interaction of random variables with gene expression data alone provides a marginal improvement in the fit, as expected by the mathematical construct of the modeling process. However, in the interaction models, when two or more random variables interact, the lack of information in each variable translates into poorer fit of the linear model to the radiation sensitivity outcome. In contrast, the interaction of the biological

variables adds more information to the linear model, as shown by the improvement in Adj-R^2 values in Table 6-1 and Figure 6-2.

However, it is intriguing that not all variables considered had similar impact in improving the model. For example RAS was significantly more important than p53 in improving the model. This observation suggests that at least part of the improvement obtained by the expanded linear models is due to a better representation of biology.

6.7 Summary

The prediction accuracy of a published model for the radiosensitivity of tumor cell lines was found to decrease when adding more cell lines. Biological indicators such Ras mutational status, p53 mutational status and the tissue of origin were included in the multivariable linear regression model in an attempt to better model the underlying biology [95]. The additive and interaction models created by including these variables at the probeset level were shown to provide a better fit for prediction of radiosensitivity than the gene expression model alone. Since the inclusion of additional variables is expected to enhance model fit, the effect of the biological indicators was compared with the effect of randomly generated variables for model fit. It was shown that the biological indicators could create more meaningful linear models than random variables.

CHAPTER 7 CONCLUSIONS AND FUTURE WORK

7.1 Conclusions

Cancer is the second leading cause of deaths in the United States. Gene expression microarrays are used to find reliable biomarkers of tumor for cancer diagnostics, treatment planning and patient management with an aim of eventually reducing fatality due to the disease.

Some of the fundamental methodological issues with successfully extracting reliable biomarkers of cancer prognosis were described using a colorectal cancer gene expression dataset as a case study. Classifier models that perform well on other datasets, e.g. identifying survival rates of lung cancer patients, performed poorly in predicting survival for the colorectal cancer dataset. Classifier performance was shown to be influenced by the intrinsic complexity of the dataset. Three measures of complexity were proposed to obtain a relative measure of the expected predictive accuracy for the classifier models. A specific measure of complexity (ϕ) was shown to correlate very closely (R^2=0.82) with expected classifier performance. The measure indicated that the survival dataset for colorectal cancer was complex, and further work was necessary to extract reliable prognostic signatures.

Data reduction using quantization methods was proposed as the first step in reducing the complexity of the colorectal cancer dataset. Since typical microarray gene expression datasets consist of very high resolution data, it was hypothesized that limiting

the numerical resolution of the data could yield simpler datasets and consequently, better predictive accuracy for classifier models. Three methods of quantization were proposed to limit the data resolution in different ways. While each method was shown to improve the predictive accuracy of the classifier models, the noise removal method was shown to maximize classifier performance. Predictive accuracy on the colorectal cancer dataset was shown to increase from 56% to 68% and the same technique was shown to enhance classifier performance from 67% to 90% accuracy on a lung adenocarcinoma dataset.

Dimensionality reduction was proposed as the second step in addressing the complexity of the heterogeneous data. A random subspace based technique using cost-sensitive analysis was proposed as a multivariable feature selection method. Multiple genes are known to be active in molecular pathways, and multiple pathways can be active in a heterogeneous tumor. The random subspace technique was designed to address this underlying biology of the tumor. Extraction of multiple sets of genes that were correlated with survival in a multivariate manner was shown to produce more accurate classifiers (68% accuracy) than a univariate feature selection method (56% accuracy).

As mentioned earlier, the goal of these gene expression microarray studies is to identify reliable biomarkers of cancer to aid in patient management. Biological indicators, for example the mutational status of certain genes such as Ras or p53, have been routinely used as clinical factors for patient selection and treatment management. A method was proposed to integrate these clinical indicators in developing a gene expression signature for predicting radiation sensitivity of tumor cells. While inclusion of biological indicators can be expected to provide meaningful information, it can also lead to over fitting of the model to the training data. Experiments using random variables were

used to demonstrate that inclusion of Ras mutational status, p53 mutational status and the tissue of origin as biological indicators in a gene expression model enhanced the correlation of the model to the underlying biology without over fitting to the data.

7.2 Future Work

Chapter 3 demonstrated that gene expression datasets may be intrinsically complex, and classifier models may fail on such datasets due to several reasons such as statistical over-fitting, high dimensionality or high resolution of the data. The measure of complexity proposed here can be used to investigate the expected classifier performance when working with models such as those described in Chapter 3. However if a different classifier model is used for analysis, the specific mathematical basis of the classification must be used to design a more applicable measure of complexity. The methodology proposed at the end of Chapter 3 may be used to develop newer measures.

Three methods of quantization reduced data complexity and enhanced classifier accuracies in the colorectal and lung adenocarcinoma datasets. These datasets contained different ranges of numerical information. The colorectal data was represented in a log-2 format and a range of 0.0-15.0 and the lung dataset had a range of 0.0-6000.0. A discussion of the methods indicated that the parameter selection for the quantization methods is dependent on the numerical information and spread of values. Further work can be done in enhancing these methods to work with datasets that have sparser ranges than the datasets used in the chapter.

The random subspace technique was designed to better model the underlying biology and functioning of genes in molecular pathways. The technique was described

using decision trees that inherently select a set of important genes from within a subspace of the dataset. Other selection methods may be investigated in the same random subspace setup, such as the Cox proportional hazards model or the multiple linear regression model to select genes from a random projection of the data. These multiple subsets of genes may then be used as an input to the multivariable models to generate predictive signatures.

A method was described to integrate biological indicators into gene expression models to enhance the modeling of the underlying biology. Three indicators were used in modeling radiation sensitivity of tumor cells. The experiments conducted to test for statistical over-fitting of the data indicated that the biological indicators provided significant information for modeling radiation sensitivity. Other indicators may be tested in the same framework to include more biological information into gene expression models thereby creating more powerful predictors for clinical use.

LIST OF REFERENCES

1. Center for Disease Control. Causes of Death. [Online] [Cited: June 15, 2010.] www.cdc.gov.

2. David G Beer, S L R Kardia, Chiang-Ching Huang, TJ Giordano, Albert M Levin, David E Misek, Lin Lin, Guoan Chen, Tarek G Gharib, Dafydd G Thomas, Michelle L Lizyness, Rork Kuick, Satoru Hayasaka, J M G Taylor, M D Iannettoni, M B Orringer, Samir Hanash. Gene-expression profiles predict survival of patients with lung adenocarcinoma. Nat Med, 2002, Vols. 8(8):816–824.

3. L. J. van 't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, K. van der Kooy, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend. Gene expression profiling predicts clinical outcome of breast cancer. Nature, Jan 31 2002, Vols. vol. 415, pp. 530-6.

4. T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science, Oct 1999, Vols. 286(5439):531–537.

5. Dalton WS, Friend SH. Cancer biomarkers - an invitation to the table. Science, 2006, 1165-1168, Vol. 312.

6. E. J. Ambrose, F. J. C. Roe. Biology of Cancer. Ellis Horwood Limited, Sussex, England, 1975.

7. Douglas Hanahan, Robert A. Weinberg. The Hallmarks of Cancer. Cell, Jan 7 2000, 57-70, Vol. 100.

8. David P. Clark, Lonnie D. Russell. Molecular biology made simple and fun. Cache River Press, A division of Quick Publishing, St Louis, MO, USA LC, 2005. ISBN 1-889899-07-0.

9. Brown PO, Botstein D. Exploring the new world of the genome with DNA microarrays.Nature Genetics, 1999, 33-37, Vol. 21 (1 Suppl).

10. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 1995, Vols. 270 (5235): 467-470.

11. Helen C. Causton, John Quackenbush and Alvis Brazma. A Beginner's Guide - Microarray Gene Expression Data Analysis. Blackwell Science Ltd, a Blackwell Publishing company, 2003. 1-4051-2735-X.

12. Greg Bloom, Ivana V Yang, David Boulware, Ka Yin Kwong, Domenico Coppola, Steven Eschrich, John Quackenbush, and Timothy J Yeatman. Multi-platform, multi-site, microarray-based human tumor classification. Am J Pathol, Jan 2004, Vols. 164(1):9–16.

13. Steven Eschrich, Ivana Yang, Greg Bloom, Ka Yin Kwong, David Boulware, Alan Cantor, Domenico Coppola, Mogens Kruhoffer, Lauri Aaltonen, Torben F. Orntoft, John Quackenbush, and Timothy J. Yeatman. Molecular staging for survival prediction of colorectal cancer patients. J Clin Oncol, May 2005, Vols. 23(15):3526–3535.

14. M J van de Vijver, Y D He, L J van't Veer, H Dai, A A M Hart, D W Voskuil, G J Schreiber, J L Peterse, C Roberts, MJ Marton, M Parrish, D Atsma, A Witteveen, A Glas, L Delahaye, T van der Velde, H Bartelink, S Rodenhuis, ET Rutgers, S H Friend, R Bernards. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med, 2002, Vols. 347(25):1999–2009.

15. V G Tusher, R Tibshirani, and G Chu. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A, 2001, Vols. 98(9):5116–5121.

16. Wessels LF, Reinders MJ, Hart AA, Veenman CJ, Dai H, He YD, van't Veer LJ. A protocol for building and evaluating predictors of disease state based on microarray data. Bioinformatics (Oxford, England), 2005, Vols. 21(19):3755-3762.

17. R, Simon. Diagnostic and prognostic prediction using gene expression profiles in high-dimensional microarray data. Br J Cancer , 2003, Vols. 89(9):1599-1604.

18. Chuang LY, Ke CH, Chang HW, Yang CH. A Two-Stage Feature Selection Method for Gene Expression Data. Omics , 2009.

19. Zervakis M, Blazadonakis ME, Tsiliki G, Danilatou V, Tsiknakis M, Kafetzopoulos D. Outcome prediction based on microarray analysis: a critical perspective on methods. BMC Bioinformatics , 2009, Vol. 10:53.

20. Wang J, Bo TH, Jonassen I, Myklebost O, Hovig E. Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data. BMC Bioinformatics, 2003, Vol. 4:60.

21. Wouters L, Gohlmann HW, Bijnens L, Kass SU, Molenberghs G, Lewi PJ. Graphical exploration of gene expression data: a comparative study of three multivariate methods. Biometrics , 2003, Vols. 59(4):1131-1139.

22. Hand DJ, Heard NA. Finding groups in gene expression data. Journal of Biomedicine & Biotechnology, 2005, Vols. 2005(2):215-225.

23. van Houwelingen HC, Bruinsma T, Hart AA, Van't Veer LJ, Wessels LF. Cross-validated Cox regression on microarray gene expression data. Stat Med , 2006, Vols. 25(18):3201-3216.

24. Rangel J, Nosrati M, Torabian S, Shaikh L, Leong SP, Haqq C, Miller JR, 3rd, Sagebiel RW, Kashani-Sabet M. Osteopontin as a molecular prognostic marker for melanoma. Cancer, 2008, Vols. 112(1):144-150.

25. Shoemaker JS, Lin SM (eds.). Methods of Microarray Data Analysis IV. Springer, 2004.

26. U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci U S A, June 8 1999, Vols. vol. 96, pp. 6745-50.

27. J. F. Torres-Roca, S. Eschrich, H. Zhao, G. Bloom, J. Sung, S. McCarthy, A. B. Cantor, A. Scuto, C. Li, S. Zhang, R. Jove, and T. Yeatman. Prediction of radiation sensitivity using a gene expression classifier. Cancer Res, Aufg 15 2005, Vols. vol. 65, pp. 7169-76.

28. S. Ramaswamy, K. N. Ross, E. S. Lander, and T. R. Golub. A molecular signature of metastasis in primary solid tumors. Nat Genet, Jan 2003, Vols. vol. 33, pp. 49-54.

29. Yeatman, T. J. Predictive Biomarkers: Identification and Verification. J. Clin. Oncol. 27, 2743-2744 , 2009, Vols. 27, 2743-2744 .

30. SH, Curry. Translational science: past, present, and future. Biotechniques, 2008 Feb, Vols. 44(2):ii-viii.

31. van't Veer LJ, Bernards R. Enabling personalized cancer medicine through analysis of gene-expression patterns. Nature, 2008 Apr 3, Vols. 452(7187):564-70.

32. Burges, Cristopher J. C. Data Mining and Knowledge Discovery, A Tutorial on Support Vector Machines for Pattern Recognition. Kluwer Academic Publishers, Boston, MA, USA, 1998.

33. Carlin JB, Doyle LW. Statistics for clinicians: 4: Basic concepts of statistical reasoning: hypothesis tests and the t-test. Journal of paediatrics and child health , 2001, Vols. 37(1):72-77.

34. Helen C. Causton, John Quackenbush and Alvis Brazma. A Beginner's Guide - Microarray Gene Expression Data Analysis. © 2003 by Blackwell Science Ltd, a Blackwell Publishing company .

35. C. Brambilla, F. Fievet, M. Jeanmart, F. de Fraipont, S. Lantuejoul, V. Frappat, G. Ferretti, P.Y. Brichon and D. Moro-Sibilot. Early detection of lung cancer: role of biomarkers. Eur Respir J, 2003, Vols. 21:36-44.

36. Ioannidis, J. P. A. Is Molecular Profiling Ready for Use in Clinical Decision Making? Oncologist. 12, 301-311 , 2007

37. Slonim, Donna K. From patterns to pathways: Gene expression data analysis comes of age. Nature Genetics Br J Cancer, 2002, Vols. 32, 502 - 508, 92(12):2114-21.

38. Steven Eschrich, Timothy J. Yeatman. DNA microarrays and data analysis: An overview. Surgery, May 2004, Vols. Vol. 136, No. 3.

39. E. J. Ambrose, F. J. C. Roe. Biology of Cancer. Sussex, England: Ellis Horwood Limited.

40. Goodsell, David S. The molecular perspective: The ras Oncogene. http://theoncologist.alphamedpress.org/cgi/content/full/4/3/263. [Online]

41. Goodsell, David S. The Molecular Perspective: p53 Tumor Suppressor. The Oncologist; AlphaMed Press , April 1999, Vols. Vol. 4, No. 2, 138-139.

42. Marieb, Elaine. Human Anatomy and Physiology. Pearson Education Inc. 2002

43. Human Genome Project Information. [Online] genomics.energy.gov, 06 2010, 23. [Cited: 06 29, 2010.] http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml.

44. M, Dufva. Introduction to microarray technology. Methods Mol Biol , 2009, Vols. 529: 1-22.

45. Affymetrix. Affymetrix. [Online] http://www.affymetrix.com/index.affx.

46. Kamath, Vidya. Master's Thesis: Use of Random Subspace Ensembles on Gene Expression Profiles to Enhance the Accuracy of Survival Prediction for Colon Cancer Patients. University of South Florida, 2005.

47. Affymetrix. Microarray data normalization techniques. [Online] 2001. [Cited: 06 29, 2010.] Affymetrix.http://www.sbeams.org/sbeams/doc/Microarray/affy_help_pages/isb_help.php?help_page=Analysis/Pipeline/Normalization_methods.xml.

48. CH, Ding. Unsupervised feature selection via two-way ordering in gene expression analysis. Bioinformatics (Oxford, England), 2003, Vols. 19(10):1259-1266.

49. Yun L, Bao-Liang L, Zhong-Fu W. A Hybrid Method of Unsupervised Feature Selection Based on Ranking. Pattern Recognition, 2006 ICPR 2006 18th International Conference on, 2006, Vols. 2006: 687-690.

50. Mitra P, Murthy CA, Pal SK. Unsupervised feature selection using feature similarity. Pattern Analysis and Machine Intelligence, IEEE Transactions, 2002, Vols. 24(3):301-312.

51. Christos B, Michael WM, Petros D. Unsupervised feature selection for principal components analysis. Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, Vol. ACM. Las Vegas, Nevada, USA

52. Yan X, Zheng T. Selecting informative genes for discriminant analysis using multigene expression profiles. BMC Genomics , 2008, Vol. 9 Suppl 2:S14.

53. Pirooznia M, Yang JY, Yang MQ, Deng Y. A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics , 2008, Vol. 9 Suppl 1:S13.

54. Jeffery IB, Higgins DG, Culhane AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC bioinformatics , 2006, Vol. 7:359.

55. Liu X, Krishnan A, Mondry A. An entropy-based gene selection method for cancer classification using microarray data. BMC bioinformatics , 2005, Vol. 6:76.

56. Glinsky, G. V. Anti-adhesion cancer therapy. Cancer Metastasis Rev, Jun 1998, Vols. vol. 17, pp. 177-85.

57. Cheng Fan, Daniel S. Oh, Lodewyk Wessels, Britta Weigelt, Dimitry S.A. Nuyten, Andrew B. Nobel, Laura J. van't Veer, and Charles M. Perou. Concordance among Gene-Expression– Based Predictors for Breast Cancer. The new England journal o f medicine, 2006, Vols. 355:560-9.

58. Basu, Tin Kam Ho and M. Complexity measures of supervised classification problems. IEEE Transactions on Pattern Analysis and Machine Intelligence, Mar 2002, Vols. 24(3):289–300.

59. Jane E. Staunton, Donna K. Slonim,Hilary A. Coller, Pablo Tamayo, Michael J. Angelo, Johnny Park, Uwe Scherf, Jae K. Lee, William O. Reinhold, John N. Weinstein, Jill P. Mesirov, Eric S. Lander and Todd R. Golub. Chemosensitivity prediction by transcriptional profiling. Proceeding of the National Academy of Sciences of the United States of America, September 11, 2001, Vols. vol. 98 no. 19 10787-10792. 10.1073/pnas.191368598PNAS .

60. Ho, Tin Kam. The random subspace method for constructing decision forests. IEEE Transactions on Pattern Analysis and Machine Intelligence, Aug 1998, Vols. 20(8):832–844.

61. Quinlan, J. Ross. C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.

62. Haykin, Simon. Neural Networks: A Comprehensive Foundation. Prentice Hall, Inc, New Jersey, 1999.

63. Tan AH, Pan H. Predictive neural networks for gene expression data analysis. Neural Networks , 2005, Vols. 18(3):297-306.

64. Eibe Frank, Mark Hall, Len Trigg, Geoffrey Holmes, and Ian H Witten. Data mining in bioinformatics using weka. Bioinformatics, 2004, Vols. 20(15):2479–2481.

65. D. C. Montgomery, G. C. Runger, N. F. Hubele. Engineering Statistics. John-Wiley & Sons, Inc., 1997.

66. Kleinbaum, D. G. Survival Analysis: A self-learning text. Springer-Verlag New York, Inc. , 1996.

67. Allison, Paul D. Survival Analysis Using SAS: A practical guide. SAS Institue Inc. , 1995. 1-55544-279-X. Cary, NC, USA

68. H. Witten, E. Frank. Data Mining. Morgan Kauggman Publishers, 2000.

69. Greenhalgh, T. How to read a paper. Papers that report diagnostic or screening tests. BMJ, Aug 30 1997, Vols. vol. 315, pp. 540-3.

70. Macready, D.H. Wolpert and W.G. No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation, Apr 1997, Vols. 1(1):67–82.

71. Klee EW, Erdogan S, Tillmans L, Kosari F, Sun Z, Wigle DA, Yang P, Aubry MC, Vasmatzis G. Impact of sample acquisition and linear amplification on gene expression profiling of lung adenocarcinoma: laser capture micro-dissection cell-sampling versus bulk tissue-sampling. BMC Med Genomics., 2009 Mar 9, Vol. 2:13.

72. Wei Keat Lim, Kai Wang, Celine Lefebvre and Andrea Califano. Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. Bioinformatics, 2007, Vols. 23(13):i282-i28.

73. R. O. Duda, P. E. Hart, D. G. Stork. Pattern Classification. John Wiley & Sons, Inc, 2001.

74. Liu Z, Tan M, Jiang F. Regularized F-measure maximization for feature selection and classification. Journal of biomedicine & biotechnology , 2009, Vol. 2009:617946.

75. Gil Chu, Balasubramanian Narasimhan , Robert Tibshirani, Virginia Tusher. SAM: “Significance Analysis of Microarrays”: Users guide and technical document.

76. Kamath, Vidya, Yeatman, Timothy and Eschrich, Steven. Toward a Measure of Classification Complexity in Gene Expression Signatures. EMBC'08, 2008. Vancouver, CA

77. Mitra P, Majumder DD. Feature selection and gene clustering from gene expression data. Pattern Recognition, 2004 ICPR 2004 Proceedings of the 17th International Conference on, 2004, Vols. 2004: 343-346 Vol.342-343-346.

78. Zhou X, Wang X, Dougherty ER. Binarization of microarray data on the basis of a mixture model. Mol Cancer Ther , 2003, Vols. 2(7):679-684.

79. Pe'er D, Regev A, Elidan G, Friedman N. Inferring subnetworks from perturbed expression profiles. Bioinformatics (Oxford, England), 2001, Vols. 17 Suppl 1:S215-224.

80. Tchagang AB, Tewfik AH. DNA Microarray Data Analysis: A Novel Biclustering Algorithm Approach. EURASIP Journal on Applied Signal Processing , 2006, Vol. 2006(Article ID 59809).

81. Rousseeuw, Leonard Kaufman and Peter J. Finding Groups in Data: An introduction to cluster analysis. John Wiley & Sons, Inc, 1990. 0-471-87876-6. Canada

82. Eschrich S, Jingwei K, Hall LO, Goldgof DB. Fast accurate fuzzy clustering through data reduction. IEEE Transactions on Fuzzy Systems, 2003, Vols. 11(2):262-270.

83. Eric J. Bernhard, W. Gillies McKenna, Andrew D. Hamilton, Said M. Sebti, Yimin Qian, Junmin Wu, and Ruth J. Muschel. Inhibiting ras prenylation increases the radiosensitivity of human tumor cell lines with activating mutations of ras oncogenes. Cancer Research, April 15 1998, vol. 58; 1754.

84. Malumbres M, Barbacid M. RAS oncogenes: the first 30 years. Nat. Rev. Cancer, June 2003, Vols. 3 (6): 459–65.

85. Eric Bair, Trevor Hastie, Debashis Paul and Robert Tibshirani. Prediction by supervised principal components. JASA, Sept 15 2004.

86. André Fujita, Luciana Rodrigues Gomes, João Ricardo Sato, Rui Yamaguchi, Carlos Eduardo Thomaz, Mari Cleide Sogayar and Satoru Miyano. Multivariate gene expression analysis reveals functional connectivity changes between normal/tumoral prostates. BMC Systems Biology, 5 Dec 2008, Vol. 2: 106.

87. Jun Liu, Sanjay Ranka and Tamer Kahveci. Classification and feature selection algorithms for multi-class CGH data. Bioinformatics , 2008, Vols. 24(13):i86-i95.

88. Shenghuo Zhu, Dingding Wang, Kai Yu, Tao Li, Yihong Gong. Feature Selection for Gene Expression Using Model-Based Entropy. IEE/ACM Transactions on Computational Biology and Bioinformatics, Jan-Mar 2010, Vol. 7 No. 1.

89. Elkan, Charles. The Foundations of Cost-Sensitive Learning. Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01), 2001.

90. Drummond, Robert C. Holte and Chris. Cost-sensitive Classifier Evaluation using Cost Curves. PAKDD'08: Proceedings of the 12th Pacific-Asia conference on Advances in knowledge discovery and data mining; 2008

91. Nitesh V. Chawla, David A. Cieslak, Lawrence O. Hall, Ajay Joshi. Automatically countering imbalance and its empirical relationship to cost. Data Mining and Knowledge Discovery, Oct 2008, Vol. 17.

92. Nitesh V. Chawla, Lawrence O. Hall, Ajay Joshi. Wrapper-based computation and evaluation of sampling methods for imbalanced datasets. International Conference on Knowledge Discovery and Data Mining, Chicago, Illinoi, 2005.

93. Kamath, Vidya P., Hall, L.O., Yeatman, T.J., Eschrich, S.A.. Multivariate Feature Selection using Random Subspace Classifiers for Gene Expression Data. IEEE International Conference on Bioinformatics and Biomedical Engineering (BIBE'07), 2007.

94. Guyon I, Weston J, Barnhill SMD, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, 2002, Vols. 46(1–3):389-422.

95. Steven A. Eschrich, Jimmy Pramana, Hongling Zhang, Haiyan Zhao, David Boulware, Ji-Hyun Lee, Gregory Bloom, Caio Rocha-Lima, Scott Kelley, Douglas P. Calvin, Timothy J. Yeatman, Adrian C. Begg, and Javier F. Torres-Roca. A Gene Expression Model of Intrinsic Tumor Radiosensitivity: Prediction of Response and Prognosis after Chemoradiation. Int J Radiat Oncol Biol Phys. , Oct 2009, Vols. 1;75(2):489-96.

96. Buffa FM, Davidson SE, Hunter RD, et al. Incorporating biologic measurements (SF(2), CFE) into a tumor control probability model increases their prognostic significance: a study in cervical carcinoma treated with radiation therapy. Int J Radiat Oncol Biol Phys, 2001, Vols. 50:1113-1122.

97. Yijun Sun, Steve Goodison, Jian Li, Li Liu and William Farmerie. Improved breast cancer prognosis through the combination of clinical and genetic markers. Bioinformatics, 2007. Vol. 23 no. 1 , pages 30–37

98.Venter JC et al. The sequence of the human genome. Science. 2001 Feb 16, 2001, Vols. 291(5507):1304-51.

99. Lander et.al. Initial sequencing and analysis of the human genome. Nature, 2001 Feb 15, Vols. 409(6822):860-921.

100. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics, 2007 Oct 1, Vols. 23(19):2507-17.

ABOUT THE AUTHOR

Vidya Kamath graduated at top of her class with a B.E. in Medical Electronics from Bangalore University in 1999. She worked at Dornier India Medical Systems, Hyderabad for 1.5 years and GE Global Research Center, Bangalore for 2.5 years. In her short professional career, she authored five technical papers and was granted two patents in the field of medical image analysis.

Vidya started her graduate work in Biomedical Engineering at USF in Fall of 2004 and has been working towards a doctorate after graduating with a masters degree in Fall 2005. As a graduate student she has authored eight technical publications including five peer reviewed papers. She has won IEEE student travel awards and presented her work at various conferences. She was the president of the student chapter of Biomedical Engineering Society at USF (2006-2007). She is a member of the Tau-Beta-Pi honor society and a student member of IEEE.