11-4-2005

Use of Random Subspace Ensembles on Gene Expression Profiles in Survival Prediction for Colon Cancer Patients

Vidya Kamath
University of South Florida

Use of Random Subspace Ensembles on Gene Expression Profiles in Survival Prediction for Colon Cancer Patients

by

Vidya Kamath

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Biomedical Engineering
Department of Chemical Engineering
College of Engineering
University of South Florida

Co-Major Professor: Rangachar Kasturi, Ph.D.
Co-Major Professor: Dmitry Goldgof, Ph.D.
Lawrence Hall, Ph.D.
Steven Eschrich, Ph.D.

Date of Approval:
November 4, 2005

Keywords: Microarray, Bioinformatics, Data Mining,
Feature Selection, Classifiers

DEDICATION

Dedicated to our battle against cancer.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to Dr. Steven Eschrich for providing me with masterful guidance through the course of this project. I am grateful to Dr. Kasturi, Dr. Goldgof, and Dr. Hall for being on my committee and guiding me with their expertise in the area of data mining. Special thanks are due to Dr. Yeatman and Dr. Gregory Bloom for insightful discussions on understanding cancer and gene expressions.

I also extend my thanks to Dr. Niranjan Pai, Kurt Kramer and Andrew Hoerter for their helpful interactions during the implementation of the project.

Finally, I am grateful to H. Lee Moffitt Cancer Center and Research Institute for allowing me to work with the microarray gene expression data, and to the colorectal cancer patients whose consent made this work possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

USE OF RANDOM SUBSPACE ENSEMBLES ON GENE EXPRESSION PROFILES IN SURVIVAL PREDICTION FOR COLON CANCER PATIENTS

Vidya Kamath

ABSTRACT

Cancer is a disease process that emerges out of a series of genetic mutations that cause seemingly uncontrolled multiplication of cells. The molecular genetics of cells indicates that different combinations of genetic events or alternative pathways in cells may lead to cancer. A study of the gene expressions of cancer cells, in combination with the external influential factors, can greatly aid in cancer management such as understanding the initiation and etiology of cancer, as well as detection, assessment and prediction of the progression of cancer.

Gene expression analysis of cells yields a very large number of features that can be used to describe the condition of the cell. Feature selection methods are explored to choose the best of these features that are most relevant to the problem at hand. Random subspace ensembles created using these selected features perform poorly in predicting the 36-month survival for colon cancer patients. A modification to the random subspace scheme is proposed to enhance the accuracy of prediction. The method first applies random subspace ensembles with decision trees to select predictive features. Then, support vector machines are used to analyze the selected gene expression profiles in cancer tissue to predict the survival outcome for a patient.

The proposed method is shown to achieve a weighted accuracy of 58.96%, with 40.54% sensitivity and 77.38% specificity in predicting 36-month survival for new and unknown colon cancer patients. The prediction accuracy of the method is comparable to the baseline classifiers and significantly better than random subspace ensembles on gene expression profiles of colon cancer.

CHAPTER 1

INTRODUCTION

1.1 Introduction

Cancer is a disease process that emerges out of a series of genetic mutations that cause seemingly uncontrolled multiplication of cells [1,2]. The progress made in the area of molecular genetics in recent years has made it possible to profile the different combinations of genetic events or alternative pathways in cells that may lead to cancer. A study of the gene expressions of cancer cells, in combination with the external influential factors has shown promise in several areas of cancer management [1,3], such as understanding the initiation and etiology of cancer, as well as detection, assessment and prediction of the progression of cancer [3].

1.2 Overview of genetics

The fascinating diversity of traits amongst living beings and the transmission of traits through generations of a species led scientists and biologists to investigate the nature of heredity since the late 1600s [4]. Use of science, reason and observation led to a series of landmark discoveries that yielded a deeper insight into the functioning of living beings. Table 1.1 shows a limited list of the contributors to classical genetics along with their contributions to the field.

Table 1.1: Landmark events in the era of classical genetics [4]

Period	*Contributor*	*Contribution to genetics*
1651	William Harvey	Identification of the egg as the basis of life
1665	Robert Hooke	Discovery of cells as the basic unit of organisms
1677	Antoni van Leeuwenhoek	Discovery of sperms
1801	Erasmus Darwin	Evolution of life based on progress, development and metamorphosis
1815	Jean Baptiste Lamarck	Evolution based on acquired characteristics
1833	Robert Brown	Description of the cell nucleus
1858	Charles Darwin	Evolution by natural selection
1865	Gregor Mendel	Law of segregation and law of independent assortment for peas
1880	Eduard Strasburger	Description of mitosis
1888	Gottfried Waldeyer	Discovery of chromosomes
1890	August Weismann	Description of meiosis
1909	Wilhelm Johannsen	Definition of “genotype”, “phenotype” and “genes”
1926	Hermann J. Muller	Proposal that the gene is the basis of life
1944	Oswald Avery, Maclyn McCarty, Colin MacLeod	Establishment of DNA as genetic material
1953	Watson, Crick	Double-helix model of DNA

The era of classical genetics focused on understanding the functional behavior of cells. Cells were identified as the basic unit of life, and chromosomes as the basis of individual traits of the cell. However, it was not until the era of molecular genetics that scientists began to investigate the structural and functional properties of the chromosomes.

The discovery of deoxy-ribonucleic acid or DNA as the molecular basis of chromosomes ushered in the era of molecular genetics [4]. In 1953, Watson and Crick [5] deduced the geometric configuration of the components of DNA along a stretch of the molecule.

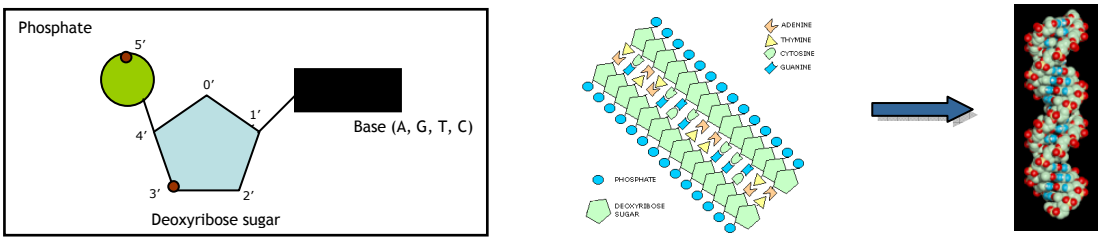
Molecular genetics involves the analysis of the exact functioning of DNA at a molecular level in the transmission of traits, and sustenance of life [2,4,6]. DNA serves as the repository of information that determines the genetic variability of an organism. It is a polymeric molecule that encodes the genetic information for an organism in an

arrangement of nucleic acid bases along the polymer chain [5,7]. A gene is a length of nucleic acids which is responsible for the transmission and expression of a hereditary characteristic [7]. Four nucleic acid bases thymine (T), adenine (A), cytosine (C) and guanine (G) are arranged in a specific sequence in a gene. This sequence determines the amino acid sequence of the polypeptide chain synthesized through the transcription or expression of the gene. A gene can be treated as a sentence that gives the step-by-step instructions for the production of the protein [7]. Each "word" in this sentence is described by a sequence of three nucleic acid bases (refer to Section 1.3).

1.3 Structure and function of DNA

Structure of DNA

The basic molecular sub-unit of DNA consists of a deoxyribose sugar, attached to a phosphate molecule on one end, and one of the four nucleic acid bases on the other[5,7].



(a) (b)

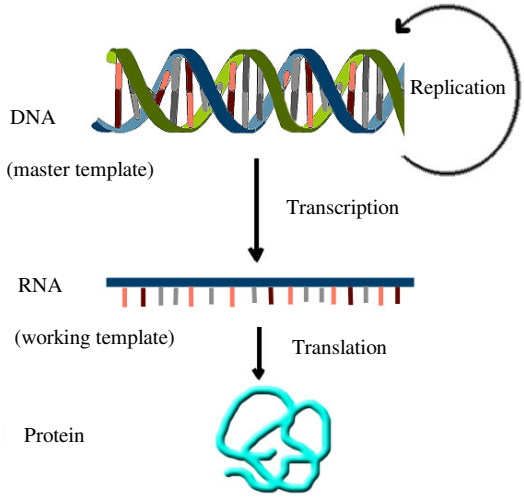
Figure 1.1: Structure of DNA (a) Basic unit of DNA (b) DNA double helix reproduced with permission from: http://www.biology-online.org

These basic molecular units attach to other such units at the 3' and the 5' position of the molecules, forming a long chain of polymeric molecules like "beads on a chain".

Further, each unit can attach to another unit at the position of the nucleic acid base. These bases cannot undergo non-specific binding: Adenine (A) bonds exclusively with Thymine (T), and Guanine (G) bonds exclusively with Cytosine (C) [7]. The bonds between the bases bring together two polymeric DNA chains like the rungs of a ladder, with the two individual strands forming the parallel sides of the ladder. Due to the oblique angle at which each of these two types of bonds can occur, the ladder twists to form a double helix structure [5,7].

Function of DNA

The central dogma of molecular genetics [7] describes the three fundamental phases in genetic information processing:



Replication:	DNA serves as a template for additional DNA synthesis
Transcription:	DNA provides a template for mRNA production
Translation:	mRNA furnishes a template for protein synthesis

Figure 1.2: The central dogma of genetic information processing

Replication

Biosynthesis of DNA occurs during cellular division or reproduction [7]. During replication, the DNA molecule functions as a template for the synthesis of two replicate

molecules which are fundamentally identical to the parent DNA. During the first stage of replication, the double stranded DNA unwinds itself to expose two single DNA strands. Each of these strands serves as a template, directing the growth of the nucleotide base sequence for the synthesis of a new complimentary strand, from the 3' to 5' end of the single-stranded template. Each of these two new complementary strands combine with one of the parent strands to form the two replicate DNA molecules. This type of synthesis is termed "semi-conservative" [7], since the parent DNA is entirely contained in the product DNAs: one of the parent strands is found in one replicate molecule, while the other strand is found in the second replicate.

The structural stability intrinsic to the formation of the base pairs reinforces the fidelity of DNA replication [7]. However, the rare occurrence of errors at the level of DNA replication could result in genetic mutations. Three basic types of errors may arise [7]:

i. Substitution, or a mismatch in base pairing during the formation of the new complimentary strands, results in the substitution of one base pair for another at a particular point in the molecule.
ii. Deletion, or the loss of a specific base pair from a particular point in the molecule
iii. Insertion, or the addition of a specific base pair at a particular point in the molecule.

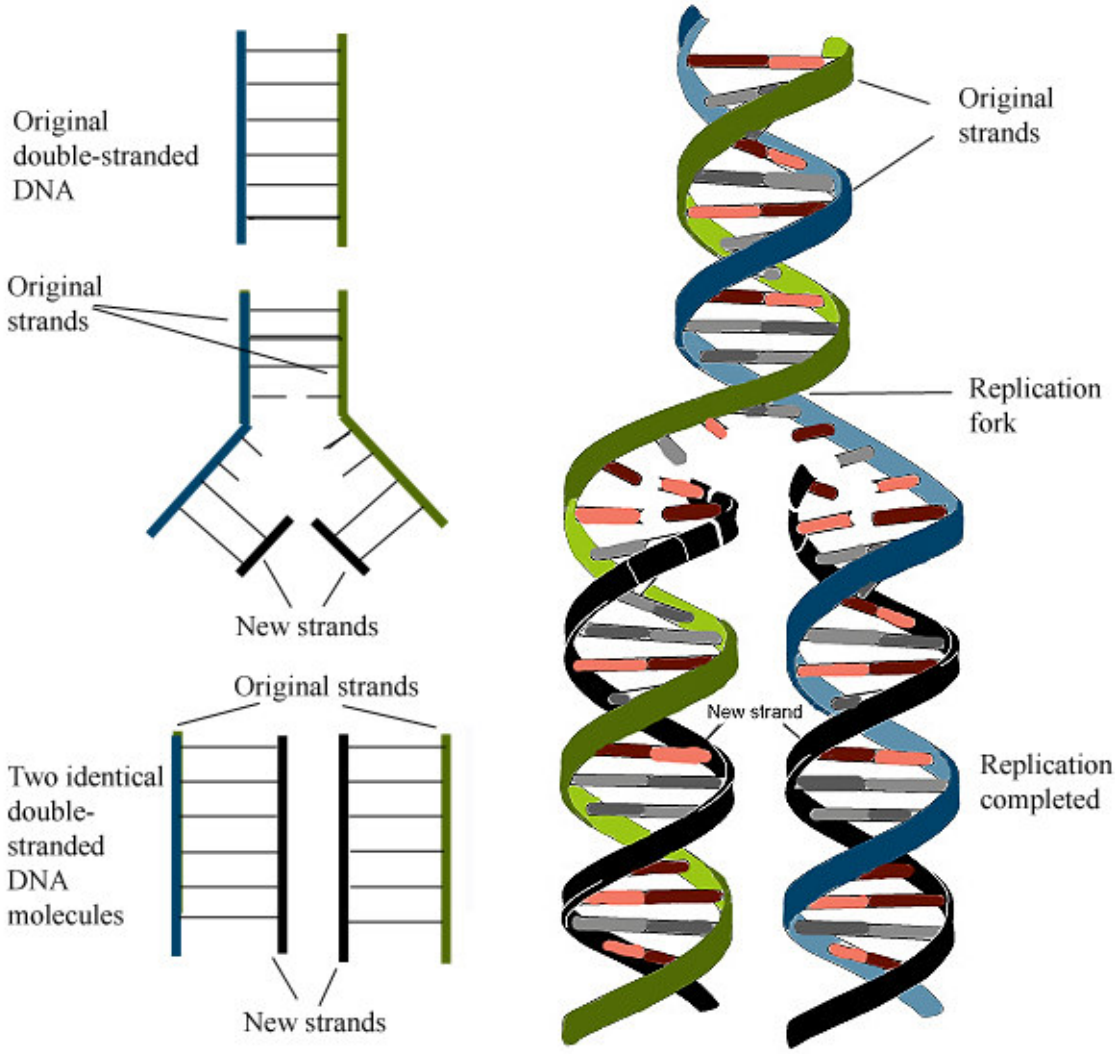


Figure 1.3: Process of DNA replication

Transcription

The genetic message encoded in a DNA molecule via the nucleotide base sequence is instrumental in the formation of a specific protein [7]. However, DNA is not directly used in the formation of a protein. Instead, mRNA (messenger RNA) is first synthesized as a working template from the DNA master template through the process of transcription [7]. Hence, the genetic information contained in DNA is transfered to the

mRNA molecule via the process of transcription. Only one of the complementary DNA strands may be used in transcription at a time, depending on the gene being transcribed [7]. Synthesis of mRNA proceeds just like in DNA replication. However, the RNA base Uracil (U) is used instead of the DNA base Thymine (T). Once synthesized, the newly formed mRNA molecule is released from the DNA template, which then resumes its right handed helical form. The newly formed mRNA is then transported to the cytoplasm of the cell, the site of translation [7].

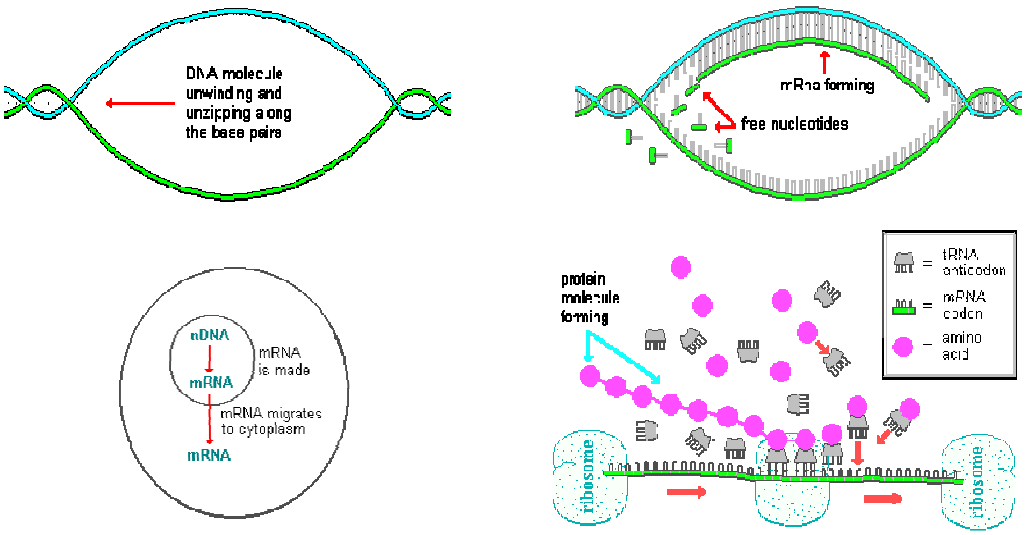


Figure 1.4: Transcription and translation; images reproduced with permission from Dennis O'Neil, copyright©2005 by Dennis O'Neil

Translation

The synthesis of a protein molecule from an mRNA template occurs through the process of translation [7]. A protein is synthesized by translating the mRNA nucleotide base sequence into the amino acid sequence of a primary polypeptide by means of a 4-letter, 64-word genetic code (Figure 1.5). Each triplet (or codon) in the mRNA sequence of bases gives instruction for one amino acid to be included into a growing polypeptide

chain [7]. Thus, the linear arrangement of codons along the mRNA template dictates the types and the linear arrangement of amino acids in the final protein product[7].

A ribosomal complex within the cell aids in setting the phase of the genetic message. Reading of the mRNA template occurs here, and protein synthesis proceeds along the 5' to 3' direction. [7]. As the genetic code is read at the ribosomal level, each codon is recognized by a particular transfer RNA (tRNA) molecule. This tRNA transports the amino acid specified by the codon to the site of protein synthesis.

1st Position (5` end)	2nd Position				3rd Position (3` end)
	U	C	A	G	
U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

Figure 1.5: The genetic code [7]

Initiation of protein synthesis occurs at the AUG codon or the start-word by activating the ribosomal complex to set the phase of translation [7]. Subsequently, the ribosome shifts one triplet down the mRNA in the 3' direction during translation and the

appropriate tRNA brings the amino acid encoded by the new codon into position. This process continues until one of the three stop-codons (UAA, UAG, or UGA) is encountered [7].

The initiator codon-ribosome complex partitions the mRNA base sequence into codons to determine the reading frame of the translation [7]. A phase-shift mutation (DNA deletions and insertions) in the gene modifies this reading frame. For example, Figure 1.6 shows three ways in which the genetic code could be read, depending on the position or phase of the first base pair. Addition or deletion of a base pair from the genetic sequence changes the sequence of the base pairs translated, and this can radically change the protein structure [7].

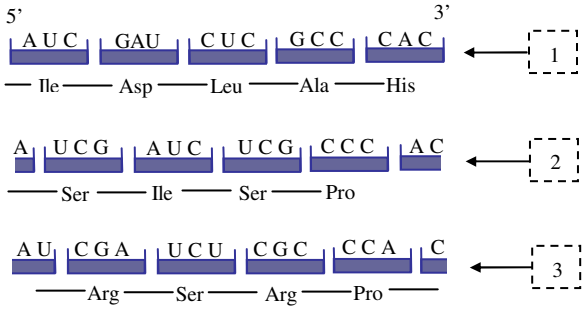


Figure 1.6: Phase shift in the reading frame of the genetic code

1.4 Cancer

1.4.1 Cancer vs normal cells

Reproduction through cell division is essential for body growth and tissue repair. Cells that are constantly sloughed off the surface, such as cells of the skin and intestinal lining, reproduce themselves almost continuously [2]. The initiation signals for cell division are not fully understood, but surface-volume relationships are deemed to be important [1,2,6]. The volume of a cell dictates the amount of nutrients needed for the

cell to survive. The need for nutrients grows in proportion to the size of the cell. However, the cell surface of the plasma membrane gradually becomes inadequate to transfer nutrients to the cell and flush the waste products out of the cell. When the cell reaches such a critical size, cell division is initiated to produce two daughter cells that are each smaller in size. Cell division is also influenced by other mechanisms including availability of space, and chemical signals such as growth factors and hormones released by neighboring or distant cells [6]. Normal cells employ the phenomenon of contact inhibition to stop proliferating when they begin touching. When cells break free from these normal controls of cell division, they begin to divide wildly thus turning into cancer cells [1,2,6].

It is estimated that four to seven mutational events must occur between an initial normal state and a final stage of malignancy of a cell [1]. For example, some epithelial cancers, such as skin cancer and colon cancer, follow a sequence that includes [2,6]:

i. Hyperplasia ("increased numbers of regularly arranged normal cells" [2])
ii. Dysplasia ("increased numbers of normal cells with some atypical cells and some abnormal arrangement of cells but with no major disturbance of tissue structure" [2])
iii. Carcinoma-in-situ ("a severe form of dysplasia, with numerous atypical cells, major disturbance of tissue structure but no invasion of surrounding tissue" [2])
iv. Invasive cancer ("spread of altered cells derived from one tissue into adjacent different tissues" [2]).

The risk of developing invasive cancer at the site of a dysplastic lesion is greater than developing cancer from normal tissue, and the risk of invasive cancer developing from a carcinoma-in-situ lesion is greater that developing it from a dysplastic lesion.

Table 1.2: Characteristics of normal vs cancer cells [6]

Normal Cells	*Cancer Cells*
• Reproduce themselves exactly	• Reproduce continuously
• Stop reproducing at the right time	• Don't obey signals from other neighboring cells
• Stick together in the right place	• Don't stick together
• Become specialized or 'mature'	• Don't become specialized, but stay immature
• Self destruct if they are damaged	• Don't self-destruct or die if they move to another part of the body

Alteration of cell behavior that transforms the cell from normal to cancerous is permanently maintained and transmitted to descendant generations of the cell through the chromosomes, the genetic component of the cell [2,6]. Normal cellular activity does not require all genes to be operational within the cell, however, a relatively intact set of chromosomes is vital. Each cell is furnished with a complete chromosome set during the process of reproduction, when each daughter cell receives a replica of the chromosomes of the parent cell. The delicate balance of the integrated genetic system may be disrupted is a chromosome or parts of a chromosome are lost from or added to the genome through some error during cell division. Such an error may have fatal consequences, not only in the cells affected but eventually in the whole organism [2,6].

Cancer cells evolve along pathways that define the fate of the tumor [1,2]. Once transformed into cancerous cells, growth becomes more rapid, and cell types of a less normal nature appear in the tissue [2,6]. The ability of the abnormal cells to invade surrounding tissues becomes more evident. The cell then undergoes a series of physiological alterations that could collectively encourage malignant growth [1,2,6]. These changes include self-sufficiency in growth signals; insensitivity to growth-inhibitory signals; evasion of programmed cell death; limitless replicative potential; sustained angiogenesis and tissue invasion and metastasis [1,2].

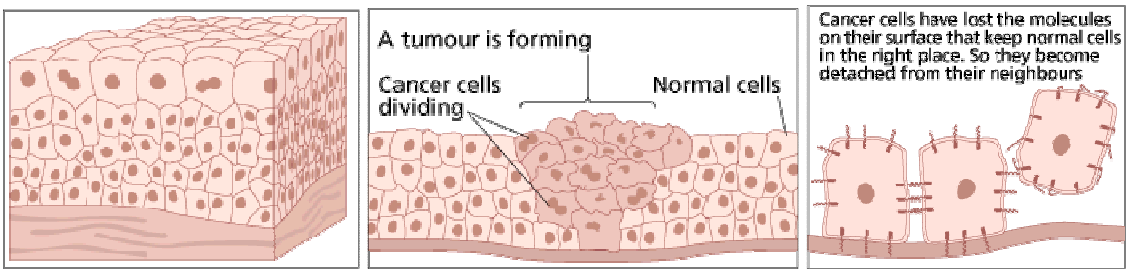


Figure: 1.6: Stages of development of cancer
images reproduced with permission from: www.cancerhelp.org.uk

The cancer cells are classified according to a grade based on how normal the cell appears to be. The more normal a cancer cell is, the lower its grade. The more abnormal or less well-developed a cancer cell is, the higher its grade. The sequence of events that cause the cells to change from dysplasia to carcinoma-in-situ to low-grade malignancy to high-grade-malignancy could possibly be programmed in the genetic material of the cell at the time of the first essential change from normal to cancerous state[1,2].

1.4.2 Causes of cancer

Mutations of genes alter the behavior of the cell that may ultimately cause the cell to become cancerous [1,2,6]. While innumerable factors, such as the DNA replicative state, the repair potential and the hormonal status of the host are likely to be promotional factors for cancer initiation, exposure of the cell to carcinogens is also a likely cause of cancer [2]. Almost all known carcinogens have been shown to be capable of irreversibly binding to genetic material in receptive animal tissues. This occurs either as a consequence of direct chemical reaction or metabolism of reactive metabolites [2]. The initial event in carcinogenesis is the introduction of certain inheritable defects causing the cells divide incorrectly. Here, genetic material is divided disproportionately between

daughter cells. This results in a mixed population of cells in both cancerous and pre-cancerous lesions that compete with each other for nutrients and survival. Selective survival of the most aggressive of these cells could lead to tumor progression [2].

At least three distinct kinds of genes are important in making a cell cancerous: oncogenes, tumor suppressor genes and DNA repair genes [2].

i. Oncogenes are genes that encourage the cell to multiply. These are normal cellular genes that when inappropriately activated cause the cell to multiply without stopping.

ii. Tumor suppressor genes are genes that stop the cell multiplication. These genes produce proteins that act to slow or regulate mitogenic activity. When these genes are impaired, the cell are not inhibited from multiplying uncontrollably and malignant progression occurs.

iii. DNA repair genes are genes that repair the other damaged genes. They genes aid in detection and facilitation of the correction of errors in the genetic code..

1.5 Microarray technology for gene expression analysis

DNA microarray chips are employed to analyze the genetic behavior of tissue [3]. An in-depth description of the behavior of cells may be obtained by analyzing the DNA of the cells. It is important to understand and locate the presence of mutations in genes that are important to the functioning of the cell, as well as to detect the genes that are active in the cell. Manual analysis of all the genes in any cell would take an extraordinarily long time. The time lag for such manual processing may be unacceptable while attempting to make decisions regarding treatment options for patients based on the possible genetic mutations in the tissue. Microarray technology alleviates this problem in

the following ways: microarray technology can follow the activity of many genes at the same time, compare the activity of genes in diseased and healthy cells, determine any mutations in a gene, and categorize diseases into subgroups while acquiring results very rapidly [3].

1.5.1 Techniques

Each cell in the body ideally owns the exact copy of the entire genome as every other cell [2,6,7]. However, only a small set of these genes are active in any one cell, and the functioning of these genes aid in understanding the functioning of the cell. Directly measuring the DNA of a cell will not aid in quantifying the level of expression of a gene. However, the number of copies of each mRNA in a cell indicates the level of activity of the gene that corresponds to that mRNA. Labeling these copies of mRNA and counting them will then directly indicate the level of activity of the corresponding genes. A sample of the tissue may be analyzed in isolation, to understand the behavior of the cell in its natural environments, or the sample tissue may be subjected to two or more different kinds of environment in order to analyze the genetic behavior of the tissue under different conditions. Further, one kind of tissue may be compared with another kind to analyze the difference in gene expressions and activity in the two tissue types. In general, gene expression analysis techniques involves three major steps: preparing the DNA chip, carrying out the reaction and collection and analysis of data [3,8].

Preparing the DNA chip

The first step to being able to analyze the gene activity is to list out the genes that need to be monitored. The sequences or parts of the sequences of these genes must be specified. A piece of each gene is synthesized as a short strand DNA, or oligonucleotide, a few base pairs long (for example, Affymetrix DNA chips [8] use 25 base pairs). Each short strand of DNA is fixed on a tiny spot on a slide. Billions of copies of each strand are affixed on the same spot of the slide. Several thousand such genes may be converted into short strands and fixed to the glass slide for analysis.

Carrying out the reaction

The next step is to convert the DNA of the target cell into mRNA under the environmental conditions being studied. If more than one environmental condition of the cell is used, the mRNA obtained under each condition is labeled with a fluorescent stain separately to make it easy to identify the level of activity under the different conditions. This is achieved by reverse transcribing the mRNA into complementary DNA or cDNA.

By introducing modified fluorescent bases into the DNA during hybridization, the cDNA can be conveniently tagged with different colors, such as red and green, for different experimental conditions. The one or more sets of cDNA are then combined and hybridized onto the DNA chip in a special chamber.

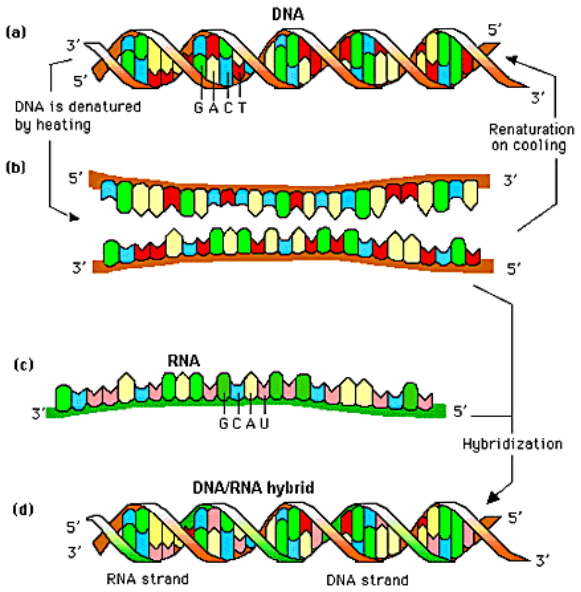


Figure 1.7: Hybridization of RNA with cDNA, image reproduced with permission: Wosik, E. cDNA-Detailed Information, Connexions Web site. http://cnx.rice.edu/content/m12385/1.2/, Sep 30, 2004

Collection and analysis of data

The final step is to measure the amount of each type of cDNA hybridized to any spot on the DNA chip. If multiple conditions are used, not just the total amount of hybridization, but the relative levels of hybridization of the two types of cDNA on any one spot of the chip are important. Color laser scanners, one for each color used in tagging the cDNA, are used to scan the DNA chip. Each color scan indicates the amount of that color cDNA hybridized to all spots on the chip. By combining the information on the color scans, one can measure the relative expressions of genes under the different experimental conditions. The results will indicate which genes are turned on, and to what level of activity under the experimental conditions. Alternately, the fluorescence of a single color will indicate the expression level of a gene under the target conditions.

Oligonucleotide arrays

Two of the several types of microarray technology available today for DNA analysis are cDNA, or complementary DNA chip and Short oligonucleotide arrays [8]. cDNA chips measure the relative abundance of a spotted DNA sequence in two DNA or RNA samples by assessing the differential hybridization of these two samples to the sequence on the array. Here, probes are defined as DNA sequences spotted on the array.

Short oligonucleotide arrays, such as Affymetrix chips [8] on the other hand, use "probesets" for measurement. Each gene is represented by 6-20 oligonucleotides of 25 base-pairs (or 25-mers). Each 25-mer is called a "probe". Two complimentary probes are created for a 25-mer that has to be analyzed: A perfect match probe is a 25-mer exact compliment of the target probe and mismatch probe is a 25-mer, same as the perfect match, but with a single homomeric base change for the middle (or 13th) base.

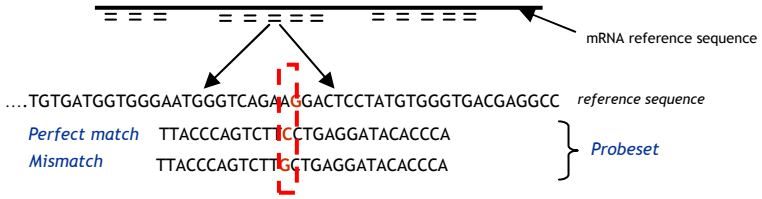


Figure 1.8: Perfect-match and mismatch probes form a probe-pair

A perfect-match and mismatch combination for a 25-mer sequence of a gene is called a "probe pair", and about 16-20 probe-pairs form a "probeset". The addition of the mismatch pair to the experiment helps in measurement of non-specific binding and the background noise [8].

1.6 Overview of bioinformatics methods for gene expression analysis

Advancements in the area of molecular genetics have enabled the mapping of the entire human genome. About 30,000 genes of the human genome have been mapped today [9], and specific information regarding the genes actively expressed in various tissue types has made it possible to identify the normal functioning of cells in the body. The introduction of microarray technology allowed the analysis of several thousands of genes in a single experiment [8]. This explosion of information makes it possible to thoroughly investigate the expression of genes in tissues [3]. The genetic activity in normal cells could be compared with the activity in tumor cells [13], and tumors of different types may be distinguished [14]. Several investigators have worked towards mining meaningful information from the thousands of genes acquired from the microarray experiments in order to distinguish between various diseased conditions of tissues [10,11,12]. In the area of cancer management, the two main areas of research have been class discovery and class prediction [13]. Class discovery involves identifying previously unrecognized tumors, and class prediction involves present or future assignment of the tumor to a previously discovered tumor type.

Golub et al [13] 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 [15,16] 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 a correlation coefficient above the optimized sensitivity threshold were classified as good prognosis, and the rest as poor prognosis.

Alon et al [17] 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 [14] 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.

Ramaswamy et al [18] analyzed a 17 gene signature to study the metastatic potential of cancer cells in solid tumors. Genes were selected based on a signal-to-noise

metric followed by a hierarchical clustering to determine the individual correlations for the selected genes. The results of the algorithm were tested using Kaplan-Meier survival analysis techniques.

Eschrich et al [19] showed that molecular staging of colorectal 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 was used with 43 genes to predict the molecular stage of a tumor sample.

1.7 Outline of the thesis

While the main goal of the study is to develop a classifier scheme using a random subspace ensemble to improve the accuracy of survival prediction for colon cancer patients, it is essential to have a fundamental understanding of the microarray gene expression data and methods generally used to analyze this data. Chapter 2 describes the microarray gene expression data used in the study.

Chapter 3 introduces the general method used to analyze gene expression data, including feature selection and classification. A brief description of some of the algorithms used at various stages of the analysis is given. The chapter concludes with a description of methods used to evaluate and measure the performance of the classifiers.

Chapter 4 introduces the concept of random subspaces and describes three methods of creating random subspace ensembles, highlighting the merits and demerits of each method.

Chapter 5 describes the experiments conducted with feature selection methods and the various baseline classifier experiments on the colon cancer gene expression dataset. A detailed description of the experiment with random subspace ensembles is presented next. The chapter is concluded with a verification of the results.

Finally, a discussion on the proposed method, its merits and potential improvements are presented in Chapter 6, followed by conclusions from the study.

In addition to the experiments with colon cancer data, the proposed method was tested on datasets with different clinical measures (leukemia and gender). A description of these experiments is presented in the Appendix Section A.

CHAPTER 2

GENE EXPRESSION DATA FOR ANALYSIS OF COLON CANCER

Colorectal cancers are the second most common cause of cancer-related deaths in developed countries and the most common GI (gastro-intestinal) cancer [20]. Colon cancer develops as polyps in the intestinal wall, and could progress slowly to a severe stage cancer if left unchecked. A common and well-accepted method of clinical staging of colon cancer is the Duke's classification (Table 2.1) of colon cancer [2,6]. However, Duke's staging system has been shown to be inadequate in determining prognosis for patients diagnosed with stages B or C of colon cancer [19]. Molecular staging, on the other hand has shown promise in predicting prognosis for patients based on the gene expression profile of the tumor [19].

Table 2.1: Dukes classification (modified by Turnbull) [2]

Stage	*Description*
A	Limited to bowel wall
B	Extension to pericolic fat; no nodes
C	Regional lymph node metastasis
D	Distant metastasis (liver, lung, bone)

The goal of this study is to analyze gene expression patterns to predict the 36 month survival rate for colon cancer patients. The samples used for the study were categorized based on the patient prognosis of cancer rather than the clinical staging. Samples were classified as good prognosis cases if the patient survived greater 36 months, and bad

prognosis if the survival was less than 36 months. Thus all patient used for the study had to have been followed for at least 36 months.

121 samples of colorectal adenocarcinoma were selected from the Moffitt Cancer Center Tumor bank and include 37 samples with bad prognosis and 84 samples with good prognosis. The samples included all four Dukes stages of colon cancer. The evidence of survival, as well as patient information such as family history of cancer, and treatment history was acquired from the cancer registry. Each tissue sample used for the microarray analysis was taken during surgical resection of the tumor from the primary site of tumor and verified as adenocarcinoma of the colon by a pathologist.

The gene expression microarray used to analyze these tumor cases was the Affymetrix Human Genome U133 Plus 2.0 Array [8]. Each microarray experiment measured the expression levels of 54675 probesets (refer to Section 1.5.1). These expression levels were normalized using the Robust Multichip Average (RMA) [21] method to yield features values in log-2 scale for analysis.

CHAPTER 3

METHODS FOR GENE EXPRESSION ANALYSIS

3.1 Introduction

DNA microarray analysis generates information about the level of expression of genes in a target cell or tissue type [3,8]. Normal and cancerous cells are expected to exhibit differential expressions of certain genes. For example, abnormally high levels of expression of oncogenes, that ordinarily regulate the multiplication of cells [1,2], could indicate a tendency of the cell to proliferate without control. The genes that are instrumental in the onset and progression of cancer are likely to be expressed differently than in normal tissue, with alterations in their expression levels as the cancer progresses. Identifying these differentially expressed genes, and analyzing their expression patterns as the cancer progresses will aid diagnosis and prognosis of cancer.

Microarray gene expression analysis involves studying the expression patterns of the genes across varying environmental conditions of the tissue, such as tumor cells treated with different types of drugs or radiation therapy, or across the different stages of development of the cancer [22]. The aim of these analyses is to identify a set of genes that are reliably expressed differently across the different stages of cancer. However, most microarray experiments yield a very large number of features for analysis [8] . In practical situations, it is reasonable to assume that only a subset of these features truly represent the distinction between the stages of cancer, as well as between cancerous and normal tissues. Hence,

methods to reduce the dimensionality of the feature set are used in the first stage of analysis, to obtain a minimum useful set for classification. This feature selection may be done in a supervised or unsupervised manner [23]. While supervised techniques use the underlying class information to select features, unsupervised techniques use empirical evidence in the data to decide whether or not a feature would aid classification.

In general, classifiers are used to analyze a set of samples and separate them into groups, such that the characteristics of each group reflect the characteristics or features of the individual samples of the group. These defining features are governed by the context of analysis. In practical situations, even when the best feature set is used, it is often difficult to identify features that unambiguously separate groups from one another, as well as predict the classes of new samples. Hence, classification methods aim at uncovering patterns that best describe the distinction between these groups. These patterns are learned from training samples and later used in predicting the class of new and unseen samples.

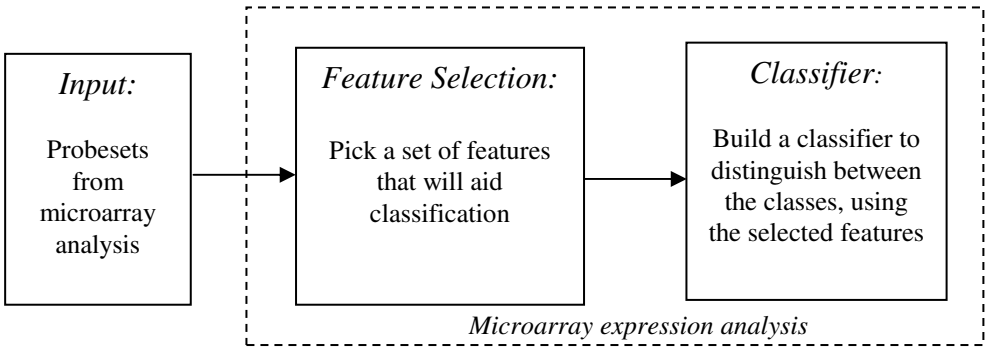


Figure 3.1: A typical setup for microarray gene expression analysis

Thus, a typical microarray gene analysis experiment would follow the steps shown in Figure 3.1. The first stage in analyzing gene expressions is to select a limited set of

features that would aid the classification stage in identifying the important patterns that distinguish between the classes. The features that may confuse the classification are dropped from consideration. A classifier is then built to learn patterns from these selected features in order to distinguish between the classes or conditions under consideration.

The knowledge gained by this classifier in learning the patterns from the training samples is evaluated to ensure that the patterns may be generalized to unseen samples. A measure of performance of the classifier will indicate the expected performance of the classifier in predicting the class of an unseen sample.

3.2 Supervised feature selection

Supervised feature selection methods use the underlying class information of all the samples to make a decision regarding the importance of a particular feature in distinguishing between the classes. Statistical techniques such as the student's t-test [24] and survival analysis [25,26], which attempt to capture the biological relevance of a feature, are used to retain a minimal set of features that would be best able to distinguish between the classes.

Student's t-test

The student's t-test is used to determine whether the means of two groups are statistically different from each other [24]. This analysis assumes normally distributed data, with mean, μ and variance, σ. The two groups for comparison are created by varying one or more features that characterize the samples. The aim of the t-test is then to

determine if the distribution of the samples changes due to the variation of the feature/s, and if so, whether the change can be detected easily.

The null hypothesis for a t-test is that the two groups are not different from each other, and hence have the same mean. A t-statistic is computed from the samples of the two groups and is treated as "evidence" for or against the null hypothesis. The computed statistic is compared to a standard measure to decide whether to accept or reject the null hypothesis. Strong evidence for being able to detect a difference between the two groups would suggest rejection of the null hypothesis.

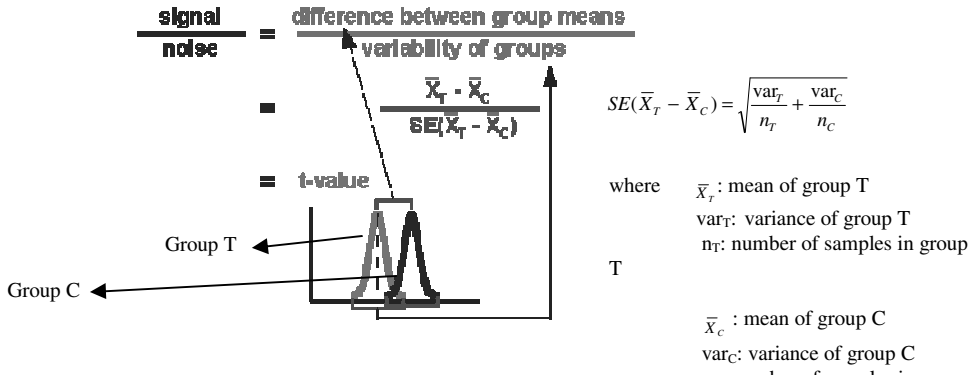
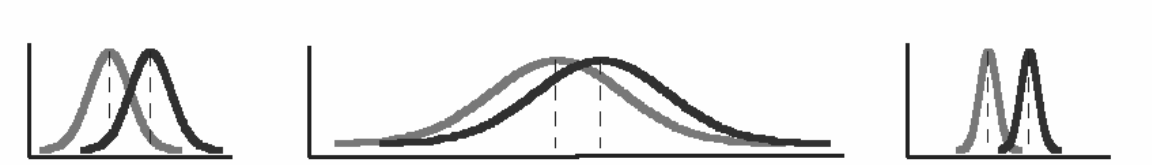


Figure 3.2: Formulation of the t-test; reproduced with permission from: http://www.socialresearchmethods.net/kb/stat_t.htm

Figure 3.2 shows the distributions of two groups with individual means. In a basic sense, the distance between the two means can be used as a measure of difference between the groups, and gives an indication of the distinction between the groups. However, as shown in Figure 3.3, the distinction between the groups may be influenced by the relative spread of the two groups.



(a) (b) (c)

Figure 3.3: Three cases with equal difference in means (a) medium variability (b) high variability (c) low variability; reproduced with permission from: http://www.socialresearchmethods.net/kb/stat_t.htm

Thus, a true measure of the difference between the two groups can be obtained by computing a score that measures the difference between their means relative to the spread or variability of their distributions.

The t-statistic is computed as the ratio of the difference between the two means and the standard error of the difference:

$$t = \frac{\overline{X}_T - \overline{X}_C}{\sqrt{\frac{\text{var}_T}{n_T} + \frac{\text{var}_C}{n_C}}}$$

where $\overline{X}_T$: mean of group T
var_T: variance of group T
n_T: number of samples in group T

$\overline{X}_C$: mean of group C
var_C: variance of group C

The t-statistic indicates the ease of distinguishing between two groups in presence of variability due to the inherent variability of the data or noise in measurement. A standard t-statistic is computed based on the degrees of freedom available and significance level desired for the test. A significance level (α), commonly set at 0.05 indicates that 5 times out of 100, a significant difference between the means could be found merely by chance, even if there was none. The computed t-statistic is compared with the standard t-statistic to obtain a p-value for the test. The p-value indicates the probability of making an error in distinguishing between the two groups.

A p-value of less than the α-level indicates that the difference between the two groups is statistically significant, and hence, the null hypothesis is rejected.

The ability of each feature in the microarray dataset to predict the classes for new samples can be examined using the described t-test. The p-value for each feature is univariately computed at a significance level $\alpha=0.05$. All features with p-values less than 0.05 are considered by convention to have statistically significant power to discriminate between the two groups of patients.

Survival analysis

When dealing with problems in cancer research, a common endpoint is determining whether a patient will survive for a certain period of time. Here, "death" is considered to be an event and survival analysis [25,26] attempts to model the time-to-event, to predict the fraction of the population that could survive past a certain time. Of those that survive, the analysis tries to predict the rate at which these patients would fail or die. Survival models may be viewed as ordinary regression models where the response variable is time.

However, this analysis also needs to account for missing data or information on patients who could not be followed for the entire duration of the study for various reasons. This introduces the concept of censoring in survival analysis [25]. It may be known that a patient had colon cancer, but died at an unknown time before data collection began. This is known as left censoring. Right censoring occurs when a patient may have a date of death at a future unknown date. When a sample is both left and right censored at the same time, the sample is said to interval censored. Another possibility of an

incomplete event is delayed-entry, when the patient does not enter the study until a certain event occurs.

Kaplan-Meier (KM) curves [25,26] are used to plot the probability of survival of the population against intervals of time. For each interval, the survival probability is computed as the ratio of the number of patients alive at that time point with the number of patients at risk. All patients who are alive and reached the time point are considered to be "at risk". Patients who either die before the time point or are "lost" for the study are not counted as "risk" patients. "Lost" patients are censored. Further, patients who have not yet reached the time point are not considered as "risk" patients.

The probability of survival to any point is estimated from the cumulative probability of surviving each of the preceding time intervals. This formula, also known as the Kaplan-Meier Product-Limit formula [25] is given by:

$$S(t) = \prod_{j=1}^{t} \left[\frac{(n-j)}{(n-j+1)} \right]^{\delta_{(j)}}$$

where n: total number of cases,

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

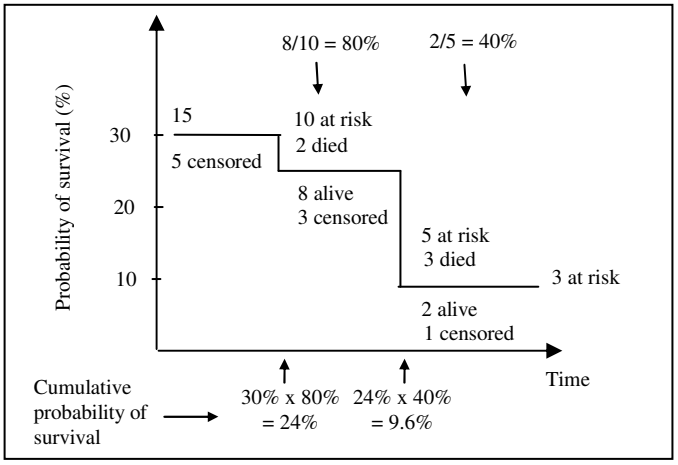


Figure 3.4: A sample Kaplan-Meier curve

The computation thus far has been shown on a single group of samples. In DNA analysis for prediction of survival for colon cancer patients, two groups of samples are used: a group that survived less than 36 months, and a group that survived greater than 36 months. Survival curves for both of these groups may be shown on a single graph. The next task is then to determine whether or not these two KM curves are statistically equivalent.

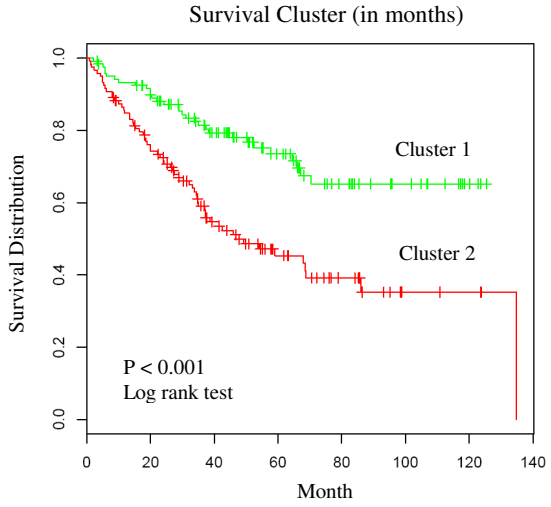


Figure 3.5: Comparison of two sample K-M curves using log-rank test

The log-rank test [25] can be used for this purpose. This test is basically a large sample chi-square test that uses as its test criterion a statistic that provides an overall comparison of the two KM curves.

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

where O_i observed score for the group i

E_i expected score for the group i

$$Var(O_i - E_i) = \sum_j \frac{n_{1j} n_{2j} (m_{1j} + m_{2j})(n_{1j} + n_{2j} - m_{1j} - m_{2j})}{(n_{1j} + n_{2j})^2 (n_{1j} + n_{2j} - 1)}$$

where n_{ij} is the number of samples in group m_{ij} ; i, =1,2

Under the null hypothesis that the two KM curves are statistically equivalent, the log-rank test statistic is approximately chi-square with one degree of freedom [25]. Thus, a p-value may be obtained at an α (say 0,05) confidence level from the chi-square distribution tables. At p-values less than the confidence level α, the null hypothesis is rejected, and hence the two curves are considered to be statistically different.

The features of the microarray gene expression data with significant p-values from the log-rank test are retained as features useful in discriminating between the two classes of patients, divided based on survival times.

3.3 Unsupervised feature selection

Unsupervised feature selection does not use any a priori information regarding the class information or distribution of samples amongst the classes in order to select features. Many unsupervised feature selection methods analyze some statistical measurement made on the samples in order to identify a small set of features that help in separating the samples into distinct groups. A feature is selected based on the strength that demonstrates its ability to separate the samples into the required number of classes. These methods may be categorized as quantitative methods and qualitative methods. The quantitative methods use statistical quantities such as the expression level or a measure of variability to reduce the feature set, while the qualitative method attempts to identify features that are relevant to the problem at hand.

Expression level threshold

While conducting a microarray gene expression analysis experiment, a minimum level of hybridization of cDNA to the DNA chip is required to reliably translate the activity of the gene to an expression level. This imposes a lower limit on the level of gene expression that may be considered useful for detection and analysis [8]. The expression level threshold method of feature selection can be used to reduce the number of features for classification based on a minimum expression level for each feature. However, it is difficult to find a crisp cut-off value for this threshold, and it is possible to find at least a few samples that have expression levels that are marginally higher than the threshold. Thus, it may not be possible to eliminate a feature based purely on the expression level below a threshold. A second limitation must then be imposed to successfully eliminate features. This limitation will only allow a feature to be eliminated if at least a pre-determined percentage of the samples display expression levels below the threshold value. Thus, feature selection by expression level threshold is parameterized by two threshold values: t, the expression level threshold and p, the threshold for minimum percentage of samples below t.

This method of feature selection will aid in identifying the features with meaningful gene expression levels and would potentially aid in classifying the samples into the relevant classes.

Measures of variability

Features that tend to have similar values for samples belonging to different classes exhibit low variance across samples of both classes. Since classifiers attempt to learn

patterns that can distinguish samples of distinct classes, features with low variance rarely aid in classification. Researchers have often used the 2-fold expression change as a measure of variability [30]. However, this approach has been questioned. In general, measures of variability are used to select a set of features that display sufficient variability between classes. Parametric or non-parametric measures of variability may be used, depending on the distribution of the expression levels of the features.

Statistical variance: Measure of variability

Statistical variance [24] is a parametric measure of variability. This measure assumes a normal distribution of the feature values, with a mean μ and variance σ [24]. Mathematically, the statistical variance is defined as: $s^2 = \sum_{i=1}^{N} (Y_i - \overline{Y})^2 / (N-1)$, where $\overline{Y}$ is the mean of the data, and N is the number of samples. It can be observed that the variance is roughly the arithmetic average of the squared distance from the mean. Squaring the distance from the mean has the effect of giving greater weight to values that are further from the mean. Thus, although the variance is intended to be an overall measure of spread, it can be greatly affected by the tail behavior, or the values at the extreme ends of the distribution.

Feature selection may be achieved using the statistical variance measure of variability by discarding all features that exhibit a variance lower than a pre-determined threshold across all the samples. The retained features would be better suited to aid classification than the discarded features.

Median of absolute deviation from the median

A non-parametric measure of variability is the median of absolute deviation from the median (MAD) [24]. Mathematically, MAD is defined as:

$$MAD = median(|Y_i - \hat{Y}|)$$

where $\hat{Y}$ is the median of the data and $|Y|$ is the absolute value of Y.

Since the median is at the middle of a distribution, the value is not as sensitive to the values at the extreme ends or tails of the distribution as are the mean and variance. Further, since the computation of MAD does not use the sample size, the MAD value is expected to be a stable measure of variability, especially in the case of small sample sizes.

Feature selection using MAD involves discarding all features that exhibit MAD values below a pre-determined threshold. As described earlier, such features with low variability across classes are considered to be ineffective in predicting the underlying classes, and hence can be safely eliminated from consideration.

Selection of biologically relevant genes

For several decades, researchers across the world have been studying the genetic behavior of cancer [1,2,3,10,11,12]. Attempts have been made to pinpoint gene mutations that may be strongly indicative of the cancerous nature of cells, as well as genes that are predictive of cancer progression [10,11,12]. It is reasonable to assume that these genes, when over or under expressed in a cancerous tissue, would exhibit expression characteristics that are significantly different from genes that are not associated with the presence or progression of cancer.

A careful analysis of the characteristics of these "cancer-related" genes could yield an insight into the behavior of genes that control the progression of cancer. Hence, these genes could be identified [27,28,29] and separated from the rest of the genes in the microarray dataset based on the expression patterns.

3.4 Classifiers for gene expression analysis

Some of the classification methods that have been used for gene expression analysis include Neural Networks [31,32,33], Support Vector Machines [31,34], and Decision Trees [31,35]. The following sections review the basic methodologies of each of these classifiers that will be used to baseline performance measurement for the analysis of the colon cancer gene expression data to predict survival.

3.4.1 Feed-forward back propagation neural network

A neural network is a massively parallel distributed processor [31,32,33] made up of simple processing units. It resembles the brain in two respects:

i. Knowledge is acquired by the network from the environment through a learning process.
ii. Inter-neuron connection strengths, known as synaptic weights, are used to store the acquired knowledge.

A feed-forward-back-propagation neural network [31]typically consists of at least three layers. The first layer is the input layer followed by one or more layers of hidden units or computational nodes, ending in a layer of output nodes. The learning algorithm employs a forward pass and a backward pass of signals through the different layers of the

network. The forward pass involves the application of an input vector to the sensory nodes of the network. The effect of this input vector is propagated through the layers of the network, producing a response at the output layer of the network. While the weights of the nodes are fixed during the forward pass, they are adjusted according to an error-correction rule during the backward pass. This error signal is computed by subtracting the actual response of the network from a desired or target response. The error signal is then propagated backward through the network against the direction of synaptic connections, adjusting the weights to make the actual response of the network closer to the desired response.

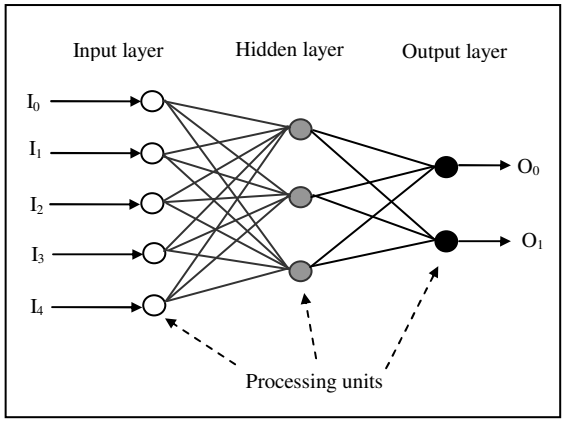


Figure 3.6: Architecture of feed-forward-back-propagation neural network

3.4.2 Support vector machines

Support vector machines are algorithms which use linear models to represent non-linear boundaries between classes [31,34]. Input feature vectors are transformed into a higher dimensional space using a non-linear mapping. Hyperplanes are defined in this high dimensional space so that data from any two class categories can always be

separated. The hyperplane that achieves the highest separation of the classes is known as the maximum margin hyperplane and generalizes the solution of the classifier.

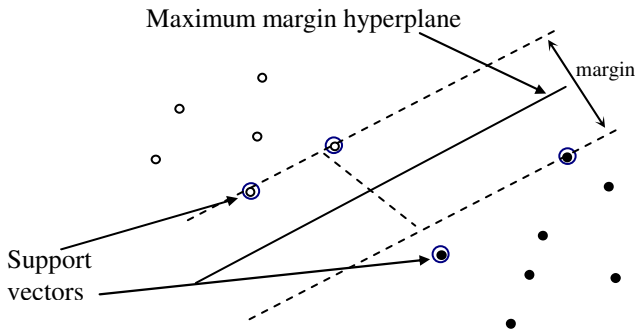


Figure 3.7: A maximum margin hyperplane in a support vector machine [31]

The maximum margin hyperplane is completely defined by specifying the vectors closest to it. These vectors are called support vectors. Since these vectors have the minimum distance to the plane, they uniquely define the hyperplane for the learning problem. Thus, the maximum margin hyperplane can be completely reconstructed given these support vectors, and all other training instances can be deleted without changing the position and orientation of the hyperplane.

Consider a simple two-class problem with two attributes or features, a_1 and a_2. A hyperplane separating the two classes may be written as:

$$x = w_0 + w_1 a_1 + w_2 a_2$$

where the three weights, w_i are to be learned.

This may be expressed in terms of the support vectors. Suppose we define a class variable y with a value 1 if it belongs to class 1 else -1.

Then, the maximum margin hyperplane is defined as:

$$x = b + \sum \alpha_i y_i a(i) \bullet a$$

where: y_i : the class value of the training instance a(i)

b, α_i : the numeric parameters that have to be determined by the learning algorithm: these parameters determine the hyperplane

a : vectors representing the test instance

$a(i)$: support vectors

$a(I) \bullet a$: dot product of the test instance with one of the support vectors

The support vectors are the training samples that define the optimal separating hyperplane and are the most difficult patterns and also the most informative patterns for the classification task. A constrained quadratic optimization technique is used to learn the parameters b and α_i [34].

3.4.3 C4.5 decision trees

Decision trees are learning algorithms that employ the "divide and conquer" strategy [31]. Decision trees are constructed by creating nodes at various levels by testing certain attributes. The first step is to select an attribute to be placed at the root node. At every node, a comparison of an attribute value with a constant is made. When using discrete attribute data, this makes one branch for each possible value of this attribute. This process splits up the samples into subsets for each value of the attribute.

The process is then repeated recursively for each branch, using only those samples that reach the branch. When a node attribute cannot split the samples into any more

subsets, a leaf node is created. Leaf nodes give a classification that applies to all samples that reach the leaf.

An unknown sample is classified by routing it down the tree according to the value of the attributes tested in successive nodes. When a leaf is reached the instance is classified according to the class assigned to the leaf.

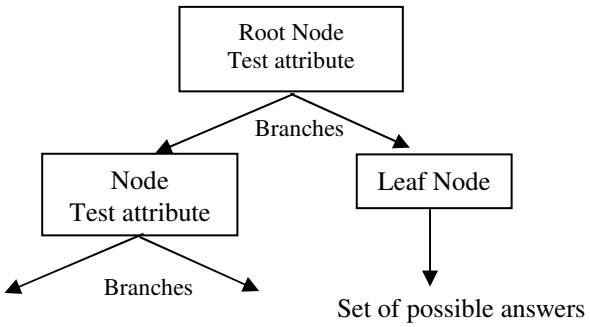


Figure 3.8: Structure of a decision tree

The structure of a decision tree is governed by the rules used to select the attribute to split on, at each node or branch. Given a set of attributes to choose from, the best choice for splitting the data is the attribute that produces subsets of samples that are most distinct from each other. This choice is made by measuring the purity of the daughter nodes at each split [31]. The best decision is made when the purest daughter nodes are created.

The C4.5 algorithm:

C4.5 [31] is a variant of the basic decision learning approach that uses the concept of information gain as a measure of purity at each node. The information gain can be

described as the effective decrease in entropy resulting from making a choice as to which attribute to use and at what level.

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

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

For each attribute that is tested as a potential splitting attribute, the entropy of the subsets created by the split is measured and compared to the entropy of the system prior to the split. The attribute that yields the maximum information gain by splitting the dataset is chosen as the best split or test attribute. By considering the best attributes for discriminating among cases at a particular node, the tree can be built up of decisions that allow navigation from the root of the tree to a leaf node by continually using attributes to determine the path to take [31]. The decision tree can be simplified using pruning techniques to reduce the size of the tree according to a user-defined level. Pruning will yield decision trees are more generalized [31].

3.5 Evaluation of classifiers

The task of machine learning is to “learn” or acquire knowledge about input data. This can be achieved by looking for and describing patterns in the input data. This acquired knowledge can be then used to predict patterns in unseen samples. The quality of knowledge gained in this process is determined by the samples used to train the system. The samples have to be representative of all characteristics that may be encountered to ensure that predictions on unknown samples are accurate. It should also be ensured that the machine learning system infers the correct patterns in the data. Methods to evaluate and predict performance on seen and unseen data help in ensuring this.

While measuring the performance of a learning algorithm, a measure of the success rate, or alternately, the error rate is used [31]. This is measured by comparing the results of classification on each of the training samples to the actual class to which the sample belongs. Thus, the success rate will indicate how well the algorithm has learnt the characteristics of the training samples. However, this gives no indication as to how the algorithm will behave when asked to predict the class of a new and unknown sample. The error in prediction may also be computed by testing the classification of test samples. If these test samples are taken from the same pool of data that was used for training, the measured success rate will be highly optimistic [31], and will not realistically indicate future performance. It is therefore necessary to set aside a set of samples that will not be used for training, but used for testing purposes only.

Generally in DNA analysis problems, the number of samples available for inferring gene activity and mutation is very small in comparison to the number of features available [3,8]. Separating a set of samples for testing will further reduce the number of samples available for training the learning algorithm. While it is beneficial to have a good size test set to rigorously test the prediction accuracy of the classifier, it is equally important to ensure that the samples used for training are representative of the population. A set of samples is generally held out as a completely independent test set while the rest are used for training. A smaller set from these training samples may be held out as a test set while training the classifier. Training and test performances are measured on this set to tune the learning algorithm. The independent test set is then used to validate the performance of the algorithm. Several methods have been used to address this issue [31].

A simple validation method is a hold-out procedure [31] that involves dividing the dataset into a fixed number of partitions. All but one partition is 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. This method is known as Cross-Validation [31], and a variable number of partitions may be used depending on the number of samples available. Ideally, the samples in each partition should represent 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.

Leave-One-Out-Cross-Validation (LOOCV) [31] leaves out a single sample for testing, while training on the rest. This method is useful when a very small number of samples are available, since it increases the number of train-test procedures that can be performed. However, this method does not ensure that the classifier learns all the classes well. Since only one sample is used to test the prediction accuracy in each fold, the classifier may predict the majority class for each sample, and still achieve high prediction accuracy. Further, this method of cross-validation may be computationally expensive.

A reasonable method of cross-validation is stratified n-fold cross-validation [31]. The sample set is divided into n partitions such that each partition is stratified in proportion to the number of samples in each class. This ensures that each the classifier is trained proportionally well to learn all the classes. Further, each test set will require the classifier to predict all classes, yielding a more realistic measure of the classifier performance. Although ‘n’ can take any value, n=10 has been experimentally shown in literature to achieve the a reasonable estimate of error [31].

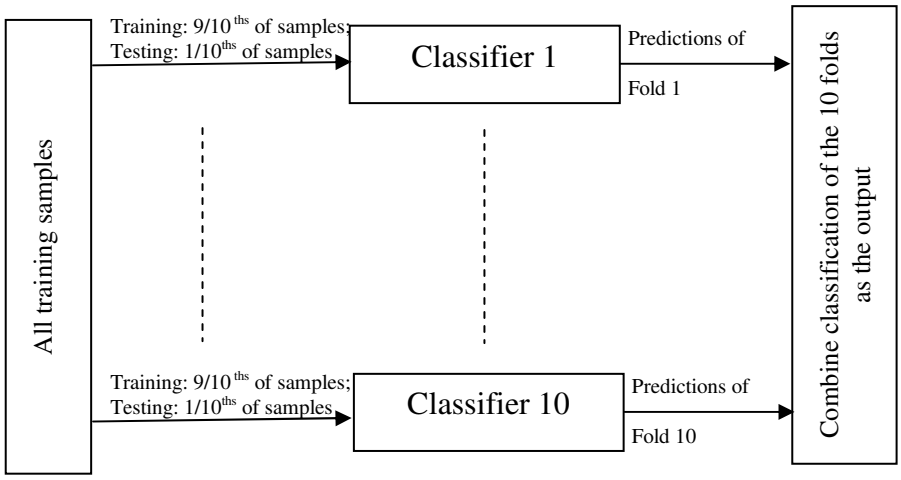


Figure 3.9: 10-fold cross-validation scheme

3.6 Accuracy of classification

In the context of predicting the survival time for patients with colon cancer, the performance of a classifier can be evaluated using a confusion matrix, as shown in the Table 3.1.

Table 3.1: Confusion matrix

Classified As / True condition	Short term survival (positive)	Long term survival (negative)
Short term survival (positive)	True Positive (a)	False Negative (b)
Long term survival (negative)	False Positive (c)	True Negative (d)

The most common measure of performance is the accuracy of classification, defined as:

$$Accuracy: \frac{a+b}{a+b+c+d}$$

Accuracy is a good measure to use if samples are distributed equally amongst both the classes. However, in cases where an unequal distribution of the samples may be expected, a weighted accuracy computation will yield a better estimate of how well the classifier performed in both the classes.

$$\textit{Weighted Accuracy}: \left(\frac{a}{a+b} + \frac{d}{c+d} \right) / 2$$

While dealing with clinical information however, measures of sensitivity and specificity [36] are used to gauge the performance in each class separately.

$$\textit{Specificity}: \frac{a}{a+b} \qquad\qquad \textit{Sensitivity}: \frac{d}{c+d}$$

Here, it can be observed that sensitivity is merely the probability that the patient will survive less than 36 months, given that the classifier predicted short term survival. Specificity is the probability that the classifier will predict long term survival given that the patient survived greater than 36 months.

Sensitivity is also the true positive rate and specificity is the true negative rate. Weighted accuracy reports the average of these rates, and hence may be used as a convenient measure to evaluate the performance of the classifier.

CHAPTER 4

RANDOM SUBSPACE ENSEMBLES FOR GENE EXPRESSION ANALYSIS

4.1 Introduction

Traditionally, classifiers have been used to uncover patterns from input features that can explain the observed characteristics of the samples, as well as make predictions on unseen samples [13-19]. The training stage uses a set of samples drawn from a larger population. If the total number of observed features that describe the population is very large, feature selection methods can be used in an attempt to pick a small set of features that adequately describe the patterns of differences between the classes in the population. Since the classifier learns these patterns from a limited set of features describing limited samples, there is a risk associated with over-training the classifier, or over-fitting the patterns to the samples at hand [31]. If the samples chosen for training do not adequately represent the population, the patterns learned would be specific in identifying these samples, and hence may not be general enough to identify or predict classes of unseen samples. The patterns learned could also be highly dependent on the features used to train the classifier. If a different feature set was used for training, a different set of patterns could be learned. With different configurations of learning parameters, different classifiers would be created. Some of these classifiers could be successful in accurately predicting classes of unknown samples, while others could have varying degrees of weaknesses depending on the feature set used to train the classifier. Further, use of

different sets of features may help in identifying different types of patterns, all of which may be important in completely describing the population. Random subspace ensembles [35] may be used to take advantage of this variation in performance due to different selection parameters in order to create a classification scheme that performs better than any single classifier [35].

4.2 Random subspace ensembles

The goal of creating ensembles of classifiers is to combine a collection of weak classifiers into a single strong classifier [35]. One way to create ensembles of classifiers is to divide the entire space of features into subspaces. Each subspace is formed by randomly picking features from the entire space, allowing for features to be repeated across subspaces. If enough such random subspaces are formed the subspaces may optimally represent all the important features in the subspaces.

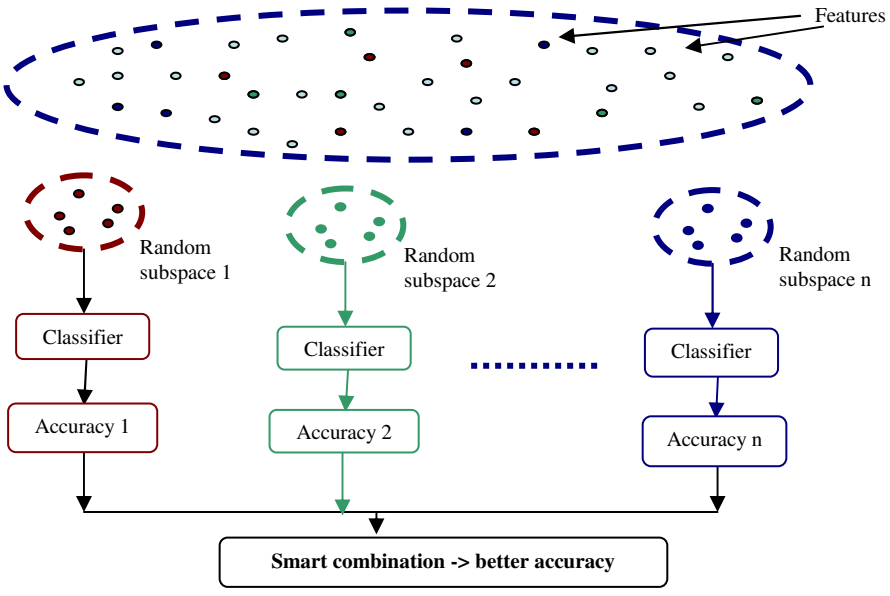


Figure 4.1: Creation of random subspace ensembles

One classifier is trained on each random subspace of features, using all the training samples. Thus, each classifier is built on one random projection of the feature space. A large number of such classifiers are created. If each classifier were tuned to learn a few characteristics of the population, then a judicious ensemble of these classifiers would be better at identifying samples from the entire population than any one classifier.

Depending on the characteristics learned by each classifier, different kinds of ensembles could be created. Voting techniques used on ensembles typically assume that all the random subspaces created are useful in some way in describing the classes. Alternately, if some of the random subspaces are found to be ineffective in describing the classes, while some others are very effective, then an ensemble could be created by using only the effective subspaces, while discarding all other subspaces.

4.3 Voting techniques to create random subspace ensembles

A general approach to the combination of random subspaces in an ensemble is the use of the majority voting technique [35]. Here, all the classifiers created are retained for use. Since each classifier is built from a random subset of the feature space, a single classifier may learn only a small section of the characteristics of the population. If the entire feature space is assumed to be important in describing the population, then each classifier created plays a role in describing the population. When a new and unknown sample from the population has to be analyzed, each classifier is considered to be equally capable in classifying the sample.

The majority voting technique uses each classifier to individually predict the class of a new sample. Then, a simple majority vote amongst the predictions is used to decide the final classification of the sample.

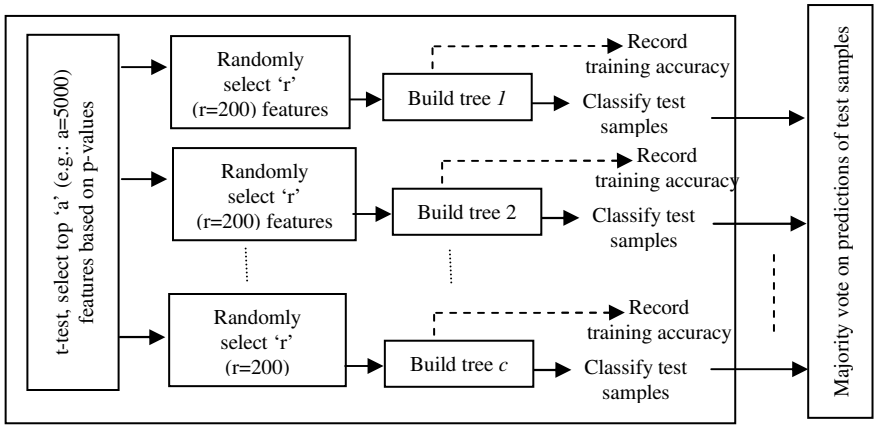


Figure 4.2: Random subspace ensemble classifier using the majority voting technique; here the number '*c*' of trees built is varied from 1 to 2000

An alternate voting technique is to use weighted majority voting instead of a simple majority. In this technique, the classifiers are not considered to learn equally well. While all classifiers are assumed to play a role in describing and distinguishing the patterns of the classes, some classifiers are deemed to be better at classifying samples than others. These "better" classifiers are given a higher weight in the voting, while the "poorer" classifiers are given a lesser weight. As in the simple majority voting case, all the random subspace classifiers are used to individually predict the class of a new and unseen sample. The quality of the classifier in predicting the class is typically defined by the field of application. The individual predictions are weighted by the quality of the classifier, before computing the majority class prediction. The majority class from the weighted vote is used as the prediction for that sample.

4.4 Selection of good subspaces

The voting techniques to create ensembles of classifiers, described in the previous section, work well when most of the features are useful in describing the characteristics of the population. In typical gene expression analysis problems, the number of features used for classification is very large. Although feature selection methods help in reducing the number of features for classification, these methods cannot ensure that every feature considered for classification is indeed important for prediction. It is reasonable to assume that in gene expression analysis problems, a subset of random subspaces may be created that are completely ineffective for classification. Including these ineffective subspaces in the ensemble may bias the classification in an undesirable manner. Hence, one approach to creating a good ensemble of classifiers is to discard these subspaces from consideration altogether. Alternatively, a small set of effective classifiers may be retained for creating the ensemble classifier.

Consider a set of random subspaces of '*r*' features selected from a set of '*a*' features. If the number '*c*' of random subspaces created is large enough to cover the feature space sufficiently, allowing for features to be repeated across subspaces, then at least a few subspaces are likely to include a majority of the "good" features. An effective classifier can be created by picking only these random subspaces with the "good" features.

A simple method to identify a good random subspace is to estimate the accuracy of the subspace in predicting the classes of a set of samples. Each random subspace is trained on a set of samples, such that the classifier learns patterns from this set in order to make class predictions on unseen samples. Hence, the quality of a random subspace can

be assessed by determining the prediction accuracy of the subspace classifier on test samples. In addition to this measure of quality, a random subspace may be assessed on the accuracy of learning the training samples. Although optimistically biased [31], training accuracy of a classifier reflects the ability of the classifier to learn the patterns of classes from the given training data. Classifiers built on subspaces that have a large number of good features should be able to learn predictive patterns better than classifiers built on poorer features. Hence, selecting a random subspace that has a combination of good classifier testing and training accuracies should ensure an overall more accurate classifier ensemble.

In order to estimate the accuracy of the selected feature subspace, the gene expression dataset is split into three separate subsets of samples. These are the independent test set (10% of the total samples), the training set (81% of the samples), and the validation set (9% of the samples). The random subspace classifiers are built on the training set (81%) and each classifier is tested on the samples in the validation set (9%). The performance of a classifier on the validation set, along with its training accuracy, is used to determine the predictive quality of the classifier. A classifier is chosen as the best random subspace classifier based on the condition of the highest validation set accuracy and in case of ties, a secondary condition of the best training accuracy is used. The features used by this classifier are selected as good features for the task.

To ensure that the selection process is relatively independent of the samples used for training, the gene expression dataset is split into the training and validation subsets in many different ways, so as to create different combinations of samples for training and validation. Consider 10 different ways of creating these subsets. The procedure of

selecting features, described above, is repeated for each of these 10 sets of data. Thus, for each of the 10 sets of training and validation samples a set of features is selected that can best describe 81% of the samples, and is tested by predicting the classes for 9% of the samples. The union of these 10 sets of selected features describe 90% of the samples (81% training samples+ 9% validation samples). The ability of this union of features to predict the class of unknown samples is tested with the help of the samples held out as the independent test set (10% of the samples). A classifier is built using these features by training on the 90% of the samples (81% training and 9% validation). This classifier is used to predict the class of each independent test sample. The weighted accuracy (see Section 3.6) of these predictions would indicate the expected performance of the classifier in predicting the class of new samples.

To further ensure that the prediction accuracy is not particularly tuned to the combination of training, validation and independent test samples, the gene expression dataset is split into these three subsets in several different ways. Consider 10 different ways of creating these subsets. Each of the 10 ways yields a weighted accuracy of prediction on the 10% of the independent samples for that set. The individual sample predictions on all 10 independent test sets are used to create the confusion matrix for the classifier scheme (see Section 3.6), and the weighted accuracy of the classifier scheme is computed from this matrix as a measure of performance.

In an experimental setup (Figure 4.3), this described procedure may be achieved by using a 10-fold cross-validation scheme. To illustrate the use of the described procedure on a gene expression dataset, consider a hypothetical gene expression dataset with 1000 samples and 50,000 features (probesets). Each fold of the 10-fold cross-

validation creates an independent test set with a distinct 100 samples (10%), and a training set with 900 samples (90%). An additional 10-fold cross-validation performed on each 900-sample training set, provides 810 samples (81% of the overall samples) for training, and 90 samples (9% of the overall samples) for validation. Therefore, for each of the 10 sets of the data (the 10 folds), 100 samples are held out as an independent test set, and 10 internal sets, each with different combinations of 810 training and 90 validation samples, are created from the 900 training samples.

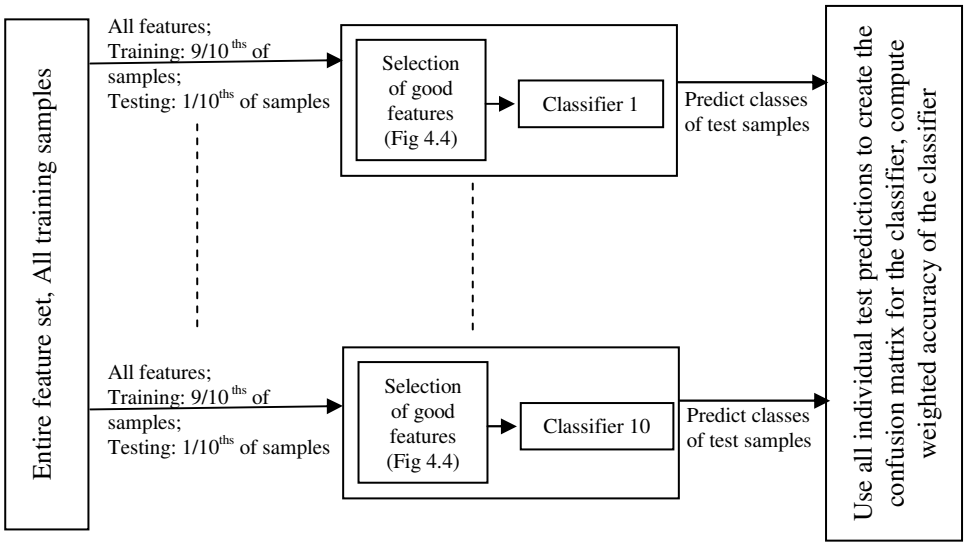


Figure 4.3: Classification scheme for selecting good subspaces

For each training set of 810 samples, a preliminary feature selection using a t-test is performed, and the best 5000 features, ranked according to significant t-test p-values (see Section 3.2), are retained for use. Random subspace classifiers are created on these 810 training samples using the selected 5000 features. Consider creating 100 random subspaces that have each been created by randomly picking 200 features from the 5000

input features. For each random subspace of 200 features, a single decision tree is built by training on all of the 810 samples. The decision tree selects, from this random subspace of 200 features, the best features to distinguish between the classes. The prediction accuracy of each decision tree (random subspace classifier) is tested on the corresponding 90 validation samples.

Each of the random subspace classifiers is trained on the same set of 810 samples and tested on the same set of 90 samples, and the classifier with the highest validation set accuracy and training accuracy is selected. The features used by the selected classifier are identified.

This procedure of selecting good features is repeated on each of the 10 sets of 810 training and 90 validation samples for a given fold of the data, yielding 10 possibly distinct sets of good features. The union of these features is then used to train a single classifier on the 900 samples (90%; the 81% training and 9% validation samples combined), and tested on the 100 held-out independent test samples (10%).

This process is repeated for all 10 folds. The individual predictions on all the independent test samples are used to create the confusion matrix (see Section 3.6) for the system. The weighted accuracy computed from this matrix estimates the expected performance of the classifier scheme in predicting the classes of new samples.

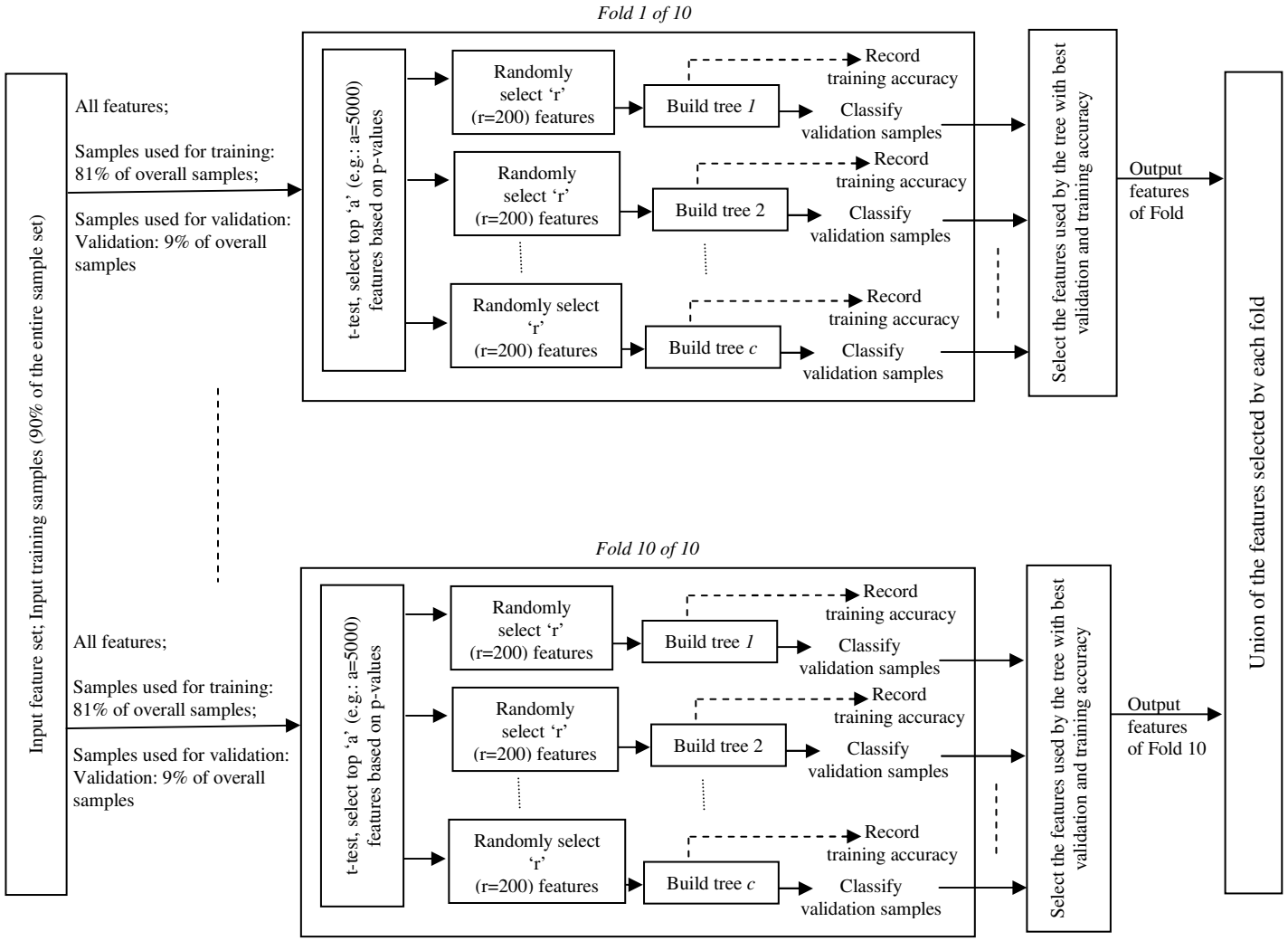


Figure 4.4: Scheme to select good features for classification with typical values of the random subspace parameters (a,r,c) for a given 10-fold cross-validation specified in parenthesis

CHAPTER 5

RESULTS

5.1 Introduction

The colon cancer gene expression dataset, described in Chapter 2, was used to create classifiers to predict survival prognosis for patients. First, supervised and unsupervised feature selection methods were explored to choose the best method for predicting survival. A series of baseline classifier experiments were conducted using the basic experimental scheme described in Section 3.1. Random subspace ensembles were created using the majority voting technique as well as the proposed technique of selecting "good" classifiers (Chapter 4). The performance of these random subspace ensembles was compared to the baseline classifier performance. Finally, the results were further tested and verified in a series of additional experiments.

5.2 Supervised feature selection

T-test

A t-test was used on each feature in the training dataset at a significance level of α=0.05. The null hypothesis for the test was that the mean expression level for the two prognosis groups was equal. A feature was considered to be significant in predicting survival for a colon cancer patient if the p-value for the feature, as determined by the t-test was less than 0.05.

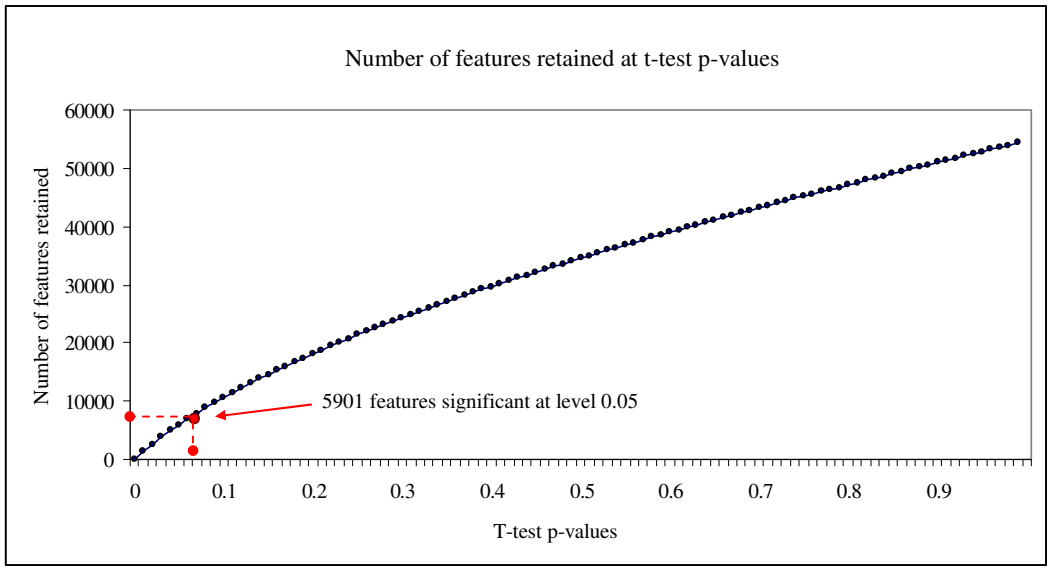


Figure 5.1: Number of features with a specified t-test p-value

Figure 5.1 shows the significance of the features in distinguishing between the two classes for all the samples, at the 0.05 level. All features with p-values less than 0.05 were considered to be significant for prediction, and features with lower p-values were considered to be stronger predictors than features with higher p-values. There were 5901 features found to be predictive features for classification. Since the t-test could aid in selecting a small number of features that were highly significant for prediction, the test was found to be a good feature selection technique for predicting survival for colon cancer patients. This test aided in reducing the number of features for classification, while ensuring that the retained features were indeed strong predictors of survival.

Survival Analysis:

The dataset was analyzed with respect to the two classes using the survival analysis techniques described in Section 3.2. Two Kaplan-Meier survival curves were

plotted for each feature, one curve for each of the two classes. A feature was considered effective in predicting survival if the survival curves for each of the two classes were statistically different. A log-rank test was used at the significance level of 0.05 to test if the survival curves were significantly different. All features with log-rank p-values less than 0.05 were considered to significant in predicting survival times for the patient.

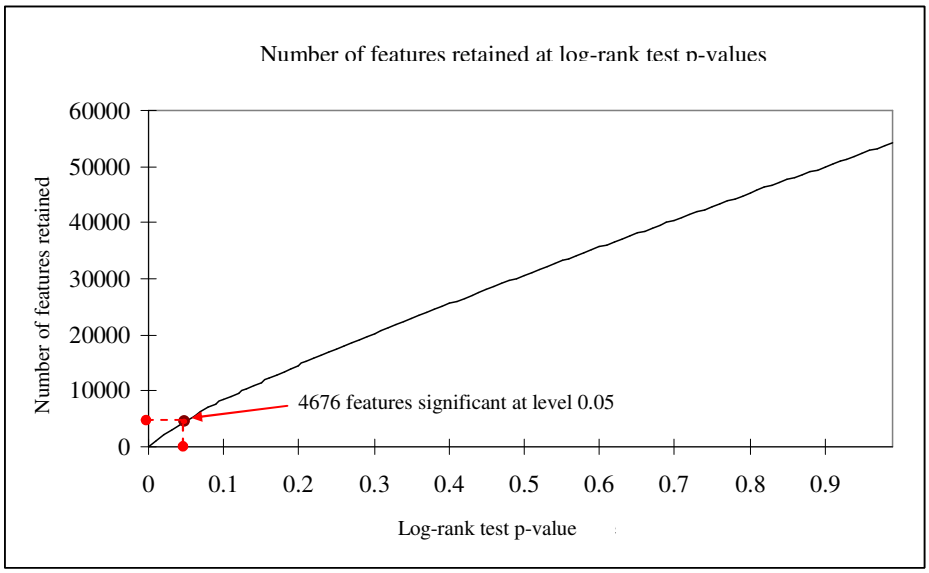


Figure 5.2: Number of features with a specified log-rank test p-value for comparing Kaplan-Meier curves of the two survival classes

It can be observed from Figure 5.2 that only 4676 features in the experiment demonstrate the ability to predict survival. Hence, when censored samples are expected to be included in the experiment, survival analysis could be a reliable feature selection tool.

5.3 Unsupervised feature selection

Quantitative Methods: Expression Level Threshold

Low expression levels recorded during microarray analysis may be attributed to noise in measurement or other undesirable effects. Feature selection by expression level threshold was used to eliminate features that seemed to arise from sources other than expression of genes. The experiment was parameterized by the threshold value t ($3.5 <= t <= 14.5$) for the expression level and the threshold p ($85\% <= p <= 100\%$) for minimum percentage of samples below t. The goal of this experiment was to identify an operating pair (t,p) such that a maximum number of non-informative features were discarded.

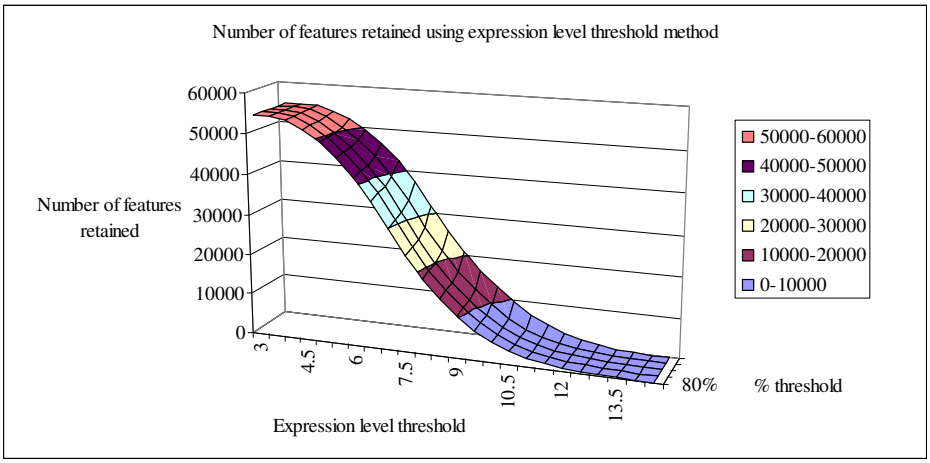


Figure 5.3: Graph of the number of features retained as the two threshold values of expression level and minimum percentage value are varied

Figure 5.3 shows the number of features retained at each point (t,p). It can be observed that the number of discarded features remains fairly constant as 'p' is varied from 85% to 100% for most values of 't'. Also, the number of discarded features drops very slowly for lower values of 't', making it difficult to clearly identify a threshold that

could distinguish between informative features and noise. Given the difficulty in selecting appropriate threshold values, as well as the insignificant drop in the number of features at low threshold values, the method was not considered as a suitable feature selection method for gene expression analysis.

Quantitative Methods: Measures of Variability

The statistical variance method of feature selection was used to eliminate features that did not have high enough variability to be useful for classification. All features with variances below a cut-off threshold *t* (0.05 <= *t* <= 8.0) were considered for elimination.

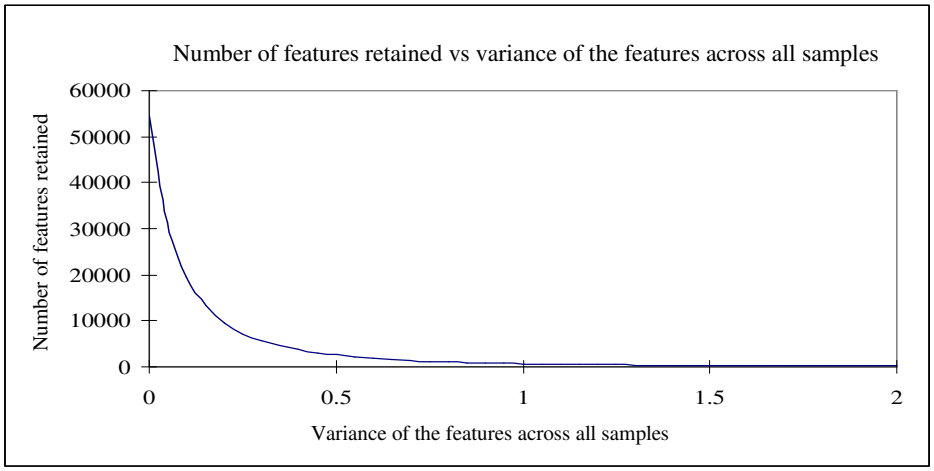


Figure 5.4: Graph of the number of features retained as the threshold for variance is varied

Figure 5.4 shows that a large number of features may be dropped with values of t <0.5. The selection of a threshold for variance could be made by either choosing the desired number of features for classification, or simply by the value of the variance. In either case, care has to be taken to ensure that truly predictive features are not dropped

from consideration. The p-values of a t-test at significance level of 0.05 were used to determine if any predictive features were eliminated. The filter based purely on variance does not take into account the effects of central tendency, such as the mean value, as the t-test does. Hence, at each threshold value of variance below 0.5, at least 25% of the eliminated features were found to be predictive, thereby rendering this feature selection method ineffective for the purpose of classification.

Feature selection with MAD was used to eliminate features with low variability. All features with MAD values less than a threshold t ($0.05 <= t <= 3.5$) were considered to be ineffective for classification and therefore removed.

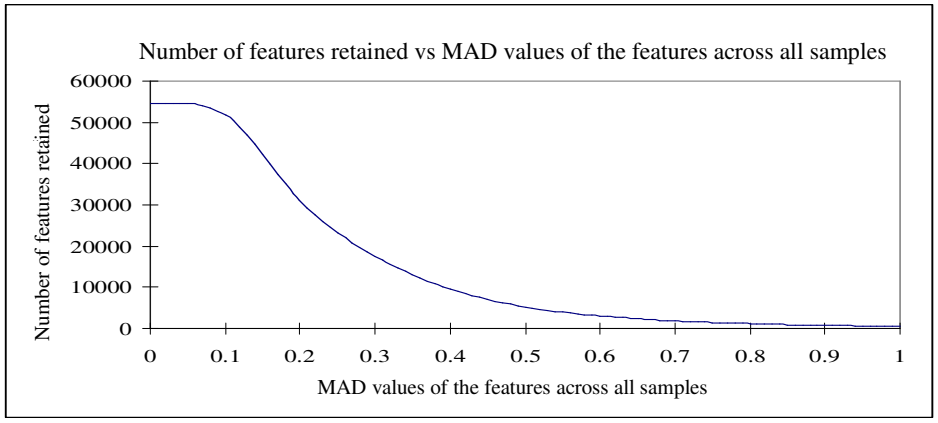


Figure 5.5: Number of features retained as the threshold for MAD values is varied

Figure 5.5 shows that a large number of features may be dropped with values of $0.1 < t < 0.5$. Here, the threshold could be chosen by specifying the desired number of features for classification, or by choosing an optimal value of variability below which the classifier would not be able to distinguish between classes. As described in the section on the experiments with statistical variance, a t-test at significance level of 0.05 was used to

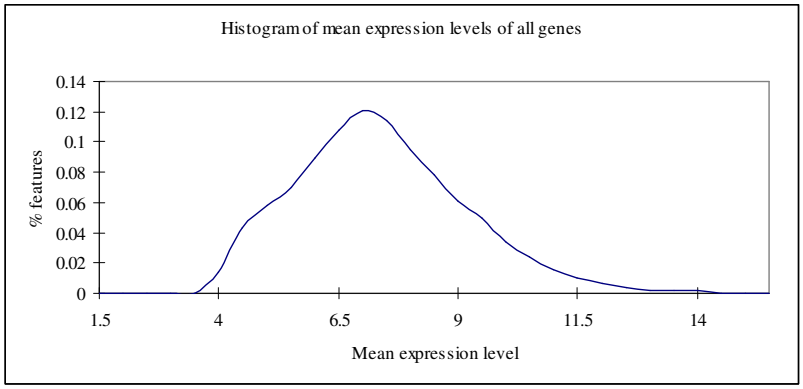
determine if predictive features were eliminated due to the MAD threshold value. At each threshold value $t < 0.5$, at least 25% of the eliminated features were found to be predictive. Thus, feature selection with MAD was not found to be useful for gene expression analysis.

Qualitative Methods: Selection of biologically relevant genes

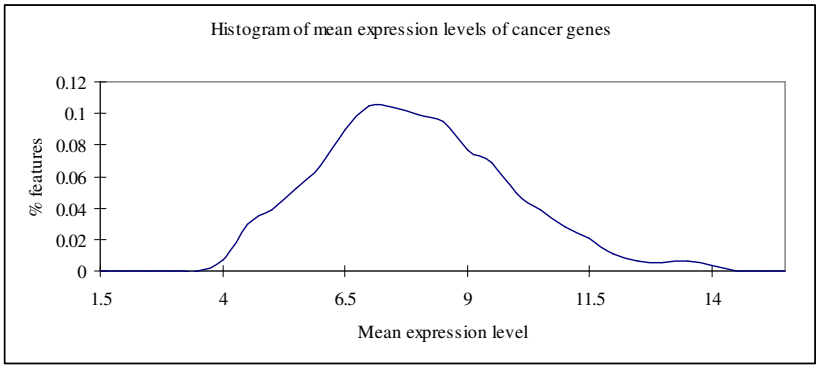
A careful list of all genes associated with cancer was created to study the characteristics of expression levels in known cancer genes [27,28,29]. A total of 5687 cancer related genes, described by 9149 probesets, were used for this experiment (refer to Section 1.5.1).

In order to determine if the cancer-related genes had any distinctive expression patterns in the colon cancer dataset, a smoothed histogram of the mean expression levels of these probesets was compared to the smoothed histogram of the mean expression levels of all the probesets in the dataset. To make a fair comparison of the curves, each histogram was normalized for the number of probesets used. If the cancer related genes were expected to have distinctive characteristics, then the two histograms would show different characteristics in terms of spread and central tendencies.

However, as shown in Figure 5.6, both histograms have very similar characteristics. It can be inferred from the graphs that the set of cancer-related genes that were used for this analysis do not display characteristics that are significantly different from genes that are not associated with cancer progression. Thus, a study of the gene expression patterns of cancer-related genes would not aid in identifying the most predictive features for classification.



(a)



(b)

Figures 5.6: Histogram of the mean level of gene expressions across all samples (a) all genes (b) cancer-related genes

5.4 Baseline experiments with colon cancer gene expression data

The gene expression data for prediction of survival for colon cancer patients was used to conduct three main experiments with three different classifiers: Neural Networks, Support Vector Machines and C4.5 Decision Trees. Each experiment was setup as a 10-fold cross-validation, with the t-test as the feature selection method. The top *a* features ($100 <= a <= 10000$) from the entire dataset were selected within each fold of the cross-validation to avoid pre-selection bias [31]. Since the distribution of the samples across the two classes was not balanced, the weighted accuracy (see Section 3.6) of each experiment was computed as a measure of success.

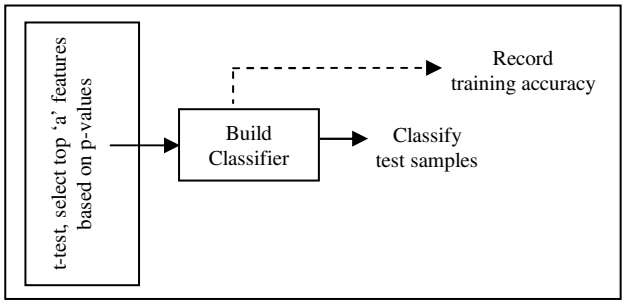


Figure 5.7: Basic classifier block for the baseline gene analysis experiment the parameter *a* was varied (100 <= *a* <= 1000)

Since the parameter *a* could take on several different values, the accuracies of the classifier for each value of *a* (100 < *a* < 10000) was explored, to determine the optimal configuration of the classifier scheme.

The Neural Network used for this experiment was Quickprop [33], a fast implementation of the Feed-forward-back-propagation network described in Section 3.4.1. The network was designed with 10 hidden units and two output nodes. The training of the classifier was designed to halt either when the net error dropped to zero, or in 500 epochs [33]. The Support Vector Machine experiments used the implementation in WEKA (31). A linear kernel was used with standard normalization. The USF implementation of C4.5 decision trees [35] was used to test the accuracies of single decision trees at the various parameter settings.

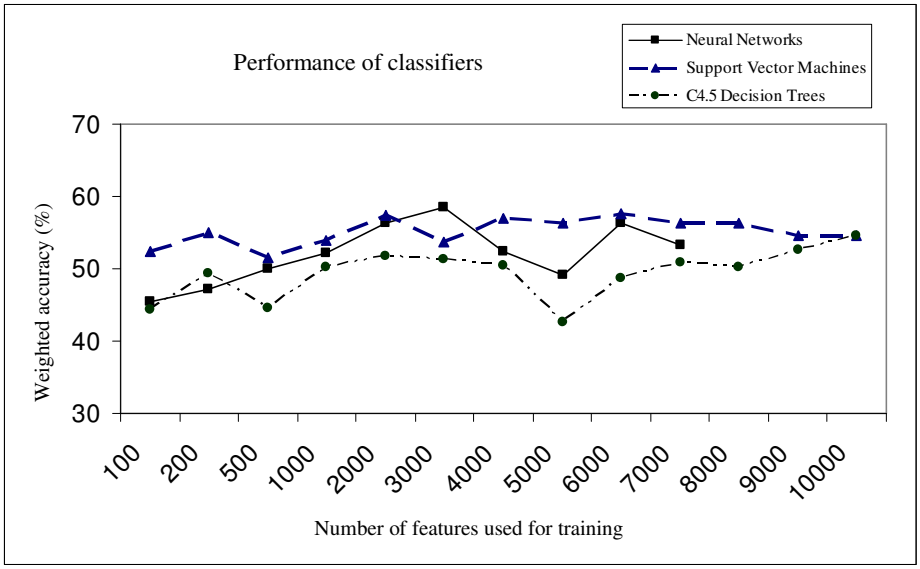


Figure 5.8: Performance of baseline classification schemes

Figure 5.8 shows that none of the baseline classifiers were able to achieve weighted accuracies higher than 58.47%. For all three classifiers, the highest accuracies were achieved when 3000-4000 features were used. This suggests that the best features for prediction are in the top 4000 features of the t-test p-values. The observation is supported by the results of feature selection with t-tests, which indicate that the top 4000 features are highly significant in prediction. As lower numbers of features are used the accuracies drop possibly due to inadequate features to represent the sample characteristics. As higher numbers of features were used, the useful features start being overwhelmed by the non-predictive features, resulting in inaccurate classifiers.

5.5 Majority voting to create ensembles

Random subspace ensembles were created using the majority voting technique described in Section 4.3. The basic classifier block shown in Figure 5.9 was used in the 10-fold cross-validation scheme (refer to Figure 3.9) to create a single ensemble classifier from a set of classifiers built on random subspaces within each fold. A t-test was used on the training samples within each fold to choose the best '*a*' features for classification (100 <= *a* <= 1000). Random subspaces were created by picking features randomly from this set of selected features. Individual decision trees were built on each random subspace and used to predict the class of each test sample. The actual class of each test sample was decided based on the majority prediction of all the trees within the fold. The confusion matrix for the final classifier was created by using the predictions of all the test samples. The weighted accuracy computed from this matrix was used as a measure of performance of the classifier.

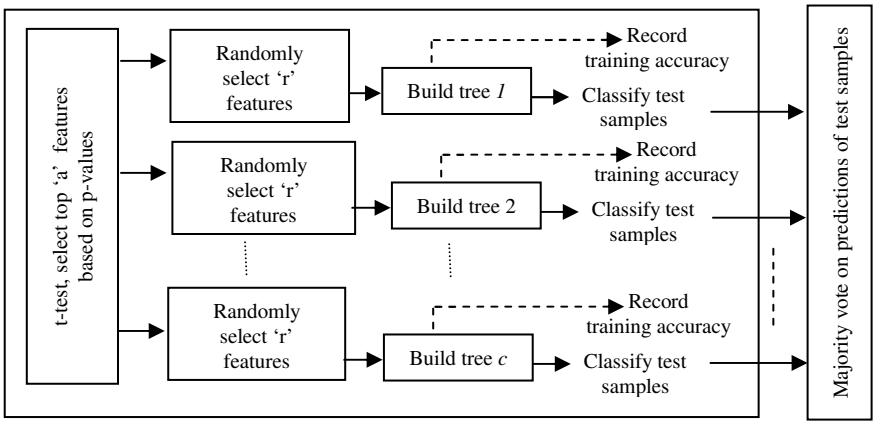


Figure 5.9: Basic classifier block to create random subspace ensembles using majority voting technique (used within the 10-fold cross-validation scheme, Figure 3.9); see Table 5.1 for experimental values of parameters *(a,r,c)*

Several experiments were conducted for the various values of the design parameters, random subspace size (r), number of random subspaces (c) and the number of features used for classification (a). The values of these parameters used for the experiment are listed in Table 5.1.

Table 5.1: Range of parameters used for majority voting technique using random subspace ensembles

Parameter	Description	Min value	Max value
a	Top features selected from t-test	5000	10000
r	Size of random subspace	50	2000
c	Number of random subspaces/trees	1	2000

If all the features selected from the t-test feature selection stage are predictive in nature, a random subspace ensemble created using a majority voting technique is expected to yield higher accuracies as a larger number of subspaces are created (refer to Appendix Sections A.1 and A.2 for details). A larger number of subspaces of a given size ensure better coverage of the feature space, and hence the ensemble of classifiers is expected to learn the patterns in the samples more accurately.

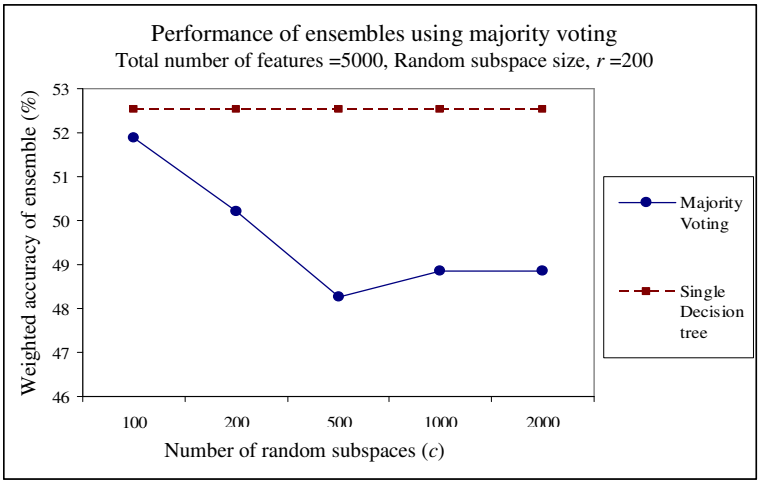


Figure 5.10: Random subspace ensembles (a=5000,r=200,c) vs single decision tree (a=5000,r=200,c=1)

The weighted accuracy of ensembles created with varying values of (a,r,c) were compared to the accuracies of single classifiers created from a single random subspace *(a,r,c=1)*. It can be observed from Figure 5.10 that, contrary to the expected outcome, there is a decrease in accuracy as the number of random subspaces is increased. This indicates that a large number of the random subspaces created are probably not very effective in describing the sample classes.

In order to investigate the nature of these subspaces, the decision tree built on each random subspace was tested on the 10% held-out test samples from a single 90%-10% split of the data. The weighted test accuracies of these subspace classifiers were analyzed to identify the subspaces that represented the sample classes well, and those that did not.

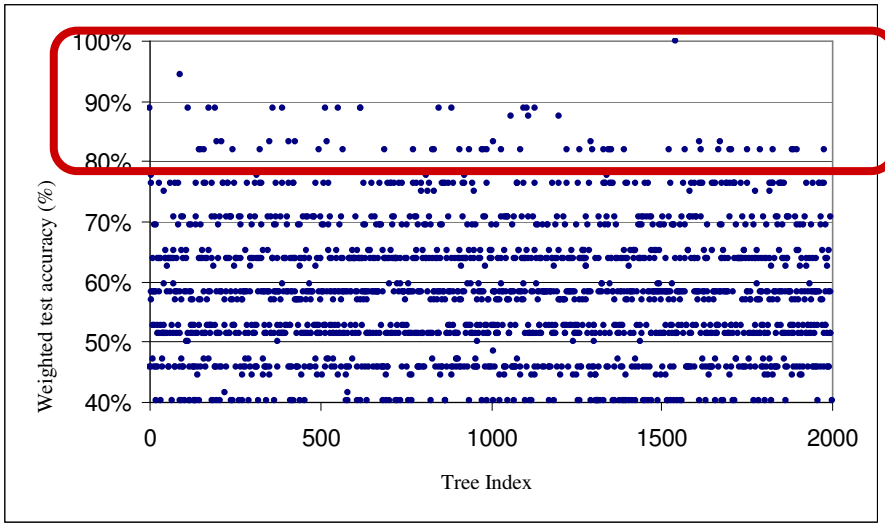


Figure 5.11: Weighted test accuracies of 2000 random decision trees

Figure 5.11 shows the spread of the weighted test accuracies of 2000 decision trees created from random subspaces of size 200 features from the top 5000 t-test features

(random subspace parameters are (*a=5000, r=200, c=2000*)). If the top 5000 features as determined by the p-values of a t-test at significance level 0.05 are predictive in nature, and the combination of these features is also predictive, then all random subspaces created from these 5000 features would be expected to perform accurately. However, less than 7% of the 2000 random subspaces created were found to have accuracy higher than 80%.

An analysis of the training accuracies for each decision tree for the corresponding test accuracies indicated that while a few random subspace classifiers seemed to have learned the training samples well, the performance on test samples was poor. Only a few subspace classifiers had been able to learn the training samples well and were able to generalize the knowledge enough to predict classes of test samples accurately (Fig 5.12).

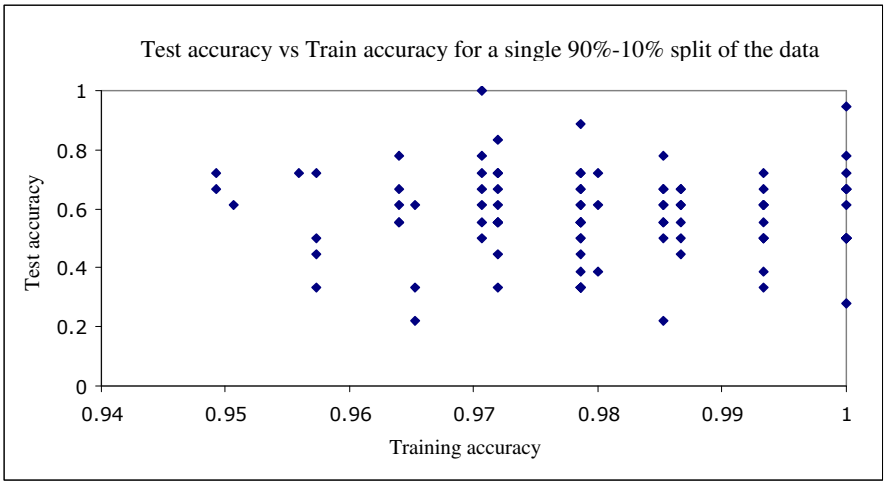


Figure 5.12: Weighted training and testing accuracies of 100 random classifiers built from random subspaces

5.6 Selection of good subspaces

Figures 5.11 and 5.12 indicate that only a small subset of the random subspaces generated on the input features are effective for prediction. The classifiers created on random subspaces that generate high test and train accuracies simultaneously are considered as good classifiers. The basic scheme to select good subspaces or features is described in Section 4.4. A 10-fold cross-validation scheme was used to split the 121-sample colon cancer gene expression dataset (see Chapter 2) into 10 sets of training (90%) and independent test (10%) sets. For each fold, an additional 10-fold cross-validation created 10 sets of training samples (81%) and validation samples (9%). A t-test was used on each of the 81% training sets to select the best 5000 features from the total of 54675 features, ranked according to the t-test p-values at the significance level of 0.05. 200 features were picked randomly from this set of 5000 features to create a random subspace. 100 such random subspaces were created. For each random subspace, a single decision tree was built on the training samples (81%), and the training accuracy was recorded. The decision tree was tested on the 9% validation set. The decision tree with the highest validation accuracy and the highest training accuracy was selected as the best classifier for that training and validation set of samples. The features used by this tree were selected as good features for the sub-fold.

Each of the 10 training and validation sets for a selected fold produces a set of good features. The union of these features sets was used to train a single classifier on the 90% training samples for that fold. This single classifier was tested using the independent test set (10%) to estimate the prediction accuracy. Ten such classifiers were built, one for each fold of the cross-validation scheme. The predictions of these classifiers on the

individual independent test samples were collectively used to create the confusion matrix for the classifier scheme. The weighted accuracy of prediction for the classifier was computed from this matrix.

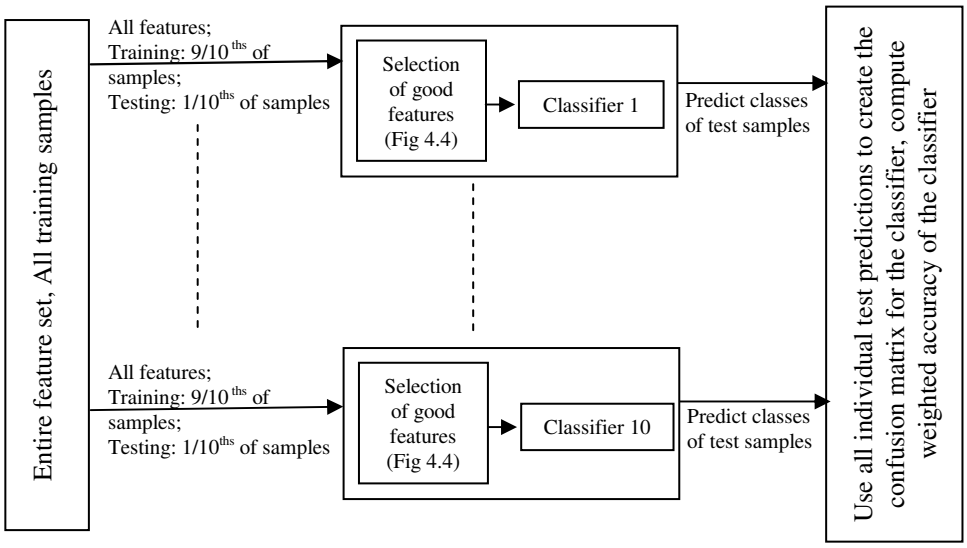


Figure 5.13: Classification by selection of good subspaces

Experiments were performed with three different classifiers (neural networks, support vector machines and decision trees) for prediction, using the scheme shown in Figure 5.13.

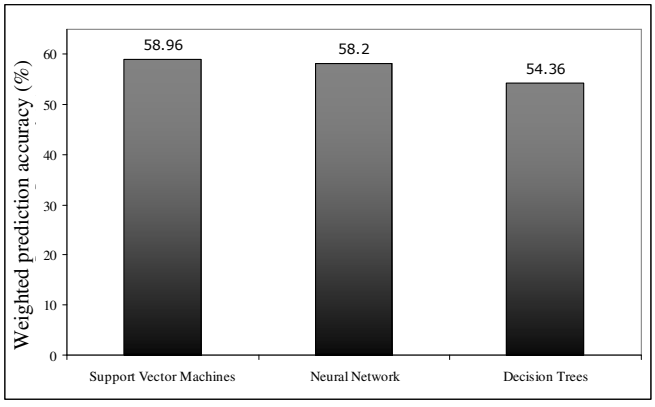


Figure 5.14: Weighted accuracies of neural networks, support vector machines and decision trees; these classifiers were trained on the union of the best features created by selecting good random subspaces (Section 5.6)

Support vector machines were found to achieve the highest accuracy and the use of decision trees resulted in the poorest accuracy of prediction. The performance measures for the best classifier are listed in Table 5.2.

Table 5.2: Confusion matrix for the performance of the support vector machine trained on the union of the features created by selecting good random subspaces (LT: survival less than 3 years, GT: survival greater than 3 years)

LT	GT	Classified as / True class
15	22	LT
19	65	GT
Weighted accuracy		58.96 %
Total accuracy		66.12 %
Sensitivity		40.54 %
Specificity		77.38 %

The classifier was used to predict one of two classes for each test sample. These two classes (survival less than 3 years and survival greater than 3 years) were used as two groups to draw survival K-M curves. The p-value for the log-rank test to compare the curves indicates that the two predicted classes are significantly different from each other. As can be observed in Figure 5.15, the percentage of patients surviving across time in the poor prognosis class (LT) decreased at a higher rate than the patients in the good prognosis category (GT). While the survival curves are significantly different when using a survival cut-off point of 36 months, a clear cut-off in the survival values cannot be observed for the two classes. Hence, a more optimal cut-off point in survival may yield better accuracy of prediction.

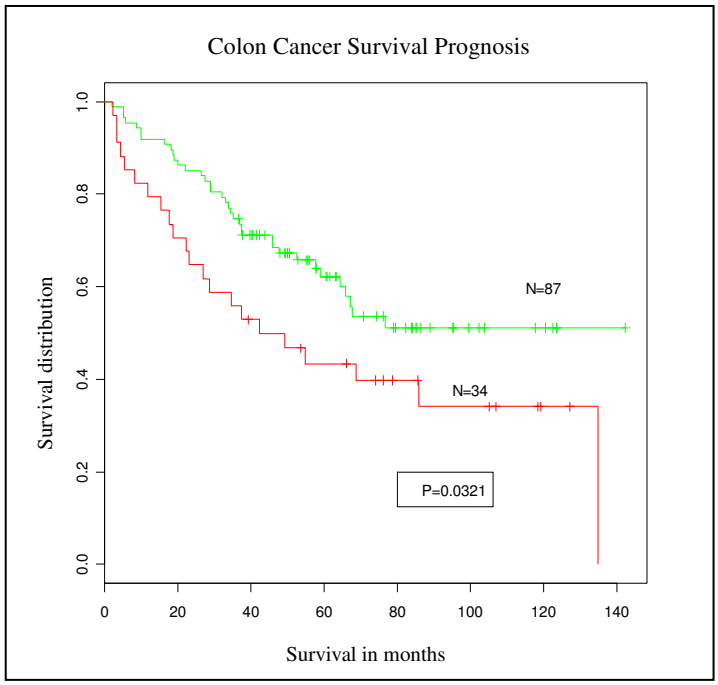


Figure 5.15: Survival curves for the predicted classes; the survival curves are statistically different at significance of 0.05 as determined by a log-rank test

Analysis of features used by the classifier

The t-test was used to select the best features for prediction on 81% of the samples. Since this selection was repeated on a different set of 81% training samples each time, a different set of features may be selected for use depending on the patterns of the training samples within the classes, with a minimum of 5000 unique features selected across the entire experiment. A larger number of unique features would indicate that the predictive strength of the features varied depending on the samples used for training. A total of 24998 unique features were selected by the t-tests across all the folds. This suggests some features were found to be predictive only when specific samples were used for training.

The random subspaces were created by picking 200 features randomly from the best 5000 features determined by the t-test. Decision trees, built on these random

subspaces, selected the best of these 200 features to create the tree. Each tree on an average used 10 features selected from the random subspace. Since the 10 features sets were used to create the union of features for classification, it is expected that between 10 and 100 unique features would be used by a single classifier. The entire classifier scheme including 10 such classifiers used a total of 744 features. Hence, each classifier used an average of 74.4 unique features. 667 of these 744 features used for classification were found to be unique.

Features that are truly predictive would ideally be found to be the best features across multiple folds of the cross-validation scheme. Figure 5.16 shows the repetition of features across three or more folds using the classifier described in Section 5.6. Since these features were selected as predictive features for various combinations of training samples, they are expected to be the most predictive for the samples used in the study.

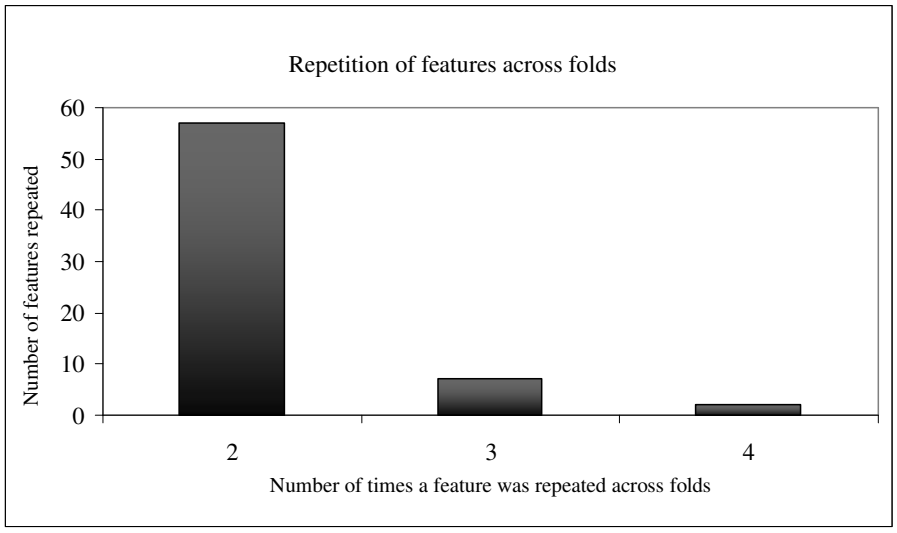


Figure 5.16: Repetition of features across two or more folds of the cross-validation scheme

5.7 Verification of results

In order to help verify the results of classification, a few additional experiments were performed to test the effect of randomization of the samples and the features subspaces on classification accuracy.

The proposed method uses randomization of the samples into train and test sets at each fold of the cross-validation, and in the creation of the training and validation sets. Further, the random subspaces select features from a pool at random, with a few repetitions of the features. The reliability of the classifier results can be ensured if the results are repeated with different random selections at every stage.

To investigate the reliability of the classifier scheme, a series of experiments were conducted to vary the configuration of the samples and feature subspaces. The first experiment was a re-run of the experiment three times with constant parameters for the entire classifier scheme. Since the parameters for partitioning of the data into the various training, validation and independent test sets split the samples into exactly the same configuration for each of these experiments, the only source of variation in results was the randomization in creation of subspaces.

Three additional experiments were conducted to change the parameters of the cross-validation schemes. In the first of these experiments, the initial random seed used to create the split of samples into the independent test and train sets was varied. The second of the experiments varied the split of samples into the training and validation sets, and the third experiment varied both the splits simultaneously.

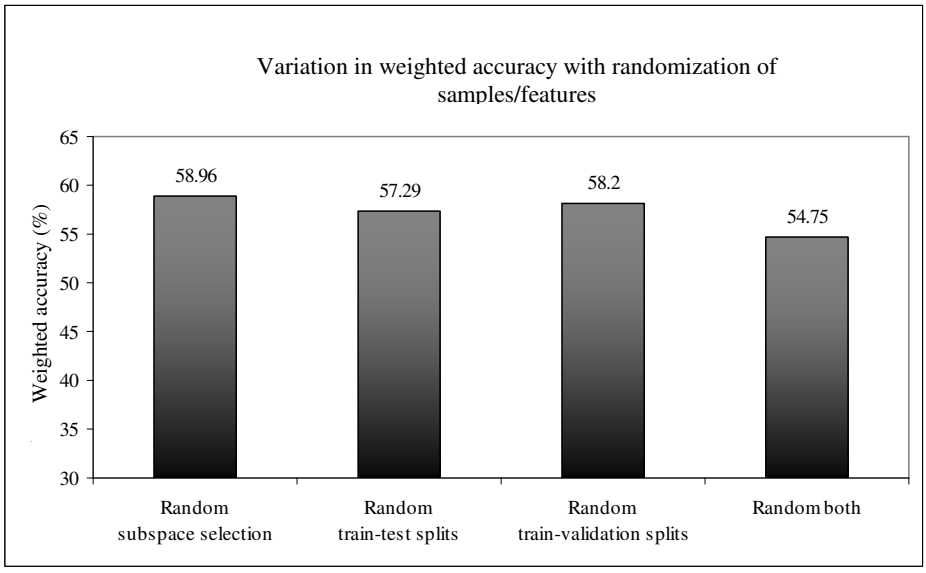


Figure 5.17: Variation in the weighted accuracy for prediction of survival for colon cancer with changes in randomization of the samples and feature subspaces; the accuracies reported here are the averages of experiments for each randomization

It can be observed from Figure 5.17 that the weighted accuracy from the verification experiments varies by 3.36%, indicating that the classification scheme is relatively stable in spite of changes in the samples used for training.

The next step in verifying the stability of the classifier is analyzing the features used for classification by each experiment. Repetition of features used would indicate that the feature selection at the random subspace stage was relatively invariant to the randomization of samples and subspace generation. Figure 5.18 shows the number of features that were used by a total of 8 experiments, including two experiments with randomization of the subspaces, and two each for the randomization for splitting the data into the various test, train and validation sets. It can be observed that several features were repeatedly picked as the best predictors in spite of the various randomization effects in the verification experiments.

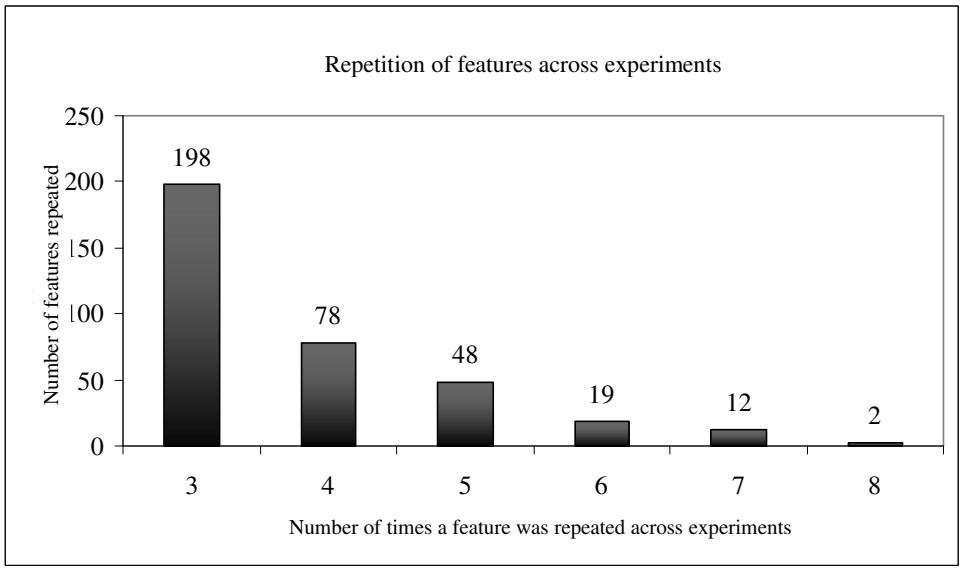


Figure 5.18: Number of features repeatedly selected as the most predictive features across all the experiments to test variability of results

CHAPTER 6

DISCUSSION AND CONCLUSION

6.1 Discussion

The proposed classifier scheme using random subspace ensembles, described in Section 5.6, achieved a weighted accuracy of prediction of 58.96% for colon cancer microarray data. Several features used by the classifier in the final prediction of samples were found to be repeated across at least three folds of the cross-validation scheme. Further, the survival times for each predicted class for this classifier were found to be significantly different (Figure 5.15), indicating that the features used for prediction are collectively predictive in nature for the colon cancer gene expression data.

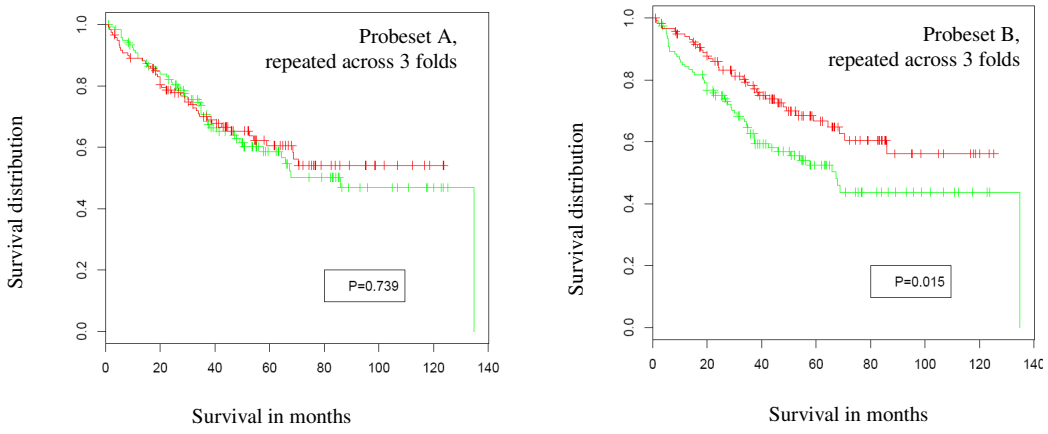


Figure 6.1: Survival curves for two genes, split on the median, repeated across three folds in the classifier scheme described in Section 5.6

Figure 6.1 shows two survival K-M curves and the corresponding log-rank tests for two of the features that were repeated across three folds of the cross-validation scheme. Since these features were picked as the most predictive features in three of the folds, they would be expected to be individually predictive. However, Figure 6.1 indicates that while the K-M curves for feature B are significantly different, indicating good predictive value for the feature, the K-M curves for feature A are not significantly different. This suggests that although the selected features may not be individually significant in prognostic value, a set of features in combination could be good for prediction of survival. The goal of the classifier is then to select an optimal set of features that can collectively predict the outcome of colon cancer in terms of survival.

Although the proposed method using random subspaces improved the weighted accuracy (Sections 5.5 and 5.6), the success of any of the classifiers generated is clearly not optimal. This could indicate that the colon cancer dataset probably includes sub-groups of patients within each of the survival groups. These sub-groups may exhibit unique characteristics that are not sufficiently described by the group as a whole. In other words, the two survival groups could likely be heterogeneous in gene expression characteristics. The voting technique to create random subspace ensembles would work well in simple cases (refer to Appendix Sections A.1 and A.2) where the groups for classification consist of homogeneous gene expression characteristics. While working with a more complex or heterogeneous set of samples, the simple voting technique would confuse the classification since all the sub-groups would not be adequately represented. In such cases, the proposed technique of selecting the best feature subspaces across multiple

folds may work well by generalizing the feature space to include a majority of the sub-groups within the training samples.

Future work

The classifier scheme of creating random subspace ensembles, by selection of the best feature subspace, has been shown to work at least as well as the baseline experiments in the task of predicting 36-month survival for colon cancer patients. However, Figure 5.15 suggests that a dichotomization on survival time points other than 36 months may lead to a more accurate classifier. The proposed method could be used with other survival time thresholds, to investigate the split of the training samples for highest accuracy of prediction.

Further, the configuration of the random subspaces used with C4.5 decision trees was selected based on the results of the baseline experiments. Experimentation with variations of the configuration would help in identifying a potentially more accurate classifier scheme and selection of features that are more robust in survival prediction. The described method used the random subspace classifier with the best validation and training accuracy for selection of features. Selection of multiple subspaces rather than a single best random subspace may enhance the accuracy of survival prediction by including better descriptors of the classes.

In the description of the proposed method, C4.5 decision trees have been used with the random subspaces to select good features for classification. Support vector machines or neural networks were used with these good features to train on the training samples. These classifiers were used for prediction of classes for new samples. Use of the

same type of classifier at the feature selection stage as well as final classification could yield a simpler model. However, the effect of such a model on the accuracy of prediction may be assessed only through experimental evidence.

6.2 Conclusion

Gene expressions of cancer tissue at different stages of development are expected to have unique signatures. Identification of these signatures would aid in prognosis of cancer, and prediction of long-term survival for the patient. Gene expressions of colon cancer tissue were studied for the purpose of predicting 36-month survival for the cancer patient. Microarray technology enables analysis of gene expressions by generating information for thousands of genes. A t-test was used to select a set of the most promising features for prediction. Random subspace ensembles created using these selected features yielded poor accuracy in survival prediction for the colon cancer data. A modification to the random subspace technique was proposed, that selected the most accurate feature subspace amongst all the random subspaces, created as the most predictive features. These predictive features were used by support vector machines to classify samples into two survival groups with a weighted accuracy of 58.96 %. The accuracy of this classifier was shown to be comparable to any of the baseline classifiers tested on the same dataset in predicting the class of new and unknown samples. Further, the method was tested on other gene expression datasets (see Appendix Sections A.1 and A.2) and shown to work with prediction accuracies comparable to the accuracies of the baseline classifiers.

REFERENCES

1. Douglas Hanahan, Robert A. Weinberg. *The Hallmarks of Cancer*. Cell, Vol. 100, 57-70, January 7, 2000. Copyright 2000 by Cell Press.

2. E. J. Ambrose, F. J. C. Roe (1975) *Biology of Cancer.* Sussex, England: Ellis Horwood Limited.

3. David J. Lockhart, Elizabeth A. Winzeler. *Genomics, gene expression and DNA arrays*. Nature Vol 405, 15 June 2000.

4. E. A. Carlson (2004). *Mendel's Legacy*. New York: Cold Spring Harbor Laboratory Press.

5. Watson, J., Crick, F.H.C. *Molecular Structure of Nucleic Acids*. Letters to Nature, Nature 171, 737-738 (1953), Macmillan Publishers Ltd.

6. Elaine Marieb. *Human Anatomy and Physiology*. Pearson Education Inc.

7. Findley, McGlynn, Findley (1989). *The Geometry of Genetics*. Wiley-Interscience Monographs In Chemical Physics.

8. Affymetrix. http://www.affymetrix.com/index.affx.

9. The Human Genome Project. http://www.ornl.gov/sci/techresources/Human_Genome/project/info.shtml

10. Edlich RF, Winters KL, Lin KY. *Breast cancer and ovarian cancer genetics*. J Long Term Eff Med Implants. 2005;15(5):533-45.

11. Murphy N, Millar E, Lee CS. *Gene expression profiling in breast cancer: towards individualising patient management*. Pathology. 2005 Aug;37(4):271-7.

12. Ueki K.Oligodendroglioma*: impact of molecular biology on its definition, diagnosis and management.* Neuropathology. 2005 Sep;25(3):247-53.

13. 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, E. S. Lander. *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*. Science, Vol. 286, 15 October 1999.

14. Gennadi V. Glinsky, Olga Berezovska, Anna B. Glinskii. *Microarray analysis identifies a death-from-cancer signature predicting therapy failure in patients with multiple types of cancer*. The Journal of Clinical Investigation, Vol. 115, No. 6, June 2005.

15. Laura J van ‘t Veer,. Hongyue Dai, Marc J van de Vijver, Yudong D. He, Hart, Augustinus A. M., Mao Mao, Hans L.Peterse, Karin van der Kooy, Marton, Matthew J., Anke T Witteveen, George J Schreiber, Ron M. Kerkhoven, Chris Roberts, Peter S. Linsley, Rene Bernards, Stephen H Friend. *Gene expression profiling predicts clinical outcome of breast cancer*. Letters to nature, Vol. 415(6871), 31 January 2002, pp 530-536.

16. M.van de Vijver, Y.He, L.van'T Veer , H.Dai, A.Hart, D.Voskuil, G.Schreiber, et al. *A gene expression signature as a predictor of survival in breast cancer*. The New England Journal of Medicine 2002; 347 (25): 1999-2009.

17. U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, A. J. Levine. *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotides arrays*. Proc. Natl. Acad. Sci. USA, Vol. 96, pp. 6745-6750, June 1999, Cell Biology.

18. Sridhar Ramaswamy, Ken N. Ross, Eric S. Lander, Todd R. Golub. *A molecular signature of metastasis in primary solid tumors*. Nature Genetics, Vol 33, January 2003.

19. Steven Eschrich, Ivana Yang, Greg Bloom, Ka Yin Kwong, David Boulware, Alan Cantor, Domenico Coppola, Mogens Kruhoffer, Lauri Aaltonen, Torben F. Orntoft, John Quackenbush, Timothy J. Yeatman. *Molecular staging for survival prediction of colorectal cancer patients*. J Clin Oncol. 2005 May 20;23(15):3526-35.

20. National Cancer Institue. *SEER Statistics*. http://seer.cancer.gov/

21. Bolstad, B.M., Irizarry R. A., Astrand, M., and Speed, T.P. *A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance*. Bioinformatics 19(2):185-193, (2003).

22. Torres-Roca JF, Eschrich S, Zhao H, Bloom G, Sung J, McCarthy S, Cantor AB, Scuto A, Li C, Zhang S, Jove R, Yeatman. *Prediction of radiation sensitivity using a gene expression classifier*. Cancer Res. 2005 Aug 15;65(16):7169-76.

23. Tusher VG, Tibshirani R, Chu G. *Significance analysis of microarrays applied to the ionizing radiation response*.Proc Natl Acad Sci U S A. 2001 Apr 24;98(9):5116-21. Epub 2001 Apr 17. Erratum in: Proc Natl Acad Sci U S A 2001 Aug 28;98(18):10515.

24. D. C. Montgomery, G. C. Runger, N. F. Hubele (1997). *Engineering Statistics*. John-Wiley & Sons, Inc.

25. D. G. Kleinbaum (1996). *Survival Analysis: A self-learning text*. New York: Springer-Verlag New York, Inc.

26. Statsoft inc. *Survival Analysis*: http://www.statsoft.com/textbook/stsurvan.html#kaplan.

27. Atlas of Genetics and Cytogenetics in Oncology and Haematology. http://www.infobiogen.fr/services/chromcancer/Genes/Geneliste.html

28. Cancer Genetics Web. http://www.cancerindex.org/geneweb/genes_a.htm

29. Tumor Gene Database. http://condor.bcm.tmc.edu/ermb/tgdb/tgdb.html

30. Ivana V Yang, Emily Chen, Jeremy P Hasseman, Wei Liang, Bryan C Frank, Shuibang Wang, Vasily Sharov, Alexander I Saeed, Joseph White, Jerry Li, Norman H Lee, Timothy J Yeatman, and John Quackenbush. *Within the fold: assessing differential expression measures and reproducibility in microarray assays*. Genome Biol. 2002; 3(11): research0062.1–research0062.12.

31. H. Witten, E. Frank (2000). *Data Mining*. Morgan Kauffman Publishers.

32. S. V. Kartalopoulos (1996). *Understanding Neural Networks and Fuzzy logic: Basic Concepts and Applications.* New Delhi: Prentice-Hall of India.

33. Simon Haykin (1999). *Neural Networks: A Comprehensive Foundation*. New Jersey: Prentice Hall, Inc.

34. Cristopher J. C. Burges (1998). Data Mining and Knowledge Discovery, *A Tutorial on Support Vector Machines for Pattern Recognition.* Boston: Kluwer Academic Publishers.

35. Tin Kam Ho.*The Random Subspace Method for Constructing Decision Forests*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 8, August 1998.

36. Trisha Greenhalgh. *How to read a paper: Papers that report diagnostic or screening tests*. *BMJ* 1997;315:540-543 (30 August).

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

BIBLIOGRAPHY

1. Findley, McGlynn, Findley (1989). *The Geometry of Genetics*. Wiley-Interscience Monographs In Chemical Physics.

2. A.H. Sturtevant (2001). *A history of genetics*. New York: Cold Spring Laboratory Press.

3. F. H. Prtugal, J. S. Cohen (1977). *A Century of DNA: A History of the Discovery of the Structure and Function of the Genetic Substance*. Cambridge: The MIT Press.

4. R. O. Duda, P. E. Hart, D. G. Stork (2001). *Pattern Classification*. John Wiley & Sons, Inc.

APPENDICES

Appendix A: Application of the proposed method on various gene expression datasets

The proposed method has been applied to the analysis of the gene expression of colon cancer in order to predict survival. The method was shown to work with prediction accuracy comparable to other classifier schemes discussed in the literature. In order to test the merit of the proposed method on analysis of gene expression profiles, the method was used to analyze two datasets with different class characteristics. The first dataset used was the publicly available leukemia dataset [13], with two main classes: ALL (acute lymphoblastic leukemia) and AML (acute myeloid leukemia). The second dataset constituted gender information, extracted from the colon cancer survival dataset. The two classes in this case were male and female patients. The description of the experiments for each of these datasets is outlined in the following sections.

A.1Analysis of leukemia data

*Data Description***:**

The leukemia gene expression dataset consists of two variants of leukemia, Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML). The dataset includes a total of 7129 normalized features or probesets. The 38 samples in the dataset include 27 samples of ALL and 11 samples of AML.

Appendix A: (Continued)

Baseline Experiments:

Basic classifier experiments were conducted to obtain a baseline performance measure on the dataset. The three classifiers used were Neural Networks, Support Vector Machines and C4.5 Decision Trees.

The t-test was used as an initial feature selection to reduce the number of features used for classification. Since the number of samples in the two classes was unequal a weighted accuracy was used to measure the success of classification.

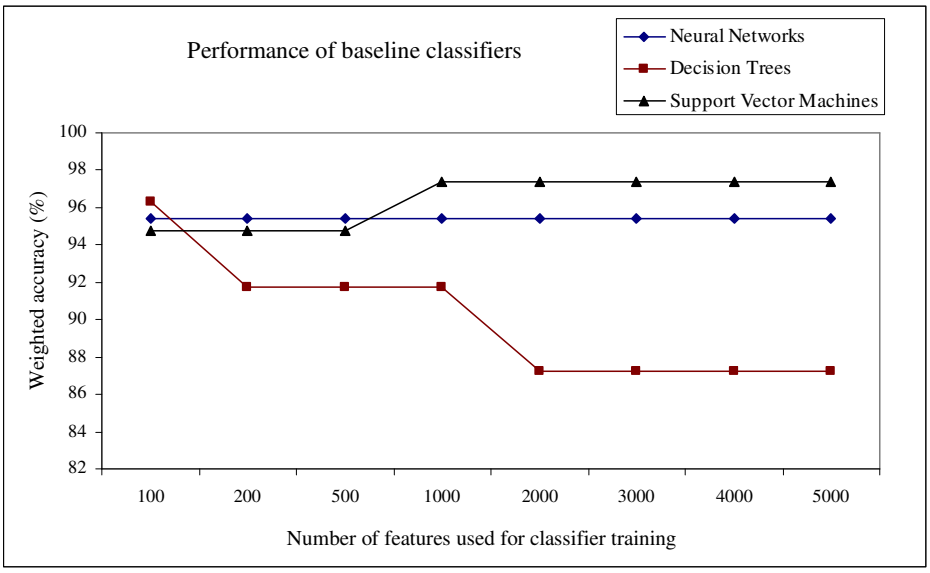


Figure A.1: Classifier performance with ALL-AML:
neural networks, support vector machines and C4.5 decision trees

The Neural Networks performed consistently, with an accuracy of 95.45%, with all values for the feature selection method ($100 <= a <= 5000$). Support Vector Machines achieved a high accuracy of 97.37%. Decision trees however, deteriorated in performance as the number of input features increased, with the maximum accuracy occurring at the lowest number of features.

Appendix A: (Continued)

Random subspace ensembles using majority voting technique:

The majority voting technique (Figure 5.9) was used in the creation of random subspace ensembles to predict classes of samples from the ALL-AML dataset. The experiment was tested at various parameters of (*a,r,c*) (Table 5.2), using the weighted accuracy to measure the performance of the ensemble. The performance of the ensembles was compared with the performance of a single decision tree built on a single random subspace selected from the same set of *a* features.

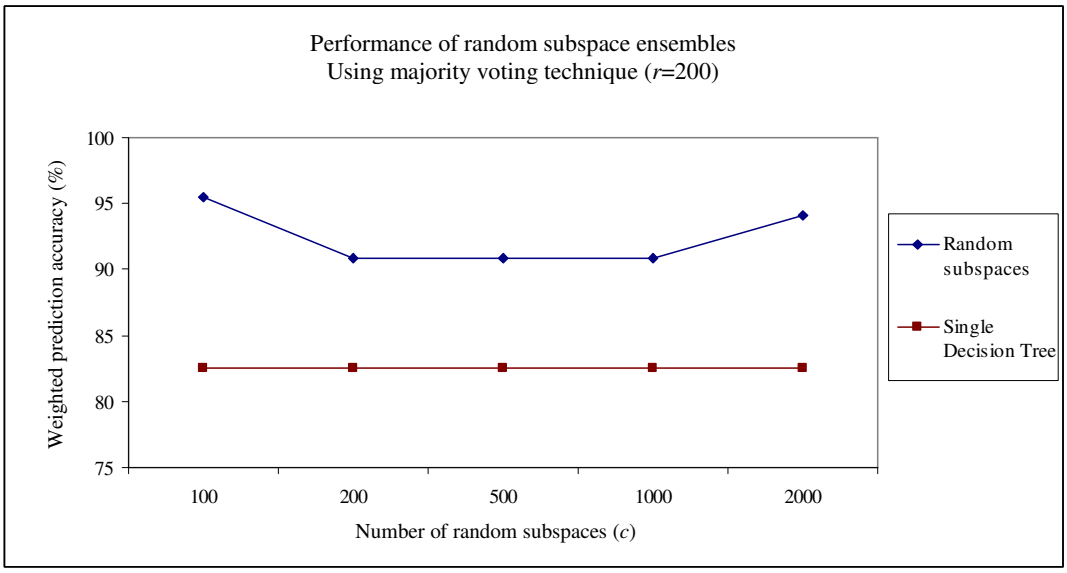


Figure A.2: Random subspace ensembles (*a*=5000,*r*=200,*c*)
vs. single decision tree (*a*=5000,*r*=200,*c*=1) on the ALL-AML dataset

The accuracy of the ensemble increases as the number of subspaces increases. The increasing number of subspaces ensures better coverage of the feature space. Since all the features seem to be predictive in nature, the accuracy of the ensemble increases as more predictive features are added to it. Each of these random subspace ensembles has a better predictive accuracy than a single decision tree (Figure A.1).

Appendix A: (Continued)

*Random subspace ensembles by selection of good classifiers***:**

The proposed method of using random subspaces ensembles by selecting the good classifiers within the cross-validation scheme (Figure 5.13) was tested on the ALL-AML dataset. 100 random subspaces, each of size 200, selected from the top 5000 t-test features, were used to create the ensembles. Support Vector Machines were used for classification. The performance of the method, assessed by computing the weighted accuracy of prediction on the 10%, held-out independent samples, is shown in Table A.1.

Table A.1: Confusion matrix for the performance of the proposed method on the leukemia gene expression dataset

ALL	AML	Classified as / True class
27	0	ALL
2	9	AML
Weighted accuracy		90.91 %
Total accuracy		94.74 %
Specificity		100.0 %
Sensitivity		81.81 %

A.2 Analysis of gender data

Data Description:

The colon cancer dataset (refer to Chapter 2) was split into two classes based on gender: MALE and FEMALE. The dataset consisted of 135 samples, with 68 female and 67 male patients. Each sample was characterized by 54675 normalized features/probesets.

Appendix A: (Continued)

Baseline Experiments:

Basic classifier experiments were conducted to obtain a baseline performance measure on the dataset. The two classifiers used were Neural Networks and Support Vector Machines. The t-test was used as an initial feature selection to reduce the number of features used for classification. Weighted accuracy was used to measure the success of classification.

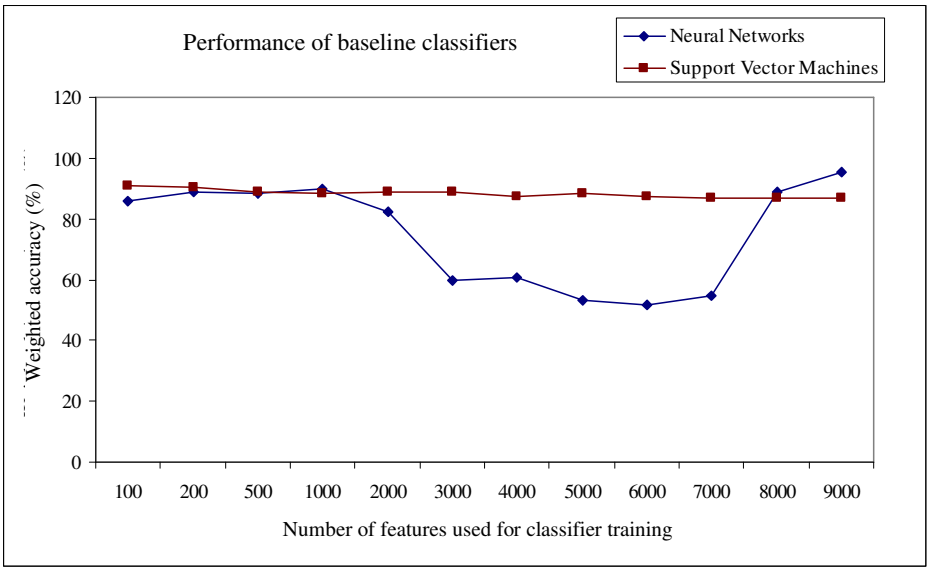


Figure A.3: Classifier performance with gender dataset: neural networks and support vector machines

Random subspace ensembles using majority voting technique:

The majority voting technique (Figure 5.9) was used in the creation of random subspace ensembles to predict gender of samples the dataset. The experiment was tested at various parameters of (*a,r,c*) (Table 5.2), using the weighted accuracy to measure the

performance of the ensemble. The performance of the ensembles was compared with the performance of a single decision tree built on a single random subspace selected from the same set of *a* features.

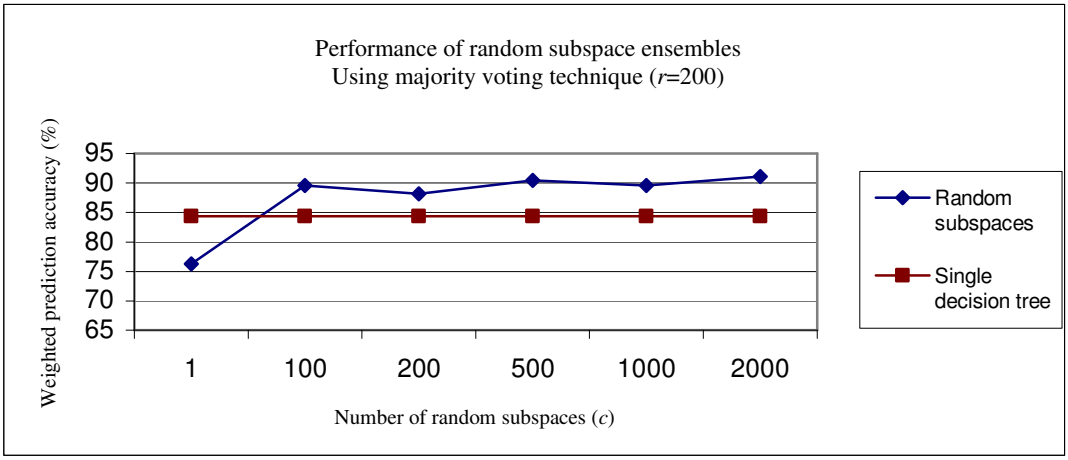


Figure A.4: Random subspace ensembles (*a*=5000,*r*=200,*c*) vs. single decision tree (*a*=5000,*r*=200,*c*=1) on gender dataset

As expected, the accuracy of the ensemble increases as the number of subspaces increases. The increasing number of subspaces ensures better coverage of the feature space. Since all the features seem to be predictive in nature, the accuracy of the ensemble increases as more predictive features are added to it. Each of these random subspace ensembles has a better predictive accuracy than a single decision tree (Figure A.1).

Random subspace ensembles by selection of good classifiers:

The proposed method of using random subspace ensembles by selecting the good classifiers within the cross-validation scheme (Figure 5.13) was tested on the gender dataset. 100 random subspaces, each of size 200, selected from the top 5000 t-test

features, were used to create the ensembles. Support Vector Machines were used for classification. The performance of the method, assessed by computing the weighted accuracy of prediction on a 10-fold cross-validation, is shown in Table A.2. It is observed that the proposed method creates a classifier that predicts classes of unknown samples with accuracy comparable to that obtained using the majority voting technique of creating random subspace ensembles.

Table A.2: Confusion matrix for the performance of the proposed method on the gender gene expression dataset

MALE	FEMALE	Classified as / True class
60	8	MALE
4	63	FEMALE
Weighted accuracy		91.13 %
Total accuracy		91.11 %
Specificity		88.23 %
Sensitivity		94.02 %