Gene and sample selection for cancer identification

Mundra Piyushkumar Arjunlal

2011

Mundra, P. A. (2011). Gene and sample selection for cancer identification. Doctoral thesis, Nanyang Technological University, Singapore.

<https://hdl.handle.net/10356/45770>

<https://doi.org/10.32657/10356/45770>

GENE AND SAMPLE SELECTION FOR CANCER IDENTIFICATION

MUNDRA PIYUSHKUMAR ARJUNLAL

School of Computer Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirement for the degree of
Doctor of Philosophy

2011

Acknowledgement

I heartily thank my advisor, Professor Jagath C. Rajapakse, for his excellent support, guidance, and technical help right from the start of my candidature. He has given me excellent research directions and helped me in developing critical scientific thinking. Without his efforts, the road to this research would not have been so enriching, fulfilling, and smooth.

I thank former research fellows in my lab, particularly Dr. Duan Kai-Bo and Dr. Minh Nguyen, for their excellent technical support and discussions to shape up my early research directions. BioInformatics Research Centre (BIRC) has given me many trustworthy friends such as Yongjin Li, Swathi, Merlin, Pooja, Zheng Bo, Divya, Liu Qian etc. I thank them for making my stay at BIRC joyful and will always cherish their memory. I thank the technical staff in BIRC, who are always ready to sort out any technical difficulty. I am also thankful to my flatmates and friends Suresh, Mahesh, Bappaditya, and Sibnath for their support.

I would like to thank School of Computer Engineering (SCE) and Nanyang Technological University for financially supporting me to pursue the doctoral studies at BIRC. Financial supports to attend various conferences by SCE, IEEE, and EvoStar are much appreciated. I express my sincere gratitude to Singapore-MIT Alliance for providing me opportunities to serve as Teaching Assistant for two consecutive years, which enhanced my skills in mathematical and statistical analysis.

I thank all the researchers who have made their datasets publically available to fight against cancer. Without them, this thesis would never have happened. I sincerely express my thanks to all others who have contributed to my research in one way or another.

Finally, my family deserves special thanks for their continuous support and love. I wouldn't have come so far without them.

Contents

Acknowledgement	i
Abstract	vii
Author's Awards and Honors	x
Author's Publications	xi
List of Figures	xii
List of Tables	xv
List of Algorithms	xix
Notations	xxi
Abbreviations	xxv
1 Introduction	1
1.1 Motivation and Objectives	2
1.2 Notations	4
1.3 Properties of Microarray Gene Expression Data	5
1.4 Classifiers	7
1.5 Gene Selection Methods	8
1.6 Filter Methods	9
1.6.1 Statistical Test based Methods	9
1.6.2 Non Parametric Test	11
1.6.3 Mutual Information based Methods	12
1.6.4 Redundancy combined with Filter Methods	13
1.7 Wrapper Methods	18
1.7.1 Evolutionary Approaches	18
1.7.2 Support Vector Machine Recursive Feature Elimination (SVM-RFE)	20
1.7.3 Other Wrapper Methods	22
1.8 Embedded Methods	23
1.9 Incorporating Biological Information	25

1.10	Contributions	27
1.11	Datasets	29
1.12	Outline	31
2	SVM-RFE with MRMR Filter	33
2.1	Motivation	33
2.2	SVM-RFE and Redundancy	36
2.3	SVM-RFE with MRMR Filter	38
2.3.1	Gene Ontology (GO) based Gene Similarity	39
2.4	Experiments and Results	40
2.4.1	Data	40
2.4.2	Preprocessing	40
2.4.3	Parameter Estimation	41
2.4.4	Implementation and Performance Evaluation	44
2.5	Results	45
2.6	Discussion	46
3	T-score with Sample Selection	55
3.1	Motivation	55
3.2	Relevant Samples	57
3.2.1	Relevant T-score	58
3.3	Support-Vectors (SV) based T-score	60
3.3.1	SV based T-score (SV t -R) with Individual Gene	61
3.3.2	SV based T-score with RFE (SV t -RFE)	61
3.4	Efficient Sample Selection Criterion	61
3.4.1	T-Score with Sample Selection (T-SS)	63
3.4.2	Margin vs. λ parameter	64
3.5	Relation of T-score with Sample Selection to SVM-RFE	65
3.6	Experiments and Results	66
3.6.1	Synthetic Data	66
3.6.2	Benchmark Datasets	67
3.6.3	Performance Evaluation	67
3.6.4	Results	68
3.6.5	Comparison of Running time	77
3.6.6	Visualization of Selected Samples	77
3.7	Discussion	79

4	Multiclass Gene Selection with Pareto Front Analysis	81
4.1	Motivation	81
4.2	Ranking with Pareto-Front Analysis	83
4.2.1	Rank Statistics	83
4.2.2	Aggregating Statistics	83
4.2.3	Siren Pitfall Problem	83
4.2.4	Multiobjective Optimization Perspective	84
4.2.5	Ranking in a given Pareto Front	86
4.2.6	F-score based Pareto Front Analysis	88
4.2.7	KW-score based Pareto Front Analysis	89
4.3	Experiments and Results	89
4.3.1	Data and Pre-processing	89
4.3.2	Performance Evaluation	90
4.3.3	Results	96
4.4	Visualization	98
4.5	Discussion	102
5	Stability, Redundancy, and Biological Analysis	105
5.1	Motivation	105
5.2	Stability Analysis	107
5.3	Redundancy Analysis	108
5.3.1	Redundancy in Two-Class Gene Selection	109
5.3.2	Redundancy in Multiclass Gene Selection	112
5.3.3	Validation of Redundancy	112
5.4	Identification of Consistent Genes	113
5.5	Results	114
5.5.1	Stability	114
5.5.2	Redundancy	115
5.5.3	Biological Validation	118
5.6	Discussion	136

6	Summary and Future Work	141
6.1	Summary	141
6.2	Future Directions	143
6.2.1	Gene Selection with Regulatory Networks	143
6.2.2	Biclustering for Simultaneous Sample and Gene Selection	143
6.2.3	Simultaneous Outlier Detection and Gene Selection	144
6.2.4	Extension of Sample Selection to Other Bioinformatics Problems	144
6.2.5	Extension to Machine Learning Problems	144
6.2.6	Analysis of Three-Dimensional Gene Expression Data	145
6.3	Conclusion	146
	References	147

Abstract

Gene-expression data gathered with microarrays play an important role in detection, classification, and understanding of many diseases including cancer. However, the numbers of samples gathered in experiments still remain in hundreds compared to the thousands of genes whose expressions are measured. One way to handle this problem is to identify relevant genes that contribute to the disease and thereafter inferring the underlying mechanisms of their functions.

This thesis focuses on identification of relevant genes, which is hindered due to several reasons. For example, relevant genes could be correlated with other genes that are biologically relevant but redundant for the classification of disease. While ranking the genes according to their relevance, it is important to consider the quality of samples as microarray samples are highly heterogeneous and multimodal in nature. This further raises an issue of stability of a gene selection method because a gene selection method should be repeatable and reproducible, giving high confidence for selected genes. For multiclass classification, sample distribution of various classes may play important role in gene selection. By considering these aspects into gene selection criteria, this research has evolved in multiple ways by introducing several novel gene ranking algorithms. To demonstrate the effectiveness of proposed approaches, the techniques are tested on cancer samples taken from various tumors such as colon, breast, lung, leukemia etc. Chapter 1 introduces the problem of gene and sample selection and a literature review on earlier work. The major contributions of the thesis are detailed in Chapters 2-5. Chapter 6 concludes the thesis with a summary and potential future work.

In Chapter 2, we propose an improved support vector machine - recursive feature elimination (SVM-RFE) technique for gene ranking by incorporating minimum redundancy maximum relevancy (MRMR) filter criteria. SVM weights are directly related to the importance of genes in sample classification, which has been demonstrated in the SVM-RFE algorithm. However, SVM-RFE does not penalize for redundant genes adequately. Therefore, we propose to combine SVM weights and the MRMR criterion

(based on mutual information) for gene ranking in a backward elimination manner. The aim is to select a minimally redundant gene subset with better performance in classification. The proposed method outperforms both the MRMR criterion and the SVM-RFE approach on benchmark cancer gene expression datasets. The method provides a general framework for compromising gene relevancy and redundancy while incorporating classifier characteristics in identifying disease samples.

A novel simultaneous sample selection and gene ranking algorithm is proposed in Chapter 3. Most of the filter based gene ranking criteria are computed using all the training samples. However, due to high variability and noise in biological samples, all the samples may not be equally important for the classification task, so the selection of samples may have a significant impact on gene ranking. We demonstrate how the T-score between gene expressions and disease categories can be decomposed into two parts, corresponding to relevant and irrelevant data samples. In the first approach, we use the SVM algorithm which automatically divides a training set into relevant points (in the form of support vectors) and irrelevant samples for classification. In the second approach, we combine the theory of SVM and modified logistic regression to select the relevant samples. A modified logistic regression is used to approximate the support vectors. We couple the proposed sample selection methods with the T-score and propose several variants for simultaneous selection of training sample points and gene ranking.

In Chapter 4, a novel gene selection method is proposed by using multi-objective optimization theory for multiclass classification. Filter methods such as F-score, Kruskal-Wallis test etc. are often used for multiclass gene selection. However, these criteria may suffer from the “siren pit-fall” problem that biases ranking towards genes belonging to easily separable classes while ignoring the genes necessary for discrimination of more difficult classes. We propose to decompose such statistics for class-specific components and apply Pareto optimization theory to optimally select genes. The performance of a Pareto-front based algorithm is evaluated on six multiclass cancer gene expression datasets and the results show significant improvement in classification accuracy. Furthermore, this method is able to remove inherent redundancy by segregating genes in to different fronts. By visualization and analysis, we also deduce several novel genes.

In Chapter 5, we perform a stability and redundancy analysis of the proposed techniques for gene selections. Even though the T-score treats each gene independently for computing the statistics, we show that sample selection can induce penalty to redundancy among genes. Using the Kuncheva index, we further demonstrate that the proposed approaches select more redundant but highly stable genes compared to SVM-RFE. With

respect to the standard T-score, the genes selected were both less redundant and stable. Furthermore, a linear weight based method is proposed to obtain consensus genes by using bootstrapped rank lists. This is a better alternative to standard frequency based approach as it includes rank information while identifying the important genes. Tests on stability, redundancy, and biological validations were also performed on multiclass gene ranking methods.

In conclusion, we have improved the SVM-RFE method for gene selection by incorporating MRMR filter criteria. In two-class filter based gene rankings, we show that the T-score can be decomposed with respect to the quality and relevance of samples. We proposed simple, yet effective algorithms to simultaneously select samples and rank genes. Such algorithms are aimed for low computational complexity and selection of highly stable genes for better performance in classification. Our research shows that indeed all the sample points are not equally important for gene selection. For multiclass gene ranking, we show that aggregation based filter ranking criteria suffer from the siren pit-fall problem. A Pareto-front based algorithm for gene ranking was proposed to consider genes relevant to weakly separable classes, which resulted in significantly improved classification performance. In summary, this research makes several novel contributions for identifying genes by computational means with microarray data, which are responsible for different types of cancers.

Author's Awards and Honors

- Best Student Paper Award, 2010 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2010, Montreal, Canada.
- Student Travel Grant, 2010 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2010, Montreal, Canada.
- Student Travel Grant, 7th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, EvoBio 2009, Tübingen, Germany.
- Nominee for Best Paper Award, 3rd Conference on Pattern Recognition in Bioinformatics, PRIB 2008, Melbourne, Australia.

Author's Publications

Journal Papers

- Jagath C. Rajapakse and Piyushkumar A. Mundra , “Multiclass Gene Selection with Pareto Front Analysis”, IEEE Transactions on Computational Biology and Bioinformatics (submitted)
- Piyushkumar A. Mundra and Jagath C. Rajapakse, “SVM-RFE with MRMR filter for gene selection”, IEEE Transactions on Nanobioscience, 2010, 9(1), 31-37.
- Piyushkumar A. Mundra and Jagath C. Rajapakse, “Gene and sample selection for cancer classification with support vectors based t -statistic”, Neurocomputing, 2010, 73(13-15), 2353-2362.

Conference Publications

- Piyushkumar A. Mundra and Jagath C. Rajapakse, “Support vectors based correlation coefficient for gene and sample selection in cancer classification ”, Proceedings 2010 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2010, 88-94
- Piyushkumar A. Mundra and Jagath C. Rajapakse, “F-score with Pareto Front Analysis for Multiclass Gene Selection”, C. Pizzuti, M.D. Ritchie, and M. Giacobini (Eds), 7th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, Lecture Notes in computer Science, LNCS 5483, 2009, 56-67
- Piyushkumar A. Mundra and Jagath C. Rajapakse, “Support Vector Based T-Score for Gene Ranking,” M. Chetty, A. Ngom, S. Ahmad (Eds), Pattern Recognition in Bioinformatics, Lecture Notes in Bioinformatics, LNBI 5265, 2008, 144-153.

- Piyushkumar A. Mundra and Jagath C. Rajapakse, “SVM-RFE with Relevancy and Redundancy Criteria for Gene Selection,” J. C. Rajapakse, B. Schmidt, G. Volkert (Eds.), Pattern Recognition in Bioinformatics, Lecture Notes in Bioinformatics, LNBI 4774, 2007, 242-252.

List of Figures

2.1	Performance of classification in various two-class datasets. Plots show average misclassification error against genes selected.	42
2.2	Comparison of class relevancy to gene redundancy ratio of gene selected by various methods in two-class datasets	43
3.1	Illustration of support vectors (relevant samples)	58
3.2	Effect of λ on the efficient sample selection criterion against margin of a sample point	65
3.3	Performance of classification against the number of selected genes in various two-class cancer datasets by various gene selection methods	69
3.4	Effect of λ parameter on classification performance of T-SS method on various two-class cancer datasets	70
3.5	Visualization of selected samples in synthetic dataset against change in λ value. The high intensity of color denotes higher selection frequency of a sample point.	78
4.1	Average misclassification error vs. number of genes on various multiclass gene expression datasets. The plots show performance of proposed approaches with their basic counterparts	93
4.2	Average misclassification error vs. number of genes on various multiclass gene expression datasets. The plots show performance of proposed approaches with their basic counterparts	94
4.3	Bar-charts indicating average and standard deviation of skewness and kurtosis for given dataset to estimate non-Gaussianity	95
4.4	Bar-charts indicating average and standard deviation of skewness and kurtosis for given dataset to estimate non-Gaussianity	96

4.5	Class-specific statistics for each class in MLL cancer dataset for top 100 genes obtained from F-score, KW-score, and PFA approaches	98
4.6	Class-specific statistics for each class in Lung cancer dataset for top 100 genes obtained from F-score, KW-score, and PFA approaches	99
4.7	Class-specific statistics for each class in 11 Tumor cancer dataset for top 100 genes obtained from F-score, KW-score, and PFA approaches	99
4.8	Class-specific statistics for each class in GCM cancer dataset for top 100 genes obtained from F-score, KW-score, and PFA approaches	100
4.9	Class-specific statistics for each class in NCI(Staunton) cancer dataset for top 100 genes obtained from F-score, KW-score, and PFA approaches	100
4.10	Class-specific statistics for each class in NCI(Ross) cancer dataset for top 100 genes obtained from F-score, KW-score, and PFA approaches	101
5.1	Stability of ranked gene lists obtained by T-score, SVM-RFE, SVt-R, SVt-RFE, and T-SS methods over two-class benchmark datasets using Kuncheva index	108
5.2	Effect of margin on parameters (a and f) of tangent approximation of sigmoid function	110
5.3	Stability of top-ranked genes obtained by F-score, KW-score, and PFA approaches over various multiclass datasets	114
5.4	Redundancy among the top-ranked genes obtained by T-score, SVM-RFE, SVt-R, SVt-RFE, and T-SS methods over various two-class datasets	116
5.5	Effect of λ on redundancy of top-ranked genes obtained by T-SS over various two-class datasets	117
5.6	Redundancy among the top-ranked genes obtained by F-score, KW-score, and PFA approaches over various multiclass datasets	118
5.7	Differentially weighted genes using proposed approaches on two-class classification datasets	122
5.8	Sample distribution using top 100 genes obtained from T-score and T-SS method on Prostate cancer dataset	128
5.9	Differentially weighted genes using proposed PFA approaches on multiclass datasets	131

List of Tables

1.1	Penalty Functions for Embedded Feature Selection	24
2.1	Number of Genes and Samples in Gene Expression Datasets	41
2.2	Performance of MRMR, SVM-RFE, and SVM-RFE with MRMR Filter methods on Gene Selection for Cancer Classification	48
2.3	Comparison of Genes Selected, Accuracies, and Significance of Difference of Performances of SVM-RFE with MRMR Filter with Other Methods .	49
2.4	Relevancy and Gene Ontology Analysis (Biological Process(BP), Molecu- lar Function(MF), and Cellular Compartment(CC)) of Selected Gene Subsets	50
2.5	Functional Descriptions of Top 10 Genes Selected from the Colon Cancer Dataset	51
2.6	Functional Descriptions of Top 10 Genes Selected from the CNS Cancer Dataset	51
2.7	Functional Descriptions of Top 10 Genes Selected from the Leukemia Can- cer Dataset	52
2.8	Functional Descriptions of Top 10 Genes Selected from the Hepato Cancer Dataset	53
2.9	Functional Descriptions of Top 10 Genes Selected from the Prostate Cancer Dataset	54
3.1	Parameters for Generating Synthetic Datasets to Evaluate Performance of T-score, SVM-RFE, SVt -R, SVt -RFE, and T-SS Methods	68
3.2	Accuracies of Classification of T-score, SVt -R, SVM-RFE, SVt -RFE, and T-SS methods on <i>synthetic</i> datasets. The average number of selected genes and sample points are indicated	71
3.3	Performance of T-score, SVM-RFE, SVt -R, SVt -RFE, and T-SS on Leukemia, Hepato, and Prostate <i>Cancer</i> Datasets	72

3.4	Performance of T-score, SVM-RFE, SVt-R, SVt-RFE, and T-SS on Colon, CNS, and Breast <i>Cancer</i> Datasets	73
3.5	Statistical Significance of improvement of T-SS method over T-score, SVt-RFE, and SVM-RFE methods	74
3.6	Comparison of Genes Selected, and Accuracies of T-SS with Other Methods	75
3.7	Comparison of Running Time(s) of T-score, SVt-R, SVM-RFE, SVt-RFE, and T-SS on Various <i>Synthetic</i> Training Datasets with Different Numbers of Features and Samples	76
3.8	Average Accuracy of T-score, SVM-RFE, SVt-R, SVt-RFE, and T-SS on Various <i>Cancer</i> Datasets using <i>k</i> NN Classifier	76
4.1	Sample Sizes of Six Multiclass Gene-Expression Datasets	90
4.2	Performance of F-score, F-PFA and Significance Analysis on Various Multiclass <i>Cancer</i> Datasets	91
4.3	Performance of KW-score, KW-PFA and Significance Analysis on Various Multiclass <i>Cancer</i> Datasets	92
5.1	Functional Description of Commonly Selected Genes from the Colon Cancer Dataset Across All Two-class Gene Ranking Methods	120
5.2	Functional Description of Exclusive Important Genes from the Colon Cancer Dataset by Proposed Methods	120
5.3	Functional Description of Commonly Selected Genes from the Leukemia Cancer Dataset Across All Two-class Gene Ranking Methods	121
5.4	Functional Description of Exclusive Important Genes from the Leukemia Cancer Dataset by Proposed Methods	121
5.5	Functional Description of Commonly Selected Genes from the Breast Cancer Dataset Across All Two-class Gene Ranking Methods	123
5.6	Functional Description of Exclusive Important Genes from the Breast Cancer Dataset by Proposed Methods	123
5.7	Functional Description of Commonly Selected Genes from the CNS Cancer Dataset Across All Two-class Gene Ranking Methods	124
5.8	Functional Description of Exclusive Important Genes from the CNS Cancer Dataset by Proposed Methods	125
5.9	Functional Description of Commonly Selected Genes from the Hepato Cancer Dataset Across All Two-class Gene Ranking Methods	125

5.10	Functional Description of Exclusive Important Genes from the Hepato Cancer Dataset by Proposed Methods	126
5.11	Functional Description of Commonly Selected Genes from the Prostate cancer dataset across T-score, SVt-RFE, and SVM-RFE methods	127
5.12	Functional Description of Exclusive Important Genes from the Prostate Cancer Dataset by T-SS Method	127
5.13	Functional Description of Commonly selected Genes from the NCI Staunton Cancer Dataset Across All Multiclass Gene Ranking Methods	128
5.14	Functional Description of Exclusive Important Genes from the NCI Staunton Cancer Dataset by PFA approaches	129
5.15	Functional Description of Exclusive Important Genes from the 11 Tumor Cancer Dataset by PFA approaches	129
5.16	Functional Description of Commonly Selected Genes from the GCM Can- cer Dataset Across All Multiclass Gene Ranking Methods	130
5.17	Functional Description of Exclusive Important Genes from the GCM Can- cer Dataset by PFA approaches	132
5.18	Functional Description of Commonly Selected Genes from the MLL Cancer Dataset Across All Multiclass Gene Ranking Methods	133
5.19	Functional Description of Commonly Selected Genes from the Lung Cancer Dataset Across All Multiclass Gene Ranking Methods	134
5.20	Functional Description of Exclusive Important Genes from the MLL Can- cer Dataset by PFA approaches	134
5.21	Functional Description of Exclusive Important Genes from the Lung Can- cer Dataset by PFA approaches	135
5.22	Functional Description of Exclusive Important Genes from the NCI Ross Cancer Dataset by PFA approaches	135
5.23	Functional Description of Commonly Selected Genes from the NCI Ross Cancer Dataset Across All Multiclass Gene Ranking Methods	136
5.24	Summary of Stability and Redundancy Analysis of T-score, SVM-RFE, SVt-R, SVt-RFE, and T-SS Methods on Various Two-class <i>Cancer</i> Datasets	137
5.25	Summary of Stability and Redundancy Analysis of F-score, KW-Score, F-PFA, and KW-PFA Methods on Various Multiclass <i>Cancer</i> Datasets .	138

List of Algorithms

1	SVM-RFE Algorithm	37
2	SVM-RFE with MRMR Filter for ranking genes	39
3	Support Vector based T-score ranking with individual gene (SV <i>t</i> -R) . . .	61
4	Support Vector Based T-score Ranking (SV <i>t</i> -RFE)	62
5	Gene ranking using T-score and sample selection	63
6	Ranking genes with Pareto Front Analysis	85
7	NSGA-II: Generating of Pareto Fronts	86
8	Crowding Distance	87

Notations

$+$	Superscript used to denote Positive class in a binary classification
$-$	Superscript used to denote negative class in a binary classification
nr	Subscript used to denote non-relevant samples
r	Subscript used to denote relevant samples
$A(t_u, t_{u'})$	Set of parent gene ontology terms shared by both t_u and $t_{u'}$
$AS_{\mathcal{P}}(i)$	Average rank score of i -th gene in a given Pareto front
α_j	Lagrange multiplier obtained from SVM for j -th sample; Sample selection criterion for j -th sample
B	Total number of subsample of a dataset
b	Index of the subsample of a dataset
β_i	Factor contributing to T-score due to non-relevant data points in i -th gene
$c_{ii', \ell}$	Correlation coefficient between gene i and i' for ℓ -th class
$cov(Y, \hat{Y})$	Covariance of corresponding columns of matrix Y and $\hat{Y}$
η	Linear SVM sensitivity parameter
D	Microarray gene-expression data matrix
$\mathcal{D}_b$	b -th subsample of the given dataset
δ	Indicator function for selection of sample

ϵ	Minimum α value to select sample points
F_i	F-score of i -th gene
G	Index set representing genes
g	Number of feature groups in generation of synthetic dataset
γ	Width of margin
Γ	Set of multiclass class labels
I	Mutual information
i	Index of a gene
i^*	gene with the least score
j	Index of a sample
K	Kernel function
k	Number of folds in cross validation evaluation
KW_i	KW-score of i -th gene
L	Total number of classes in multiclass classification
$\mathcal{L}$	Loss function
l	Set of class labels; $+1, -1$ for binary class and $1, 2, \dots, L$ for multiclass
λ	Tuning parameter in modified logistic regression
m	Total number of samples measures in a given microarray experiment
M	Subset of sample points in a given microarray dataset, $M \subset D$
μ	Mean vector of gene expressions
n	Total number of genes measures in a given microarray experiment
n_{nr}	Number of irrelevant random noise features in synthetic dataset
n_r	Number of relevant features in synthetic dataset

ν	Prescribed constant for the Gaussian in generation of synthetic dataset
Ω_i	Final weight to gene i for biological validation
ω_{ib}	Weight to each gene i by b -th ranking list
$p(t)$	Probability of finding child of GO term t
$\mathcal{P}^z$	z -th Pareto front
π	Prior probability of a given class
φ	Ratio of number of <i>relevant</i> points to total number of training sample points
$Q(S, i)$	Redundancy of i -th gene with subset S
R	Ranked gene list
R_S	Relevancy of gene subset S
ρ	Similarity between two variables
r_i	Ranking score of i -th gene
$r_{i\ell}$	Discriminatory statistic of i -th gene in ℓ -th class
$\mathcal{R}_i$	Rank of i -th gene in rank list
S	Subset of genes, $S \subset G$
$\overline{\mathcal{S}}_D$	Average stability of a gene ranking method over dataset D
$\mathcal{S}(R^b, R^{b'})$	Kuncheva index between two gene rank list R^b and $R^{b'}$
σ^2	Variance vector of gene expressions
Σ	Gene covariance matrix of a given dataset
T_i	Set of GO terms annotating i -th gene
t	Gene Ontology term
τ	Convex combination weight of SVM weight and MRMR score

u	Index of GO term
$WS_{\mathcal{P}}(i)$	Weighted sum score of i -th gene in a given Pareto front
w_0	Bias of a classifier
w_i	SVM weight of gene i
x_{ij}	Expression value of gene i in sample j
$\hat{Y}$	Actual observation matrix for multiclass classification
Y	Predicted observation matrix for multiclass classification
y_j	Class label of j -th sample
z	Index of Pareto front

Abbreviations

CFS	Correlation based Feature Selection
EA	Evolutionary Algorithm
FDR	Fisher Discriminant Ratio
F-PFA	F-score with Pareto-Front Analysis
GO	Gene Ontology
k NN	k Nearest Neighbour Classifier
LOO	Leave One Out
LR	Logistic Regression Classifier
MRMR	Minimum Redundancy Maximum Relevancy
MSVM-RFE	Multiclass or Multiple Support Vector Machine -Recursive Feature Elimination
NSGA	Nondominated Sorting Genetic Algorithm
OVA	One vs All
PFA	Pareto Front Analysis
SNR	Signal-to-Noise Ratio
SVM	Support Vector Machine
SV t -RFE	Support Vector based T-score - Recursive Feature Elimination
SVM-RFE	Support Vector Machine - Recursive Feature Elimination
T-SS	T-score with Sample Selection

Chapter 1

Introduction

In a space of a decade or more, DNA microarray technology has become one of the most vital advances in gathering data about biological systems. It has enabled us in measuring gene-expressions of thousands of genes simultaneously, giving a snapshot of the underlying state of a cell at a given time. DNA microarray technology hence provides unprecedented opportunities to unravel cellular events and reshaping life sciences. Particularly in cancer research, microarray technologies are often used to identify genes that may have important functional role in tumor growth, and to build gene regulatory networks for making inferences on biological and disease mechanisms. This will facilitate development of effective drug to combat cancer. Furthermore, identification of important genes helps to develop a computer based diagnostic prediction systems.

Parallel to such biotechnological advancements, development of computational techniques is of utter importance because systematic computational analysis of microarray gene-expression datasets is a vital and unavoidable step in extracting useful information from gene expression data. Such techniques involve classification, regression, outlier detection, differential gene identification, gene selection, clustering and many more. The classification models can be broadly divided into two categories: supervised and unsupervised. In the supervised approach, the process of classification entails placing objects into the pre-existing category, in which the classification model has been previously learnt from data accompanying the class information. On the other hand, unsupervised approaches deal with developing a set of categories (classes) while placing the samples in the classes, for example, for different stages of cancer. Supervised approaches of classification include support vector machines, artificial neural networks, logistic regression, etc. In most of bioinformatics problems, such as cancer classification, the information about the category of a particular cell line or tissue exists. Since the class information

is already available, DNA microarray analysis for cancer classification is formulated as a supervised classification problem in this research.

In cancer classification using microarray datasets, we need to diagnose whether a sample is from a cancer tissue or a benign one, or to differentiate between various kinds of cancer tumors. Microarray gene-expression datasets pose formidable challenges to computational techniques due to computational intricacies as well as biological complexities. For example, high cost of microarray experiments allows only a few numbers of samples to be gathered in an experiment. This results in datasets with a very few samples (in hundreds), compared to the thousands of genes whose gene-expressions are measured simultaneously. Microarray gene expressions form high dimensional dataset and therefore standard techniques suffer from computational instabilities and the curse of dimensionality. This makes gene selection (also referred to as the feature selection or variable selection) an important and challenging problem in microarray data analysis. From biological point of view, scientists are interested in a small set of markers that can be tested experimentally in laboratory, ultimately leading to effective drugs for a particular disease.

First, in Section 1.1, we will discuss the motivation and key objectives of this thesis in improving the gene selection methods for microarray gene-expression datasets, especially in cancer identification, when using supervised classifiers. After briefly describing notations in Section 1.2, as it is important to understand the key properties of microarray datasets in order to develop better gene selection algorithm, we will review the properties of high-dimensional microarray datasets in Section 1.3. A brief review of relevant gene selection methods and classifiers is presented in number of relevant sections. This will be followed by summarizing the major contributions of this research on both two-class as well as multiclass datasets in Section 1.10. The organization of thesis is given in Section 1.12.

1.1 Motivation and Objectives

Gene selection (also known as the feature selection or variable selection) is an important step and a widely addressed research topic in microarray data analysis. Genes from microarray datasets may be relevant or irrelevant for the classification of samples. While relevant genes are essential for the classification, irrelevant genes may degrade the performance of the classifier and are unlikely to have any biological relevance. Though much research work has already been focussed on gene selection and each method has its

own pros and cons; Many methods yet fail to identify non-redundant, reproducible, and true biomarkers to effectively classify cancerous tumors. The specific motivations and objectives of this thesis are as follows:

- (i) Most of the existing filter methods rank redundant genes/features high for classification. Moreover, the definition of relevance of feature is incomplete if the classifier characteristics are not considered [152]. Recently, many algorithms have been developed in order to minimize redundancy and maximize the relevance of top ranked genes [31], [54], [96], [115], [177]. However, classifier characteristics have not been considered for ranking genes in filter methods. On the other hand, support vector machine - recursive feature elimination (SVM-RFE) is one of the most popular wrapper methods [53] but it does not adequately penalize for redundancy among the genes. A proper compromise between gene relevancy and redundancy along with the characteristics of a classifier is essential to achieve a gene subset with minimal redundancy among the top-ranked genes while achieving good performance of classification. Therefore, we investigate how a linear combination of the classifier characteristics (such as SVM) along with a filter (such as MRMR) criteria in gene selection can improve performance of classification.

In this thesis, we intend to develop a backward elimination based gene selection method to improve both the MRMR and SVM-RFE algorithm. The objective is to have a better compromise between gene relevancy and redundancy along with classifier characteristics so that better classification performance is achieved with a fewer genes.

- (ii) In filter methods of gene ranking, the usual approach is to use all the given training samples for gene ranking. But classification methods such as SVM and boosting utilize only the boundary points to determine the discriminant hyperplane. Biologically, the multimodal nature of gene expressions poses high variability of sample points. Such variability may have a significant impact on gene ranking algorithms. In fact, sample selection is inadequately addressed in classification of microarray data.

In this thesis, we ask the fundamental question whether all training points are important for the filter based gene selection or not. Using only subset of samples, the second aim of this thesis is to develop an algorithm for simultaneous selection of samples and genes.

- (iii) Gene selection methods such as SVM-RFE suffer from low stability or reproducibility of gene ranking and have high computational complexity. Efficient sample selection techniques are needed to minimize the computational time. On the other hand, use of filter methods along with wrapper/embedded methods may improve the stability without compromising the classification performance. Therefore, we are motivated to investigate the combination of a filter method with a sample selection criterion to derive a stable and computationally inexpensive gene selection method with good performance in classification.
- (iv) Most of the feature ranking methods suffer from the siren-pitfall problem [40]: “Where it is hard to get good predictive features for some class(es), existing feature scoring mechanisms will focus on the features that are useful for other, easier classes, and will ignore the difficult classes”. Such a siren-pitfall problem is prevalent in microarray analysis but largely neglected.

Therefore, for multiclass gene ranking, the objective is to develop a filter based gene ranking method that can tackle the siren-pitfall problem. Along with good features for easily separable classes, the method should be able to identify adequate number of genes that can separate difficultly separable classes.

- (v) Generally, the gene selection methods are only evaluated for improvement in performance of classification. However, the stability and redundancy analysis are equally important to critically evaluate merits of the method. Along with biological validation using gene ontologies, we intend to critically evaluate the proposed approaches and their basic counterparts for both two-class and multiclass gene ranking methods.

1.2 Notations

In this thesis, we will resort to the following mathematical notations: Let us consider a microarray dataset $D = \{x_{ij}\}_{i=1, j=1}^{n, m}$ consisting of m gene expression samples obtained on n genes where x_{ij} is the expression of gene i in sample j . The vector $x_j = (x_{ij})_{i=1}^n$ denotes gene expressions on sample j and $x_i = (x_{ij})_{j=1}^m$ denotes the expressions of gene i across all the samples. Let the binary target class label of sample j be $y_j \in l = \{+1, -1\}$ taking values $+1$ or -1 for being a benign or cancerous, respectively and $\Gamma = \{l\}_{l=1}^L$ be the set indexing $L(> 2)$ multiclass labels.

1.3 Properties of Microarray Gene Expression Data

Analysis of high-dimensional datasets presents significant challenges to conventional pattern recognition and statistical learning techniques, mainly because of the large number of genes(n) involved, compared to the number of samples(m): i.e., $n \gg m$. Below we summarize important characteristics of microarray data, which should be considered while designing their analysis methods:

(i) *Curse of Dimensionality:*

The performance of a statistical model or approach depends on the interrelationship among the sample size, data dimensionality, model complexity, and outcome measures. However in microarray datasets with $n \gg m$, data analysis suffers from the curse of dimensionality. For example, a naive learning technique (dividing the feature space into cells and associating class label of cell type) requires the number of training samples to be an exponential function of the feature dimension [23], [55], [160]. In the context of finding a nearest neighbour in the Euclidean space, the distance may eventually converge to an equidistant with an increase of dimensionality. Hence, it is important to reduce the dimensionality of microarray datasets to avoid incorrect, incomplete, or suboptimal solutions.

(ii) *Variability and Noise:*

Microarray gene expression datasets exhibit high amount of variability and noise due to the variability of biological samples and experimental errors. Proper normalization of data is needed to reduce such variability. For example, many parametric statistical methods assume normality, which may not be applicable after certain preprocessing of data [104]. A good detailed review of microarray normalization techniques can be found in [120].

Along with biological variability, it is important to dissociate actual gene expression values from experimental noise. This noise in gene expressions can be decomposed into two parts: (a) the noise in sample preparation, and (b) the hybridization noise [153]. The sample preparation noise initiates from the process of RNA amplification, and the hybridization noise refers to the randomness in the process of RNA binding to the probes. Tu *et al.* quantitatively analyzed such noise and concluded that noise in sample preparation is generally much smaller than hybridization noise [153]. To reduce the effect of such noise, replication of microarray experiments is necessary [2], [81].

(iii) *Multimodality:*

Microarray data is influenced by the presence of multiple interrelated biological processes that can hinder identifying true relationships between gene subsets and outcomes, and create statistically correct (but biologically false) relationships. This multimodality can arise from heterogeneity of tissue and cell samples. For example, a breast tumor may contain normal and neoplastic epithelial cells, adipocytes, fibroblastic and/or many other different types of cells [22]. Gene expression data measured from such tumor will consist of signals from all these different kinds of cells. Stability of a gene ranking method is significantly affected by the small sample size of such multimodal data. Though the number of sample points is very small compared to the number of genes in a typical microarray dataset, it is important to consider the quality of samples in developing the algorithms.

(iv) *Collinearity or Redundancy:*

If genes act together in a particular pathway, the expression values of those genes may have a strong correlation. This property is useful for clustering in identifying gene functions from expression data. However, the performance of sample classification of many disease predictive algorithms deteriorates because of the presence of strong correlations among the inputs.

For additional information on the characteristics of gene expression datasets, readers are referred to [23]. A detailed review on properties of gene expression datasets and how to overcome the challenges posed by such data are presented in [160].

Gene selection is an important and vital step in microarray data analysis and in the development of relevant models. The above properties pose significant challenges to microarray data analysis in general and for gene selection in particular. Collinearity makes gene selection problem very challenging because an algorithm should select a biologically relevant gene subset with minimal redundancy. Variability and multimodality of datasets affect the stability of the gene ranking algorithm. The selection of samples may be necessary to handle the multimodal nature of data and should be considered while ranking genes. Detecting outliers and excluding them from building classifiers improve the performance of classification [134]. In multiclass datasets, some classes may lack good predictive genes and such classes are often dominated by other classes, which may result in selection of a sub-optimal gene subset. A desirable gene subset should be minimally redundant and adequately represent each class in multiclass problem such that good classification accuracy and better generalization capabilities are achieved.

Many gene ranking algorithms have been presented in the last decade to tackle specific challenges posed by properties of microarray data. At the same time, researchers have also studied the effect of such high dimensional data on various classifiers. Both the selection of the classifier as well as gene selection/ranking method are vital. Several reviews on feature selection and classification methods have recently appeared [14], [38], [128], [175]. In next section, we overview the details of the earlier works and highlight their key aspects.

1.4 Classifiers

The choice of a classifier plays an important role in performance of gene ranking algorithms. Supervised classifiers like support vector machines (SVM), artificial neural network (ANN), logistic regression (LR), decision trees, random forests (RF), k -nearest neighbours (k NN), Bayesian techniques, etc., and unsupervised techniques like K -means, expectation-minimization (EM), self-organizing feature map (SOFM), etc., have been used for microarray data analysis [33], [77], [80], [116], [138], [139]. Very recently, Freyhult *et al.* compared a cluster analysis methods, considering various gene selection techniques, missing value imputation, and normalization [43]. They concluded that a large amount of genes are indeed required to achieve good performance with clustering techniques.

Among various techniques, Pirooznia *et al.* compared different classification algorithms along with gene selection using support vector machine - recursive feature elimination (SVM-RFE), Chi-square and correlation-based feature selection (CFS) on 8 datasets [116]. The results showed that without gene selection, SVM and ANN generally perform better than other classifiers while the performance of decision tree was the worst. Further, with proper gene selection, performance of all classifier improved. A combination of gene selection with SVM-RFE and either SVM or ANN as classifier outperformed other gene selection and classifier combinations on all eight datasets. Performance of unsupervised clustering techniques was generally below the supervised techniques when class labels are available. They concluded that the choice of gene selection method, number of genes while testing and numbers of available samples influence the classification performance. Among various reports discussed in the review by Lancashire *et al.*, the performance of ANN was comparable with other classifiers [77]. However, they stressed that the regularization of ANN is extremely important to prevent over-fitting of ANN model.

Recently, RF and SVM have gained lots of attention in bioinformatics community. The advantages of RF over other methods are that it has embedded property of selection of gene and is unaffected by high dimensions and irrelevant features. This is an ensemble based training algorithm, which has a very few parameters to tune and is applicable to both two-class as well as multiclass problems. On the other hand, SVM and gene selection techniques based on SVM for microarray analysis [53], [67], [116], [139], [146] has become popular due to several reasons: the dependence of its solution on a number of samples rather than their dimensionality, inherent ability to select data points important for classification, and good generalization capabilities [160]. Moreover, high-dimensionality of the dataset does not affect its performance.

To compare the effectiveness of RF with SVM, Statnikov *et al.* recently tested both the algorithms on 22 datasets [139]. The results indicated that while the performance was comparable in two-class problems, the SVM outperformed RF in multiclass datasets when all the genes are used. With gene selection, SVM was again found to be superior to RF in multiclass datasets. These results were in contrast to the results published in [30]. However, Statnikov *et al.* showed significant potential bias in the experimental procedures adopted by Diaz-Uriarte and Andres [30], which might have resulted in opposite trends.

Particularly for multiclass datasets, Dudoit *et al.* showed that simple classifiers like k NN and diagonal linear discriminant analysis (LDA) outperformed tree based methods and Fisher LDA [33]. Lee *et al.* extended such comparisons on more datasets and on different machine learning methods [80]. The results showed that the performance of SVM was superior to any other machine learning method. The performance of k NN was found to be better than that of logistic regression and generalized partial least squares. They also recommended bagging and boosting techniques to improve the performance compared to a single tree. Considering the advantages, we resort to SVM as the base classifier for evaluation of gene selection methods in this thesis. Our findings also confirmed that SVM shows superior performance in most cancer classification problems.

1.5 Gene Selection Methods

Designing a sample classification system needs training samples from all the relevant classes. In cancer classification using gene expressions, each sample is represented by a set of genes (features) whose expressions are measured using microarrays. These genes can be classified into two categories: (i) relevant and (ii) irrelevant, with respect to the classification of samples. The relevant genes can be further classified into strongly relevant

and weakly relevant [71], [173]. The strongly relevant genes are absolutely necessary in the optimal gene subset while weakly relevant genes are conditionally important. However, weakly relevant genes may also be redundant. Highly correlated genes are generally referred as redundant genes as they do not provide any additional information for building a classifier. For the purpose of gene expression analysis, the aim of any gene selection methods should be to select maximally relevant genes while removing both redundant as well as irrelevant genes. From a purely computational view, gene selection helps in the following aspects:

- (i) reduction of effective data
- (ii) reduction of input features
- (iii) improvement of performance of classification or subsequent analysis

In general, feature selection is a well studied problem and various approaches can be classified into: (1) filter, (2) wrapper, and (3) embedded methods. Similarly, based on how the feature space is searched, methods are classified into forward or/and backward methods, sequential floating search, and optimization based methods. In this section, we will discuss various aspects of existing gene selection methods succinctly.

1.6 Filter Methods

Filter methods use a ranking criteria based on the relationship between features and class labels without taking into account the functionality of the classifier. These methods have been attractive because of their simplicity and speed. Various statistical tests and entropy based criteria are termed as filter methods which are discussed below.

1.6.1 Statistical Test based Methods

Using gene-expressions as continuous variables, statistical test based ranking criteria are mainly quantified by the mean and/or variance of gene-expressions in the respective classes. The underlying principle behind these criteria is that genes with small variances but large differences in class means are ranked high. Apart from these, statistical correlation between gene-expressions and the class labels is also used as filter criteria. To reduce the effects of outliers, non-parametric methods like Wilcoxon Rank sum test have

also been used [177].

T-score

Parametric tests such as t -test [61], F -test [31], [33] have been employed as gene filtering criteria. With simple modifications to these statistical tests, T-score and F-score have become very popular because of their ability to make statistical inference and selection of inputs. Similarly, correlation coefficients have been used in gene ranking [161]. Here we discuss the standard T-score and its relevance in classification in detail.

Suppose two classes of samples are distributed in multivariate Gaussian distributions with different means but with an equal covariance matrix. In such cases, it is sufficient to look at the log-odds ratio to determine which class the sample belongs to:

$$\log \frac{P(+|x)}{P(-|x)} = x^T \Sigma^{-1} (\mu^+ - \mu^-) - \frac{1}{2} \mu^{+T} \Sigma^{-1} \mu^+ + \frac{1}{2} \mu^{-T} \Sigma^{-1} \mu^- + \log \frac{\pi^+}{\pi^-} \quad (1.1)$$

where μ^+ and μ^- denote the mean of features for positive (+) and negative (-) classes respectively; π^+ and π^- represent the prior probabilities of respective classes; and Σ represents the covariance matrix.

Equation (1.1) can be used to discriminate different classes and discrimination is dependent only on the first term. If we assume a statistical independence between genes (i.e., $\Sigma = \sigma^2 \mathbf{I}$, where $\mathbf{I}$ is the identity matrix). From (1.1) the relevance (r_i) of gene i for discrimination is given by:

$$r_i = \frac{(\mu_i^+ - \mu_i^-)}{\sigma_i^2} \quad (1.2)$$

When σ_i^2 is replaced by pooled variance, the relevance (r_i) of gene i is given by the T-score:

$$r_i = \frac{|\mu_i^+ - \mu_i^-|}{\sqrt{\frac{m^+ \sigma_i^{+2} + m^- \sigma_i^{-2}}{m^+ + m^-}}} \quad (1.3)$$

where m^+ and m^- denote the number of samples in positive and negative classes; μ_i^+ and σ_i^{+2} represent mean and variance of expression values of gene i in the positive class. A gene, having a higher T-score between two classes, is assumed to have higher class separability [61].

The T-score, by assuming independence among genes and normality of expression values, determines the rankings and is computationally very efficient. However, the T-score performs poorer in classification in comparison to more advanced techniques mainly

because it does not take the characteristics of the classifier into consideration while ranking genes. It also uses all the given training samples for computing the T-score (r_i). The performance of such an approach may degrade due to the presence of outliers and irrelevant data points for classification.

F-score

The F-score is used for ranking genes in multi-class problems. The F -test based selection criterion, F-score, is the ratio of between-class sum of squares to within class sum of squares [33] and is given by

$$F_i = \frac{\sum_j \sum_l \delta(y_j = l)(\bar{x}_{il} - \bar{x}_i)^2}{\sum_j \sum_l \delta(y_j = l)(x_{ij} - \bar{x}_{il})^2} \quad (1.4)$$

where $\bar{x}_i$ is the average expression level of gene i across all samples, $\bar{x}_{il}$ is the mean expression level of gene i across samples belonging to l -th class, and $\delta(\cdot)$ denotes indicator function which is equal to 1 if the argument is true and 0 otherwise. The F-score value, F_i , assigns the importance of the gene i for classification.

1.6.2 Non Parametric Test

KS-test

Similar to F-score, Kruskal-Wallis (KW) test is also popular for gene selection. It is a non-parametric test to compare the distributions of different classes. The test statistic for the KW-test is given by

$$KW_i = \frac{12}{m(m+1)} \left(\sum_l m_l (\bar{\mathcal{R}}_{l.} - \bar{\mathcal{R}}_{i.})^2 \right) \quad (1.5)$$

where m_l is the number of observations in class l ; For gene i , $\bar{\mathcal{R}}_{l.}$ denotes the average rank in the l -th class and $\bar{\mathcal{R}}_{i.}$ is overall average of ranks.

Other methods are signal-to-noise ratio [49], Fisher discriminant ratio [10], symmetric divergence [103], correlation coefficient etc. All such methods compare mean and variance (or standard deviation) of gene expression values between classes. Troyanskaya *et al.* compared the performance of nonparametric methods for identifying differentially expressed genes and showed that the rank-sum method is more conservative than the nonparametric t -test based approach [150]. Based on the concept of ‘relative expression reversals’, Tan *et al.* proposed a k -top scoring method which generated simple rules for classification of gene expression data [144].

In multiclass domain, Chen *et al.* studied the effects of various test statistics, namely Brown-Forsythe test statistic, Welch test statistic, adjusted Welch test statistic, Cochran test statistic, and Kruskal-Wallis (KW) test statistic, for multiclass gene selection [19]. Xuan *et al.* proposed to use a weighted Fisher criterion (wFC) for gene selection instead of standard Fisher criterion. The wFC measures the separability of clusters and approximates the true means of Bayes prediction error [166]. It deemphasizes the contribution from well-separated classes while improving the representation of the neighbouring classes. A generalization of signal-to-noise ratio for multiclass gene selection was developed by defining gene dominant index and gene dormant index [151]. The final classification was performed using a few top ranked dominant and dormant genes.

1.6.3 Mutual Information based Methods

For some criteria, gene-expression values are discretized into V different symbolic categories for computational reasons. For example, if expression values are discretized into three categories, those categories may represent under-expressed, not expressed, and over-expressed states of genes. Though there is an information loss due to conversion of continuous valued gene-expressions into discrete feature values, it may indirectly help in handling noise. Most of the criteria using discretized genes have entropy/mutual information or distance measures as underlying principle. A few popular measures used in gene selection methods using discretized expression values in two class problems are presented as follows:

Shannon Entropy:

$$SH(x_i) = - \sum_{v=1}^V p(x_i = v|l = 1) \log_2 p(x_i = v|l = 1) + p(x_i = v|l = -1) \log_2 p(x_i = v|l = -1) \quad (1.6)$$

where $p(x_i|l)$ gives the likelihood of feature x_i given the class label l .

Kolmogorov-dependence:

$$KO(x_i) = \sum_{v=1}^V (|p(x_i = v|l = 1) - p(x_i = v|l = -1)|) p(x_i = v) \quad (1.7)$$

Kullback-Leibler Divergence

$$\text{KL}(x_i) = \sum_{v=1}^V p(x_i = v|l = 1) \log_2 \frac{p(x_i = v|l = 1)}{p(x_i = v)} + \sum_{v=1}^V p(x_i = v|l = -1) \log_2 \frac{p(x_i = v|l = -1)}{p(x_i = v)} \quad (1.8)$$

Euclidean Distance

$$\text{Eu}(x_i) = \sqrt{\sum_{v=1}^V (p(x_i = v|l = 1) - p(x_i = v|l = -1))^2} \quad (1.9)$$

where gene x_i has V different discrete values.

Apart from these, criteria like infogain, Chi-square, gini-index and many others are also extensively used for feature selection in various fields, including gene-expression data analysis.

1.6.4 Redundancy combined with Filter Methods

Filter criteria discussed in the previous section rank genes solely based on their relevance with respect to the class labels. Relevancy of a feature suggests the importance of that feature for the classification and is estimated using a criterion like mutual information or correlation coefficient. Two highly correlated features do not contribute additional information for classification and therefore carry redundant information that deteriorate the performance of the classifier [173]. Therefore, it is necessary to adequately compromise the relevancy and redundancy in order to achieve optimal performance.

Several methods have been developed, which attempt to maximize relevancy of a feature set while keeping redundancy of the subset at a minimum level in other application domains. Battiti first developed a mutual information based method to minimize redundancy in neural network learning, without formally discussing the relevancy and redundancy terms [7]. Kohavi and John introduced the concept of strong and weak feature relevancy [71]. Yu and Liu then recommended that an optimal feature set should include strongly relevant features as well as weakly relevant (but non-redundant) features and proposed a correlation based feature selection algorithm [173]. They in fact used the concept of “Markov Blanket” [72] in their correlation based relevancy-redundancy method. Very recently, Estevez *et al.* has given a good brief description of mutual information based methods which attempt to maximize the relevancy while minimizing the redundancy [36]. They also proposed a normalized mutual information based approach with a genetic algorithm based wrapper search for feature selection.

Particularly in microarray domain, correlation based feature selection [54], minimum redundancy maximum relevancy [31], [115], entropy based relevancy and redundancy [96] and Bayes error based relevancy redundancy [177] method have gained popularity for finding minimally redundant gene subset in binary classification. For multiclass gene ranking, Ooi *et al.* proposed a differential degree of prioritization (DDP) for relevancy and redundancy [111]. We discuss each of these methods briefly.

Correlation Feature Selection (CFS)

Using standard statistical correlation coefficient, Hall proposed a gene subset selection algorithm while minimizing the redundancy [54], [161]. The CFS criterion evaluates a gene subset by

$$CFS_{S,l} = \frac{|S| \bar{\rho}_{S,l}}{\sqrt{|S| + |S|(|S| - 1) \bar{\rho}_{i,i'}}}, \quad \forall i, i' \in S \quad (1.10)$$

where $CFS_{S,l}$ is a score of gene subset S , containing $|S|$ number of genes; $\bar{\rho}_{S,l}$ is the average correlation between genes in the subset S and l class labels while $\bar{\rho}_{i,i'}$ is the average pair-wise correlation coefficient for all the genes $i, i' \in S$. The denominator term represents the redundancy in the subset. This criterion may be used with forward, backward or random search of the gene subset.

Minimum Redundancy Maximum Relevancy (MRMR)

The MRMR criterion aims at selecting genes that are maximally relevant to the prediction of tissue classes while keeping the redundancy among the genes to a minimum level. The mutual information between class labels l and gene i will quantify the relevance of gene i for the classification task. The class relevancy R_s of a gene subset $S \subset \{1, 2, \dots, n\}$ is measured by

$$R_S = \frac{1}{|S|} \sum_{i \in S} I(l, i) \quad (1.11)$$

where $I(l, i) = \sum_l \sum_i p(l, i) \log \frac{p(l, i)}{p(l)p(i)}$ is the mutual information between class labels l and gene i . Similarly, the redundancy of a gene subset is determined by the mutual information among the genes. If two genes have high mutual information, they represent similar information and hence are redundant for classification. The redundancy of gene i with the other genes in the subset S is given by

$$Q_{S,i} = \frac{1}{|S|^2} \sum_{i, i' \in S, i' \neq i} I(i, i') \quad (1.12)$$

In the MRMR approach, gene ranking is either obtained by optimizing the ratio of class relevance to gene redundancy of the set S :

$$i^* = \arg \max_{i \in S} \frac{R_S}{Q_{S,i}} \quad (1.13)$$

or by optimizing the difference between relevancy and redundancy:

$$i^* = \arg \max_{i \in S} (R_S - Q_{S,i}) \quad (1.14)$$

In general, the quotient based criterion penalizes more for redundancy than the difference based criterion. As a variant, F -statistic between the genes and the class label l can be reckoned as the score of maximum class relevance for continuous value gene expressions for multiclass datasets. F -statistic generalizes t -test for two-class classification to multiclass classification problem. The minimum redundancy condition is specified by the Pearson correlation coefficient.

In MRMR criterion, the trade-off between relevancy and redundancy was not optimized. As suggested by [51], small redundancy may in fact improve the classification performance. Nevertheless, this simple quotient and difference criteria seems to perform reasonably well in most cases.

Liu *et al.* proposed an **entropy** based approach in which the normalized mutual information was used for gene relevancy and redundancy estimation [96]. The relevancy of gene i with class labels l is given by,

$$R_S = \sum_{i \in S} U(l, i) \quad (1.15)$$

where U is the normalized mutual information. The redundancy between two genes i and i' can again be computed using U . Parzen window based estimation of mutual information for continuous data has also been used [115]. However, for computational efficiency, mutual information is estimated using discretized data.

They combined the relevancy and redundancy into a binary quadratic programming based problem, which can be solved by using a heuristic simulated annealing or greedy method. However, these methods may lead to a local optimum, resulting in sub-optimal gene set with inferior prediction accuracy. To avoid that, an iterative procedure was proposed in which data was initially clustered and partitioned into different groups using K-means clustering algorithm [96]. The greedy method or simulated annealing method was then applied to select a subset of genes from each cluster such that selected genes

have low mutual relevance with respect to each other but have maximum relevance to class targets. The genes selected from each subset were then combined to obtain a single gene set. In this method, number of clusters has to be pre-determined, which may not be optimum.

Bayes Error Analysis

Zhang and Deng proposed a two stage method where the relevant genes were selected by a criterion function, like Wilcoxon test, and subsequently followed by the criterion controlling the upper bound of Bayes error to remove redundant genes [177]. They proposed that by controlling the Bayes error, it is possible to find the optimal or sub-optimal gene set for a given classification problem without designing the classifier.

In other recent developments, Hu *et al.* proposed a covariance distance score which statistically exploits the intergene correlation structure [57]. Osl *et al.* developed a demoting strategy for redundant genes as a post processing step after deriving a rank list [113]. The redundant genes are demoted based on their performance in classification. In a very recent work, the behavior of random-forest permutation based feature importance measure with redundant features was studied [108]. The analysis shows that an unconditional RF gives higher weights to redundant features.

Differential Degree of Prioritization (DDP)

Ooi *et al.* proposed the DDP approach where a parameter was tuned to determine relative importance between relevancy (with the F-score) and redundancy (with the correlation coefficient) [111]. The relevancy of gene i is measured by,

$$R_S = \frac{1}{|S|} \sum_{i \in S} F_i \quad (1.16)$$

where F_i is the F-score of gene i , measured using (1.4).

The redundancy of a gene subset is determined by using pair-wise correlation coefficients among the genes. If two genes are highly correlated, they represent similar information and hence are redundant for classification. The redundancy of gene i with the other genes in the subset S is given by

$$Q_S = \frac{1}{|S|^2} \sum_{i, i' \in S, i' \neq i} |\rho(i, i')| \quad (1.17)$$

where $\rho(i, i')$ gives correlation coefficient between two genes i and i' . To trade-off between relevancy and redundancy, following ranking criterion was proposed [111] :

$$W_S = \frac{R_S}{Q_S^{1-\tau}} \quad (1.18)$$

where $\tau \in (0, 1]$ is a DDP parameter which denotes prioritization between maximum relevancy and minimum redundancy. They also introduced another ranking criterion based on anti-redundancy. This criterion is given by,

$$W_S = (R_S)^\tau (Q'_S)^{1-\tau} \quad (1.19)$$

where, τ denotes a DDP between maximizing relevancy and maximizing anti-redundancy. Q'_S is anti-redundancy in subset S which can be computed using,

$$Q'_S = \frac{1}{|S|^2} \sum_{i, i' \in S, i' \neq i} 1 - |\rho(i, i')| \quad (1.20)$$

Numerical experiments on various datasets showed that anti-redundancy based criterion in (1.19) outperformed the redundancy based criterion in (1.18). One of the plausible reason for such result is the fact that the criterion in (1.18) penalizes more for redundancy compared to the anti-redundancy criterion.

Relief

The Relief algorithm randomly samples the data sample points from the training set and updates the relevancy weights based on the differences among selected samples and two nearest samples of same and different classes (known as near-hit and near-miss) [70]. “ReliefF” is a generalized “Relief” algorithm that handles the multiclass problems [73]. The RReliefF algorithm was proposed for continuous value target function [125]. Although popular and having good theoretical foundations, the algorithm suffers from redundancy. Recently, a gene selection algorithm by combining ReliefF and MRMR methods was proposed [180]. This is a two stage method in which ReliefF is applied to find a candidate gene set in the first stage. This filters out many unimportant genes and reduces the computational load for MRMR. In the second stage, MRMR is directly applied to reduce the redundancy and select effective and compact gene set from the given candidate gene set.

1.7 Wrapper Methods

Though filter methods are simple and computationally very efficient, they have two main disadvantages:

- (i) Though simple, their classification performances are inferior compared to more advanced methods such as wrapper methods.
- (ii) Most basic filter methods do not account for redundancy among the top-ranked genes. Hence, the number of genes, required for achieving a desired classification performance is generally high.

Wrapper methods refer to algorithms that take into account the classifier characteristics for gene selection. The gene selection algorithms conduct a search for the optimal subset of genes, incorporating the classifier characteristics as part of the evaluation function [71]. This search can be conducted in various ways such as an evolutionary optimization based approach, forward/backward selection algorithm, etc.

1.7.1 Evolutionary Approaches

Evolutionary algorithms (EA) facilitate a parallel search strategy to find asymptotic global solution for optimization by keeping a population of solutions. They have been recently applied for gene selection. Evolutionary algorithms operate on a population of candidate solutions (here, different gene subsets), select individuals from a population based on their performance (classification accuracy, number of genes, etc.), and apply various operators to obtain new individuals in a manner that the solutions will gradually move towards user-specified objectives. Genetic algorithms (GA) are the most popular among evolutionary algorithms in bioinformatics. Wahde and Szallasi reviewed such approaches for classification of gene-expression data [157].

Li *et al.* proposed a GA/ k -nearest neighbour (k NN) approach for gene subset selection with GA and then subsequent classification by using k NN [84]. Further, they studied the sensitivity and reproducibility of such approaches [85]. They recommended that 20-50 genes should be embedded in chromosome (individuals in populations) to achieve better results. Also, such an approach gave many different subsets with similar classification performance and hence, further biological analysis on such subsets was recommended. Deb and Raji-Reddy applied a multi-objective genetic algorithm for minimizing classification error (in training and test data) as well as minimizing the number of genes in

the subset [27]. Because of the multi-objective approaches, a set of solutions with varying classification performance and number of genes was obtained. In another approach, Wahde and Szallasi proposed an evolutionary approach in which gene selection as well as classification task were embedded in the genetic algorithm itself [156]. Hence, unlike other approaches [27], [84], this approach did not need classifiers.

Apart from genetic algorithms, Deutsch proposed a simulated annealing based gene selection approach and results showed that satisfactory classification performance was achieved with a very few genes [28]. A hybrid of particle swarm optimization (PSO) and GA was proposed for gene selection, resulting in better performance than using either PSO and GA alone [86]. Recently, Zhu *et al.* proposed a memetic algorithm (and genetic algorithm) based framework with wrapper-filter approach for gene selection [184]. The addition or deletion of a gene in a subset was based on the score of a univariate filter. Similarly, they also proposed a Markov-blanket embedded evolutionary strategies for gene selection [183]. Yang *et al.* proposed an interesting approach of multitask learning using genetic algorithms [168]. Using GA, they used redundant and discarded genes for classification and achieved marginal improvement in performance. Huang *et al.* developed an intelligent genetic algorithm based approach to select non-redundant genes for multiclass classification [59].

Evolutionary algorithms have also been used for multiclass gene selection [63], [93], [95], [111]. A genetic algorithm based approach with automatic determination of members for predictive gene group and optimal number of genes with better classification performance by using maximum likelihood was proposed [111]. Liu *et al.* combined EA based approach with SVM and achieved better performance [93]. Jirapech-Umpai and Aitken compared a simulated annealing with GA for multiclass gene selection and k NN based classification [63]. They also studied various gene ranking methods for initial filtering of genes and reported that there can be notable dependencies between initial filtering and final gene subset obtained by EA approaches. Very recently, Liu and Xu presented a genetic programming based one-vs.-all approach for multiclass gene selection [95]. Genetic programming (GP) has an advantage over other EA approaches as it can discover underlying data relationship. Zhu *et al.* proposed a Markov-blanket embedded memetic algorithm based method to identify both class relevant and fully relevant genes for multiclass classification, which achieved better performance of classification [185].

Though evolutionary optimization based approaches present a kind of parallelization for selection of gene subsets and heuristically lead towards optimal subsets, they suffer

from high computational cost and poor performance when the number of genes involved gets large. Most of the approaches initially filter genes based on some filtering criteria before the selection process. Such initial filtering may also have an effect on the evolution of final gene subset [63].

1.7.2 Support Vector Machine Recursive Feature Elimination (SVM-RFE)

SVM weights represent the importance of the corresponding input or feature for the classification. Using such weights and starting from full set of genes, SVM-RFE eliminates genes recursively in a backward elimination manner from gene expression data for identification of relevant genes [53], [67].

The SVM weight vector is given by

$$w = \sum_j \alpha_j y_j x_j \quad (1.21)$$

If w_i represents the component of the weight vector corresponding to the gene i , SVM-RFE uses w_i^2 as the ranking criteria [53], [67]. The smallest such weight corresponds to the least significant gene for classification. Using support vector weights and starting with all the genes, the gene which has the minimum weight for classification is removed recursively in a backward elimination manner. And then again SVM is learned and the least significant gene is removed and so on until all the genes are ranked. In what follows, we describe a few advancements of SVM-RFE.

Multiple SVM-RFE

Duan et al. presented a bootstrapped version of SVM-RFE, referred to as multiple SVM-RFE algorithm for gene selection [67]. Multiple linear SVM were trained on different sub-samples of the training data at each iteration. A statistical method was then proposed to find the ranking score by aggregation on multiple SVM weights. Using these weights with a backward elimination strategy, ranking of genes is performed. This algorithm incorporated the bootstrapping of samples with SVM-RFE. Results showed a significantly improved classification performance on benchmark datasets. A similar method was recently used for biomarker selection of cancer [1]. It was also noted that multiple bootstrapping based SVM significantly improves the stability of the SVM-RFE.

Two-Stage SVM-RFE

In standard SVM-RFE algorithm, a small number of genes are removed (in a backward elimination process) in each of the iteration. Such a strategy is presumed to result in a better gene subset. However, Tang *et al.* showed that this is not a valid assumption with a real dataset and the performance of SVM-RFE becomes unstable with the number of genes removed in each iteration [146]. To remove such instability, they proposed a two-stage approach in which irrelevant and redundant genes are removed in the first stage with a minimum loss of relevant genes. The second stage finetunes the gene rankings with elimination of a single gene in each iteration.

Fuzzy Granular SVM-RFE

Tang *et al.* proposed a three stage hybrid algorithm, which is based on principles of fuzzy clustering, granular computing and SVM-RFE [147]. First, two variance based relevance indices were used to eliminate the most irrelevant genes and rebalance the distribution of positively and negatively related genes. Second, recursive groupings of genes in different fuzzy granules were selected by using fuzzy k -means clustering. Such clustering was used to remove redundant genes. Finally, the SVM-RFE was applied to obtain the final informative gene subset.

Multiclass SVM-RFE

The standard SVM-RFE algorithm was originally proposed for two-class classification problem. Zhou and Tuck recently proposed an extension of SVM-RFE algorithm for multiclass problem [182]. They proposed to sum SVM weights obtained from different class-wise decision functions for gene ranking. The optimization of SVM objective function is possible with any one approach to multiclass problem: one-vs-all, one-vs-one, and all-at-one. These weights were then used for computation of weights of each gene. Earlier, in a similar approach, Duan *et al.* extended multiple SVM-RFE [67] algorithm for multiclass gene selection [32]. Weights obtained from different one-vs.-one or one-vs.-all SVM in backward elimination manner were combined to define the rankings of genes.

Using entropy measure over SVM weights, Furlanello *et al.* speeded up the SVM-RFE method by removing large chunk of genes in a single iteration [47]. In another algorithm, irrelevant features were first filtered and then redundancy among the features were taken into consideration by sequential forward search by using SVM and correlation coefficient [97]. Li and Yang analyzed if the success of RFE approaches is dependent on

the particular choice of a classifier or a type of recursive procedure [83]. The algorithm was tested on SVM, ridge regression(RR), and Rocchio classifier. And RR-RFE was found to perform better than the other two on three datasets. This is due to the higher penalty on redundant features in RR algorithm. It was also concluded that SVM is not as effective as RR in RFE because SVM only penalizes for redundant features of samples on the margin. However, RR is a regression technique and hence not widely employed for classification of samples using gene expression datasets.

1.7.3 Other Wrapper Methods

Several approaches suggest to reduce the dimensionality of data as a preprocessing step. Nguyen and Rocke reduced the dimensions of gene expression datasets by using partial least squares (PLS) and then performed classification using logistic regression [107]. The dimensionality reduction was better with PLS than the principal component analysis (PCA). Similarly, Fort and Lambert-Lacroix [41] proposed to use dimensionality reduction with the PLS followed by ridge based logistic regression. The dimensionality reduction coupled with filter and wrapper methods minimized the error in classification and reduce the computational time. Leung and Hung proposed a multiple filter and multiple wrapper based approach to improve the robustness of a single filter-wrapper approach [82].

A Bayesian framework was proposed to identify significant genes using the hypothesis that genes are either under-expressed or over-expressed in one class compared to other class [169]. The method requires a small number of genes to achieve reasonable performance in classification. A Bayesian stochastic approach for gene selection by using probit regression model with a generalized singular g -prior distribution model for regression coefficients was recently proposed by Yang and Song [167]

Louw and Steel proposed a kernel Fisher discriminant analysis (KFDA) based criterion to rank features recursively [101]. Niijima and Kuhara introduced maximum-margin criterion (MMC) based RFE approach for gene selection and argued that performance of MMC-RFE is comparable with SVM-RFE [109]. Yousef *et al.* proposed recursive cluster elimination (RCE) based approach for gene selection [171]. They combined K -means clustering algorithm to identify correlated gene clusters and used SVM to rank those gene clusters, by iteratively eliminating the cluster with the least SVM rank. Tibshirani *et al.* [149] developed a shrunken centroid based method for diagnosis of multiple cancer types. The shrunken centroid automatically selected genes as well as performed

classification. Tuv *et al.* proposed a tree-based ensemble approach to select a compact non-redundant features using artificial irrelevant feature [154]. By combining the eight breast cancer datasets, Kim studied the effect of sample size on performance and stability of gene signature [69]. The results suggest that more samples can significantly improve both stability and performance.

Zhou and Mao proposed a LS bound measure, which is derived by using leave-one-out (LOO) procedure of least square SVM (LS-SVM) [181]. The LS bound measure is given by

$$\sum_{j=1}^m \left(\alpha_j^0 \left[(d_{min}^j)^2 + 2/\eta \right] - 1 \right)_+ \quad (1.22)$$

where α_j^0 is a Langrangian multiplier of sample x_j , obtained from solving LS-SVM equations; d_{min}^j is the minimum distance between sample point x_j and its nearest neighbour, and η is LS-SVM training parameter. This criterion is then used in a sequential forward search (SFS) and sequential floating forward search (SFFS), and hence it is computationally less expensive than other RFE based methods. Similarly, Tang *et al.* also proposed a SFS and SFSS gene selection algorithm using leave-one-out (LOO) error of LS-SVM [145]. They also proposed a gradient descent gene selection algorithm using logistic LOO error criterion. Though the number of genes selected using such LOO error criteria like LS bound and LOO error based method is small, the classification performance is found to be inferior to other RFE based methods.

Generally, the wrapper approaches are found to perform better than filter approaches. However, they suffer from a few drawbacks:

- (i) high complexity and computationally expensive as each iteration needs a training of the classifier.
- (ii) lacks biological interpretability of selected genes.

1.8 Embedded Methods

Embedded methods refer to a set of methods/algorithms in which gene selection is embedded within the construction of the classifier. Therefore, the structure of a classifier and discriminant function play a crucial role in embedded methods of gene selection. Embedded methods include gene selection in the classification such as LASSO. A recent review on various penalized feature selection methods can be found in [102].

Table 1.1: Penalty Functions for Embedded Feature Selection

Name	Penalty Function	Remarks
Lasso	$\sum_{i=1}^n w_i $	–
Bridge	$\sum_{i=1}^n w_i ^\gamma$	–
Adaptive Lasso	$\sum_{i=1}^n \frac{1}{ w_i^* } w_i $	w_i^* is the weight adjustment for each coefficient
Elastic Net	$\sum_{i=1}^n w_i ^\gamma + (\sum_{i=1}^n (w_i^2))^\eta$	γ and η are real-valued regularization parameter

Generally a penalized method can be written as an optimization criterion:

$$\widehat{w}^* = \arg \min_w \{ \text{Obj}(w^* : D) + \eta \times \text{Pen}(w) \} \quad (1.23)$$

where $\text{Obj}(w^* : D)$ is the objective function of classification and $\text{Pen}(w)$ represents ‘Penalty’ function and it in fact controls the complexity of the model through η . Linear discriminant analysis, support vector machine, logistic regression or even a standard regression minimization function can be used as an objective function for classification. The parameter η controls the goodness-of-fit and penalty for the model complexity. When $\eta \rightarrow \infty$, the model has a few variables and when $\eta \rightarrow 0$, the model will have a high goodness-of-fit but with almost all the variables used, which will result in a model with higher complexity. In literature, various penalty schemes have been proposed [37], [46], [130], [148], [176], [186], [187], some of which are summarized in Table 1.1.

Penalty functions shown in Table 1.1 suggest that as $|w_i|$ or complexity increases, the penalty will increase. This may result in bias while estimating a large parameter. To circumvent this problem, Fan and Li proposed a smoothly clipped absolute deviation (SCAD) penalty [37] which was adopted in gene selection [176]. Based on hinge loss function, a hybrid huberized support vector machine for gene selection and classification was proposed [159]. Cawley and Talbot [17] and Liu *et al.* [98] in two separate studies developed a penalized Bayesian logistic regression method to select genes as well as to perform classification where regularization parameter was analytically determined by using a prior. De Mol *et al.* argued that a two-stage strategy for gene selection, in which

the first step consists of an elastic net to select genes [106]. The second step consists of ridge regression using the selected genes at the first step. By introducing the second step, they achieved stable genes with good performance in classification. Very recently, Cheng presented a sparse learning machine for simultaneous gene selection and classification where SVM based constraints were incorporated in $L1$ -norm linear regression [20]. Using coordinate descent algorithm, Friedman *et al.* proposed a computationally efficient algorithm to estimate parameters of generalized linear models with convex penalties [44]. The associated software can be used for linear regression, two-class logistic regression, and multinomial regression models with Lasso, ridge, or elastic net penalty. A review of mathematical properties of penalized methods can be found in [38].

Random forests and Recursive Feature Elimination

Apart from penalized methods, random forest based gene selection approach has recently gained popularity. Random forest is an ensemble of classification trees which uses both bagging (bootstrap aggregation) and random feature selection for tree building. Diaz-Uriarte and Andres [30] studied a random forest based gene importance method. For each tree, the candidate set of features is a random subset of features from a subset of samples. The feature importance measure is based on the decrease of classification accuracy when values of a feature in a node of a tree are permuted randomly.

To select genes, Diaz-Uriarte and Andres also proposed a random forest based recursive feature elimination (RF-RFE) and showed that classification performance could be improved using such approach. Starting with all genes, they built a forest and gene importance was measured using permutation based criterion. The importance of a gene is based on the decrease of classification accuracy when values of a gene-expression in a node of a tree are permuted randomly. The genes with smallest permutation based importance are discarded and a new forest was constructed. The procedure was continued till the minimum number of genes are achieved which have out-of-bag error within the one standard error of minimum error. Though successful, RF based approaches are ensemble approaches which are not within the scope of this thesis and remain as possible future work.

1.9 Incorporating Biological Information

The goal of gene selection is to identify the best subset of genes from microarray data, which can best classify the disease samples and make inferences on the underlying biological mechanism. Therefore, it is desirable to incorporate biological knowledge in the gene

selection algorithms so that selected genes becomes biologically more relevant. With the developments in Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) databases, a few techniques have been developed to integrate such biological knowledge in the process of gene ranking.

Network structures derived using hidden Markov random fields have been used to identify differentially expressed genes [163]. Information from gene networks was also integrated into statistical analysis of genomic data [162]. By integrating pathway knowledge, Johannes *et al.* very recently developed a reweighted RFE approach for risk stratification of cancer patients by using modified version of Google's page rank algorithm [64]. The hypothesis was that a gene with a low fold change should have an increased influence on a classifier if it is connected to a differentially expressed gene. Moving from a network information, Helleputte and Dupont proposed a partially supervised gene selection method, in which two genes for a given dataset were used to bias the gene rankings, resulting in good classification performance as well as stability [56].

Qi and Tang used GO annotated terms of genes in the ranking algorithm [119]. The hypothesis was that it is highly unlikely for the genes annotated with the same GO terms to correlate with the class labels by chance. In an iterative procedure, they computed the discriminative scores of GO terms by averaging the discriminative scores of genes annotated by them and selected a gene with highest ranked GO term. Qi and Tang proposed another approach with a convex combination of GO based semantic similarity with gene's expression similarity to remove redundancy in gene selection methods [118]. They also statistically assessed that the discriminative power and semantic similarity are not related with each other and the best results can only be obtained when GO based similarity is used conjunctively with some expression profile based similarity criterion.

Tai and Pan proposed a gene function group based regularized Linear discriminant analysis with in-built gene selection algorithm [143]. Basically, they regularized covariance matrix based on gene function data such that it encouraged between-class gene group independence and within-group gene correlations. This was based on the hypothesis that genes in same functional group or biological process/pathway should be correlated and should be independent of genes from other groups. They proposed a block covariance matrix, in which, genes with in a group form a block if they share same pathway (from KEGG data). The genes without a KEGG pathway annotation forms a block of size one, i.e., gene is independent. Performance result had significance in a way that if more than one gene selected from a particular group, corresponding pathway to those genes play an important role in classification and should be investigated further.

Though integration of biological knowledge in a gene ranking algorithm is highly desirable, the incompleteness of the biological information limits the development of such approaches. For example, only the top-ranked genes were used for integration of biological knowledge [118], [119]. Among those genes, many did not have GO terms. With such assumptions, potential (but unknown) biomarkers are likely to be ignored. Furthermore, many of the earlier techniques described can't use biological data. For example, one of the biomarker already found for colon cancer from gene-expression dataset may be unutilized as GO terms corresponding to it was unavailable at that time [174]. Nevertheless, integration of biological knowledge promises more robust and biologically relevant results that are interpretable to biologists.

1.10 Contributions

Gene selection is a very widely studied problem in gene-expression data analysis. Hence, we extensively reviewed the available literature on various filter, wrapper, and embedded methods for gene ranking earlier in this chapter. After considering the pros and cons of various methods and desirables for final solutions, we identified three important resolutions: (i) to minimize the redundancy among top-ranked genes without deteriorating the classification performance, (ii) to study the effects of sample point selection on filter based gene ranking, and (iii) to propose a solution to the siren-pitfall problem in multiclass gene ranking. Here are the major contributions of this thesis:

(i) SVM-RFE with MRMR filter to minimize redundancy

Our aim is to achieve a gene subset with a very good performance of classification. This requires elimination of redundant genes as they deteriorate classifier performance. To do so, in this thesis, we propose a SVM-RFE based formulation in which the ranking criterion is a linear combination of a SVM weight of genes and MRMR based relevancy to redundancy ratio. Further, we determine the trade-off between SVM weights and relevancy to redundancy ratio using a tuning parameter. Extensive experiments on five benchmark cancer datasets and using various performance metrics indicates a statistically significant improvement in classification performance with the proposed approach.

(ii) Simultaneous gene and sample selection

Filter methods, for example the T-score, use all the available samples that may not be useful for classification. We show analytically that the T-score can be

decomposed into two-parts: one corresponds to relevant and other to non-relevant samples for classification. First, we investigate support vectors to compute the T-score. As the determination of boundary point is dependent on the dimensions of data, we performed gene ranking in backward elimination manner with the T-score as the ranking criterion. The results on synthetic as well as six benchmark datasets show significant performance improvement.

(iii) Efficient Algorithm for gene and sample selection

The use of support vectors is computationally expensive and leads to relatively unstable biomarkers. We propose a criterion to determine support vectors approximately. Such criterion coupled with the T-score to rank genes are much more effective and less computationally expensive. Without compromising performance of classification, the stability of gene selection is significantly improved while computational complexity is reduced.

(iv) Multiclass gene ranking with Pareto front analysis

For multiclass gene ranking, F-score and KW-score are widely used filter criteria. However, these criteria may suffer from siren-pitfall problem and become biased towards easily separable classes, resulting in poor classification performance due to unequal representation of all classes through the top ranked genes. We propose a Pareto front based algorithm to alleviate this problem. F-score and KW-score are decomposed into individual class-specific statistics. By applying Pareto front analysis over such class specific statistics, we divide genes into various fronts and final gene ranking is obtained using ranking criteria such as the crowding distance. Experiments on six multiclass benchmark datasets show significant improvement of performance in classification.

(v) Stability, Redundancy, and Biological Analysis

The stability of technique is important for reproducibility of gene rankings. At the same time, we are also interested in non-redundant genes. Along with the performance in classification, we also analyze the gene rankings, obtained from both two-class and multiclass methods, for their stability and redundancy as evaluation criterion. We mathematically analyze the effect that sample selection has on the redundancy among genes. Instead of frequency based approach, we propose a linear weight based method to identify consensus genes from different bootstrapped trials. The consensus genes are further validated using biological knowledge available from gene ontology.

1.11 Datasets

In this thesis, we consider six widely used two-class and multiclass datasets to benchmark the proposed techniques. For two-class gene selection, the datasets used are colon [3], leukemia [49], prostate [137], CNS [117], hepato [60] and breast [164] cancer gene-expressions. For multiclass problem, we perform our analysis on GCM [123], Lung [9], MLL [4], NCI Ross [126], NCI Staunton [140], and 11 Tumor [142] dataset. These are widely used benchmark datasets to evaluate gene ranking methods, having varying numbers of classes and genes. The detail of each dataset is given as follows:

Colon

This dataset contains gene-expressions of 40 colon adenocarcinoma tumor samples and 22 normal colon tumor samples. Hence, the objective is to identify a tissue as being cancerous or noncancerous. The dataset was obtained using Affymetrix oligonucleotide array over more than 6500 human genes. However, based on the highest minimal intensity, only 2000 genes were selected for further analysis by the previous studies.

Acute Leukemia

This two-class dataset contains 25 samples of acute myeloid leukemia (AML) patients and 47 samples of acute lymphoblastic leukemia (ALL) patients. An Affymetrix system was used and a total of 7129 gene-expressions were measured. Here, AML and ALL are two types of leukemia and the task is to classify AML samples from ALL.

Prostate

Prostate tumors are among the most heterogeneous of cancers and the goal here is to classify tumor and nontumor samples. To do so, a dataset with 59 nontumor and 77 tumor samples were collected using oligonucleotide microarrays, containing probes for approximately 12600 genes and ESTs.

Hepato

Due to high intrahepatic recurrence rate, the prognosis of hepatocellular carcinoma is relatively poor. With a dataset of 20 tumors obtained from patients who suffered from early intrahepatic recurrence, and of 40 tumors without early recurrence, the task is to predict early intrahepatic recurrence or non-recurrence. The number of gene-expression

levels in this dataset is 7129.

Central Nervous System

Using Affymetrix system of probes, gene-expressions were gathered over 39 medulloblastoma survivors tumors and 21 treatment failures. The task is to identify genes that can improve identification of survivors from treatment failures. The gene-expression levels were quantified into 7129 probes.

Breast

The goal is to discriminate breast tumors based on estrogen receptor (ER) status. To do so, a dataset consists of 25 samples with ER+ status and 24 samples with ER- status have been gathered using Affymetrix platform over 7129 probes. The aim is to investigate the potential of the ER status as a possible biomarker.

GCM

This multiclass dataset has 198 tumor samples of 14 different classes. The task is to classify: leukemia (30 samples), lymphoma (22 samples), CNS (20 samples), prostate (14 samples), lung (12 samples), breast (12 samples), colorectal (12 samples), ovary (12 samples), mesothelioma (11 samples), bladder (11 samples), renal (11 samples), pancreas (11 samples), melanoma (10 samples), and uterus (10 samples) tumors. The data were derived using Affymetrix oligonucleotide microarrays (Hu6800 and Hu35KsubA) containing a total of 16063 probe sets.

11 Tumor

This 11 Tumor dataset contains expression levels of 12533 genes of 174 tumor tissues. It used an Affymetrix system to gather the gene-expressions. The dataset comprises 27 serous papillary ovarian adenocarcinomas, 26 prostate adenocarcinomas, 26 infiltrating ductal breast adenocarcinomas, 23 colorectal adenocarcinomas, 14 lung adenocarcinomas, 14 lung squamous cell carcinomas, 12 gastroesophageal adenocarcinomas, 11 clear cell carcinomas of the kidney, 8 bladder carcinomas, 7 hepatocellular carcinomas and 6 pancreatic adenocarcinomas.

NCI Ross and NCI Staunton

These datasets were part of the NCI 60 cell lines that were derived from tumors from

a variety of tissues and organs. It contains 9 renal, 9 non-small-cell lung carcinomas (NSCLC), 8 breast, 8 melanomas, 7 colon, 6 central nervous system (CNS), 6 leukemias, 6 ovarian and 2 prostate tumors. The NCI Ross dataset was derived using spotted cDNA arrays while NCI Staunton was developed using Affymetrix microarrays. Due to small class size, the prostate cell lines were excluded in the analysis.

Lung

In general, this multiclass dataset includes various lung tumors (186 samples) and normal lung tissues (17 samples). The tumor tissues belong to different types of lung cancers, 127 samples are lung adenocarcinomas, 21 samples are squamous cell lung carcinomas, 20 samples are pulmonary carcinoids, and 6 samples are small-cell lung carcinomas (SCLC). The original dataset contains 12600 gene probe sets, derived using an Affymetrix system.

MLL

This dataset contains the gene-expression profiles of three classes of leukemia: acute lymphoblastic leukemia (ALL), acute myelogenous leukemia (AML), and mixed-lineage leukemia (MLL). The number of ALL, AML, and MLL samples are 24, 28, and 20 respectively. The data were also derived using Affymetrix oligonucleotide arrays containing 12582 gene sets.

In this thesis, we have developed several two-class and multiclass gene ranking techniques. We have used the above mentioned datasets for evaluating the performance as well as analyzing their stability, and redundancy analysis. Appropriate comparisons with other state-of-the-art methods are presented in the respective chapters.

1.12 Outline

This thesis is organized as follows: Chapter 2 presents the details of SVM-RFE with MRMR filter method that combines SVM characteristics with properties of MRMR filter for gene selection. In Chapter 3, we describe a support vector based T-score computation algorithm for simultaneous selection of sample points and genes. Chapter 4 deals with a multiclass gene ranking method using Pareto front analysis in order to study the effect of multiclass data distributions on gene ranking. In Chapter 5, we analyze the stability and redundancy issues of the proposed gene ranking methods and perform biological validation of the top ranked genes. Finally, in Chapter 6, we summarize the research work done and discuss future directions of this research.

Chapter 2

SVM-RFE with MRMR Filter

Support vector machine recursive feature elimination (SVM-RFE) does not adequately take into account the redundancies among input features. In view of this, this chapter enhances the SVM-RFE method for gene selection by incorporating a minimum redundancy maximum relevancy (MRMR) filter. The aim is to find a gene subset with minimal redundancy among the top-ranked genes, which will lead to a better performance of classification. The relevancy and redundancy of the set of genes are measured by the mutual information between genes and class labels, and among the genes, respectively. The method improves the identification of cancer tissues from benign tissues on several benchmark datasets as it takes into account the redundancy among the genes during their selection by SVM-RFE. It also lays a framework for combining filter methods and wrapper methods for gene selection, such as MRMR and SVM-RFE methods. Gene Ontology analysis reveals that the method selected genes that are important in distinguishing cancerous samples and of similar functional properties.

2.1 Motivation

In microarray experiments, a large number of genes have correlated gene expression values because of their involvement in similar biological processes and molecular functions. Such genes usually tend to degrade the performance of classification and become redundant as they do not provide any additional information for classification task [173]. Moreover, a feature consisting of redundant genes do not add to the representation of the characteristics of target phenotypes [31]. Therefore, a gene selection method should provide a set of maximally relevant but minimally redundant set of genes to achieve optimal performance in classification task.

Several approaches have been developed to maximize the relevancy of a gene subset while minimizing the redundancy among the genes for disease classification [31], [96], [111], [115], [177], [180]. For example, MRMR criterion based on mutual information is developed for this purpose [31], [115]. Though a good set of genes should contain genes that are highly relevant and non-redundant, weakly relevant (but non-redundant) genes help correlation based feature selection algorithms [71], [173]. A proper trade-off between relevancy and redundancy of genes may be useful for classification [51]. MRMR method does not allow a compromise between relevancy and redundancy of genes. Greedy algorithms and simulated annealing have been attempted to determine the optimal trade-off between the relevancy and the redundancy among a set of genes [7], [96]. In another study, the relevancy-redundancy criterion was attempted in two stages [177]: using Wilcoxon test or F-test, the relevant genes set were obtained from original microarray dataset and the redundant genes were subsequently removed from the selected gene set by controlling the upper bound of Bayes error. Ooi *et al.* studied the trade-off between relevancy and redundancy in multiclass gene selection problem by introducing a data dependent tuning parameter called differential degree of prioritization [111]. Recently, ReliefF and MRMR algorithms were combined in a two stage strategy for gene selection in a large scale study [180]. First, a small subset of genes was selected using ReliefF and then, MRMR method was applied to select non-redundant genes into the subset. All the above methods are filter approaches and do not incorporate the characteristics of the classifier into the gene selection process.

Wrapper approaches usually achieve higher classification accuracies by embedding classifier characteristics in the gene selection process [30], [53], [67], [122], [127], [145], [146], [147], [157], [171], [181]. Least-square SVM based leave-one-out sequential forward selection algorithm was proposed for gene selection [145], [181]. A recursive cluster elimination based approach has been introduced to rank gene clusters in classification [171]. Using the principles of bagging and random gene selection for tree building, a random forest based approach was proposed to measure the importance of a gene in classification [30]. Wahde and Szallasi reviewed evolutionary algorithms based wrapper methods, in which gene selection is achieved by optimizing a selection criterion by using genetic operations [157]. Though the above methods seem to outperform filter methods, each of them has its own pros and cons, and most of them suffer from high computational cost and instabilities.

The support vector machine recursive feature elimination (SVM-RFE) approach to gene selection has attracted many researchers recently [53], [67], [122], [146], [147]. SVM-RFE is a multivariate gene ranking method that uses SVM weights as the ranking criterion. Earlier, Duan *et al.* introduced multiple SVM-RFE (MSVM-RFE) where the SVM was trained on multiple subsets of data and the genes were ranked using statistical analysis of gene weights in multiple runs, and demonstrated its applications in selecting genes in microarray data [67] and identifying peaks in mass spectrometry data [121]. Performance of SVM-RFE becomes unstable at some values of the gene filter-out factor, i.e., the number of genes eliminated in each of the iteration. To overcome this, two-stage SVM-RFE algorithm was proposed in which initial gene subset was selected using several MSVM-RFE models with different gene filter-out factors and in the second stage, genes were selected by eliminating one gene at each iteration [146]. A fuzzy granular SVM-RFE algorithm was recently proposed by incorporating statistical learning, fuzzy clustering, and granular computing to select highly informative genes [147].

In wrapper methods, the classifier characteristics, such as SVM weights in SVM-RFE, are incorporated into gene ranking based on their relevancy. But SVM-RFE still do not account for the redundancy among the genes adequately [83], [165]. Our aim is to combine classifier characteristics with a filter criterion in order to minimize the redundancy among selected genes, resulting in a selection of a small subset of genes and improved classification accuracy. In this Chapter, we propose an approach that incorporates mutual information based MRMR filter in SVM-RFE to minimize the redundancy among the selected genes. As seen later, our approach, referred to as SVM-RFE with MRMR Filter, improved the accuracy of classification and yielded a smaller gene sets on several benchmark cancer gene expression datasets. Experiments showed that the method outperformed MRMR and SVM-RFE methods, as well as other popular methods on most datasets. Furthermore, the selected genes were biologically relevant in discriminating cancer samples and have properties of belonging to the same pathway.

This chapter is organized as follows: Section 2.2 briefly describes redundancy analysis with SVM-RFE. A detailed description of the proposed algorithm, SVM-RFE with MRMR filter, is given in Section 2.3. Numerical experiments on five gene expression datasets are demonstrated in Section 2.4. The results and comparisons of performances of the proposed algorithm with earlier approaches are discussed in Section 2.5. Section 2.6 concludes the chapter with a discussion.

2.2 SVM-RFE and Redundancy

SVM-RFE has been successfully used to rank genes from gene expression data for cancer classification [53], [67]. It uses SVM weights to determine gene ranking and starting with whole gene set, remove the least relevant gene in a backward elimination manner. The objective function in SVM defines the margin of separation between two classes [55]. The optimal margin of separation is obtained by maximizing over weights:

$$\min_{\xi, w, w_0} \frac{1}{2} \|w\|^2 + \eta \sum_{j=1}^m \xi_j \quad (2.1)$$

$$\text{subject to } y_j (w^T x_j + b) \geq 1 - \xi_j, \forall j = 1, \dots, m \quad (2.2)$$

$$\text{and } \xi_j \geq 0, \forall j = 1, \dots, m \quad (2.3)$$

where w is the weight vector; w_0 is the bias of the linear decision function, η is SVM sensitivity parameter, and ξ_j is a slack variable for sample point j . Total number of training samples is m and set $\{x_j, y_j\}_{j=1}^m$ represents the training set. By Lagrange formulation and some manipulation, the dual problem equivalent to (2.1)-(2.3) is obtained as follows:

$$\max_{\alpha} W(\alpha) = \sum_{j=1}^m \alpha_j - \frac{1}{2} \sum_{j=1}^m \sum_{j'=1}^m \alpha_j \alpha_{j'} y_j y_{j'} K(x_j, x_{j'}) \quad (2.4)$$

$$\text{subject to } 0 \leq \alpha_j \leq \eta, \text{ for all } j = 1, \dots, m \quad (2.5)$$

$$\text{and } \sum_{j=1}^m \alpha_j y_j = 0 \quad (2.6)$$

where α_j are the parameter (Lagrange multiplier in the primal problem) obtained by training the SVM. *Support vectors* are the training samples whose corresponding α_j values are nonzero. SVM formulation depends on the support vectors to define boundaries, which are the sample points at the discriminatory boundary and within the separating margins of the two classes. It can be shown that α_j are zero for non-support vector points.

Using support vector weights and starting with all the genes, the gene which is the least significant for classification (i^*) is removed recursively in a backward elimination manner. The ranking score is computed from the coefficients of weight vector w of a linear SVM. The SVM weight vector is given by

$$w = \sum_j \alpha_j y_j x_j \quad (2.7)$$

If w_i represents the component weight vector of gene i , SVM-RFE uses w_i^2 as the ranking criteria [53], [67]. Therefore, the smallest weight corresponds to the least significant gene for classification.

Algorithm 1 SVM-RFE Algorithm

begin : Selected gene set $S = \{i\}_{i=1}^n$; Ranked gene set $R = \emptyset$
repeat
 Train linear SVM with gene set S
 Calculate the weight of each gene w_i
 Ranking score of gene i , $r_i = w_i^2$
 Select the gene with smallest ranking score $i^* = \arg \min_i(r)$
 Update $R = [R; i^*]$; $S = S \setminus \{i^*\}$
until all genes are ranked
end : output R

The objective function of SVM is given by $J = (1/2) \|w\|^2$ neglecting the misclassification. The approximate change in objective function due to the removal of a feature by expanding the objective function in Taylor series to second order is given by

$$\Delta J(i) = \frac{\partial J}{\partial w_i} \Delta w_i + \frac{\partial^2 J}{\partial^2 w_i} (\Delta w_i)^2 \quad (2.8)$$

As $J = (1/2) \|w\|^2$ and first order term equals zero at the optimum, (2.8) is written as

$$\Delta J(i) = (\Delta w_i)^2 \quad (2.9)$$

where Δw_i corresponds to removal of i -th gene.

In other words, if w_i represents the component of SVM weight vector corresponding to gene i , SVM-RFE uses w_i^2 as the ranking criteria. The smallest weight corresponds to the least significant gene for classification. This is explained by the optimal brain damage (OBD) algorithm [79].

As seen, the whole process of gene removal does not take gene redundancy into consideration. The genes are removed based on the minimal change of the objective function. In a more direct way, SVM weight for n genes can be computed by (2.7), that is,

$$[w_1, w_2, \dots, w_n] = [\alpha_1 y_1, \alpha_2 y_2, \dots, \alpha_m y_m] \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nm} \end{bmatrix}^T \quad (2.10)$$

For example, consider the weights of gene i and i' :

$$\begin{aligned} w_i &= \alpha_1 y_1 x_{i1} + \alpha_2 y_2 x_{i2} + \dots + \alpha_m y_m x_{im} \\ w_{i'} &= \alpha_1 y_1 x_{i'1} + \alpha_2 y_2 x_{i'2} + \dots + \alpha_m y_m x_{i'm} \end{aligned} \quad (2.11)$$

From the above equations, it is clear that if gene i and i' are correlated (i.e., redundant), their SVM weights will be correlated too. Thus, the SVM-RFE algorithm will be unable to handle the redundancy among genes. Moreover, Li and Yang analyzed the Lagrange multipliers of SVM for between gene redundancy analysis [83]. Using the loss function of modified logistic regression to approximate SVM [178], and further simplifying the analysis with sigmoid function and tangent of sigmoidal function, SVM models were shown to penalize for the redundancy among genes for only those samples whose margin is exactly one, $y_j w^T x_j = 1$. That is, only sample points at the boundary of the margin account for the redundancies among genes. Other samples could be useful for exploiting the redundancy among genes.

2.3 SVM-RFE with MRMR Filter

The MRMR filter when used alone may not yield the optimal subset of genes because the classifier performs independently and is not involved in the selection of genes. On the other hand, SVM-RFE does not take into account the redundancy among genes except for those at the boundary of the margin. Our aim is to remove redundancy among relevant genes by introducing a MRMR filter to the gene selection in SVM-RFE. As seen later, this improves classification accuracy because redundant and relevant genes compromise to classify cancer samples.

In the SVM-RFE with MRMR filter approach, genes are ranked by a convex combination of the relevancy given by SVM weights and the MRMR criterion. For i th gene, the ranking measure r_i is given by

$$r_i = \tau |w_i| + (1 - \tau) \frac{R_{S,i}}{Q_{S,i}} \quad (2.12)$$

where w_i denote SVM weight and $\frac{R_{S,i}}{Q_{S,i}}$ contributes the MRMR criterion of gene i in the subset S . The parameter $\tau \in [0, 1]$ determines the trade-off between SVM ranking and MRMR ranking. The relevancy $R_{S,i}$ of gene i in the set S on classification is given by

$$R_{S,i} = \frac{1}{|S|} \sum_l I(l, i), \quad \forall i \in S \quad (2.13)$$

To facilitate backward selection, a gene-wise MRMR criterion is used for ranking in the present approach. Also, we use $|w_i|$ as the gene relevancy measure from SVM to better compromise with the redundancy of genes.

Algorithm 2 illustrates an iteration of SVM-RFE with MRMR Filter method of ranking genes: the least important gene at a time is identified after ranking the genes in the subset S . In each iteration, one (or a set) of least significant genes is removed and the remaining genes train SVM and go through the removal process again until converges. The iteration stops when removal of any more genes does not improve the performance of classification.

Algorithm 2 SVM-RFE with MRMR Filter for ranking genes

```

begin
Set  $\tau$ 
Ranked set of genes,  $R = \emptyset$ 
Given set of genes,  $S \subset G$ 
repeat
  Train linear SVM with gene set  $S$ 
  Calculate the weight of each gene  $w_i$ 
  for each gene  $i \in S$  do
    Compute  $R_{S,i}$  and  $Q_{S,i}$ 
    Compute  $r_i$  using (2.12)
  end for
  Select the gene with smallest ranking score,  $i^* = \arg \min\{r_i\}$ 
  Update  $R = [R; i^*]$ ;  $S = S \setminus \{i^*\}$  ;
until all genes are ranked
end : output  $R$ 

```

2.3.1 Gene Ontology (GO) based Gene Similarity

Gene Ontology (GO) [5] provides a structured and controlled vocabulary to annotate gene and gene products of an organism. Functions of genes are distributed in terms of three hierarchies (or taxonomies): molecular function (MF), biological process (BP), and cellular compartment (CC). The inter-relationships among the terms are represented in a directed acyclic graphs (DAG) where each node of the graph represents a term and an edge represents *is-a* or *part-of* relation between the two terms.

GO helps to evaluate functional similarities among a set of genes [67], [158] as there exists a direct relationship between semantic similarity of gene pairs and their structural and sequence based similarity [99], [129]. The semantic similarity represents the biological

similarity among a set of genes. In order to find biological relevancy and redundancy of genes selected by our method, we investigate GO based biological semantics.

The concept of semantic similarity among genes defines how GO terms describing different genes relate to one another. We use Lin's semantic similarity measure that estimates similarity on the basis of parent commonality of two query terms, incorporating the information content of the queries [89]. For each term t , let $p(t)$ be the probability of finding a child of t in the annotation database. If t_u and $t_{u'}$ are two query terms and $A(t_u, t_{u'})$ represents the set of parent terms shared by both t_u and $t_{u'}$, then the similarity is given by

$$\rho(t_u, t_{u'}) = \frac{2 \times \max_{t \in A(t_u, t_{u'})} \log(p(t))}{\log(p(t_u)) + \log(p(t_{u'}))} \quad (2.14)$$

To compute semantic similarity between two genes i and i' , let T_i and $T_{i'}$ be the sets of terms annotating the two genes, respectively. The semantic similarity between the two genes is defined as the average inter-set similarity of terms in sets T_i and $T_{i'}$. The biological similarity of the genes in the set S is given by

$$\rho(S) = \frac{2}{|S|(|S| - 1)} \sum_{i, i' \in S, i \neq i'} \frac{1}{|T_i| |T_{i'}|} \sum_{t_u \in T_i, t_{u'} \in T_{i'}} \rho(t_u, t_{u'}) \quad (2.15)$$

The similarity measures how selected genes are close in their functions or work in the same pathway. However, this does not take into account the multimodality of the selected gene subset.

2.4 Experiments and Results

2.4.1 Data

We evaluated the performance of the proposed method on five microarray gene expression cancer datasets, namely, colon [3], leukemia [49], hepato [60], prostate [137], and CNS [117]. The numbers of samples and genes in the datasets are given in Table 2.1. These datasets have been widely used to benchmark gene selection algorithms. Where there is no separate testing set is available, the original dataset was randomly divided into separate training and testing datasets.

2.4.2 Preprocessing

The gene expressions were normalized to zero mean and unit variance, based on gene expressions of a particular sample. Gene expression values were directly used as input

Table 2.1: Number of Genes and Samples in Gene Expression Datasets

Dataset	Genes	# Training samples	# Testing samples
Colon	2000	40	22
Leukemia	7129	38	34
Hepato	7129	33	27
CNS	7129	40	20
Prostate	12600	102	34

features to SVM but were discretized in order to compute mutual information to evaluate MRMR values. The discretization of expression values were done to represent over-, no- and under-expression of the gene. The discretization $\tilde{x}$ of the gene expression x was performed as follows:

$$\tilde{x} = \begin{cases} +2 & \text{if } x > \mu + \sigma/2 \\ -2 & \text{if } x < \mu - \sigma/2 \\ 0 & \text{otherwise} \end{cases} \quad (2.16)$$

where the discretized value $\tilde{x} = +2$, 0 , or -2 , represents an over-expression, no expression, and an under-expression of the gene, respectively. μ and σ denote the mean and standard deviation of gene expressions of the sample.

2.4.3 Parameter Estimation

A linear SVM model was trained with the selected subset of genes S in each iteration of Algorithm 2. The sensitivity (η) of linear SVM was estimated from the set of $\{2^{-20}, 2^{-19}, \dots, 2^0, \dots, 2^{15}\}$, giving the maximum Matthew's Correlation Coefficient (MCC^1) on 10-fold cross-validation on training data. MCC was chosen because of small sample sizes and the imbalances of labels in most datasets. The value of η was used with linear SVM model to select genes in SVM-RFE and SVM-RFE with MRMR Filter methods.

The τ value for the SVM-RFE with Filter method was then empirically determined from a set $\{0.2, 0.4, 0.5, 0.6, 0.8\}$ based on the best 10-fold cross-validation accuracy.

$$^1MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

(a) Colon

(b) Leukemia

(c) Hepato

(d) CNS

(e) Prostate

Figure 2.1: Performance of classification in various two-class datasets. Plots show average misclassification error against genes selected.

(a) Colon

(b) Leukemia

(c) Hepato

(d) CNS

(e) Prostate

Figure 2.2: Comparison of class relevancy to gene redundancy ratio of gene selected by various methods in two-class datasets

2.4.4 Implementation and Performance Evaluation

Starting with all the genes whose expression values were measured, as given in Algorithm 2, genes were gradually removed in order to select the most relevant and redundant subset of genes. To expediate the gene selection with both SVM-RFE and SVM-RFE with MRMR Filter methods, we use following strategy:

$$\text{Number of genes removed} = \begin{cases} 100 & \text{if } n' \geq 10000 \\ 10 & \text{if } 1000 \leq n' < 10000 \\ 1 & n' < 1000 \end{cases} \quad (2.17)$$

where n' is the number of genes in the gene set S .

After parameter estimation with 10-fold cross-validation, genes were selected using the optimal parameters realised with training data. Using the selected genes, the performances of the methods were evaluated on test samples. Because of the small number of test samples and imbalances of training and testing sets, the test errors were evaluated on bootstrapped samples: training and testing datasets were merged and all samples were then partitioned for training and testing sets for 100 times. The performance measures such as accuracy, sensitivity, and specificity were averaged over 100 trials. Genes were selected from the gene subset giving the least average test error. MCC was used to evaluate the performances on test data. We performed pair-wise t -test to determine whether there is statistically significant difference in performances of the present method over the other methods.

The area under receiver operating characteristics (ROC) curve (or AUC) was also used as performance evaluation metric, particularly considering the skewed class distributions and unequal misclassification cost [39]. ROC curves are 2-dimensional graphs in which *true positive* rates were plotted against *false positive* rates. For a classifier, such as SVM, which just indicate the class labels as output, the unbiased AUC is just the average of sensitivity and specificity [15].

Implementation of MRMR in the MATLAB toolbox from the original authors was used for experiments [31]. The classification accuracy of the gene subset was evaluated using SVM. We have also compared our results with the SVM-RFE method. For gene selection and testing with SVM, we used LIBSVM - version 2.84 software [18]. For implementation of LS-bound SVM method, the original source codes (available as supplementary material with [181]) were utilized. The performance of classification with the proposed technique was also compared with Random Forest - Recursive Feature Elimination (RF-RFE) technique [30]. As a standard practice of random forest, the accuracy

is reported with 100 times out-of-bag criterion. For GO analysis, *GOSim* [45] package developed in *R* was used to compute the Lin's similarity measures. Gene Entrez Ids were obtained from NCBI database. The genes without Human Genome Organization (HUGO) symbols were excluded from GO analysis.

2.5 Results

Table 2.2 shows a comparison of classification performances of MRMR, SVM-RFE, and SVM-RFE with MRMR filter methods on five datasets. Error convergence with the selected number of genes are shown in Figure (2.1). As seen, the performance of SVM-RFE with MRMR Filter was significantly better as indicated by the most of the performance measures on leukemia, hepato, CNS and prostate datasets, compared to MRMR method. The performance is significantly improved than SVM-RFE in prostate cancer dataset. SVM with MRMR Filter showed significant improvement of sensitivity against SVM-RFE method on colon data. In all datasets, the number of genes required to achieve better classification is less compared to the SVM-RFE.

In addition to MRMR and SVM-RFE methods, we compared our method with several other widely used methods: Bayes Error based filter with k NN and SVM [177], t -test with Fisher Discriminant Analysis [75], Correlation based feature selection (CFS) [161], RF-RFE [30] and Least Square-bound with SVM [181].

Table 2.3 shows a comparison of performances of the present method with other existing methods for gene selection. As seen, SVM-RFE with MRMR Filter achieved better classification performance on colon, leukemia, CNS, and prostate cancer datasets and equal performance for Hepato on Bayes + SVM. The classification performance is statistically improved compared to RF-RFE in prostate and CNS dataset and is comparable in other datasets. The numbers of genes selected were however higher than the other methods.

Table 2.4 shows mutual information based relevancies of selected gene subsets with different methods on five datasets. For all five datasets, the present method selected comparable or slightly more relevant genes compared to other methods, but the differences were statistically insignificant. Figure (2.2) depict the change in mutual information based class relevancy and redundancy with gene subset. To evaluate biological similarity among genes, GO based similarity analysis with Lin's similarity measure were performed on the selected gene sets. Table 2.4 also gives gene ontology based similarities for the best gene subsets selected by different methods. There were no statistically significant

differences among GO measures of genes selected by different methods. This indicates that the present method selected genes that are functionally similar to those selected by other methods. Except for the MRMR selected genes on prostate cancer data, the present method always selected a smaller subset of genes, yet functionally similar and coherent to those sets selected by other methods.

Tables 2.5-2.9 represent the gene name, description and pathway information about top 10 genes obtained by the proposed method. As seen in Tables, many top ranked genes are involved in cytokine receptor pathway, epithelial signalling pathway, cell adhesion molecules, p53 signalling pathway, and MAPK pathway. These pathways are generally considered to be involved in cancer [29], [110].

For example, mitogen-activated protein kinase (MAPK) pathways are conserved three-tier kinase modules. Such modules link extracellular signals to the cellular machinery in order to control cell growth, proliferation, differentiation, migration, and apoptosis. As mentioned in [29], cancerous mutations in MAPK pathways mostly affects Ras and B-Raf in the extracellular signal-regulated kinase pathway while stress-activated pathways, such as Jun N-terminal kinase and p38, largely seem to counteract malignant transformation. The balance and integration between these signals is extremely important for success of drug therapy. Similarly, a few reports indicate that the changes in the adhesion properties of neoplastic cells play a pivotal role in the development of recurrent, invasive, and distant metastasis and hence progression of cancer [110].

2.6 Discussion

The redundancy among gene expressions in microarray data hinders identification of cancerous tissues from benign tissues. In order to remove the redundancy or collinearity among a set of genes, we introduced MRMR filter in the ranking of genes by the SVM-RFE method. The efficacy of embedding of MRMR filter in SVM-RFE was evidenced with the improved classification performance on benchmark datasets. The introduction of MRMR criterion enhanced the gene selection based on SVM weights. On most benchmark datasets, the proposed approach outperformed other methods because it was able to reduce the redundancy among the genes selected.

Furthermore, the present method selected a fewer genes compared to MRMR and SVM-RFE methods on most datasets. We also observed improvements of standard deviations of performance measures (accuracy, sensitivity, specificity, MCC, and AUC) over

bootstrap samples in the experiments. This indicates that the present method is more stable than the other gene selection methods.

The proposed algorithm is computationally more expensive than SVM-RFE or MRMR methods. By removing a set of genes instead of one gene at a time, some improvement on computational speed was made. The use of discretized data for computation of mutual information leads to a loss of information but could improve the robustness to noise. But other methods that quantize the data, based on optimal performance could further improve the performance of the present method. Parameter τ and the sensitivity η of the SVM were separately and empirically determined. Investigation of other methods that could estimate both parameters together may yield better estimates of parameters.

It is not advisable to determine relevant and redundant genes independent of the classifier, for example using only MRMR criterion. On the other hand, a simple filter might improve the selection of genes if it understands how the filtered genes are used by the classifier. We proposed to incorporate MRMR criterion into the ranking scheme of SVM-RFE so that the filter takes care of the redundancies among genes.

Functional similarity exists among the genes involved in a specific biological process but usually hinders computational methods of tissue classification. The GO analysis shows that functional similarities among the genes in terms of biological pathways, molecular functions, and colocalization of the top ranked genes by the present method were comparable to those selected by MRMR or SVM-RFE methods. However, SVM-RFE with MRMR Filter selects a fewer number of genes and attempts to eliminate redundant genes some of which may have biological functions important to cancer. Therefore, once a good classification is achieved, the original gene set needs to be scanned to find genes that are biologically similar to those selected. Ideally, one could integrate the biological measures of relevancy and redundancy, using information from GO, into the gene selection process so that the results are directed by biological underpinnings. However, this is not feasible at this time as many genes measured by microarrays do not have entries in the GO database for example.

This work can be extended to investigate into other filter criteria. Similarly important is to investigate the enhancement of introducing filters, such as MRMR, into the recent development of SVM-RFE, such as MSVM-RFE, two-stage SVM-RFE, and Fuzzy-Granular SVM-RFE. However, the scope of this work is to demonstrate how the incorporation of redundancy removal, such as by MRMR filter, into a classifier-based gene selection method could improve the process of gene selection.

Table 2.2: Performance of MRMR, SVM-RFE, and SVM-RFE with MRMR Filter methods on Gene Selection for Cancer Classification

Dataset	Measure	MRMR	SVM-RFE	SVM-RFE with MRMR filter	Significance of Difference with	
					MRMR	SVM-RFE
Colon	# Genes	90	90	78		
	Accuracy	91.00 $\pm$ 5.85	91.00 $\pm$ 5.17	91.68 $\pm$ 5.13	...	...
	Sensitivity	87.38 $\pm$ 11.72	86.75 $\pm$ 10.18	89.50 $\pm$ 9.18	...	$p < 0.05$
	Specificity	93.07 $\pm$ 6.05	93.43 $\pm$ 5.53	92.93 $\pm$ 5.79	...	...
	MCC	0.81 $\pm$ 0.13	0.81 $\pm$ 0.11	0.83 $\pm$ 0.11	...	...
	AUC	0.90 $\pm$ 0.07	0.90 $\pm$ 0.06	0.91 $\pm$ 0.06	...	...
Leukemia	# Genes	91	47	37		
	Accuracy	97.18 $\pm$ 2.86	97.88 $\pm$ 2.07	98.35 $\pm$ 1.93	$p < 0.01$	$p < 0.05$
	Sensitivity	93.36 $\pm$ 6.98	95.00 $\pm$ 5.13	96.36 $\pm$ 4.37	$p < 0.001$	$p < 0.05$
	Specificity	99.85 $\pm$ 0.86	99.90 $\pm$ 0.70	99.75 $\pm$ 1.31	...	...
	MCC	0.94 $\pm$ 0.06	0.96 $\pm$ 0.04	0.97 $\pm$ 0.04	$p < 0.01$	$p < 0.05$
	AUC	0.97 $\pm$ 0.04	0.97 $\pm$ 0.03	0.98 $\pm$ 0.02	$p < 0.001$	...
Hepato	# Genes	63	79	11		
	Accuracy	84.78 $\pm$ 5.38	89.19 $\pm$ 5.88	88.15 $\pm$ 4.59	$p < 0.001$	...
	Sensitivity	73.00 $\pm$ 19.96	80.25 $\pm$ 12.20	88.12 $\pm$ 10.26	$p < 0.001$	$p < 0.001$
	Specificity	89.74 $\pm$ 6.91	92.95 $\pm$ 6.95	88.16 $\pm$ 5.72	...	$p < 0.001$
	MCC	0.64 $\pm$ 0.13	0.75 $\pm$ 0.13	0.74 $\pm$ 0.10	$p < 0.001$	...
	AUC	0.81 $\pm$ 0.07	0.87 $\pm$ 0.07	0.88 $\pm$ 0.05	$p < 0.001$	...
Prostate	# Genes	9	85	10		
	Accuracy	96.47 $\pm$ 3.07	96.24 $\pm$ 3.37	98.29 $\pm$ 2.30	$p < 0.001$	$p < 0.001$
	Sensitivity	96.08 $\pm$ 3.68	95.88 $\pm$ 4.08	98.16 $\pm$ 2.69	$p < 0.001$	$p < 0.001$
	Specificity	97.56 $\pm$ 5.6	97.22 $\pm$ 5.56	98.67 $\pm$ 3.96	...	$p < 0.05$
	MCC	0.92 $\pm$ 0.07	0.91 $\pm$ 0.08	0.96 $\pm$ 0.05	$p < 0.001$	$p < 0.001$
	AUC	0.97 $\pm$ 0.03	0.97 $\pm$ 0.03	0.98 $\pm$ 0.02	$p < 0.001$	$p < 0.001$
CNS	# Genes	37	89	77		
	Accuracy	88.15 $\pm$ 6.84	92.35 $\pm$ 5.79	93.00 $\pm$ 4.82	$p < 0.001$	...
	Sensitivity	81.86 $\pm$ 15.31	81.57 $\pm$ 14.68	86.00 $\pm$ 12.18	$p < 0.001$	$p < 0.01$
	Specificity	91.54 $\pm$ 7.21	98.15 $\pm$ 7.21	96.77 $\pm$ 4.91	...	$p < 0.05$
	MCC	0.75 $\pm$ 0.15	0.83 $\pm$ 0.13	0.85 $\pm$ 0.10	$p < 0.001$	...
	AUC	0.87 $\pm$ 0.08	0.90 $\pm$ 0.08	0.91 $\pm$ 0.06	$p < 0.001$	...

Table 2.3: Comparison of Genes Selected, Accuracies, and Significance of Difference of Performances of SVM-RFE with MRMR Filter with Other Methods

Method	Measure	Colon	Leukemia	Hepato	Prostate	CNS
Bayes	# Genes	10	7	9	16	11
+ KNN	Accuracy	88.23	95.71	86.00	97.29	77.6
	Significance	$p < 0.001$	$p < 0.001$	$p < 0.01$	$p < 0.01$	$p < 0.01$
Bayes	# Genes	10	8	10	3	42
+ SVM	Accuracy	86.27	97.12	88.96	95.41	81.75
	Significance	$p < 0.001$	$p < 0.001$	...	$p < 0.001$	$p < 0.001$
t-test	# Genes	3	2	3	11	8
+ FDA	Accuracy	82.68	90.68	73.00	87.29	76.30
	Significance	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
LS-Bound	# Genes	18	5	4	4	5
+ SVM	Accuracy	85.23	94.74	84.74	96.06	77.70
	Significance	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
CFS	# Genes	4	28	76	98	100
	Accuracy	87.41	98.26	87.82	94.76	87.25
	Significance	...	...	...	$p < 0.001$	$p < 0.001$
RF-RFE	# Genes	3	6	3	5	5
	Accuracy	88.71	98.62	90.00	95.59	88.34
	Significance	...	...	...	$p < 0.001$	$p < 0.001$
SVM-RFE	# Genes	78	37	11	10	77
with MRMR	Accuracy	91.68	98.35	88.12	98.29	93.00
	Significance	...	...	...	...	...

Table 2.4: Relevancy and Gene Ontology Analysis (Biological Process(BP), Molecular Function(MF), and Cellular Compartment(CC)) of Selected Gene Subsets

Dataset	Method	Selected Genes	Relevancy	BP	MF	CC
Colon	MRMR	90	0.27	0.40	0.38	0.55
	SVM-RFE	90	0.21	0.41	0.38	0.56
	SVM-RFE with MRMR	78	0.28	0.41	0.38	0.56
Leukemia	MRMR	91	0.39	0.42	0.37	0.59
	SVM-RFE	47	0.38	0.42	0.38	0.60
	SVM-RFE with MRMR	37	0.40	0.42	0.38	0.61
Hepato	MRMR	63	0.28	0.40	0.36	0.56
	SVM-RFE	79	0.20	0.41	0.35	0.58
	SVM-RFE with MRMR	11	0.28	0.48	0.35	0.63
Prostate	MRMR	9	0.28	0.41	0.32	0.57
	SVM-RFE	85	0.22	0.41	0.37	0.57
	SVM-RFE with MRMR	10	0.30	0.50	0.36	0.61
CNS	MRMR	37	0.19	0.40	0.34	0.57
	SVM-RFE	89	0.10	0.44	0.36	0.56
	SVM-RFE with MRMR	77	0.17	0.43	0.36	0.57

Table 2.5: Functional Descriptions of Top 10 Genes Selected from the Colon Cancer Dataset

Gene Name	Description	Pathway
IL8	Interleukin 8	Bladder cancer; Cytokine-cytokine receptor interaction; Toll-like receptor signaling pathway; Epithelial cell signaling in Helicobacter pylori infection
DARS	Aspartyl-tRNA synthetase	Alanine and aspartate metabolism ; Aminoacyl-tRNA biosynthesis
Unknown	...	...
CDH3	Cadherin 3, type 1, P-cadherin (placental)	Cell adhesion molecules (CAMs)
HSPD1	Heat shock 60kDa protein 1 (chaperonin)	Type I diabetes mellitus
Unknown	...	...
Unknown	...	...
TSPAN1	Tetraspanin 1	...
MT1G	Metallothionein 1G	...
Unknown	...	...

Table 2.6: Functional Descriptions of Top 10 Genes Selected from the CNS Cancer Dataset

Gene Name	Description	Pathway
LOC442747	Hypothetical LOC442747	...
INA	Internexin neuronal intermediate filament protein, alpha	...
PCOLCE	Procollagen C-endopeptidase enhancer	...
Unknown	...	...
PYGL	Phosphorylase, glycogen; liver (Hers disease, glycogen storage disease type VI)	Starch and sucrose metabolism; Insulin signaling pathway
POU4F1	POU class 4 homeobox 1	Tight junction
NID1	Nidogen 1	...
Unknown	...	...
Unknown	...	...
MMP16	Matrix metalloproteinase 16 (membrane-inserted)	...

Table 2.7: Functional Descriptions of Top 10 Genes Selected from the Leukemia Cancer Dataset

Gene Name	Description	Pathway
YAF2	YY1 associated factor 2	...
BMP7	Bone morphogenetic protein 7 (osteogenic protein 1)	Cytokine-cytokine receptor interaction; Hedgehog signaling pathway; TGF-beta signaling pathway
MAGEA5	Melanoma antigen family A, 5	...
ATM	Ataxia telangiectasia mutated	Cell cycle; p53 signaling pathway; Apoptosis
SUMO1	SMT3 suppressor of mif two 3 homolog 1 (S. cerevisiae)	...
MEF2C	Myocyte enhancer factor 2C	MAPK signaling pathway
SFRS3	Splicing factor, arginine/serine-rich 3	...
PTPRZ1	Protein tyrosine phosphatase, receptor-type, Z polypeptide 1	Epithelial cell signaling
MLLT3	Myeloid/lymphoid or mixed-lineage leukemia (trithorax homolog, Drosophila); translocated to, 3	...
APBB1	Amyloid beta (A4) precursor protein-binding, family B, member 1 (Fe65)	Alzheimer's disease

Table 2.8: Functional Descriptions of Top 10 Genes Selected from the Hepato Cancer Dataset

Gene Name	Description	Pathway
OAS2	2'-5'-oligoadenylate synthetase 2, 69/71kDa	...
Unknown	...	...
DDX17	DEAD (Asp-Glu-Ala-Asp) box polypeptide 17	...
TMSL3	Thymosin-like 3	Regulation of actin cytoskeleton
PTPN4	Protein tyrosine phosphatase, non-receptor type 4 (megakaryocyte)	...
CUGBP2	CUG triplet repeat, RNA binding protein 2	...
PDIA5	Protein disulfide isomerase family A, member 5	...
MEF2C	Myocyte enhancer factor 2C	MAPK signaling pathway
GRIA1	Glutamate receptor, ionotropic, AMPA 1	Neuroactive ligand-receptor interaction; Long-term potentiation; Long-term depression

Table 2.9: Functional Descriptions of Top 10 Genes Selected from the Prostate Cancer Dataset

Gene Name	Description	Pathway
HPN	Hepsin (transmembrane protease, serine 1)	...
HSPD1	Heat shock 60kDa protein 1 (chaperonin)	Type I diabetes mellitus
SERPINB5	Serpin peptidase inhibitor, clade B (ovalbumin), member 5	p53 signaling pathway
GSTM1	Glutathione S-transferase M1	Glutathione metabolism; Metabolism of xenobiotics by cytochrome P450
LMO3	LIM domain only 3 (rhombotin-like 2)	...
JUNB	Jun B proto-oncogene	...
S100A4	S100 calcium binding protein A4	...
ATP1A1	ATPase, Na ⁺ /K ⁺ transporting, alpha 1 polypeptide	...
MAF	V-maf musculoaponeurotic fibrosarcoma oncogene homolog (avian)	...
CHD9	Chromodomain helicase DNA binding protein 9	...

Chapter 3

T-score with Sample Selection

The T-score is widely used for gene ranking with microarray gene expressions. It finds separability of expressions in different classes by measuring the correlation between input expressions and output class labels. Such a filter based criterion is usually computed using all the training samples, all of which however may not be equally important for the classification task. In this chapter, we decompose the T-score into two parts:- corresponding to *relevant* and *irrelevant* samples. With this decomposition, two approaches are proposed to compute the T-score for simultaneous of gene and sample selection: (i) designating support vectors as the *relevant* sample points, and (ii) an efficient and approximate selection of *relevant* samples by using a modified logistic regression loss function.

The proposed approaches allow simultaneous selection of sample and genes in backward elimination manner. We demonstrate that by selecting genes as well as samples, performance of the classification is improved significantly on synthetic as well as on several benchmark cancer datasets. Furthermore, comparison with other gene selection approaches and k NN classifier indicate that the performance improves consistently across majority of datasets.

3.1 Motivation

In the traditional filter approach, the standard practice is to consider all the samples for gene ranking. The advances of SVM and boosting algorithms have alluded that all the samples may not be equally relevant for disease classification [42], [55]. Also, a standard logistic regression based model gives weights to sample points according to their importance in classification. Removal of samples that are noisy and do not provide

useful information for classification may improve the classification [78], [134]. Due to the heterogeneity of biological tissues and cell samples, microarray datasets are multimodal by nature [23] and therefore quality of samples vary. Most of the filter based gene ranking methods do not pay attention to usefulness and quality of individual data sample.

Though the selection of sample points is largely ignored in filter-based approaches, a few algorithms have been proposed in other fields for simultaneous selection of samples and features. Blum and Langley had presented a review on the methods for selecting feature and samples [12]. Littlestone and Mesterharm proposed a perceptron algorithm and an edited nearest-neighbor based method to learn from a data point only when it violated the hypothesis and ignored points which were correctly classified [91]. Liu *et al.* [92] and Pechenizkiy *et al.* [114] proposed to use a *kd*-tree for selective sampling of data points for feature selection. Liu *et al.* partitioned all data points into clusters by using a *kd*-tree, selected random data point from each cluster, and then performed feature selection by using Relief [70] which looks for boundary points to estimate feature weights. On the other hand, Pechenizkiy *et al.* employed the principal component analysis for dimensionality reduction after partitioning large datasets with *kd*-tree [114]. Recently, Cavill *et al.* proposed an evolutionary approach for simultaneous feature and sample selection, in which GA and *k*NN were combined for metabolic profiling data classification [16]. Similarly, Shen *et al.* proposed a modified particle swarm based approach for simultaneous sample and gene selection [131]. Many of the above algorithms apply a specific method (e.g. *kd*-tree) to partition large datasets and select samples points for feature selection or use evolutionary algorithm based heuristics for simultaneous selection of samples and features.

To select both sample points and genes, it is necessary to find important samples. SVM theory provides an elegant approach because samples points within the margins are the most important ones for determining the discriminatory function. Such points would be the *relevant* data points in our analyses. However these points are not embedded into a gene selection algorithm. Therefore, we propose a backward elimination based iterative approach to rank genes by using *relevant* sample points only and thereby compute T-score. We also study an independent and single gene based selection of the samples and its effect on performance of classification.

However, these approaches still involve the use of computationally expensive algorithms like SVM for sample selection. Using the principles of modified logistic regression model, we derive a criterion for sample selection. This is further coupled with the T-score

and then we propose a backward elimination based algorithm to simultaneously select sample points and genes.

Experiments on synthetic data as well as on several of benchmark datasets, with various performance metrics and comparison to existing gene ranking methods, indicate superior performance of the proposed backward elimination method. An efficient algorithm to determine the relevant points not only results in superior computational complexity and stability but also improves the classification.

This chapter is organized as follows: In Section 3.2, we describe the T-score method, SVt-RFE, and efficient algorithm for sample and gene selection: T-score with sample selection (T-SS). Numerical experiments and results are discussed in Section 3.6 and the final section concludes this chapter with a discussion.

3.2 Relevant Samples

Data samples can be grouped into two categories according to their importance for classification: *relevant* and *irrelevant*. Below, we define *relevant* training samples in relation to the margin of separation of two classes:

For two perfectly separable classes, the *Margin of separation* is defined as a summation of maximum distance of the nearest sample in each classes to the optimal hyperplane.

The margin of separation can be generalized for the non-separable classes using slack variables [155]. SVM is the optimum margin classifier which is used to find the *relevant* samples in this chapter. Theoretically, only the support vectors (SV) or the points on the boundary of margin of separation are used in the determination of discriminant function of the classifier [55]. Figure (3.1) illustrates SV and the margin of separation (γ) between two classes. Using margin of separation and assuming that γ is the width of the optimum margin, we define relevant samples as: sample points within and on this margin of separation of classifier are *relevant* sample points. Selection of such samples is largely dependent on the size of the margin of separation. Geometrically, the samples that are nearer to the discriminant hyperplane form the subset of *relevant* samples. Specifically, the points which are located at more than $\gamma/2$ distance from the separating hyperplane are ignored and termed as irrelevant samples.

In a linear SVM, to select SV of training data, a dual problem given by (2.4) with the constraints (2.5) and (2.6) is to be solved. The weights of SVM are given by (2.7). It can be shown that the Lagrange multiplier α_j are non-zero for support vector points [55]. In other words, those are the only samples used for the computation of SVM weights. In

this work, we use such SV or samples for which α_j is non-zero for the computation of the T-score. That is, we indirectly use the SVM optimization function for selection of the data samples. In this chapter, we will use relevant data points and “support vectors” interchangeably.

Figure 3.1: Illustration of support vectors (relevant samples)

3.2.1 Relevant T-score

The T-score is widely used for ranking of genes for sample classification. In the following, we decompose the T-score into two terms, one corresponding to the *relevant* and the other to the *irrelevant* samples.

Let φ denote the ratio of number of *relevant* points to total number of training sample points. Let φ^+ represents the ratio of number of *relevant* points in positive class to total number of positive training samples. Hence, we have $\varphi = \frac{m_r}{m}$ and $\varphi^+ = \frac{m_r^+}{m^+}$. Similarly we define φ^- for negative class.

The T-score value (r_i) of gene i is given by,

$$r_i = \frac{|\mu_i^+ - \mu_i^-|}{\sigma_i^2} \quad (3.1)$$

Since φ^+ is the ratio of number of *relevant* points to total number of points in $+$ class, and φ^- is the ratio of number of *relevant* points to total number of points in $-$ class,

$$\mu_i^+ = \varphi^+ \mu_{i,r}^+ + (1 - \varphi^+) \mu_{i,nr}^+ \quad (3.2)$$

$$\mu_i^- = \varphi^- \mu_{i,r}^- + (1 - \varphi^-) \mu_{i,nr}^- \quad (3.3)$$

where $\mu_{i,r}^+$ and $\mu_{i,nr}^+$ denote the mean values computed with relevant and irrelevant samples in positive class. Using (3.2) and (3.3) in (3.1), we have,

$$r_i = \frac{\varphi^+ \mu_{i,r}^+ - \varphi^- \mu_{i,r}^-}{\sigma_i^2} + \frac{(1 - \varphi^+) \mu_{i,nr}^+ - (1 - \varphi^-) \mu_{i,nr}^-}{\sigma_i^2} \quad (3.4)$$

Assuming that the number of sample points are equal in both classes (i.e., $m^+ = m^- = \frac{m}{2}$), and the same number of samples are selected in both classes (i.e., $m_r^+ = m_r^-$, and hence $\varphi^+ = \varphi^- = \varphi$),

$$r_i = \varphi \left(\frac{\sigma_{i,r}}{\sigma_i} \right)^2 r_{i,r} + (1 - \varphi) \left(\frac{\sigma_{i,nr}}{\sigma_i} \right)^2 r_{i,nr} \quad (3.5)$$

where $r_{i,r} = \frac{\mu_{i,r}^+ - \mu_{i,r}^-}{\sigma_{i,r}^2}$. Similarly we define $r_{i,nr}$. With (3.5), the T-score is decomposed into two parts: $r_{i,r}$ which is a score due to relevant samples, and $r_{i,nr}$ is due to irrelevant samples

As $1 - \varphi$ is a ratio of *irrelevant* (i.e. non-selected) data points to total number of instances, the (3.5) can be re-written as,

$$r_i = \varphi \left(\frac{\sigma_{i,r}}{\sigma_i} \right)^2 r_{i,r} + \beta_i \quad (3.6)$$

where,

$$\beta_i = \frac{2}{m\sigma_i^2} \left(\sum_{j=1}^{m_{nr}^+} x_{ij} - \sum_{j=1}^{m_{nr}^-} x_{ij} \right)$$

With $y_j = \{+1, -1\}$,

$$\beta_i = \frac{2}{m\sigma_i^2} \sum_{j=1}^{m_{nr}} y_j x_{ij} \quad (3.7)$$

The term β_i is associated with the *irrelevant* sample points, that is, those samples that lie outside the margin γ . Due to this component in the standard T-score, relevant genes may be ranked differently because of *irrelevant* but correctly classified sample points, that are far from the separation boundary. By forcing such β_i to zero, the rankings may be improved and henceby performance of classification.

3.3 Support-Vectors (SV) based T-score

As shown in (3.6), the T-score criterion is affected if *irrelevant* samples are used in determination of the ranking of a gene. Moreover, as shown in machine learning theory, the generalization of a model may be poor and overfitting may occur when all the training samples (both *relevant* and *irrelevant* samples) are used in determination the discriminant hyperplane [42], [155]. Intuitively, this suggests that gene selection with T-score is likely to result in a model that has poor generalization capabilities.

To improve the performance of the T-score, we propose to use only the SV points in T-score computation. The use of SV points for T-score computation can be achieved with following modifications to the computation of mean and variance,

$$\mu_i^+ = \frac{1}{m_r^+} \sum_{j=1}^{m^+} \delta(\alpha_j \neq 0) x_{ij} \quad (3.8)$$

$$\sigma_i^+ = \sqrt{\frac{\sum_{j=1}^{m^+} \delta(\alpha_j \neq 0) (x_{ij} - \mu_i^+)^2}{m_r^+ - 1}} \quad (3.9)$$

where μ_i^+ and σ_i^+ represent mean and standard deviance of expression values of gene i in positive class, and m_r^+ denote the number of SV points in the positive class. The $\delta(\cdot)$ is an indicator function, that is 1 when the argument is a SV with non-zero Lagrange multiplier of given class and zero otherwise. Similarly, we define μ_i^- and σ_i^- for the negative class. Referring to (3.8) and (3.9), instead of taking only m_r^+ and m_r^- points (which are support vector points), the standard mean and variance calculations in T-score based methods use all points in positive and negative class to compute mean and variance of the gene expressions [61].

The SV are determined either with respect to each of the gene independently or by using all the genes in the subsets. Based on how SV are identified, we propose two strategies to rank genes with sample selection: (i) Support Vector based T-score ranking with individual gene (SVt-R) and (ii) Support Vector based T-score ranking with Recursive Feature Elimination (SVt-RFE)

Algorithm 3 Support Vector based T-score ranking with individual gene (SVt-R)

Begin : Gene subset $S = \{i\}_{i=1}^n$
for Gene $i = 1$ to n **do**
 Train SVM model with expressions x_i of gene i as input
 Remove the non-support vector points
 Compute T-score ($r_{i,r}$) of gene i with SV points
end for
 R = Ranked genes based on T-score values for all genes
end : Ranked gene list R

3.3.1 SV based T-score (SVt-R) with Individual Gene

In this method, samples are selected based on the gene expression values of an individual gene across the samples. SVM is trained with the expression of each gene. The resulting SV points are used to compute its T-score. The genes are then ranked based on the computed T-score. The method is given in Algorithm 3.

Like the standard T-score method, the proposed method also assumes independence among features. Because of SVM training of individual gene, the method is computationally expensive.

3.3.2 SV based T-score with RFE (SVt-RFE)

In this approach, starting from all genes in the subset, we select SV points by using SVM algorithm. The selected samples are then used to compute the gene ranking by using T-score. The gene with the least T-score value is eliminated from the gene subset. The iterations continues until all the genes are ranked. The method is given in Algorithm 4.

The algorithm determines the support vectors based on all genes in the subset and ranks genes in a backward elimination manner. Compared to SVt-R described in Algorithm 3, the assumption of independence of genes is relaxed and thus the SVM algorithm need not be trained for each of the genes independently.

3.4 Efficient Sample Selection Criterion

The margin of separation of SVM is defined by support vectors or the samples on the margin. The support vectors are the samples that define the discriminant function. In this section, a computationally efficient method is presented for using the samples not only on the margins but also within the margin for gene selection. Determining the

Algorithm 4 Support Vector Based T-score Ranking (SVt-RFE)

Begin : Ranked gene list $R = \emptyset$, and gene subset $S = \{i\}_{i=1}^n$
repeat
 Train a linear SVM with the gene set S as input variable
 Obtain support vector points and compute the ranking statistic r_i
 Select the gene with smallest ranking statistic $i^* = \arg \min(r_i)$
 Update $R = [R; i^*]; S = S \setminus \{i^*\}$
until all genes are ranked
end : output R

margin of separation has a computational complexity of $O(\max(n, m)m^2)$. This becomes even more costlier for SVM-RFE as each iteration needs retraining of the SVM. Therefore, a simpler model that selects sample points on and within the margins, is computationally inexpensive, and presents good biological interpretability, is needed.

An approximate loss function for SVM using the concepts of logistic regression has been proposed by [178]. This function uses a sequence of smooth functions for iterations to uniformly converge to SVM objective function. This approximate loss function is given by,

$$\mathcal{L}(x, y : w) = \frac{1}{\lambda} \ln (1 + \exp (-\lambda (yw^T x - 1))) \quad (3.10)$$

where λ is a tuning parameter and w denotes the weights. Instead of using a standard 0-1 loss function in SVM, the use of (3.10) leads to the following penalized objective function:

$$\mathcal{L}_P(x, y : w) = \sum_{j=1}^m \frac{1}{\lambda} \ln (1 + \exp (\lambda (1 - y_j w^T x_j))) + \eta \|w\|^2 \quad (3.11)$$

where η denotes the sensitivity parameter.

Setting the partial derivation of (3.11) with respect to each gene i to zero,

$$w = \sum_{j=1}^m \frac{1}{2\eta} \frac{\exp (\lambda (1 - y_j w^T x_j))}{1 + \exp (\lambda (1 - y_j w^T x_j))} y_j x_j \quad (3.12)$$

In (3.12), the multiplication factor to $y_j x_j$ incorporates the margin information while computing weights. For example, if margin $y_j w^T x_j$ is greater than one, the multiplication factor becomes zero for large value of λ . In a sense, it rejects the contribution of that particular sample point. Hence, based on this property and considering that $\frac{1}{2\eta}$ is a multiplicative factor, we propose the following sample selection criterion:

Algorithm 5 Gene ranking using T-score and sample selection

BeginGene set $S = \{i\}_{i=1}^n$, data D , and ranked list $R = \emptyset$;set λ ; $\epsilon = 0.001$ **repeat**Find the set of samples $M \subset D$ using (3.13) with $\alpha_j > \epsilon$ **if** $|M_l| < 2$ **then** $M_l = D$ **end if**Compute the ranking r_i using samples in M Select the gene $i^* = \arg \min \{r_i\}$ Update $R = [R; i^*]$; $S = S \setminus \{i^*\}$ **until** all genes are ranked**end** : output R

$$\alpha_j = \frac{\exp(\lambda(1 - y_j w^T x_j))}{1 + \exp(\lambda(1 - y_j w^T x_j))} \quad (3.13)$$

With respect to SVM-RFE, the standard 0-1 loss function gives the following SVM weight vector [53], [155]:

$$w = \sum_{j=1}^m \alpha_j y_j x_j \quad (3.14)$$

Comparing (3.12) and (3.14), we can represent the SVM induced weight to a particular sample point with (3.13).

3.4.1 T-Score with Sample Selection (T-SS)

The margin of a sample point is defined as the distance from the sample point to the discriminant boundary. The margin of j th data sample is given by the term $y_j w^T x_j$. Zhang *et al.* proposed an gradient descent based algorithm to determine the margin of separation [178]. In order to simplify the computations, we propose to use T-score of each gene as an estimate of w^T and thereby remove the optimization step. With this idea, we propose an algorithm for simultaneous sample and gene selection, which is described in Algorithm 5.

Sample points are selected using (3.13) with a small threshold ϵ . Let M_l denote the set of selected sample points in class l . With a given λ value, the samples are selected using the margin information based on T-score. Using only selected samples, the genes

are ranked using the T-score. A gene with least absolute score is then removed from the gene set and the whole process is iterated again until all genes are ranked. In other words, the proposed method ranks genes in a backward elimination manner with selection of samples. The T-score with sample selection method fails whenever there are less than two relevant sample points in any class. In such cases, we revert to all the sample points and compute the ranking scores in that iteration with all training samples.

The margin is determined by using the T-score of individual genes. It has a direct relation with log-odds ratio which is given by

$$\log \frac{P(+|x)}{P(-|x)} = x^T \Sigma^{-1} (\mu^+ - \mu^-) + w_0 \quad (3.15)$$

Here, w_0 is a bias term and was computed using

$$w_0 = \log \frac{\pi^+}{\pi^-} - \frac{1}{2} \mu^{+T} \Sigma^{-1} \mu^+ + \frac{1}{2} \mu^{-T} \Sigma^{-1} \mu^- \quad (3.16)$$

where π^+ and π^- represent the prior probabilities of respective classes; the Σ represents the covariance matrix. As we assume independence among genes, in (3.15) and (3.16), the covariance matrix becomes $\Sigma = \sigma^2 \mathbf{I}$, where $\mathbf{I}$ is an identity matrix. Such an approximation results in optimization less algorithm. In computing the sample margin $yw^T x$, the bias term is included in w . We normalized w before computing the margin of a sample:

$$w = \frac{w}{\|w\|} \quad (3.17)$$

3.4.2 Margin vs. λ parameter

The selection of samples in Algorithm 5 depends on the selection of λ . The determination of this parameter involves information of the margin of a sample point. Figure 3.2 shows how the α value varies with sample margin for a given λ value. The λ parameter determines the cut-off point to select the samples. More importantly, as shown in Figure (3.2), increase in λ value results in much sharper boundary in selecting the sample points.

Figure 3.2: Effect of λ on the efficient sample selection criterion against margin of a sample point

3.5 Relation of T-score with Sample Selection to SVM-RFE

Similar to SVt-RFE, genes are ranked using backward feature elimination in SVM-RFE. Starting with all genes, recursively least important gene is removed. The weight for an individual gene is computed by using non-zero α_j value (obtained by solving (2.4) with constraints (2.5) and (2.6)) and (3.14).

The SVM-RFE also uses only the points within optimum margin γ for the rank computation. The α_j value obtained from SVM training has zero lower bound and η upper bound. The points with $\alpha_j = \eta$ are known as *essential* support vectors. If all the *relevant* sample points are *essential* support vectors, and if both classes have same number of sample points and equal number of points are selected from each class, we can derive the relation of proposed approaches with SVM-RFE.

With essential support vectors ($\alpha_j = \eta$) and $y_j = \{+1, -1\}$, the SVM weight for gene i can be given by

$$\frac{w_i}{\eta} = \sum_{j=1}^{m_r^+} x_{ij} - \sum_{j=1}^{m_r^-} x_{ij} \quad (3.18)$$

$$\frac{w_i}{\eta} = m_r^+ \mu_{i,r}^+ - m_r^- \mu_{i,r}^- \quad (3.19)$$

With $m_r^+ = m_r^- = \frac{m_r}{2}$,

$$w_i = \eta \frac{m_r}{2} (\sigma_{i,r}^2) r_{i,r} \quad (3.20)$$

This could be easily computed depending on how $r_{i,r}$ is computed. Nevertheless, this clearly demonstrates the relation between SVM-RFE criterion and the proposed simultaneous gene and sample selection algorithms.

3.6 Experiments and Results

3.6.1 Synthetic Data

We generate a synthetic dataset that attempts to mimic multimodal and correlative structure of gene expression data. Recently, large scale simulation studies on synthetic datasets have been performed to understand and benchmark various feature selection methods [58], [112], [136]. In order to study the effectiveness of the T-SS method on the performance of classification, a bimodal multivariate dataset with each class modelled with an equally likely prior and Gaussian densities and the same covariance matrices Σ is generated. Each class is modelled by two different means: one for the majority of sample points and other for the generation of the *assumed* irrelevant samples. For generation of the feature group, we followed the method presented in [136].

Let n_r be the number of relevant features and g be the total number of feature groups. The mean of the relevant feature i is computed from,

$$\mu_i = \nu \left(1 - \frac{1}{g} \left\lfloor \frac{(i-1)g}{n_r} \right\rfloor \right) \quad (3.21)$$

where ν is a prescribed constant for the Gaussian in given class and $\lfloor . \rfloor$ is the floor function. We set the mean such that for any $i < i'$, mean $\mu_i \geq \mu_{i'}$.

We presume that the correlation between any two relevant features are the same in all Gaussians. The n_{nr} number of the irrelevant random noise features were also added, hence the total number of the features (n) in the dataset equals to $n_r + n_{nr}$. The noisy features were generated from zero mean and unit variance Gaussian distributions.

The performance of the training dataset was evaluated on dataset with both single and bimodal sample points in each class. The bimodal test set (*test-b*) was generated similar to the training data points. The single modal test set (*test-s*) consisted of sample points from only one multivariate Gaussian distribution in each class. Each test set had

1000 data points with equal class probability. The training set was generated 50 times, and each set was evaluated on both *test-s* and *test-b* datasets. All the parameter settings are given in the Table 3.1.

3.6.2 Benchmark Datasets

Similar to Chapter 2, we performed extensive experiments on various microarray gene expression datasets, namely, colon [3], leukemia [49], prostate [137], CNS [117], hepato [60] to evaluate the performance of the proposed methods. Additionally, we also performed experiments on the breast [164] cancer datasets which has 25 positive and 24 negative samples. Details of the datasets are given in section 1.11.

As discussed in previous Chapter, we use similar steps for preprocessing to select samples. SVM sensitivity parameter was also determined similarly for SVt-R and SVt-RFE methods.

The sample selection criterion in Algorithm 5, derived from modified logistic regression, needs λ to be optimized. It was selected from a set of $\{1, 3, 5, 7, 10\}$. Genes were ranked using each of the λ value. The final λ corresponds to the maximum five-fold external cross validation.

3.6.3 Performance Evaluation

Ranking of genes in each dataset was achieved using T-score, SVM-RFE, SVt-R, SVt-RFE, and T-SS methods. Only the training data was used to rank genes using a linear SVM. Using the ranked genes, the gene subset was tested starting from the top ranked gene and then successively adding one gene at a time in testing the subset till the total number of genes in subset equals 100.

Unlike SVM-RFE with MRMR, the proposed approaches are relatively computationally less expensive. Hence, we evaluate the performance with more stringent external cross-validation method. The test errors were evaluated on five fold cross validation for 20 times. In such cross validation, all samples are partitioned into k -folds. The ranking of genes is only obtained from merging $(k-1)$ -folds into a set, known as training set. The remaining fold is used only for performance evaluation. The performance measures such as accuracy, sensitivity, specificity, MCC, and area under ROC curve (AUC) [39] were averaged over the total of 100 trials. For a classifier (e.g., SVM), which outputs only the class labels, the unbiased AUC is the average of sensitivity and specificity [15]. In such cross validation, the cardinality of the gene subset giving the minimum average test error

Table 3.1: Parameters for Generating Synthetic Datasets to Evaluate Performance of T-score, SVM-RFE, SVt-R, SVt-RFE, and T-SS Methods

No. of relevant features (n_r)	50
No. of noise features (n_{nr})	200
No. of biomarker types (g)	5
No. of training sample points	60
Subclass distribution of training points	25 and 5
Parameter ν in + class training data and <i>test-b</i>	+1 and +5
Parameter ν in - class training data and <i>test-b</i>	1 and -5
Parameter ν in <i>test-s</i> for respective classes	+1 and -1
Correlation among relevant features	0.4

was reported as the number of genes corresponding to the best classification performance. We also performed pair-wise one-sided *t*-test to determine whether the performance of the T-SS method is statistically better than other methods.

We also compared the proposed approaches with SVM-RFE method. This method was performed in an exact similar way to that of the SVt-RFE method except the ranking criteria. In all gene selection methods and testing the classifier, we used LIBSVM - 2.84 software [18]. In addition to standard T-score and SVM-RFE methods, we also compared the proposed methods with several other widely used methods: Bayes error based filter with *k*NN and SVM [177], minimum redundancy maximum relevancy criterion (based on the T-score) [31], correlation based feature selection (CFS) [161], elastic net logistic-regression (EN-LR) [44], and least square-bound with SVM [181]. Further, we compared the performance of the standard T-score, SVM-RFE, SVt-R, SVt-RFE, and T-SS methods using the *k*NN classifier (with 5 nearest neighbours).

3.6.4 Results

We have generated synthetic datasets to evaluate the performances of the proposed methods with T-score and SVM-RFE. The results are shown in Table 3.2. In SVt-RFE setting,

(a) Colon

(b) Leukemia

(c) Hepato

(d) CNS

(e) Prostate

(f) Breast

Figure 3.3: Performance of classification against the number of selected genes in various two-class cancer datasets by various gene selection methods

(a) Colon

(b) Leukemia

(c) Hepato

(d) CNS

(e) Prostate

(f) Breast

Figure 3.4: Effect of λ parameter on classification performance of T-SS method on various two-class cancer datasets

the selection of data points improves the performance of the T-score method on both *test-s* and *test-b* datasets. The results with SVt-R method were comparable with the standard

Table 3.2: Accuracies of Classification of T-score, SVt-R, SVM-RFE, SVt-RFE, and T-SS methods on *synthetic* datasets. The average number of selected genes and sample points are indicated

Method		Accuracy		Selected irrelevant genes	Selected samples
		<i>test-s</i> data	<i>test-b</i> data		
T-score		89.65 ± 2.02	90.37 ± 2.09	0.2	60
SVt-R		89.45 ± 1.22	90.39 ± 1.29	5.86	56.58
SVM-RFE		86.96 ± 1.96	88.19 ± 1.78	20.44	39.38
SVt-RFE		90.45 ± 0.82	91.50 ± 0.99	9.80	38.73
T-SS	$\lambda = 1$	90.28 ± 1.39	91.09 ± 1.23	1.66	43.65
	$\lambda = 3$	90.07 ± 1.50	90.93 ± 1.41	14.04	29.29
	$\lambda = 5$	88.83 ± 2.16	89.82 ± 1.83	26.12	21.31
	$\lambda = 7$	88.32 ± 2.15	89.61 ± 1.71	32.18	17.62
	$\lambda = 10$	87.29 ± 2.31	88.44 ± 2.19	35.08	15.50

T-score. The T-SS method improved the performance of classification, particularly when λ is set to 1. As the λ value is increased, the classification accuracy decreased in the synthetic dataset.

Comparing the rankings of relevant features, the T-score method ranks the most number of relevant genes. The use of the SVM algorithm in RFE methods results in noisy gene selection. The results were consistent with the findings of [135]. On the other hand, SVM enables selection of the sample points. In SVM-RFE and SVt-RFE, the average number of training sample points used for the computation of gene ranking criteria is significantly less than the standard T-score method and SVt-R method. In the T-SS method, as the λ value is increased, less number of relevant genes were selected. Increase in λ value also resulted in a fewer number of samples for the T-score calculation.

Though the SVt-RFE and T-SS methods showed superior performance on synthetic datasets, it is important to compare the results on various benchmark gene-expression datasets. Tables 3.3 and Table 3.4 show a comparison of classification performances of T-score, SVM-RFE, and proposed approaches on benchmark datasets. The corresponding pair-wise *t*-test results are shown in Table 3.5. Here, the symbol ‘...’ denotes statistically

Table 3.3: Performance of T-score, SVM-RFE, SVt-R, SVt-RFE, and T-SS on Leukemia, Hepato, and Prostate *Cancer* Datasets

Dataset	Measu- rement	T-score	SVM-RFE	SVt-R	SVt-RFE	T-SS
Leukemia	# Genes	65	43	100	63	49
	Accuracy	96.36 ± 4.72	96.65 ± 4.14	97.00 ± 4.35	97.01 ± 4.09	97.00 ± 4.10
	Sensitivity	94.40 ± 11.04	95.00 ± 9.16	95.20 ± 9.04	95.20 ± 9.04	95.40 ± 8.46
	Specificity	97.43 ± 4.83	97.54 ± 4.52	97.98 ± 4.20	97.99 ± 4.18	97.88 ± 4.50
	MCC	0.92 ± 0.10	0.93 ± 0.09	0.94 ± 0.09	0.94 ± 0.09	0.94 ± 0.08
	AUC	0.96 ± 0.06	0.96 ± 0.05	0.94 ± 0.09	0.97 ± 0.05	0.97 ± 0.05
Hepato	# Genes	82	71	100	44	88
	Accuracy	68.58 ± 9.16	65.83 ± 10.82	68.00 ± 10.51	66.67 ± 10.19	69.25 ± 9.53
	Sensitivity	42.50 ± 22.33	43.00 ± 23.32	39.50 ± 24.14	43.50 ± 23.47	45.25 ± 22.39
	Specificity	81.63 ± 11.58	77.25 ± 13.69	82.25 ± 13.43	78.25 ± 12.52	81.25 ± 12.24
	MCC	...	...	...	...	...
	AUC	0.62 ± 0.11	0.60 ± 0.12	0.61 ± 0.12	0.61 ± 0.12	0.63 ± 0.11
Prostate	# Genes	11	3	42	18	97
	Accuracy	85.73 ± 5.24	89.34 ± 6.81	84.82 ± 6.00	88.24 ± 8.94	88.06 ± 5.64
	Sensitivity	88.15 ± 7.48	91.82 ± 7.18	88.37 ± 8.70	90.47 ± 8.61	89.47 ± 8.25
	Specificity	82.52 ± 9.61	86.04 ± 12.99	80.13 ± 11.62	85.27 ± 14.42	86.20 ± 10.43
	MCC	0.71 ± 0.10	0.79 ± 0.14	0.70 ± 0.13	0.76 ± 0.18	0.76 ± 0.11
	AUC	0.85 ± 0.05	0.89 ± 0.07	0.84 ± 0.07	0.88 ± 0.09	0.88 ± 0.06

insignificant change in performance. As seen, the performance of SVt-RFE method is better than the standard T-score for most of the performance measures in CNS, prostate, breast, and leukemia datasets. On the other hand, T-SS outperforms standard T-score in all datasets. It is also significantly better than SVM-RFE in colon, breast, and hepato datasets, as shown in Table 3.5. The performance of SVt-R approach is comparable with standard T-score. The backward elimination methods (SVt-RFE and T-SS) performs

Table 3.4: Performance of T-score, SVM-RFE, SVt-R, SVt-RFE, and T-SS on Colon, CNS, and Breast *Cancer* Datasets

Dataset	Measurement	T-score	SVM-RFE	SVt-R	SVt-RFE	T-SS
Colon	# Genes	83	97	94	32	91
	Accuracy	86.53 $\pm$ 9.00	83.47 $\pm$ 9.37	86.98 $\pm$ 9.20	86.08 $\pm$ 9.44	87.12 $\pm$ 9.59
	Sensitivity	80.10 $\pm$ 19.08	74.35 $\pm$ 18.97	79.75 $\pm$ 19.85	78.25 $\pm$ 17.84	80.90 $\pm$ 18.44
	Specificity	90.00 $\pm$ 10.05	88.50 $\pm$ 11.47	90.87 $\pm$ 9.86	90.37 $\pm$ 10.02	90.50 $\pm$ 10.22
	MCC	0.72 $\pm$ 0.19	0.65 $\pm$ 0.20	0.72 $\pm$ 0.20	0.70 $\pm$ 0.20	0.73 $\pm$ 0.20
	AUC	0.85 $\pm$ 0.10	0.81 $\pm$ 0.10	0.85 $\pm$ 0.11	0.84 $\pm$ 0.10	0.86 $\pm$ 0.11
CNS	# Genes	88	73	1	100	84
	Accuracy	63.25 $\pm$ 10.20	69.25 $\pm$ 11.71	61.42 $\pm$ 8.59	65.00 $\pm$ 12.76	65.50 $\pm$ 12.14
	Sensitivity	41.00 $\pm$ 23.54	54.65 $\pm$ 22.41	8.90 $\pm$ 17.51	49.50 $\pm$ 24.04	47.25 $\pm$ 23.66
	Specificity	75.27 $\pm$ 15.28	77.04 $\pm$ 16.34	89.98 $\pm$ 14.84	73.18 $\pm$ 16.90	75.30 $\pm$ 16.58
	MCC	...	...	...	...	...
	AUC	0.58 $\pm$ 0.11	0.66 $\pm$ 0.12	0.49 $\pm$ 0.07	0.61 $\pm$ 0.14	0.61 $\pm$ 0.13
Breast	# Genes	57	97	100	87	99
	Accuracy	86.17 $\pm$ 11.78	87.67 $\pm$ 11.02	86.58 $\pm$ 10.81	87.97 $\pm$ 11.35	89.30 $\pm$ 10.77
	Sensitivity	88.80 $\pm$ 13.13	91.20 $\pm$ 13.43	88.80 $\pm$ 13.73	89.80 $\pm$ 13.48	89.20 $\pm$ 12.85
	Specificity	83.45 $\pm$ 19.07	84.15 $\pm$ 17.99	84.20 $\pm$ 18.24	86.05 $\pm$ 18.82	89.45 $\pm$ 15.42
	MCC	0.74 $\pm$ 0.23	0.77 $\pm$ 0.21	0.75 $\pm$ 0.21	0.78 $\pm$ 0.22	0.80 $\pm$ 0.21
	AUC	0.86 $\pm$ 0.12	0.87 $\pm$ 0.11	0.87 $\pm$ 0.11	0.87 $\pm$ 0.11	0.89 $\pm$ 0.10

significantly better than the SVt-R approach. The performance of the SVt-RFE method is comparable with SVM-RFE method.

In T-SS, we compute α parameter on all sample points as we select the samples based on it which indirectly affects the T-score of genes. Apart from the margin information, calculation of α is dependent on the parameter λ . Effect of this parameter on the performance of classification of benchmark datasets is shown in Figure (3.4). As depicted

Table 3.5: Statistical Significance of improvement of T-SS method over T-score, SVt-RFE, and SVM-RFE methods

Dataset	Method	Performance Metric				
		Accuracy	Sensitivity	Specificity	MCC	AUC
Colon	T-score	...	...	...	...	...
	SVt-RFE	...	$p < 0.05$	...	$p < 0.05$	$p < 0.05$
	SVM-RFE	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$	$p < 0.001$
Leukemia	T-score	$p < 0.05$	...	...	$p < 0.05$	...
	SVt-RFE	...	...	...	...	...
	SVM-RFE	...	...	...	...	...
Hepato	T-score	...	$p < 0.05$	...	...	...
	SVt-RFE	$p < 0.05$	...	$p < 0.05$	...	$p < 0.05$
	SVM-RFE	$p < 0.001$	...	$p < 0.01$	...	$p < 0.01$
Prostate	T-score	$p < 0.001$	...	$p < 0.01$	$p < 0.001$	$p < 0.01$
	SVt-RFE	...	...	...	...	...
	SVM-RFE	...	...	...	...	...
CNS	T-score	$p < 0.05$	$p < 0.01$	...	...	$p < 0.05$
	SVt-RFE	...	...	...	...	...
	SVM-RFE	...	...	...	...	...
Breast	T-score	$p < 0.001$	...	$p < 0.001$	$p < 0.001$	$p < 0.001$
	SVt-RFE	$p < 0.05$	...	$p < 0.01$	...	$p < 0.05$
	SVM-RFE	$p < 0.01$	...	$p < 0.001$	$p < 0.05$	$p < 0.01$

in Figure, the λ parameter has a significant impact on final classification performance. Breast, colon, and hepato cancer datasets need low λ value, i.e. the algorithm uses more number of training samples.

We compared the performance of SVt-RFE and T-SS with various state of the art gene

Table 3.6: Comparison of Genes Selected, and Accuracies of T-SS with Other Methods

Method	Measure	Colon	Leukemia	Hepato	Prostate	CNS	Breast
Bayes +	# Genes	21	5	15	6	24	27
k NN	Accuracy	82.75	91.99	62.83	92.28	62.83	82.10
Bayes +	# Genes	5	5	14	19	28	29
SVM	Accuracy	82.27	92.35	64.08	91.52	61.17	84.03
MRMR	# Genes	19	52	98	6	81	88
	Accuracy	83.02	95.38	69.00	81.78	66.50	86.68
LS-Bound	# Genes	3	98	83	45	94	99
	Accuracy	81.71	93.51	61.75	84.23	61.00	77.77
CFS	# Genes	92	62	92	8	28	72
	Accuracy	85.01	94.92	68.75	81.68	58.92	86.09
EN-LR	# Genes	48	95	15	99	87	96
	Accuracy	83.72	96.22	68.33	85.18	66.42	85.37
SV t -RFE	# Genes	32	63	44	18	100	87
	Accuracy	86.08	97.01	66.67	88.24	65.00	87.97
T-SS	# Genes	91	49	88	97	84	99
	Accuracy	87.12	97.00	69.25	88.06	65.50	89.30

ranking methods. Table 3.6 shows the results of comparison of classification accuracy. As seen, the proposed methods achieved better classification performance on most of the cancer datasets, including elastic net logistic regression though the numbers of genes selected were higher than the other methods.

The results in Tables 3.3 and 3.4 were obtained using SVM as a base classifier. For a fair comparison of T-score, SVM-RFE, SV t -RFE and T-SS, k NN was used as the base classifier. Table 3.8 gives a comparison of classification performances on top ranked genes obtained from various methods. The results show that SV t -RFE outperforms T-score on

Table 3.7: Comparison of Running Time(s) of T-score, SVt-R, SVM-RFE, SVt-RFE, and T-SS on Various *Synthetic* Training Datasets with Different Numbers of Features and Samples

Method		Training with 250 Features			Training with 140 Samples		
		60 Samples	100 Samples	140 Samples	250 features	500 features	1000 features
T-score		0.025	0.026	0.027	0.027	0.054	0.11
SVt-R		4.03	6.19	8.83	8.83	18.19	36.34
SVM-RFE		11.31	26.92	39.31	39.31	161.85	778.75
SVt-RFE		10.76	25.61	39.65	39.65	159.99	776.26
T-SS	$\lambda = 1$	1.59	1.77	1.86	1.86	4.11	13.31
	$\lambda = 3$	1.63	1.78	1.80	1.80	3.98	13.05
	$\lambda = 5$	1.58	1.69	1.77	1.77	3.94	12.81
	$\lambda = 7$	1.61	1.71	1.76	1.76	3.92	12.73
	$\lambda = 10$	1.59	1.69	1.77	1.77	3.95	12.63

Table 3.8: Average Accuracy of T-score, SVM-RFE, SVt-R, SVt-RFE, and T-SS on Various *Cancer* Datasets using k NN Classifier

Dataset	T-score	SVM-RFE	SVt-R	SVt-RFE	T-SS
Colon	85.96	83.17	86.29	86.17	86.27
CNS	65.58	68.33	63	64.33	64.58
Leukemia	94.75	94.13	95.82	96.87	96.87
Breast	86.49	87.68	87.02	88.92	89.24
Hepato	69.75	68.08	69.00	66.25	70.08
Prostate	82.01	87.40	84.41	86.58	87.35

all benchmark datasets while performance is comparable with SVM-RFE.

3.6.5 Comparison of Running time

The running time of the gene selection algorithms depend on number of genes and number of training data points. We generated five synthetic datasets by using different combinations of training sample points and dimensionality. The running times of various gene selection methods were estimated on such synthetic datasets and average results are shown in Table 3.7. In all experiments, the linear SVM sensitivity parameter (η) was set to 1 for fair comparison. The running times were recorded using MATLAB codes on single core, Intel Core(TM)2 CPU with 2.40 GHz processor and 1 GB RAM.

As shown in Table 3.7, the running time of the SVt-R, SVM-RFE, and SVt-RFE increased much faster than the standard T-score method with increase in dimensionality and training data points. The significant increase in the running time is due to the SVM algorithm, which is used for selection of the data points. For the SVM, complexity of the order of $\max(n, m)m^2$ can be assumed [52]. For the SVt-R method, a SVM will be trained for each of the genes. For RFE like algorithms, SVM will be retrained each time with the reduced number of genes, and hence, running time of SVM-RFE and SVt-RFE are comparable. However, increase in computational time with the increase in the number of training samples is marginal in the T-SS method. Due to backward elimination nature of the algorithm, the computational time of the T-SS method increases with the number of features. Yet, compared to SVM-RFE and SVt-RFE methods, the increase in computational time is insignificant. The λ parameter does not have any effect on computational time with varying number of samples or features. Computation of the standard T-score is proportional to the number of examples (m), and needed to be evaluated for each of the gene. This results in much smaller running time.

3.6.6 Visualization of Selected Samples

In the proposed algorithms, not only genes are ranked, but the sample points are also selected. Visualization of selected samples will help to interpret why some points are often selected, more importantly, why not selected. Visualization of selected sample points for high dimensional dataset is a challenging task.

To alleviate the problem of high dimensionality, we first computed principal components of a typical synthetic dataset with top 50 (i.e. informative) genes. We used only informative genes so that the samples from both classes can be effectively visualized without overlapping data points. We then plotted the first two principal components of dataset, with two different markers for both classes. Differential shade of color was then

Figure 3.5: Visualization of selected samples in synthetic dataset against change in λ value. The high intensity of color denotes higher selection frequency of a sample point.

applied to indicate the selection frequency of a sample point. The intensity of darkness indicates how frequently a sample point is selected.

One such plot for a typical synthetic dataset is shown in Figure (3.5). The plot shows how increase in λ value affects the selection frequency of sample points. As seen, for low λ value, all the points have influence on computing the T-score. As the λ value increases, less number of samples are used to compute the T-score. The proposed method is a backward elimination method. Hence the removal of genes also influence the sample selection frequency as removal of genes will have direct effect on computation of α .

3.7 Discussion

We have studied the effect of selection of sample points on the performance of T-score based feature selection criteria. In one approach, *relevant* sample points for gene selection were selected using SVM algorithm. In the second approach, we derived a sample selection criterion using the modified logistic regression. The proposed methods of gene selection used genes either individually or in a subset. The performance was evaluated on synthetic as well as on variety of benchmark datasets.

The synthetic dataset has been designed with bimodal multivariate Gaussians for each of the class. Each class has some sample points with very high mean values as compared to the majority of samples, so that a SVM training model may recognize such points as *irrelevant* sample points. The dataset was also mixed with large number of noisy genes. The standard T-score method was able to rank most of the relevant features as top ranking features. In comparison, the backward elimination methods, particularly based on SVM algorithm, identified less number of relevant genes. The RFE like algorithm partially penalizes redundancy among features [83]. Compared to randomly generated genes, the multivariate relevant features may still have redundancy among themselves. Hence, the RFE algorithm tends to select some of such genes in the top list. Interpretation of sample selection was aptly visualized in a typical synthetic dataset by using principal components.

In the SVt-R method, the selection of sample points was done by considering each gene independently. Although the results were not consistent across all datasets, the removal of non-support vector points improved classification performance, as observed in the Leukemia cancer dataset. However, the selection of data points may not be optimal because of the assumption of independence among genes. In order to relax the independence assumption, we proposed the backward elimination approach where determination of support vector points was based on the genes in a subset. This approach achieved consistently better performance in selection of genes and classification of samples. The algorithm using the approximate support vectors further improved the performance.

In the backward elimination, removal of one gene changes the support vectors. Unlike in standard T-score, with the change of selected data points in each of the iteration, the mean and the variance of expression of a gene change, resulting in the change of the T-score value. Due to the iterative nature in selection of samples, the proposed SVt-RFE method is computationally more expensive than the T-score method, but comparable to SVM-RFE. Though we didn't experience in our experiments, the proposed SVt-RFE

method will fail whenever there is only a single support vector in either of the classes. Similarly, the T-SS method fails whenever there is less than two relevant data points. In such cases, we used all the data points in that particular iteration as indicated in the algorithm.

In order to enhance the performance of the T-score based approach for gene ranking, we select support vectors as the most *relevant* samples. Testing on benchmark cancer classification datasets revealed that the proposed methods, particularly T-SS, performed significantly better than the standard T-score approach. This study suggests that all the sample points are not important for gene ranking and selection of samples results in improving the performance of classification of cancer samples.

Chapter 4

Multiclass Gene Selection with Pareto Front Analysis

The previous chapters deal with gene selection for two-class cancer identification. In this chapter, we discuss the multiclass gene selection problem for identification of cancer tumors. In such problem, the F-score, KW-score etc. are widely used filter criteria for gene selection. However, due to data intrinsic characteristics, these ranking criteria may become biased towards few classes that are easy to identify from other classes. Gene selection under such influence of some classes may result in inferior classification performance. We propose a Pareto front based approach for gene selection to alleviate such bias. We demonstrate the improvement in classification performance using F-score and KW-score. The genes are ranked using class specific rank statistics obtained from decomposition of the F-score and the KW-score. We evaluate our approach on six multiclass cancer gene expression datasets and the results show significant improvement in classification performance.

4.1 Motivation

Filter methods are very popular for multiclass gene ranking as they are simple, easy to use, and computationally very efficient. They evaluate a gene subset by looking at its data characteristics with respect to class labels. For multiclass classification problem, some of the mostly used filter methods include F-statistics, Kruskal-Wallis test, mutual information, entropy, etc. [76]. For comparative review on multiclass classification and feature selection methods, the reader is referred to [87].

F-statistics is based on the classical F-test, which uses the ratio of between-class sum of squares to within-class sum of squares as the test statistic [33]. On the other hand,

the Kruskal-Wallis test is a non-parametric method and is computed using the square of difference between average rank within-class to overall mean of ranks. Advancing further, Chen et al. proposed to use the Brown-Forsythe test statistics, and the Welch test statistics instead of the F-test and KW-test [19]. Unlike F-test where equal variance is assumed, these tests relax such assumption. In other heuristic approaches, genes were ranked using mean and standard deviation of expression values with the assumption that the expressions have short distances from class centroids and simultaneously have small within class variations [21]. Similarly, Shen et al. also proposed a centroid based ranking score, which is multiplication of difference between class mean to overall mean with the disparity between between-class sum of squares to within-class sum of squares [132]. A generalization of signal-to-noise ratio for multiclass gene selection was developed by defining gene dominant index and gene dormant index. The final classification was performed using few top ranked dominant and dormant genes [151]. Based on the observation that potential marker genes tend to have high expression levels, a weighted average difference method was proposed which used the average difference and relative average signal intensity to rank highly expressed genes at the top [66]. To remove redundancy from top ranked genes, specific algorithms like minimum redundancy maximum relevancy (MRMR) [31] and differential degree of prioritization (DDP) [111] were developed using F-test and correlation coefficient. However, most of the above filter based ranking criteria do not consider the separability of individual classes between each other which biases gene ranking towards easily separable classes.

Generally, a solution to any multiclass classification problem is class specific. Some of the classes are more difficult to classify compared to others while some are easier to distinguish from the others. Scarcity of good features that can equally distinguish the classes is one of the primary reasons behind this phenomena. This data intrinsic characteristic leads to lesser known problem called ‘siren pitfall’ in selection of features for multiclass classification [40]. In such a pitfall, features of easily separable classes dominate the ranking and many such features occupy the top rankings while features required for difficult to separate classes are ranked low.

In gene expression analysis, many filter based multiclass gene selection criteria are generalizations of the two-class gene selection criterion. Such generalization may involve aggregation of each two-class test to give a final ranking criterion. However, with aggregation, the components of some classes may dominate the overall ranking. A ranking criterion with aggregation selects genes that are relevant to easily separable classes, ignoring difficult to separate classes, and ultimately again leads to a siren-pitfall in feature

selection. The motivation of remove such pitfall is that the obtained gene subset represent all the classes equally while improving the performance of classification.

We look at this problem from a multi-objective perspective. Instead of aggregation, we treat the score of discrimination of two-classes as an objective value. Using principles of multi-objective evolutionary algorithms, we divide genes into various Pareto-fronts. Final gene ranking is obtained using various criterion such as crowding distance.

4.2 Ranking with Pareto-Front Analysis

4.2.1 Rank Statistics

F-score [33] and KW-score [19] are two popular gene ranking methods using microarray data, which are based on F -statistics and Kruskal-Wallis statistics. The former presumes that the data are normally distributed and independent while the latter is a non-parametric statistic. We show below that both statistics are aggregated statistic of a particular gene belonging to each class.

4.2.2 Aggregating Statistics

Let a separability statistic r_{il} is a *one-against-all* rank statistic of gene i in class l . It gives a measure of how discriminative gene i in class l with respect to all other classes. The ranking statistic r_i is an aggregating statistic of separability statistics r_{il} of gene i :

$$r_i = \sum_{l \in \Gamma} r_{il} \quad (4.1)$$

The rank of gene i is denoted by $\mathcal{R}_i$ and determined by the rank statistic r_i , and $\mathcal{R}_i \prec \mathcal{R}_{i'}$ indicates that the rank of gene i is better than that of gene i' overall (that is, considering all the classes). In other words,

$$r_i > r_{i'} \Leftrightarrow \mathcal{R}_i \prec \mathcal{R}_{i'} \quad (4.2)$$

Genes with high aggregation statistics are ordered early in the sequence of ranking.

4.2.3 Siren Pitfall Problem

For a particular gene, if class-specific statistics r_{il} are high for a few classes, it may overshadow the gene's contribution for discriminating other classes. As a result, many

genes will be selected as relevant, based on their ability to discriminate strong classes and a very few genes will be selected for their relevance to discriminating the difficult classes. For example, for the particular gene i , if $r_{il} \gg r_{il'}$ for all $l \neq l'$, r_i is dominated by r_{il} and $r_i \approx r_{il}$. That is, the gene will be dominated by class l and selected only because of its relevance to the class l .

When aggregating statistics are used, gene selection is dominated by a few features of easily separable classes and blinded by difficult to separate classes. This siren pitfall is caused by the imbalance of strong features and data samples present in different classes. Because of the nature of aggregation based statistics, correlated genes are likely to rank in a similar manner. In such a scenario, if redundancy is not properly addressed, redundant genes are likely to be highly ranked and deteriorate the performance of classification.

4.2.4 Multiobjective Optimization Perspective

In order to remove the pitfalls of aggregation of statistics of different classes, we approach the multiclass gene ranking problem with perspective of multiobjective optimization. Instead of adding up the statistics, we take each ranking statistic component as one objective statistic. By doing so, we try to select genes in such a way that the separability statistic of each class is maximized by virtue of gene selection.

In general, each element in the set $\{r_{il}\}_{i=1, l=1}^{n, L}$ of separability statistics defines the importance of gene i in separability of class l . For L classes, we have L such objective functions for each gene i . To achieve better classification, genes having maximum objective statistic value in each class, $\max_l r_{il}$, need to be selected.

However, most of the genes are only relevant to classification of a few classes. Hence, we need to select a subset of genes, which has relevant genes for all classes. We experiment with the separation statistic r_{il} based on the F-score and the KW-score. We resort to Pareto-front analysis, a popular theory of multiobjective evolutionary approach (MOEA) for selecting genes.

The idea is to first divide all the genes into different Pareto-fronts using r_{il} as objective values. It is similar to dividing genes in a given population into different Pareto-fronts, as done in an iteration of any standard multi-objective evolutionary optimization procedure. The genes in each Pareto-fronts may then be ranked by using different strategies, for example crowding distance criterion.

For Pareto-front analysis, we define following gene dominance criteria based on multi-objective optimization theory.

Algorithm 6 Ranking genes with Pareto Front Analysis

```

Gene set  $S = \{i\}_{i=1}^n$ 
Compute class-specific statistics  $\{r_{il}\}_{i=1, l=1}^{n, L}$  of genes in  $S$ 
Determine Pareto fronts set  $\mathcal{P} = \{\mathcal{P}^z\}_{z=1}^Z$  with NSGA-II
for Pareto fronts  $\mathcal{P}^z \in \mathcal{P}$  do
    Compute the ranking criterion (such as crowding distance) of each gene
    Sort and rank the genes in the front
end for
return Ranked Genes

```

Definition 4.1 A gene i is said to dominate the other gene i' , if both the following conditions 1 and 2 are true:

- (1) The objective statistic value of gene i is no worse than that of gene i' in all classes
- (2) At least one of the objective statistic value of gene i is better than that of gene i'

Mathematically, if the objectives are to be maximized,

$$i \succ i' \Leftrightarrow \forall l (r_{il} \geq r_{i'l}) \wedge \exists L (r_{il} > r_{i'l})$$

A pair of genes where neither dominates the other are known as *mutually non-dominating* genes with respect to each other, while if a gene is not dominated by any other gene of the set is called *non-dominated* gene.

Definition 4.2 Among a set of genes S , the non-dominated set of genes S' are those genes that are not dominated by any gene of the set S .

After Pareto-front analysis, we obtain Z numbers of ranked fronts $\mathcal{P}$, each front containing several genes. To rank genes within a front, we use three methods (i) crowding distance, (ii) weighted average sum, and (iii) average rank, and then sort it in descending order. The details of the method of gene ranking is described in Algorithm 6.

To determine the Pareto-fronts of a given set of genes with their objective statistic values, we perform all pair-wise comparisons and determine non-dominated genes. If these genes are non-dominated by any other genes in the whole set, they form a Pareto-optimal set. In the context of gene selection, top ranked Pareto-front gene sets represent highly important genes for classification. NSGA-II is one of the very popular algorithms to obtain Pareto-optimal set [26].

In NSGA-II, first the Pareto nondominated individual genes from gene set S are identified, and are grouped into $\mathcal{P}^{(1)}$, the Pareto-front with rank 1. The method for

Algorithm 7 NSGA-II: Generating of Pareto Fronts

```

1 Input: set  $S$  of genes
   $z = 1$ 

2 Set gene counter  $i = 1$  and  $\mathcal{P}^{(z)} = \emptyset$ 

3 For each gene  $i' \in S \setminus i$ 
  if  $(i' \prec i)$ 
    go to step 5

4 if  $i' \neq |S|$ 
   $i' = i' + 1$ 
  go to step 3
  else  $\mathcal{P}^{(z)} = \mathcal{P}^{(z)} \cup \{i\}$ 

5  $i = i + 1$ 
  if  $i \leq |S|$ 
    go to step 3
  else  $\mathcal{P}^{(z)}$  is  $z$ -th non-dominated set of genes
     $S = S \setminus \mathcal{P}^z$ 
    If  $S \neq \emptyset$ 
       $z = z + 1$ 
      go to step 1

6 Output: Pareto fronts  $\mathcal{P} = \{\mathcal{P}^z\}_{z=1}^Z$ 

```

minimization is shown in Algorithm 7. After identifying first Pareto-front, the second non-dominated pareto genes are identified from the subset $S \setminus \mathcal{P}^{(1)}$. The same procedure is continued till $S \setminus \mathcal{P}^{(z)} = \emptyset$, where $\mathcal{P}^{(z)}$ is the subset that contains the worst genes.

4.2.5 Ranking in a given Pareto Front

In a given Pareto-front, we have many non-dominated genes. To rank genes within it, we employed three different methods.

4.2.5.1 Weighted Sum Approach

The most natural way to rank the genes within a given Pareto front is by combining all the class-wise objective statistics values into single one by using aggregation. However, the aggregation is done after dividing genes into various Pareto fronts by using class-wise

Algorithm 8 Crowding Distance

begin: Genes in the Pareto front $\mathcal{P}^z$ and let $n^* = |\mathcal{P}^z|$
 Let $d_{i^*} = 0$ for all $i^* = 1, 2, \dots, n^*$
for $l = 1$ to L **do**
 Sort all objective values $r_{.l}$ in $\mathcal{P}^z$ in worse order
 Set $d_1 = d_{n^*} = \infty$
 for $i^* = 2$ to $n^* - 1$ **do**
 $d_{i^*} = d_{i^*} + (r_{i^*+1,l} - r_{i^*-1,l}) / (r_{.l}^{\max} - r_{.l}^{\min})$
 end for
end for

scores. The aggregation can also be done using weighted coefficients. Here, we will give equal weights to all the classes. In a given Pareto front, the ranking criteria for gene i is given by,

$$WS_{\mathcal{P}}(i) = \sum_l r_{il} \quad (4.3)$$

4.2.5.2 Crowding-Distance

The computation of the crowding-distance measure needs sorting of the genes in a given front according to each objective function value in ascending order. The genes with largest or smallest objective value were assigned with infinite crowding distance value. The intermediate genes were assigned a distance value equal to the absolute normalized difference in the function values of the two adjacent solutions. This procedure was continued over all objective functions. The sum of individual distance values corresponding to each objective denotes the overall crowding distance value. Algorithm 8 describes details to compute crowding distance.

The crowding-distance is basically a diversity preserving mechanism. The smaller value of crowding-distance represents that a gene is crowded with many genes having similar objective function values. Unlike weighted sum approach, the crowding distance increases the diversity of genes in a given Pareto front.

4.2.5.3 Average Ranking (AR)

Bentley and Wakefield proposed average ranking method [8]. Here in a given objective, ranking list is obtained using the fitness of each solution in this particular objective. Same procedure is then followed for all the objectives. Finally, ranks in each objectives are aggregated. Though this method was developed for global ranking, we use it only

in a given Pareto front. In this case, we compute rank of genes in a given Pareto front by using r_{il} . Let $\mathcal{R}_{il,\mathcal{P}}^*$ be the rank of gene i in class l and Pareto front $\mathcal{P}$. The average ranking is computed by

$$AR_{\mathcal{P}}(i) = \sum_l \mathcal{R}_{il,\mathcal{P}}^* \quad (4.4)$$

Below we present how ranking statistic $\mathcal{R}_i$ is obtained by partitioning the F-score and KW-score by considering their belongingness to different classes.

4.2.6 F-score based Pareto Front Analysis

F-score measures the ratio between the intraclass distance and the interclass distance. Let F_i is the ranking score of gene i given data X :

$$F_i = \frac{\sum_l \sum_j \delta(y_j = l)(\bar{x}_{il} - \bar{x}_{i.})^2}{\sum_l \sum_j \delta(y_j = l)(x_{ij} - \bar{x}_{il})^2} \quad (4.5)$$

where $\bar{x}_{i.}$ is the average expression level of gene i across all samples and $\bar{x}_{il}$ is the average expression level of the gene i of the samples belonging to l -th class, and $\delta(\cdot)$ denotes the indicator function that is equal to 1 if the argument is true and 0 otherwise.

The component F_{il} that corresponds to class l of F_i is given by

$$F_{il} = \frac{(\bar{x}_{il} - \bar{x}_{i.})^2}{\sum_j \sum_l \delta(y_j = l)(x_{ij} - \bar{x}_{il})^2} \quad (4.6)$$

where F_{il} can be considered as a *class-specific statistic* which is given by the ratio between the variance of gene expressions of gene i of the samples belonging to class l and the variance of expression of the gene in all the samples. In other words, statistics in $\{F_{il}\}_{l=1}^L$ give ranking statistics of gene i in classes l .

In this chapter, we propose to decompose the F -statistics into its class-wise scores and then apply the Pareto front analysis. In such formulation, we will have L objective statistic values for each gene. To achieve better classification, genes having maximum objective statistic value in each class need to be selected. However, most of the genes are only relevant to classification of a few classes. Hence, we need to select a subset of genes, which has relevant genes for all classes. Using F_{il} (i.e. rank statistic value) and Pareto analysis presented in Algorithm 6, we select such subset of genes. We present three variants of F-score, weighted sum approach (Fsum), average rank approach (Favg), and crowding distance based method (F-PFA).

4.2.7 KW-score based Pareto Front Analysis

Similar to the F-score, the Kruskal-Wallis (KW) score is a non-parametric statistic based on the rankings of the gene expressions. The KW-score of gene i is given by

$$KW_i = \frac{12}{m(m+1)} \left(\sum_l m_l (\bar{\mathcal{R}}_{il} - \bar{\mathcal{R}}_i)^2 \right) \quad (4.7)$$

where $\bar{\mathcal{R}}_{il}$ is the average of ranking of expression values of gene i considering the samples belonging to l -th class and $\bar{\mathcal{R}}_i$ is the overall average rank of the expressions of gene i considering all the samples and m_l is the number of samples belonging to class l .

By decomposing KW_i into class-wise components, class-specific statistics of gene i of class l is given by

$$KW_{il} = m_l (\bar{\mathcal{R}}_{il} - \bar{\mathcal{R}}_i)^2 \quad (4.8)$$

We present three variants of KW-score, weighted sum approach (KWsum), average rank approach (KWavg), and crowding distance based method (KW-PFA).

4.3 Experiments and Results

4.3.1 Data and Pre-processing

To evaluate the performance of the proposed approaches, we performed extensive experiments on six microarray gene expression datasets, namely, GCM [123], Lung [9], MLL [4], NCI Ross [126], NCI Staunton [140], and 11 Tumor [142]. These are widely used benchmark datasets to evaluate gene ranking methods with varying number of classes and genes.

All six datasets contain a large number of genes, out of which many genes have constant gene expression levels. Therefore, it becomes necessary to pre-process data. For NCI Ross dataset, we have employed the preprocessed dataset as given by the supplementary information in [25]. For other datasets, we employ the following steps for pre-processing:

- (i) Absolute gene expression values were thresholded with floor values of 100 and ceiling of 16000.
- (ii) From training dataset, we excluded genes with absolute gene expression value change less than 100 across all training samples

Table 4.1: Sample Sizes of Six Multiclass Gene-Expression Datasets

Dataset	# Classes	# Samples	# Genes after preprocessing
GCM	14	198	14122
MLL	3	72	10930
NCI(Ross)	8	58	5643
NCI(Stau)	8	58	3144
Lung	5	203	5343
11 Tumor	11	174	9700

(iii) Transformation with log10 base

We used above pre-processed datasets to determine the rankings of genes. For evaluation of classification performance, genes were further standardized to zero mean and one standard deviation in all datasets. The number of samples and genes after pre-processing are given in Table 4.1.

4.3.2 Performance Evaluation

Genes in each dataset have been ranked using F-score, KW-score, and the proposed Pareto front based approaches. As seen from Table 4.1, the number of samples in datasets are small and it can cause a peculiar problem while dividing into a training and testing sets. The performance evaluation is not reliable if only one testing set is used. Ranking with each method has been performed 100 times and performance measures reported are averages over 100 test runs. In each test, the data is split into a training and testing dataset and only the training dataset is used to learn the parameters and rank the genes by using four-fold cross-validation, giving 400 trials in total. The method is only tested on the corresponding testing set. In each run, training and test partitioning in cross-validation are randomly performed. The above procedure minimizes any bias of performance towards the data sampling.

Performance measure such as test accuracy and Gorodkin Correlation Coefficient(GCC) are computed for such 400 trials. The mean and standard deviation of these performance measures are shown in Table 4.2 and Table 4.3. In datasets with small sample size with

Table 4.2: Performance of F-score, F-PFA and Significance Analysis on Various Multiclass Cancer Datasets

Dataset	Measurement	F-score	F-PFA	p-value
GCM	Gene	385	217	
	Accuracy	70.23 ± 5.13	76.26 ± 5.03	$p < 0.001$
	GCC	0.68 ± 0.05	0.75 ± 0.05	$p < 0.001$
MLL	Gene	397	294	
	Accuracy	95.11 ± 4.48	95.82 ± 3.95	$p < 0.01$
	GCC	0.93 ± 0.07	0.94 ± 0.06	$p < 0.01$
NCI(Ross)	Gene	323	386	
	Accuracy	67.23 ± 11.36	70.18 ± 10.11	$p < 0.001$
	GCC	0.64 ± 0.13	0.67 ± 0.11	$p < 0.001$
NCI(Staunton)	Gene	315	397	
	Accuracy	59.63 ± 11.31	61.81 ± 10.66	$p < 0.01$
	GCC	0.56 ± 0.12	0.58 ± 0.12	$p < 0.001$
11 Tumor	Gene	395	400	
	Accuracy	90.79 ± 3.88	90.85 ± 3.87	...
	GCC	0.90 ± 0.04	0.90 ± 0.03	...
Lung	Gene	399	50	
	Accuracy	93.56 ± 3.07	94.69 ± 2.61	$p < 0.001$
	GCC	0.87 ± 0.06	0.89 ± 0.05	$p < 0.001$

unbalanced class distribution it is not reliable enough to evaluate performance based on accuracy only. In such cases, Matthew's correlation coefficient (MCC) is used for two-class classification. We employ GCC performance measure, which is a generalization of MCC for multiclass classification problem [50], [94] .

Table 4.3: Performance of KW-score, KW-PFA and Significance Analysis on Various Multiclass *Cancer* Datasets

Dataset	Measurement	KW-score	KW-PFA	<i>p</i> -value
GCM	Gene	394	327	
	Accuracy	62.42 ± 5.12	76.65 ± 4.95	$p < 0.001$
	GCC	0.60 ± 0.06	0.75 ± 0.05	$p < 0.001$
MLL	Gene	398	43	
	Accuracy	94.40 ± 5.17	95.68 ± 4.19	$p < 0.001$
	GCC	0.92 ± 0.08	0.94 ± 0.06	$p < 0.001$
NCI(Ross)	Gene	283	397	
	Accuracy	67.92 ± 10.53	69.13 ± 10.57	$p < 0.1$
	GCC	0.65 ± 0.12	0.66 ± 0.12	$p < 0.05$
NCI(Staunton)	Gene	372	382	
	Accuracy	61.54 ± 10.51	62.64 ± 10.36	$p < 0.1$
	GCC	0.58 ± 0.12	0.59 ± 0.11	$p < 0.1$
11 Tumor	Gene	397	335	
	Accuracy	93.73 ± 3.32	93.55 ± 3.09	...
	GCC	0.93 ± 0.04	0.93 ± 0.03	...
Lung	Gene	189	229	
	Accuracy	95.59 ± 2.45	95.59 ± 2.45	...
	GCC	0.91 ± 0.05	0.91 ± 0.05	...

Assuming the actual observation matrix $\hat{Y}$ and prediction matrix Y are two $m \times L$ matrices (with m number of samples and L number of classes), GCC computes correlation between both matrices. An element in $\hat{Y}_{jl}$ is equal to 1 if j th sample belongs to class l and 0 otherwise. Similarly in prediction matrix Y_{jl} , an element is 1 if j th sample is

(a) GCM Dataset with F-score

(b) GCM Dataset with KW-score

(c) 11 Tumor Dataset with F-score

(d) 11 Tumor Dataset with KW-score

(e) MLL Dataset with F-score

(f) MLL Dataset with KW-score

Figure 4.1: Average misclassification error vs. number of genes on various multiclass gene expression datasets. The plots show performance of proposed approaches with their basic counterparts

predicted to belong to class l and 0 otherwise. GCC is defined as

(a) NCI(Ross) with F-score

(b) NCI(Ross) with KW-score

(c) NCI(Staunton) Dataset with F-score

(d) NCI(Staunton) Dataset with KW-score

(e) Lung Dataset with F-score

(f) Lung Dataset with KW-score

Figure 4.2: Average misclassification error vs. number of genes on various multiclass gene expression datasets. The plots show performance of proposed approaches with their basic counterparts

$$GCC = \frac{COV(\hat{Y}, Y)}{\sqrt{COV(\hat{Y}, \hat{Y})} \sqrt{COV(Y, Y)}} \quad (4.9)$$

Figure 4.3: Bar-charts indicating average and standard deviation of skewness and kurtosis for given dataset to estimate non-Gaussianity

where $COV(\hat{Y}, Y)$, $COV(\hat{Y}, \hat{Y})$, and $COV(Y, Y)$ are covariances of the corresponding matrices, defined as arithmetic average of the covariance of corresponding columns of the matrices. Similar to the Pearson's correlation coefficient and MCC, GCC values vary between $[-1, 1]$ range. Higher the GCC value, better the prediction is.

To determine the number of genes needed for optimum classification with given gene rank list, we tested gene subsets starting from top ranked gene and then successively adding one gene at a time in the testing subset till the total number of genes in subset equals 400. The number of genes corresponds to the subset where the classification performance is maximum.

In all testing methods, we have used LIBSVM - 2.84 software [18] with OVA formu-

Figure 4.4: Bar-charts indicating average and standard deviation of skewness and kurtosis for given dataset to estimate non-Gaussianity

lation. We employed linear SVM for testing top ranked gene subsets. Performance of linear SVM depends on sensitivity parameter η values, hence for results shown in Table 4.2 and 4.3, η was tuned from finite set $\{2^{-20}, \dots, 2^0, \dots, 2^{15}\}$.

4.3.3 Results

In this chapter, we present two Pareto-front approaches with three ranking criteria for multiclass gene selection. Table 4.2 and Table 4.3 show the performance of classification with proposed approach over six widely used benchmark datasets. For the sake of clarity, we only show the F-score and KW-score with crowding distance based method. We employed one-sided t-test to determine if the improvement in classification performance

is statistically significant. By evaluating top 400 genes in forward manner from each method in each of train/test split of dataset, we obtain average missclassification error. Such curves are depicted in Figure (4.1) and (4.2).

As seen in Table 4.2, F-PFA approach significantly improves classification performance except 11 Tumor dataset. As depicted in error curves, the F-PFA converges to minimum error rate faster than other criteria in all datasets. The error curves also shows that except Lung and NCI(Staunton) datasets, performance of Fsum and Favg is similar to F-score method. Performance of Favg is worsened in GCM dataset compared to F-score method.

Similar to F-PFA, KW-PFA also significantly improves the performance of classification compared to its counterpart, KW-score criterion in GCM, NCI(Ross), NCI(Staunton), and MLL dataset. The performance with KWavg and KWsum is only marginally better compared to KW-score in MLL dataset. As seen from the error curves in most dataset, KW-PFA also converges faster to minimum average misclassification error.

Comparing results shown in Table 4.2 and Table 4.3, KW-score performs better classification than F-score in 11 Tumor, Lung, and NCI(Staunton) datasets. KW-score is a non-parametric criterion and hence robust. On the other hand, F-score assumes Gaussianity of the data in a given class. It is indeed important to validate this assumption as it can have significant impact on the performance of classification. We test Normality assumption using skewness and kurtosis.

Gene expression datasets are high dimensional. Measuring multivariate skewness and kurtosis of such datasets is difficult due to singularity of the covariance matrix [88]. Hence we measure the skewness and kurtosis of top 500 genes individually for each class. Top 500 genes were obtained using F-score over whole dataset. For a given dataset, average skewness and kurtosis with standard deviation in each class are depicted in Figure (4.3) and Figure (4.4).

Close to zero skewness implies that the gene expression values are relatively evenly distributed on both side of the mean. The Figure (4.3e)-(4.3f) and Figure (4.4a)-(4.4b) suggest that NCI(Ross), MLL, and to some extent GCM dataset relatively follows the Gaussianity assumption. However, close to zero skewness does not necessarily mean the distribution is symmetric. In case of heavy-tailed or low-tailed or too peaked or too flat, the distribution can depart from normality even though skewness is almost zero [124]. Such departures may be detected by kurtosis, the fourth moment of normal distribution. Higher skewness and kurtosis value can together indicate the departure from normality.

Figure 4.5: Class-specific statistics for each class in MLL cancer dataset for top 100 genes obtained from F-score, KW-score, and PFA approaches

In our analysis, the 11 Tumor and Lung dataset has higher skewness as well as kurtosis for some classes. This suggests that these datasets do not follow the Gaussianity assumption and hence better performance is achieved by KW-score based methods than F-score and related methods. We applied t -test to validate if KW-score is better than F-score in Lung and 11 Tumor dataset, and the p -value for such comparison for both datasets were less than 0.001.

4.4 Visualization

In the proposed approaches, we treat each component of the F-score and the KW-score individually. To visualize the effect of such approach on top ranking genes, we employ the heat maps. Heatmaps help to visualize the over-representation and under-representation of genes for given class. As F-score values varies in different range for different classes, we scale each of the component and then plot the heatmap. Note that F-score and KW-score are aggregation criteria and hence, we have plotted their components. In Figure(4.5) to Figure(4.10), we depict objective values of top 100 genes obtained from F-score, KW-score, F-PFA and KW-PFA gene ranking methods for all datasets.

Figure 4.6: Class-specific statistics for each class in Lung cancer dataset for top 100 genes obtained from F-score, KW-score, and PFA approaches

Figure 4.7: Class-specific statistics for each class in 11 Tumor cancer dataset for top 100 genes obtained from F-score, KW-score, and PFA approaches

Figure 4.8: Class-specific statistics for each class in GCM cancer dataset for top 100 genes obtained from F-score, KW-score, and PFA approaches

Figure 4.9: Class-specific statistics for each class in NCI(Staunton) cancer dataset for top 100 genes obtained from F-score, KW-score, and PFA approaches

Figure 4.10: Class-specific statistics for each class in NCI(Ross) cancer dataset for top 100 genes obtained from F-score, KW-score, and PFA approaches

In MLL dataset, the task is to classify three classes, acute lymphoblastic leukemia (ALL), mixed-lineage leukemia (MLL), and acute myeloid leukemia (AML). These classes are denoted by 1, 2, and 3 respectively. As seen from the Figure (4.5), top 100 genes are under-representing MLL class while over represented particularly in AML class. The F-PFA and KW-PFA try to balance the gene representation from each class, which may be resulting in improvement in classification performance.

The Lung cancer dataset has 203 samples, which has five classes, adenocarcinoma (AD), normal lung (NL), small cell lung cancer (SCL), squamous cell carcinoma (SCC), and pulmonary carcinoid (PC). Figure (4.6) suggests that AD and PC dominate gene ranking over other classes which were balanced by the proposed PFA approaches.

The GCM dataset is the most challenging one as it contain 14 classes, namely, breast adenocarcinoma (BR), prostate adenocarcinoma (PR), lung adenocarcinoma (LU), colorectal adenocarcinoma (CR), lymphoma (LY), bladder transitional cell carcinoma (BL), melanoma (ML), uterine adenocarcinoma (UT), leukemia (LE), renal cell carcinoma (RE), pancreatic adenocarcinoma (PA), ovarian adenocarcinoma (OV), pleural mesothelioma (ME) and central nervous system (CNS) and area denoted from 1 to 14. As seen the plots, BR, BL, and CNS are dominating the F-score criterion. The F-PFA and KW-PFA plots look much more balance for each class.

In case of 11 Tumor dataset, as seen from Figure (4.7), there is no significant over-representation and under-representation of genes in F-score or KW-score. Hence, the PFA approaches performs similar to the F-score and KW-score. As shown in Figure (4.9) and Figure (4.10), there is no class clearly dominating other classes in NCI (Ross) and NCI (Staunton) datasets. However the relevant PFA plots of top-100 genes objective statistics values represent better balance of genes for each classes, resulting in better performance in classification.

4.5 Discussion

In this chapter, we have investigated the hypothesis that easy class separation of one class from others can affect gene ranking, and may reduce overall classification performance. The primary reason for low classification performance is that the genes responsible for difficultly separable classes may not be represented adequately. To alleviate this problem, we presented Pareto-front analysis based F-score and KW-score ranking method for gene selection in multiclass problem. To rank genes within given Pareto front, we investigated performance of three criteria, weighted sum of objectives, average rank, and crowding distance.

The proposed approach explores the “One-against all” components of given ranking criteria, such as F-score and KW-score. The Pareto-front analysis is performed using such components. As depicted by visualization plots, the proposed approach balances the representation of each class in objective space, often resulting in better classification performance. Moreover, the proposed approaches have inherent characteristics of removing highly correlated genes in objective space. With PFA approach, the genes which have same class-wise objective statistics value with top ranking gene will be downgraded to lower Pareto-fronts. This redundancy removal is different from earlier methods, which computes redundancy directly from the gene expression values.

Computing multivariate Gaussianity of a high dimensional dataset is a challenge. Hence using marginal skewness and kurtosis, we investigated if normality assumption is not true or not. In some datasets, KW-score based approaches were better than F-score based methods, which can be attributed to non-Gaussianity of data.

In the proposed method, increase in number of classes, i.e. number of objectives, may result in large number of genes in first front itself. To avoid this, we used one-vs.-all approach for individual class wise objective computation. If there are L classes, one-vs.-all approach will result in L objective values, compared to $L(L - 1)/2$ objective values

with one-vs.-one approach. Due to this strategy, we were able to reduce number of genes in each front.

As discussed in [26], non-dominated sorting with NSGA-II has computational complexity of $O(Ln^2)$, where L is number of objectives and n is population size. In gene ranking context, n is number of genes in dataset. Though another fast algorithm with computational complexity of $O(n \log^{L-1} n)$ had been proposed [62], because of very large number of genes in microarray datasets, the computational complexity still remains high with the proposed methods compared to standard F-score and KW-score.

In conclusion, we have presented a new F-score and KW-score based Pareto-front methods for gene ranking in multiclass problem. This method was developed by treating class-wise F-score and KW-score as multiple objectives and comparing genes with Pareto analysis. Though the performance was not consistent across all datasets, it seems to select fewer genes and/or performs better on most of the datasets.

Chapter 5

Stability, Redundancy, and Biological Analysis

Recent developments in evaluation of performance of a gene selection algorithm stress for estimation of error by either bootstrapping or unbiased cross-validation. Variations in samples result in significantly different gene lists. The reproducibility or stability of gene selection algorithms is extremely important if the selected genes are to undergo further biological experiments. At the same time, gene selection methods should give minimally redundant genes for better classification performance. In this chapter, we evaluate stability and redundancy of both two-class and multiclass top-ranked genes obtained from various gene selection methods developed in this thesis. The gene lists are obtained from the standard T-score, SVM-RFE, SV t -RFE, and T-SS for two-class problems and the F-score, KW-score, and PFA approaches for multiclass classification. We also present a method to identify consensus genes and biologically validate them with gene ontology.

5.1 Motivation

Identification of a true gene marker is one of the ultimate goals of a gene selection method. A true gene marker should be able to distinguish cancer samples from a benign samples for example. Non-reproducibility of gene ranking methods casts doubts over the computational results and hence does not give confidence in further biological validation. The instability of a gene selection method could be due to the existence of multiple number of weakly relevant genes and the small number of available training samples. These samples may be further complicated if they are multimodal. For reliable and

stable gene selection, one needs sample points in order of thousands [35]. On the other hand, all the samples may not be equally relevant for classification. The problem we address here is how relevant genes are stably selected while having a good performance in classification.

Boulesteix and Slawski presented an excellent review on stability of gene ranking methods [13]. There are three ways to measure the stability over various rank lists: (i) considering all the rank lists simultaneously, (ii) measuring the pair-wise stability and averaging over total number of rankings, and (iii) comparing with the reference rank list, if available. Many stability criteria measure the ratio of intersection to the union of two sets [100], [133], [188]. A variant called the normalized Percent of Overlapping Genes (nPOG) criterion additionally includes highly correlated genes while measuring stability [179]. Dice-Sorensen's index [100], [188] and Ochiai's index [188], relative Hamming distance [34], Tanimoto distance metric [141] are a few examples of various distance based stability criteria. The Kuncheva index [74] assumes that both the gene set have same number of genes. The stability measure by Kuncheva index also corrects a bias of a common gene in two list occurring randomly [1], [74]. These criteria only measure the stability of a gene subset, ignoring the rank information of genes in a given list. For a typical feature ranking list, Spearman's correlation coefficient [68], Canberra distance [65], or overlap score [170] uses the rank information as well for the estimation of stability. Hyper-geometric distribution can be used as well to derive an overlapping probability of top ranking gene lists [48]. The advantage of such method is that it provides a p -value cutoff for top ranking genes.

In earlier chapters, we have introduced several novel algorithms for gene selection. The study of their stability will be investigated together with the relevance of selected genes. It is of no practical value if we get different sets of genes due to variations of samples. Also the performance of classification should remain high irrespective of variations of samples. Hence, we are interested in deriving a gene set which is not only very stable but also has desirable classification performance.

The third important issue in evaluating the gene subset is the redundancy analysis. Genes involved in similar biological functions, pathways, and processes tend to have similar expressions, giving rise to redundancy among them. Given a gene expression dataset, our aim is to identify all the relevant genes and their associated processes. Due to redundancy, some of the important, but weakly expressed compared to highly expressed genes, can not be detected. Apart from not providing any new biological information,

such genes many a times deteriorate the performance of classification. Hence, it is in both biological and computational interest to have a stable, yet minimally redundant, gene subset that can accurately identify cancerous tumors. However, removal of redundant genes may miss important genes of biological relevance.

In this chapter, we evaluate previously developed gene selection techniques for both two-class and multiclass problems for their stability and redundancy, along with biological validation. This is necessary part of this research as without such analysis the methods are not applicable for large scale clinical studies. For stability analysis, we use the Kuncheva index as we are interested in true intersection of two ranking lists without considering correlation. Pair-wise absolute correlation coefficient is used for the redundancy analysis. Also, biological validation of genes is necessary to verify that computational method indeed deliver biologically relevant results. To do so, we present a method to identify consensus genes and validate them biologically.

5.2 Stability Analysis

We take a similarity based approach to determine the stability of a gene selection method. It is measured by the average over all pair-wise similarity comparisons among all the ranked gene lists obtained by a particular method over different subsets of training samples [74].

Let $\{D^b\}_{b=1}^B$ be a set of B sub-samplings of the dataset and R^b be the b -th rank list of genes. The stability $\mathcal{S}_D$ of the method over the dataset D is given by

$$\mathcal{S}_D = \frac{2}{B(B-1)} \sum_{b=1}^B \sum_{b'=b+1}^B \mathcal{S}(R^b, R^{b'}) \quad (5.1)$$

where $\mathcal{S}(R^b, R^{b'})$ is a similarity measure between the rank gene lists R^b and $R^{b'}$ for top n^* genes in both lists. One of the popular measures to find similarities between two gene lists is the Kuncheva index [74] which is given by

$$\mathcal{S}(R^b, R^{b'}) = \frac{|R^b \cap R^{b'}| - n^{*2}/n}{n^* - n^{*2}/n} \quad (5.2)$$

where n denotes total number of genes in a dataset and n^* is the set of the top genes.

The Kuncheva index is similar to the Jaccard index and ranges between $[-1, 1]$ with a large value indicating a large number of common genes between subsets. A negative Kuncheva index denotes that overlap between the two subsets is due to chance. The term n^{*2}/n corrects for a bias due to the chance of selecting common features between two randomly chosen subsets.

(a) Colon

(b) Leukemia

(c) Hepato

(d) CNS

(e) Prostate

(f) Breast

Figure 5.1: Stability of ranked gene lists obtained by T-score, SVM-RFE, SVt-R, SVt-RFE, and T-SS methods over two-class benchmark datasets using Kuncheva index

5.3 Redundancy Analysis

The redundancy among top-ranked genes of the proposed novel approaches are analyzed here. In what follows, we describe analysis of redundancy for two-class and multiclass

problems.

5.3.1 Redundancy in Two-Class Gene Selection

Gene ranking with T-SS and SVt-RFE is determined by the T-score but by using only the selected samples. The standard T-score criterion treats each gene independently from others, resulting in zero penalty for redundant genes. However, unlike the standard T-score, the SVt-R, SVt-RFE, and T-SS depend on selection of samples. The rank statistics does not remain constant and changes according to the selected samples in recursive ranking. The sample selection itself may induce some penalty for redundancy among genes.

The standard T-score is given by

$$r_i = \frac{(\mu_i^+ - \mu_i^-)}{\sigma_i^2} \quad (5.3)$$

Assuming an equal number of samples in both classes $m^+ = m^- = m/2$, (5.3) is written as,

$$r_i = \frac{2}{m\sigma_i^2} \left(\sum_{j=1}^m y_j x_{ij} \right) \quad (5.4)$$

In (5.4), all the sample points are used to compute gene ranking criterion, independent of other genes. However, in Chapter 3, we proposed to use a sample selection criterion, which first selects sample, and uses only those to compute the T-score. With sample selection, the standard T-score can be represented as

$$r_i = \frac{2}{m\sigma_i^2} \left(\sum_{j=1}^m \alpha_j y_j x_{ij} \right) \quad (5.5)$$

Though we use the α_j as either zero or one, corresponding to sample being relevant or irrelevant, the α_j can be given by,

$$\alpha_j = \frac{\exp(\lambda(1 - y_j r_i^T x_j))}{1 + \exp(\lambda(1 - y_j r_i^T x_j))} \quad (5.6)$$

which can be re-written as

$$\alpha_j = 1 - \frac{1}{1 + \exp(\lambda(1 - y_j \sum_i r_i x_{ij}))} \quad (5.7)$$

Figure 5.2: Effect of margin on parameters (a and f) of tangent approximation of sigmoid function

To demonstrate the effect of such sample selection criteria on r_i , we follow the work of [83] in which (5.7) was first simplified to a non-linear equation using sigmoid function $\text{sig}(e) = 1/[1 + \exp(-e)]$. The new equation is given by,

$$\alpha_j = 1 - \text{sig} \left(\lambda \left(y_j \sum_i r_i x_{ij} - 1 \right) \right) \quad (5.8)$$

However, (5.8) is still a non-linear equation. Using tangent lines such as $\text{sig}(e) = a_e e + f_e$, we replace (5.8) with simpler equation where, $a_e = [\text{dsig}(e)]/de$ and $f_e = \text{sig}(e) - a_e e$. Hence, $\text{sig}(\lambda(y_j \sum_i r_i x_{ij} - 1))$ is written as $a_j \lambda y_j \sum_i r_i x_{ij} - a_j \lambda + f_j$. As a_j is function of $y_j \sum_i r_i x_{ij}$, it is given by

$$a_j = \frac{\exp(\lambda(1 - y_j \sum_i r_i x_{ij}))}{[1 + \exp(\lambda(1 - y_j \sum_i r_i x_{ij}))]^2} \quad (5.9)$$

The Figure (5.2) shows the effect of margin value $y_j \sum_i r_i x_{ij}$ on a_j . As the margin of a sample varies from negative to positive, the value of a_j changes from 0 to 0.25 and then back to 0. The Figure (5.2) also depicts effect of margin value $y_j \sum_i r_i x_{ij}$ on f_j .

With sigmoid approximation with tangent lines, (5.8) is re-written as

$$\alpha_j = 1 - a_j \lambda \left(y_j \sum_i r_i x_{ij} \right) + a_j \lambda - f_j \quad (5.10)$$

Substituting (5.10) into (5.5), we obtain

$$r_i = \frac{2}{m\sigma_i^2} \left(\sum_{j=1}^m \left(1 - a_j \lambda \left(y_j \sum_{i'=1}^n r_{i'} x_{i'j} \right) + a_j \lambda - f_j \right) y_j x_{ij} \right) \quad (5.11)$$

which is rearranged as,

$$r_i = \frac{2}{m\sigma_i^2} \left(\sum_{j=1}^m (1 + a_j \lambda - f_j) y_j x_{ij} - a_j \lambda \left(\sum_{j=1}^m \sum_{i'=1}^n x_{ij} x_{i'j} r_{i'} \right) \right) \quad (5.12)$$

Even though (5.12) is derived using standard T-score, it clearly suggests that the T-SS method indeed penalizes for redundancy. Considering that $1 - f_j \geq 0$ and assuming that λ is sufficiently large, (5.12) has some interesting properties that are discussed as follows:

Case 1

As shown in Figure (5.2), for samples with margin less than one, both a_j and f_j are close to zero. Substituting these values in (5.10), we obtain that $\alpha \cong 1$. It suggests that all the samples with less than one margin are selected and such points does not effect the redundancy. It also means that (5.4) and (5.12) are equivalent and such samples will not have any effect on gene redundancy.

Case 2

If the samples have margin close to one, the a_j reaches maximum value of 0.25 while f_j nears 0.5. The (5.12) suggests that samples with close to one margin will maximally penalize the redundancy among genes.

Case 3

The third case is with sample having greater than one margin value. For such cases, the a_j becomes close to zero. On the other hand, f_j becomes close to one. The (5.10) denotes that for such cases $\alpha \cong 0$. Obviously, these points do not take part in computation of T-score, and hence have no effect on redundancy.

In previous Chapter, we do not distinguish between points on the margin and points with in the margin. However, the relevant sample set still penalize partially for the redundancy among top-ranked genes.

5.3.2 Redundancy in Multiclass Gene Selection

Similar to two-class problems, a gene (feature) is considered to be redundant if its gene expression values are significantly correlated with another gene. A selection of highly redundant gene is undesirable. The aggregation of class-specific scores discussed in Chapter 4 leads to the selection of redundant genes.

Due to aggregation based ranking criterion, redundant genes are ranked higher. For example, if genes i and i' are only relevant to a class ℓ , the aggregation based criterion may not be able to detect it and rank both genes closer. However, a PFA approach compares the class-specific statistics and divide genes into various fronts based on definition of non-dominance. Demoting one of the gene in higher Pareto front may result in removal of redundancy.

5.3.3 Validation of Redundancy

To validate the analysis on both binary and multiclass problems, we evaluate the redundancy among the top-ranked genes. The concept of redundancy is derived from how genes are correlated to one another in a given dataset. We use the absolute value of Pearson's correlation coefficient to estimate the redundancy among top-ranked genes in a given dataset. In a gene rank list R^b , we first measure a pair-wise correlation coefficient of top n^* genes, resulting in a $n^* \times n^*$ correlation matrix with each element representing pair-wise similarity. Using the upper triangular matrix, we obtained average of absolute pair-wise correlations, which represents redundancy among those n^* top-ranked genes in rank list R^b . This value is averaged over the total number of gene rankings, i.e., number of bootstrapped trials (B).

Mathematically, the average redundancy among top n^* genes over B trials can be given by,

$$\overline{Q} = \frac{1}{B} \sum_{b=1}^B \frac{2}{n^* (n^* - 1)} \sum_{i=1}^{n^*-1} \sum_{i'=i+1}^{n^*} |\rho(x_i, x_{i'})| \quad (5.13)$$

where $|\rho(x_i, x_{i'})|$ is absolute value of Pearson's correlation coefficient between expression values of gene i and i' . While evaluating the performance of classification, we obtained 100 and 400 ranked gene lists in total for each two-class and multiclass dataset respectively. We computed redundancy among genes with 100 top ranked genes for two-class and upto 200 genes for multiclass gene selection.

5.4 Identification of Consistent Genes

In Chapter 3 and 4, we employed five fold cross-validation for 20 times to evaluate the performance of two-class gene selection methods and four fold cross validation for 100 times to evaluate multiclass gene selection methods, resulting in a large number of gene ranking lists for each datasets. Ideally, a biologist is interested in same genes listed from all the trials with good classification performance. However, due to variations in sampling and in methodological considerations, we obtain significantly different gene rankings for each run of a cross-validation. This is also highlighted in stability analysis. As a biologist can afford to perform actual biological validation on a few genes only, it is of paramount importance to select top-ranked consensus genes consistently across different rank lists.

One of the usual and commonly used approaches is to obtain important consensus genes from their frequency of selection in the top n^* number of genes. The shortcoming of this approach is that it completely disregards the rank information of the top n^* genes. For example, if we are looking into top 100 genes, the gene ranked at 1 and at 100 are treated same. To incorporate the rank information in frequency based methods, we propose following simple approach for obtaining important genes for further biological validation.

Let $\mathcal{R}_{i,R^b}$ be the rank of i -th gene in R^b rank list. We are interested in finding consensus genes from top n^* ranked gene from each rank list. The weight to each gene i is given by

$$\omega_{ib} = \begin{cases} n^* - \mathcal{R}_{i,R^b} + 1 & \text{if } n^* \geq \mathcal{R}_{i,R^b} \\ 0 & \text{otherwise} \end{cases} \quad (5.14)$$

The (5.14) suggests that in a given rank list, the top most ranked gene will get weight of n^* , and the subsequent genes in that list will have weights $n^* - 1, n^* - 2, \dots, 0$. After computing such weights for all genes from each ranked list, we average these values over all bootstrapped lists (B). The final weight of consistency is then given by,

$$\Omega_i = \frac{1}{B} \sum_{b=1}^B \omega_{ib} \quad (5.15)$$

The benefit of using (5.15) is that it can incorporate the rank information along with the frequency of being in top-ranked gene subset. A gene with higher average weight indicates that it is often found in the top rankings and present in the most of the rank lists. This gives more confidence to validate it further. For biological validation, we analyzed top 100 genes for two-class problem and top 400 genes for multiclass problem.

Figure 5.3: Stability of top-ranked genes obtained by F-score, KW-score, and PFA approaches over various multiclass datasets

5.5 Results

5.5.1 Stability

For two-class benchmark datasets, Figure (5.1) depicts the stability performance of the standard T-score, SVM-RFE, and the proposed approaches: SVt-RFE, SVt-R, and T-

SS using the Kuncheva index. From the Figure, we observe few interesting trends. The stability of gene lists obtained from standard T-score is the highest across all the six benchmark datasets. Second, the stability of SVM-RFE gene lists is significantly low for top-ranked genes, which is undesirable. The proposed approaches have stability ranging in between the standard T-score and the SVM-RFE. While the stability of gene rank lists obtained from T-SS nears standard T-score, *SVt*-RFE gene list results are more similar to SVM-RFE. Except Prostate cancer dataset, *SVt*-RFE and T-SS have more stable top-ranked genes. The stability at top rank is of greater significance for biological validation.

Figure (5.3) show the stability of the gene lists obtained from F-score, KW-score, and their Pareto front based counterparts over each benchmark multiclass datasets used in Chapter 4. The general trend is that F-score and KW-score give more stable genes compared to PFA approaches. While comparing PFA approaches, KW-PFA has resulted in more stable gene lists than FPFA in all six datasets. This is due to the fact that KW-PFA is based on non parametric method and hence less susceptible to sample outliers.

5.5.2 Redundancy

We measure the average redundancy among top-ranked genes in a given rank list using the absolute value of Pearson's correlation coefficient. For two-class datasets, we have hundred gene lists obtained from standard T-score, SVM-RFE, *SVt*-RFE, *SVt*-R, and T-SS methods. The average redundancy among top hundred genes across hundred gene list is shown in Figure (5.4). The results show that (i) the standard T-score produces highly redundant top-ranked genes among all five methods, (ii) the SVM-RFE results in less redundant top-ranked genes, and (iii) the proposed approaches with sample selection have intermediate redundancy to that of T-score and SVM-RFE. Among proposed approaches, the T-SS gave highly redundant genes and approaches standard T-score for redundancy. The performance of *SVt*-RFE with respect to redundancy is similar to SVM-RFE, but on higher side.

Particularly for T-SS, we also studied the effect of λ parameter on redundancy among top-ranked genes. The results are shown in Figure (5.5). Apart from Prostate and Breast cancer datasets, the results show that increase in λ results in decrease in redundancy among top-ranked genes. As seen from the figures, an increase in λ resulted in less number of selection of relevant sample points. The decrease in redundancy may be attributed by less number of relevant sample points.

Figure 5.4: Redundancy among the top-ranked genes obtained by T-score, SVM-RFE, SVt-R, SVt-RFE, and T-SS methods over various two-class datasets

For multiclass benchmark datasets, the absolute value of Pearson's correlation coefficient was used to measure the redundancy among top-ranked genes obtained from F-score, KW-score, and the proposed PFA approaches. As shown in Figure (5.6), the top-ranked

Figure 5.5: Effect of λ on redundancy of top-ranked genes obtained by T-SS over various two-class datasets

genes obtained by using FPFA has significantly low redundancy in all datasets. The KW-PFA also has similar trend except MLL dataset. The F-score and KW-score result in highly redundant top-ranked genes.

Figure 5.6: Redundancy among the top-ranked genes obtained by F-score, KW-score, and PFA approaches over various multiclass datasets

5.5.3 Biological Validation

Cancer is generally caused by unregulated cell divide and growth, which results in lumps of tissues, or tumors. Cell growth and divide are dependent on some key processes

and pathways such as the MAPK pathway, the TGF- β pathway, cell adhesion, post transcriptional regulation etc.

For example, the cytokine transforming growth factor β (TGF- β) signaling pathway stops the cell cycle at the G1 stage to stop proliferation, induce differentiation, or promote apoptosis in normal cells. In cancer cells, this pathway is mutated and hence it can not control the cell resulting in cancer cells proliferation. This ultimately causes immunosuppression and angiogenesis, which makes the cancer more invasive [11]. Similarly, in many cancers, a defect in the MAPK pathway leads to uncontrolled growth [29]. Alterations in regulatory sequences, RNA-binding proteins, or in upstream signalling pathways affect the stability and/or translational efficiency of mRNAs encoding proto-oncogenes, cytokines, cell cycle regulators and other regulatory proteins to promote tumorigenesis and cancer progression [6], [105]. In our analysis, we assume that a gene is biologically relevant and important if it is involved in such pathways.

The commonly selected genes for colon cancer dataset are shown in Table 5.1 which includes three unknown expressed sequence tags (EST). These clones are identified by Hsa.6814, Hsa.5392, and Hsa.37937, which are described as Collagen Alpha 2(XI) Chain, 60S Ribosomal Protein L24, and Myosin Heavy Chain Nonmuscle respectively. Another gene known as SREBF1, which is important in sterol regulation, cellular response to starvation, cholesterol, lipid and steroid metabolic process, was only identified with SVM-RFE. We found that CSRP1 has three variants in the datasets, and our method was able to identify all of them. SVM-RFE was able to select only one of them as they might have similar expression values. Some of the biologically important genes which were identified with proposed approaches are listed in Table 5.2 and Figure (5.7a) shows the differential weights of many such genes.

In the leukemia dataset, some of the top-ranked genes include SUMO1, MLLT3, PT-PRN2 etc., which are directly involved in many important processes such as transcription regulation and DNA repair. Table 5.3 also shows three commonly selected mRNA sequences which are again important in DNA transcription and various ion bindings. Table 5.4 represents some of top-ranked genes obtained by using proposed approaches. These genes got significantly higher weights but ignored in SVM-RFE. Figure (5.7b) shows the differential weights of these genes. The genes RPSA, KRT16, and ALDH3B2 are important in cell adhesion, cell proliferation, and lipid metabolic process, oxidation reduction pathway respectively.

Table 5.1: Functional Description of Commonly Selected Genes from the Colon Cancer Dataset Across All Two-class Gene Ranking Methods

Gene Name	Description	Pathway/Process
GUCA2B	H.sapiens mRNA for GCAP-II/uroguanylin precursor	body fluid secretion; cGMP biosynthetic process; excretion; regulation of cyclic nucleotide biosynthetic process
Unknown	...	...
CSRP1	Human cysteine-rich protein (CRP) gene; metal ion and protein binding	...
Unknown	...	...
Unknown	...	...
MYL9	myosin, light chain 9, regulatory	muscle contraction; regulation of muscle contraction

Table 5.2: Functional Description of Exclusive Important Genes from the Colon Cancer Dataset by Proposed Methods

Gene Name	Description	Pathway/Process
DES	Human desmin gene	cytoskeleton organization; muscle contraction; regulation of heart contraction
KCNMB1	Human MaxiK potassium channel beta subunit mRNA	ion transport;potassium ion transport; response to calcium ion; synaptic transmission
HSPD1	Mitochondrial matrix protein P1 precursor;	cell death; T-cell activation; activation of caspase activity; positive regulation of apoptosis;positive regulation of interferon-alpha, interferon-beta, interferon-gamma, interleukin-10 production

Table 5.3: Functional Description of Commonly Selected Genes from the Leukemia Cancer Dataset Across All Two-class Gene Ranking Methods

Gene Name	Description	Pathway/Process
SUMO1	Human ubiquitin-homology domain protein PIC1 mRNA	DNA repair; negative regulation of DNA binding, transcription, and transcription factor activity; positive regulation of proteasomal ubiquitin-dependent protein catabolic process and of protein complex assembly; protein sumoylation; regulation of protein localization
MLLT3	myeloid/lymphoid or mixed-lineage leukemia	anterior/posterior pattern formation; regulation of transcription, DNA-dependent; segment specification
Unknown	Human endogenous retrovirus envelope region mRNA	DNA integration; DNA Binding
Unknown	Homo sapiens 5,10-methenyltetrahydrofolate synthetase mRNA	folic acid and derivative biosynthetic process; formate metabolic process; magnesium ion binding; ATP binding
Unknown	Human proto-oncogene (BMI-1) mRNA	regulation of transcription, DNA-dependent
PTPRN2	protein tyrosine phosphatase, receptor type	protein dephosphorylation

Table 5.4: Functional Description of Exclusive Important Genes from the Leukemia Cancer Dataset by Proposed Methods

Gene Name	Description	Pathway/Process
CHIC1	cysteine-rich hydrophobic domain 1	...
RPSA	ribosomal protein SA	ribosomal protein SA; cell adhesion; cell surface receptor linked signal transduction
ALDH3B2	aldehyde dehydrogenase 3 family, member B2	alcohol metabolic process; cellular aldehyde metabolic process; lipid metabolic process; oxidation reduction
KRT16	keratin 16(focal non-epidermolytic palmoplantar keratoderma)	cell proliferation; epidermis development

(a) Colon

(b) Leukemia

(c) Prostate

(d) Breast

(e) Hepato

(f) CNS

Figure 5.7: Differentially weighted genes using proposed approaches on two-class classification datasets

Table 5.5: Functional Description of Commonly Selected Genes from the Breast Cancer Dataset Across All Two-class Gene Ranking Methods

Gene Name	Description	Pathway/Process
GMPS	Guanosine 5' - Monophosphate Synthase	glutamine metabolic process; purine base and purine nucleotide biosynthetic process
HMGA1	high mobility group AT-hook 1	DNA unwinding involved in replication; initiation of viral infection; loss of chromatin silencing; nucleosome disassembly; positive regulation of transcription
ACOT2	acyl-CoA thioesterase 2	acyl-CoA metabolic process; lipid metabolic process
GSTP1	glutathione S-transferase pi 1	anti-apoptosis; central nervous system development; Drug metabolism - cytochrome P450 pathway; KEGG cancer pathway
CTSC	cathepsin C	aging; immune response; proteolysis; Lysosome pathway

Table 5.6: Functional Description of Exclusive Important Genes from the Breast Cancer Dataset by Proposed Methods

Gene Name	Description	Pathway/Process
XPB1	X-box binding protein 1	Protein processing in endoplasmic reticulum pathway; immune response; regulation of transcription, DNA-dependent
FOXA1	Human hepatocyte nuclear factor-3 alpha (HNF-3 alpha)	branching morphogenesis of a tube; chromatin remodeling; embryo development; lung development; regulation of hormone levels;
YBX1	Y box binding protein 1	RNA splicing; regulation of transcription, DNA-dependent
COX6C	cytochrome c oxidase subunit	Alzheimer's disease; Cardiac muscle contraction; Metabolic pathways; Oxidative phosphorylation; generation of precursor metabolites and energy
TMBIM6	transmembrane BAX inhibitor motif containing 6	lung development; negative regulation of apoptosis

Table 5.5 shows the commonly selected important genes in Breast cancer dataset. These genes are involved in some of the key biological processes such as nucleotide biosynthetic process, DNA unwinding, loss of chromatin silencing, nucleosome disassembly, positive regulation of transcription, aging, anti-apoptosis etc. A small value of λ has improved the performance of T-score, which suggests that many genes will have similar weights to T-score. However, it is equally important to look into some genes whose weight is reduced in T-SS. Figure (5.7d) shows that the first two genes, XBP1 and FOXA1, are highly weighted using T-score method, but are low weighted by sample selection approaches including SVM-RFE. On the other hand, other genes in Table 5.6 are equally weighted by T-score and T-SS method.

Many commonly selected genes across all methods in CNS datasets are shown in Table 5.7. These genes are involved in some key cancer related pathways and processes such as activation of Kinase, cell differentiation, Neurotrophin signaling pathway, positive regulation of transcription. Some of the exclusively highly weighted genes obtained from proposed approaches include COL6A3, PCOLCE, MMP2 etc. These genes take part in important processes such as focal adhesion process, muscle organ development, Bladder cancer pathways, GnRH signaling pathway etc.

Table 5.7: Functional Description of Commonly Selected Genes from the CNS Cancer Dataset Across All Two-class Gene Ranking Methods

Gene Name	Description	Pathway/Process
HMGA1	high mobility group AT-hook 1	DNA unwinding involved in replication; initiation of viral infection; loss of chromatin silencing; nucleosome disassembly; positive regulation of transcription
NTRK3	neurotrophic tyrosine kinase, receptor, type 3	Neurotrophin signaling pathway; cell differentiation; mechanoreceptor differentiation; nervous system development; transmembrane receptor protein tyrosine kinase signaling pathway
GAP43	growth associated protein 43	activation of protein kinase C activity by G-protein coupled receptor protein signaling pathway; cell differentiation; nervous system development
APOD	apolipoprotein D	lipid metabolic process

Table 5.8: Functional Description of Exclusive Important Genes from the CNS Cancer Dataset by Proposed Methods

Gene Name	Description	Pathway/Process
COL6A3	collagen, type VI, alpha 3	ECM-receptor interaction; Focal adhesion; muscle organ development
PCOLCE	procollagen C-endopeptidase enhancer	multicellular organismal development
MMP2	matrix metalloproteinase 2	Bladder cancer; GnRH signaling pathway; Leukocyte transendothelial migration; angiogenesis; blood vessel maturation; bone trabecula formation

Table 5.9: Functional Description of Commonly Selected Genes from the Hepato Cancer Dataset Across All Two-class Gene Ranking Methods

Gene Name	Description	Pathway/Process
MKNK1	MAP kinase interacting serine/threonine kinase 1	MAPK signaling pathway; Insulin signaling pathway; intracellular protein kinase cascade; protein phosphorylation; regulation of translation
TGFBR1	transforming growth factor, beta receptor 1	TGF-beta signaling pathway; Pancreatic cancer; activation of MAPKK activity; induction of apoptosis etc.
HAX1	HCLS1 associated protein X-1	interleukin-1 binding
HLA-DPB1	major histocompatibility complex, class II, DP beta 1	Cell adhesion molecules pathway ; Intestinal immune network for IgA production pathway; immune response process
FADD	Fas (TNFRSF6)-associated via death domain	Apoptosis pathway; activation of proapoptotic gene products; induction of apoptosis by extracellular signals; signal transduction
ADAM3B	ADAM metalloproteinase domain 3B	...

Table 5.9 represents some of the highly weighted genes across all methods for Hepato dataset. The genes have shown activity in MAPK signaling pathway, protein phosphorylation, interleukin-1 binding, TGF-beta signaling pathway, Pancreatic cancer, Apoptosis

pathway, etc. As shown in previous chapter, the T-score has better performance in classification. Hence, as shown in Figure (5.7e), some differentially selected genes have similar weights in both T-score and T-SS, but the weights are significantly reduced for SVM-RFE. These genes are also important in metabolism of nucleotides, NOD-like receptor signaling pathway, anti-apoptosis, negative regulation of I-kappaB kinase/NF-kappaB cascade, MAPK signaling pathway, B cell proliferation, apoptosis, etc.

Table 5.10: Functional Description of Exclusive Important Genes from the Hepato Cancer Dataset by Proposed Methods

Gene Name	Description	Pathway/Process
UCK2	uridine-cytidine kinase 2	Pyrimidine metabolism pathway; Metabolism of nucleotides
TNFAIP3	tumor necrosis factor, alpha-induced protein 3	NOD-like receptor signaling pathway; anti-apoptosis; negative regulation of I-kappaB kinase/NF-kappaB cascade
CNC05480	cytochrome C oxidase liver-specific subunit	Oxidative phosphorylation; aerobic respiration
MEF2C	myocyte enhancer factor 2C	MAPK signaling pathway; Myogenesis; Signaling in Immune system; Signalling by NGF; B cell proliferation; apoptosis

We obtained surprising results in biological validation of the Prostate cancer datasets. In top-ranked genes, very few have similar weights across all methods. Particularly, T-SS did not select any overlapping highly weighted gene with other methods. The commonly selected genes across T-score, SVM-RFE, and SVt-RFE are shown in Table 5.11. The differential weights are shown in Figure (5.7(c)), the top weighted genes by T-SS are not selected in T-score and SVM-RFE, and are given very low weights with SVt-RFE method. However, genes in both Table 5.11 and Table 5.12 shows involvement in important biological processes. To investigate the discrepancy in results, we plotted the first two principal components of the samples using top 100 genes obtained from T-score and T-SS. The plots shown in Figure (5.8) clearly shows that there exists bimodal distribution for one class. For this class, the samples away from the boundary are not selected in process of T-SS, resulting in totally different gene list.

Table 5.11: Functional Description of Commonly Selected Genes from the Prostate cancer dataset across T-score, SVt-RFE, and SVM-RFE methods

Gene Name	Description	Pathway/Process
CHRD1	chordin-like 1	BMP signaling pathway; cell differentiation; multicellular organismal development; nervous system development; ossification
UBE2C	ubiquitin-conjugating enzyme E2C	Ubiquitin mediated proteolysis pathway; cell division; cell cycle
ELAVL3	ELAV (embryonic lethal, abnormal vision, Drosophila)-like 3 (Hu antigen C)	cell differentiation

Table 5.12: Functional Description of Exclusive Important Genes from the Prostate Cancer Dataset by T-SS Method

Gene Name	Description	Pathway/Process
TNK1	tyrosine kinase, non-receptor, 1	negative regulation of cell growth; protein autophosphorylation
IDO1	indoleamine 2,3-dioxygenase 1	Tryptophan metabolism pathway; cytokine production involved in inflammatory response; negative regulation of T cell apoptosis; negative regulation of activated T cell proliferation; negative regulation of interleukin-10 production; positive regulation of apoptosis; positive regulation of chronic inflammatory response; positive regulation of interleukin-12 production
PKN2	protein kinase N2	signal transduction
PITRM1	pitrilysin metalloproteinase 1	proteolysis
CYB5B	cytochrome b5 type B (outer mitochondrial membrane)	electron transport chain process

Figure 5.8: Sample distribution using top 100 genes obtained from T-score and T-SS method on Prostate cancer dataset

Table 5.13: Functional Description of Commonly selected Genes from the NCI Staunton Cancer Dataset Across All Multiclass Gene Ranking Methods

Gene Name	Description	Pathway/Process
ATP7A	ATPase, Cu++ transporting, alpha polypeptide	cell death; cellular copper ion homeostasis; Copper ion import and export; T-helper cell differentiation
NF1	neurofibromin 1	MAPK signaling pathway; cell communication; negative regulation of MAP kinase activity, MAPKKK cascade, and endothelial cell proliferation etc.
THBD	thrombomodulin	Complement and coagulation cascades pathway; negative regulation of blood coagulation, fibrinolysis, and platelet activation
MYL6	myosin, light chain 6, alkali, smooth muscle and non-muscle	Vascular smooth muscle contraction pathway

Similar to two-class datasets, we analyzed the multiclass datasets. As the evaluation of classification performance involves hundred time four fold cross-validation, the maximum gene weight achieved for a gene of such datasets is 400. Table 5.13 represents some of the top-weighted genes of NCI Staunton dataset across all methods. These genes are involved in key processes such as cell death, cell proliferation, MAPK pathway etc. Table 5.14 shows some of the top-ranked differentially selected genes obtained using PFA approach.

Interestingly, there are two mRNA sequences and one enzyme for which the gene names are unknown but are involved in Kinase and apoptosis pathway. The differential weights for such genes, enzyme, and mRNA sequences are depicted in Figure (5.9c).

Table 5.14: Functional Description of Exclusive Important Genes from the NCI Staunton Cancer Dataset by PFA approaches

Gene Name	Description	Pathway/Process
GJA1	gap junction protein, alpha 1, 43kDa	Arrhythmogenic right ventricular cardiomyopathy (ARVC) pathway; Gap junction pathway; negative regulation of cell proliferation; cell-cell signaling; apoptosis etc.
GATA3	GATA binding protein 3	anatomical structure morphogenesis; defense response
mRNA seq1	FUSE binding protein 2 (FBP2) mRNA	...
Enzyme 1	Casein kinase II subunit beta (EC 2.7.1.37)	...
mRNA seq2	Caspase-like apoptosis regulatory protein 2 (clarp) mRNA, alternatively spliced	...

Table 5.15: Functional Description of Exclusive Important Genes from the 11 Tumor Cancer Dataset by PFA approaches

Gene Name	Description	Pathway/Process
WT1	Wilms tumor 1	RNA splicing; epithelial cell differentiation; germ cell development; negative regulation of apoptosis, cell growth, and cell proliferation
POSTN	periostin, osteoblast specific factor	cell adhesion; skeletal system development

As shown in previous Chapter, 11 Tumor dataset suffers from non-Gaussianity. Due to it, we found no gene consistently selected in top-ranks across all four approaches, as gene weights obtained from F-score and KW-score approaches differed significantly.

In Table 5.15, we represent two genes which were differentially selected and present in top-ranks. These are involved in pathways such as epithelial cell differentiation, germ cell development, negative regulation of apoptosis, cell growth, cell adhesion, and cell proliferation.

For GCM dataset, the commonly selected top-weighted genes are shown in Table 5.16, where two unknown genes were also selected consistently. The known genes are involved in various cancer pathways. Table 5.17 represents some of the highly differentially selected genes across standard methods and PFA approaches. From the Figure (5.9a), it is clear that these genes are less frequently selected using F-score or KW-score. In fact, KW-score has the lowest performance in classification, and it did not select most of these genes. Biological information shows that these genes are indeed important in cancer related pathways.

Table 5.16: Functional Description of Commonly Selected Genes from the GCM Cancer Dataset Across All Multiclass Gene Ranking Methods

Gene Name	Description	Pathway/Process
GAL4	Galectin-4	cell adhesion; sugar binding
Unknown	Homo sapiens mRNA for KIAA0874 protein, partial cds	...
PTP4A2	protein tyrosine phosphatase type IVA, member 2	protein dephosphorylation; cell signalling molecule
Unknown	Homo sapiens mRNA for KIAA0874 protein, partial cds	...
APCL	adenomatosis polyposis coli 2, mRNA	Basal cell carcinoma pathway; Colorectal cancer pathway; Endometrial cancer pathway
Unknown	Carcinoembryonic Antigen Precursor	immune response; cell adhesion

(a) GCM

(b) 11 Tumor

(c) NCI Stau

(d) NCI Ross

(e) MLL

(f) Lung

Figure 5.9: Differentially weighted genes using proposed PFA approaches on multiclass datasets

Table 5.17: Functional Description of Exclusive Important Genes from the GCM Cancer Dataset by PFA approaches

Gene Name	Description	Pathway/Process
Unknown	HE4 mRNA for extra-cellular proteinase inhibitor	HE4 responsible for Ovarian and Endometrial Cancer
SFTPA1	Pulmonary Surfactant-associated protein A precursor	Lung disease and respiratory problem
MGB2	Mammaglobin 2	androgen binding; esophagus, stomach, colon, pancreas, and gall bladder cancer
GSTM5	Glutathione transferase M5	S-cytochrome P450 pathway; Glutathione metabolism pathway; Metabolism of xenobiotics by cytochrome P450

The commonly selected top-ranked genes for MLL dataset are shown in Table 5.18. These genes are biologically very important as they take part pathways and processes like Cytokine-cytokine receptor interaction, cell surface receptor linked signaling pathway, metabotropic glutamate receptor signaling pathway, Myogenesis, signaling in Immune system, apoptosis, cell differentiation, cell cycle and differentiation, MAPK signaling pathway, cell death and proliferation, etc. The differentially selected gene weights are depicted in Figure (5.9e), and their biological importance is shown in Table 5.20. These genes are related with Kinases, C450 pathway, cell signalling etc. which plays significant role in cancer tumor development.

As like 11 Tumor dataset, lung dataset also suffers from non-Gaussianity. Hence there are very few commonly selected top-ranked genes across F-score and KW-score approaches. However, Table 5.19 shows some of the such genes which are involved in DNA/protein binding, ion transport etc. Table 5.21 represents differentially selected genes which are involved in lung alveolus development, MAPK signaling pathway, Arrhythmogenic right ventricular cardiomyopathy (ARVC) pathway etc. The differential gene weights are shown in Figure (5.9f).

Table 5.18: Functional Description of Commonly Selected Genes from the MLL Cancer Dataset Across All Multiclass Gene Ranking Methods

Gene Name	Description	Pathway/Process
IL17RA	interleukin 17 receptor A	Cytokine-cytokine receptor interaction pathway; cell surface receptor linked signaling pathway
HOMER3	homer homolog 3	metabotropic glutamate receptor signaling pathway; protein targeting
RASA3	RAS p21 protein activator 3	GTPase activator activity; calcium-release channel activity; metal ion binding
MEF2A	myocyte enhancer factor 2A	Myogenesis; Signaling in Immune system; apoptosis; cell differentiation; multicellular organismal development; muscle organ development; nervous system development; regulation of transcription, DNA-dependent
PAFAH1B1	platelet-activating factor acetylhydrolase, isoform Ib, subunit 1 (45kDa)	Metabolic pathways; Cell Cycle, Mitotic; cell cycle and differentiation; nervous system development
MYCBP	c-myc binding protein	regulation of transcription; spermatogenesis
MAP3K11	mitogen-activated protein kinase kinase kinase 11	MAPK signaling pathway; G1 phase of mitotic cell cycle; JNK cascade; activation of MAPK activity; cell death and proliferation;

The commonly selected top-ranked genes of NCI Ross dataset are shown in Table 5.23 and are important in the biological process such as epithelial cell signaling, regulation of cell cycle progression, ECM-receptor interaction pathway, Focal adhesion pathway, small cell lung cancer pathway etc. The differentially selected gene weights are depicted in Figure (5.9d) and their biological importance is shown in Table 5.22. It contains two hypothetical protein which are yet to be identified biologically.

Table 5.19: Functional Description of Commonly Selected Genes from the Lung Cancer Dataset Across All Multiclass Gene Ranking Methods

Gene Name	Description	Pathway/Process
ATP1B2	ATPase, Na ⁺ /K ⁺ transporting, beta 2 polypeptide	Aldosterone-regulated sodium reabsorption pathway; Cardiac muscle contraction pathway; Gastric acid secretion pathway; Proximal tubule bicarbonate reclamation pathway; Salivary secretion; ATP biosynthetic process
POLR2E	polymerase (RNA) II (DNA directed) polypeptide E, 25kDa	Huntington's disease pathway; Metabolic pathways; Purine and Pyrimidine metabolism pathway; RNA splicing; RNA elongation from RNA polymerase II promoter
TOX4	TOX high mobility group box family member 4	DNA binding; protein binding

Table 5.20: Functional Description of Exclusive Important Genes from the MLL Cancer Dataset by PFA approaches

Gene Name	Description	Pathway/Process
FLT3	fms-related tyrosine kinase 3	Acute myeloid leukemia; Cytokine-cytokine receptor interaction; Hematopoietic cell lineage; hemopoiesis; lymphocyte differentiation; myeloid progenitor cell differentiation; negative regulation of B cell differentiation; regulation of cell proliferation
IGF1R	insulin-like growth factor 1 receptor	Adherens junction pathway; Endocytosis pathway; Focal adhesion pathway; Oocyte meiosis pathway; brain development; negative regulation of MAPKKK cascade;
EFNA3	ephrin-A3	cell-cell signaling
YES1	v-yes-1 Yamaguchi sarcoma viral oncogene homolog 1	Adherens junction pathway; Tight junction pathway; Hemostasis pathway; Signaling in Immune system; glucose transport
CYP1A1	cytochrome P450, family 1, subfamily A, polypeptide 1	Metabolism of xenobiotics by cytochrome P450; Retinol metabolism pathway; Steroid hormone biosynthesis pathway; Tryptophan metabolism pathway; amine metabolic process

Table 5.21: Functional Description of Exclusive Important Genes from the Lung Cancer Dataset by PFA approaches

Gene Name	Description	Pathway/Process
SOX2	SRY (sex determining region Y)-box 2	cell cycle arrest; epithelial tube branching involved in lung morphogenesis; lung alveolus development etc.
CACNB1	calcium channel, voltage-dependent, beta 1 subunit	Arrhythmogenic right ventricular cardiomyopathy (ARVC) pathway; Cardiac muscle contraction pathway; Dilated cardiomyopathy pathway; Hypertrophic cardiomyopathy (HCM) pathway; MAPK signaling pathway; calcium ion transport
DNAJC4	DnaJ (Hsp40) homolog, subfamily C, member 4	protein folding; response to unfolded protein; protein binding
MAP1B	microtubule-associated protein 1B	central nervous system development; embryo development etc.
MT1B	metallothionein 1B	Metal such as Zinc, cadmium etc. binding
TST	thiosulfate sulfurtransferase (rhodanese)	cyanate catabolic process; sulfate transport

Table 5.22: Functional Description of Exclusive Important Genes from the NCI Ross Cancer Dataset by PFA approaches

Gene Name	Description	Pathway/Process
SDC3	syndecan 3 (N-syndecan)	Cell adhesion molecules (CAMs) pathway; ECM-receptor interaction pathway
GDA	guanine deaminase	Metabolic pathways; Purine metabolism pathway; nervous system development
LOC92906	hypothetical protein	...
W01329	hypothetical protein	...

Table 5.23: Functional Description of Commonly Selected Genes from the NCI Ross Cancer Dataset Across All Multiclass Gene Ranking Methods

Gene Name	Description	Pathway/Process
TJP1	tight junction protein 1 (zona occludens 1)	Adherens junction pathway; Epithelial cell signaling in Helicobacter pylori infection pathway; Gap junction pathway; cell-cell junction assembly
SNK	serum-inducible kinase	Highly similar to cytokine-inducible serine/threonine kinase (CNK); may be involved in the regulation of cell cycle progression
BMPR1A	bone morphogenetic protein receptor, type IA	Cytokine-cytokine receptor interaction pathway; TGF-beta signaling pathway; bone morphogenetic protein signaling pathway
ITGA6	integrin, alpha 6	Arrhythmogenic right ventricular cardiomyopathy (ARVC) pathway; Cell adhesion molecules pathway; ECM-receptor interaction pathway; Focal adhesion pathway; Hematopoietic cell lineage pathway; Hypertrophic cardiomyopathy (HCM) pathway; Small cell lung cancer pathway

5.6 Discussion

In two-class classification, as the standard T-score ranks genes independently, it has highest stability of gene selection compared to other methods. However, the downside of such independent ranking is that it can't penalize the redundant genes and therefore result in poor classification performance. This was aptly shown in redundancy analysis where the T-score gave highly redundant genes. On the other hand, the reverse trend was observed in SVM-RFE method, which has both low stability and redundancy among the top-ranked genes. This is because SVM-RFE partially penalizes for redundant genes. With proposed sample selection criterion, we analytically determine how selection of the samples penalizes for the redundancy. Though we consider all the relevant points equally in our approaches, the analysis shows that selection of relevant samples penalizes the redundant genes. In Table 5.24, we summarize the results of stability and redundancy analysis on two-class datasets, where 1 to 5 denotes low to high values.

Table 5.24: Summary of Stability and Redundancy Analysis of T-score, SVM-RFE, SVt-R, SVt-RFE, and T-SS Methods on Various Two-class *Cancer* Datasets

Criterion	Dataset	Methods				
		T-score	SVM-RFE	SVt-R	SVt-RFE	T-SS
Stability	Colon	4	1	2	3	5
	Leukemia	5	2	1	3	4
	Prostate	5	3	2	1	4
	CNS	4	1	5	2	3
	Hepato	5	1	3	2	4
	Breast	5	1	2	3	4
Redundancy	Colon	5	1	3	2	4
	Leukemia	5	1	2	3	4
	Prostate	3	1	1	4	5
	CNS	5	1	3	2	4
	Hepato	4	1	3	2	5
	Breast	5	1	3	2	4

For multiclass problem, we observe that generally both F-score and KW-score have high stability as well as redundancy. The PFA approaches compromised on stability, but gave less redundant genes, which improved the performance in classification significantly. As with the standard T-score for two-class classification, the F-score and the KW-score rank genes independently. However, PFA approaches decompose the F-score and the KW-score and genes are then compared with each other by using their class-specific scores. Due to such comparison, the method becomes more prone to sample variations and hence relatively less stable. On the other hand, such comparison results in reduction in redundancy among the top ranked genes. Table 5.25 summarizes the results of stability and redundancy analysis on multiclass datasets, where 1 to 4 denotes low to high values.

In an ideal solution, the gene subset should be highly stable with respect to sample variations, minimally redundant, and should have excellent performance of classification. The proposed approaches in this thesis for both two-class as well as multiclass gene selec-

Table 5.25: Summary of Stability and Redundancy Analysis of F-score, KW-Score, F-PFA, and KW-PFA Methods on Various Multiclass *Cancer* Datasets

Criterion	Dataset	Methods			
		F-score	KW-score	F-PFA	KW-PFA
Stability	GCM	3	4	1	2
	11 Tumor	4	3	1	2
	NCI (Ross)	4	3	1	2
	NCI (Stau)	3	4	1	2
	MLL	3	4	1	2
	Lung	4	3	1	2
Redundancy	GCM	3	4	2	1
	11 Tumor	4	3	1	2
	NCI (Ross)	4	3	1	2
	NCI (Stau)	4	3	1	2
	MLL	3	4	1	2
	Lung	4	3	1	2

tion compromises the stability and redundancy. The T-SS method has higher stability than SVM-RFE, but also gave relatively highly redundant genes. The PFA approaches have relatively low stability but relatively low redundancy as well, compared to their basic counterparts. However, in most of the cases, such a compromise has resulted in significantly improved performance in classification, which is of the far most importance. If a highly stable and minimally redundant gene subset has a very poor performance in classification, it is of no practical use to biologists.

Using the proposed rank based gene weights, many genes were differentially weighted in consensus gene list. Out of which, many have similar weights compared to standard T-score, but significantly different from SVM-RFE. The discrepancy in gene weights may be attributed to redundancy penalization by SVM-RFE. However, many of these genes were involved in important pathways and biological processes and finally resulted in improved performance in classification. This shows that our method was able to identify many

biologically relevant, though redundant, genes.

In this chapter, we have evaluated the results of gene selections, obtained in previous chapters, with three criteria: stability of gene selection method, redundancy among top-ranked genes, and biological validation of genes. We used the Kuncheva index and absolute value of the Pearson's correlation coefficient to measure stability and redundancy respectively. Further, we showed analytically how margin based sample selection penalizes the redundancy among top-ranked genes. For biological validation, we proposed a new approach to assign weights to genes based on their ranks in given gene rank list. This method incorporates rank information into frequency based selection to obtain the consensus genes from various bootstrapped results.

Chapter 6

Summary and Future Work

6.1 Summary

Gene selection is a vital step in analyzing and classifying microarray gene expression datasets for both biological and computational reasons. In this thesis, we have concentrated on finding a stable and/or minimally redundant gene subset with good accuracy in identification of cancer tumors in a given gene expression dataset. However, search for redundant genes may eliminate relevant genes. In this research, (i) we have improved an earlier known and popular method SVM-RFE and MRMR, (ii) proposed novel methods: SVt-RFE, SVt-R, and T-SS with selection of samples for two-class classification, and (iii) incorporated concepts of Pareto fronts in filter criteria to alleviate undesirable effects of the siren-pitfall problem which is frequent in multiclass gene ranking.

In Chapter 2, we have introduced MRMR filter criterion in the SVM-RFE formulation to improve the relevancy to redundancy of genes with the aim of obtaining better performance of classification with minimally redundant genes. In the present approach, we have employed a convex combination of the SVM-RFE and MRMR criterion. With extensive experiments on five benchmark datasets, we have demonstrated improved classification performance of the proposed approach than both MRMR and SVM-RFE approach. We have also performed Gene-ontology (GO) analysis to biologically validate our findings, which shows that selected genes have similar biological properties. Though the performance is not consistent across all datasets, this method provides a better framework for incorporating classifier characteristics while compromising relevancy to redundancy of genes.

In Chapter 3, we have studied the effect of selection of samples for computing a filter criterion such as T-score. We have analytically demonstrated that the T-score can be

decomposed into two parts, corresponding to relevant and non-relevant sample points. To force the term related with irrelevant samples to zero, we have first used support vectors as selected sample points. However, such approach is computational expensive and has relatively poor stability. Using concepts of modified logistic regression, we have proposed a sample selection criterion to tackle computational complexity issues. Coupling with T-score, we have proposed a backward elimination based simultaneous selection of samples points and gene ranking algorithm. Performance analysis on both synthetic as well as benchmark datasets shows statistically significant improvement in results.

In Chapter 4, we have studied multiclass gene selection problem. Filter criteria, such as the F-score, KW-score etc., suffer from ranking bias towards a few classes which are easy to separate from others, a problem known as siren-pitfall. The gene selections without considering such pitfall may significantly detriment the performance of classification. In this thesis, we proposed a solution to circumvent such problem. We have first decomposed the ranking statistics into class specific statistics and then performed a multiobjective optimization theory based Pareto front analysis to divide genes into various fronts. We have used various criteria to rank genes within a given Pareto front. With decomposition of the F-score and KW-score into class-specific scores, we have achieved significantly better performance of classification in six benchmark datasets. Using heat maps, we have demonstrated the effectiveness of PFA approaches for gene selection.

Along with performance of classification, stability of gene selection methods and redundancy among genes are important criteria to computationally evaluate a gene ranking method. Biological validation of top-ranked genes is equally important. In Chapter 5, we have critically evaluated the proposed approaches for both two-class and multiclass gene selection using these parameters. Using Kuncheva index, compared to SVM-RFE, we show that the proposed T-SS method gives significantly better stability with improved or similar performance of classification in majority of datasets. However, the gene subset has higher redundancy compared to SVM-RFE. We have mathematically shown the effect of sample selection on redundancy among genes using T-score. Comparing with standard T-score, top-ranked genes obtained from T-SS have slightly inferior stability, but at the same time, they have low redundancy. More importantly, the classification performance is improved. To validate the results biologically, we have proposed a linear weight based method to incorporate the rank information in frequency approach to identify the true consensus genes. Such genes are further validated with biological function and pathway information, obtained using GO.

6.2 Future Directions

6.2.1 Gene Selection with Regulatory Networks

In this thesis, we have concentrated on identifying individual genes that can accurately classify the cancer samples. However, genes rarely work independent of other genes. A particular biological activity is due to many genes regulating one another in a gene regulatory network (GRN). Such networks are causal networks meaning that activity of one gene causes one or more genes to react in positive or negative manner. Function of such networks are defined by hubs or motifs, which are basically core genes of the network. Identification of such core genes is of great biological value as these small local regulatory network has tremendous potentials for drug development. The building of GRNs may again be affected by sample quality.

Recently, Yousef *et al.* also presented a support vector machine - recursive network elimination based method [172]. The gene network modules were identified using GXNA software, and then their importance in classification was measured using SVM. To integrate a priori network based information, Li and Li proposed a network constrained penalty method for gene selection [24]. The KEGG pathway network was represented as a graph and its corresponding Laplacian matrix and was integrated to predict the survival time of glioblastoma. With the concept of margin based sample selection proposed in this thesis, one can build better networks by identifying the relevant sample points. These networks can then be used with techniques like recursive network (cluster) elimination or network constrained gene selection to identify the core genes. Further, the stability and redundancy analysis could be improved using network information instead of a statistical criterion like Pearson's correlation.

6.2.2 Biclustering for Simultaneous Sample and Gene Selection

In this thesis, we have achieved simultaneous sample and gene selection by using the theory of SVM and T-score. The biclustering approaches can also be modified for simultaneous sample and gene selection. Using a representative bicluster for each of the class, a suitable classification scheme can be developed where genes in a given bicluster help discriminate the samples of that class from the rest. This will need to define an appropriate optimization function to select the genes and samples.

6.2.3 Simultaneous Outlier Detection and Gene Selection

A direct outlier detection algorithm based on a statistical criterion along with gene selection may be an interesting future direction. However, the high dimensional nature of such a dataset presents a huge challenge for developing an outlier detection criterion. Similar to T-SS, once a suitable criterion is developed, it can be coupled with gene ranking criterion.

6.2.4 Extension of Sample Selection to Other Bioinformatics Problems

As demonstrated in this thesis, selection of sample is indeed important as all the samples are not equally relevant for the classification task. The proposed algorithm T-SS is extremely efficient. However, we used only 0-1 weights for the selection of samples. As an alternative, one can even consider samples beyond the margin with varying weights liked to the distance to the discriminant function and incorporate such weights in gene selection. In a way, development of such a boosting classifier with in-built gene selection will be interesting.

The proposed approaches in this thesis are not limited to cancer identification. In recent developments, similar to gene expression datasets, a plethora of single nucleotide polymorphism (SNP) datasets are available which are high dimensional in nature. On the other hand, high content screening (HCS) image datasets have large number of samples where SVM-RFE becomes extremely computational intensive method. One can apply the T-SS algorithm on such biological datasets to achieve good classification performance as well as stability. In this thesis, we have only employed T-score as gene ranking criterion. However, there are other filter-based gene ranking criteria, such as signal-to-noise ratio, Fisher discriminant analysis, symmetric divergence, which could be applied for gene-expressions, SNP, or HCS datasets. On biological front, it is important to understand the biological significance of exclusion of a sample points. Currently, there are no method or database which validates the biological importance of a sample points. In future, it will be of great significance to develop an analysis method to validate samples.

6.2.5 Extension to Machine Learning Problems

In this thesis, we have used SVM as a base classifier which achieved reasonably good performance. To further improve the performance in classification, SVM can be combined

with other classifier such as ANN. Development of a classification algorithms for high dimensional dataset can be an interesting research direction.

Many multiclass gene selection algorithm including wrapper methods perform the aggregation in deriving the final ranking score. For example, recently proposed multiclass SVM-RFE algorithm aggregates the SVM weights of each of the one-vs-all or one-vs-one SVM gene weight. The proposed Pareto front analysis can be applied to improve such wrapper algorithm. Using Pareto analysis, a multiclass boosting classifier could be improved by detecting which classes are difficult to separate and put higher weights to samples from such classes. Not only gene selection, a Pareto front analysis may in fact improve the performance of some aggregation based classifiers. Moreover, there are many two-class filter gene ranking criteria that can not be directly applied to multiclass gene selection problem. Considering such criteria as multiple-objectives, one can effectively convert these two class criteria into multiclass criteria by using our proposed Pareto-front based gene ranking method.

6.2.6 Analysis of Three-Dimensional Gene Expression Data

The gene selection techniques, developed for tissue classification thus far, serves static gene expression data. With the advancement of microarray technologies, a new kind of data have emerged with time as an additional dimension [90]. These datasets have patient's gene expressions measured at different time points. Such datasets represent the dynamics of the disease or tumor growth. Hence, it poses unique challenges as patients in same class can have different disease growth rates. Existing static gene expression based gene selection methods have to be modified to suit such instances.

In such datasets, it is important to consider the dynamics of gene expression while developing gene selection technique for efficient classification. Gene expressions have to be considered as vectors. In addition to clustering samples, the clustering to identify similar temporal patterns is a necessity, and therefore genes could be identified based on similarity of gene expression profiles. Biclustering may present an interesting solution to such a problem. One can develop a supervised biclustering based gene selection algorithm, in which, a gene will be represented by a bicluster of different samples (of same class) and time points. Such gene specific biclusters will then be used for the classification. Such approaches may have an interesting property of gene expression time series alignment and sample selection.

6.3 Conclusion

This research has proposed several novel contributions and recommendations for gene and sample selection from microarray data in identification of cancer. We proposed to introduce MRMR filter to SVM-RFE to handle the redundancy problem. The presence of irrelevant samples deteriorate selection of genes and hence performance of classification. We efficiently incorporated sample selection in T-score (T-SS) method. In multiclass gene selection, we proposed a Pareto-front based analysis to handle the siren-pitfall problem. We believe that the gene selection techniques developed in this thesis could help identify cancer better and give stable/minimally-redundant gene subsets with biological relevance.

References

- [1] T. Abeel, T. Helleputte, Y. Van de Peer, and P. Dupont *et al.* Robust biomarker identification for cancer diagnosis with ensemble feature selection methods. *Bioinformatics*, 26(3):392–398, 2010.
- [2] D. Allison, X. Cui, G. Page, and M. Sabripour. Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews: Genetics*, 7:55–65, 2006.
- [3] U. Alon, N. Barkai, D. Notterman, and K. Gish *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of National Academy of Sciences*, 96:6745–6750, 1999.
- [4] S.A. Armstrong, J.E. Staunton, L.B. Silverman, and R. Pieters *et al.* MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia. *Nature Genetics*, 30(1):41–47, 2002.
- [5] M. Ashburner, C. Ball, J. Blake, and D. Botstein *et al.* Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nature Genetics*, 25(1):25–29, 2000.
- [6] Y. Audic and R. Hartley. Post-transcriptional regulation in cancer. *Biology of the Cell*, 96:479–498, 2004.

-
- [7] R. Battiti. Using mutual information for selecting features in supervised neural net learning. *IEEE Transactions on Neural Network*, 5:537–550, 1994.
- [8] P.J. Bentley and J.P. Wakefield. Finding acceptable solutions in the pareto-optimal range using multiobjective genetic algorithms. In P.K. Chawdhary, R. Roy, and R.K. Pant, editors, *Soft computing in engineering design and manufacturing*, Part 5, pages 231–240, 1997.
- [9] A. Bhattacharjee, W.G. Richards, J. Staunton, and C. Li *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proceedings of National Academy of Sciences*, 98(24):13790–13795, 2001.
- [10] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford Press, 1996.
- [11] G. Blobel, W. Schiemann, and H. Lodish. Role of transforming growth factor β in human disease. *New England Journal of Medicine*, 342:1350–1358, 2000.
- [12] A. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271, 1997.
- [13] A-L Boulesteix and M. Slawski. Stability and aggregation of ranked gene list. *Briefings in Bioinformatics*, 10(5):556–568, 2008.
- [14] A.-L. Boulesteix, C. Strobl, T. Augustin, and M. Daumer. Evaluating microarray-based classifiers: An overview. *Cancer Informatics*, 6:77–97, 2008.
- [15] S.B. Cantor and M.W. Kattan. Determining the area under the ROC curve for a binary diagnostic test. *Medical Decision Making*, 20:468–470, 2000.
-

-
- [16] R. Cavill, H. Keun, E. Holmes, and J. Lindon *et al.* Genetic algorithms for simultaneous variable and sample selection in metabonomics. *Bioinformatics*, 25(1):112–118, 2009.
- [17] G Cawley and N. Talbot. Gene selection in cancer classification using sparse logistic regression with bayesian regularization. *Bioinformatics*, 22(19):2348–2355, 2006.
- [18] C. Chang and C. Lin. LibSVM: A library for support vector machines. www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.
- [19] D. Chen, Z. Liu, X. Ma, and D. Hua. Selecting genes by test statistics. *J. Biomedicine and Biotechnology*, 2:132–138, 2005.
- [20] Q. Cheng. A sparse learning machine for high-dimensional data with application to microarray gene analysis. *IEEE Transactions on Computational Biology and Bioinformatics*, 7(4):636–646, 2010.
- [21] J-H. Cho, D. Lee, J. H. Park, and I-B. Lee. New gene selection method for classification of cancer subtypes considering within-class variation. *FEBS Letters*, 551:3–7, 2003.
- [22] R. Clarke, R. Dickson, and M. Lippman. Hormonal aspects of breast cancer: growth factors, drugs and stromal interactions. *Critical Review on Oncology and Hematology*, 12:1–23, 1992.
- [23] R. Clarke, H.W. Ransom, A. Wang, and J. Xuan *et al.* The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8:37–49, 2008.
- [24] C.Li and H. Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 2008.
-

-
- [25] A. Culhane, G. Perriere, and D. Higgins. Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, 4(1):59, 2003.
- [26] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2):182–197, 2002.
- [27] K. Deb and A. Raji Reddy. Reliable classification of two-class cancer data using evolutionary algorithms. *BioSystems*, 72:111–129, 2003.
- [28] J. Deutsch. Evolutionary algorithms for finding optimal gene sets in microarray prediction. *Bioinformatics*, 19(1):45–52, 2003.
- [29] A. Dhillon, S. Hagan, O. Rath, and W. Kolch. MAP kinase signalling pathways in cancer. *Oncogene*, 26:3279–3290, 2007.
- [30] R. Diaz-Uriarte and S. Andres. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:3, 2006.
- [31] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *J Bioinformatics Computational Biology*, 3:185–205, 2005.
- [32] K-B. Duan, J.C. Rajapakse, and M. Nguyen. One-versus-one and one-versus-all multiclass SVM-RFE for gene selection in cancer classification. In E. Marchiori, J. H. Moore, and J.C. Rajapakse, editors, *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics, EVOBIO*, volume 4447 of *LNCS*, pages 47–56, 2007.
- [33] S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *J American Statistical Association*, 97(457):77–86, 2002.
-

-
- [34] K. Dunne, P. Cunningham, and F. Azuaje. Solutions to instability problems with sequential wrapper-based approaches to feature selection. In *Department of Computer Science, Trinity College, Dublin, Tech. Report TCD-CD-2002-28*, 2002.
- [35] L. Ein-Dor, O. Zuk, and E. Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of National Academy of sciences*, 103(15):5923–5928, 2006.
- [36] P. Estevez, M. Tesmer, C. Perez, and J. Zurada. Normalized mutual information feature selection. *IEEE Transactions on Neural Network*, 20(2):189–201, 2009.
- [37] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. American Statistical Association*, 96:1348–1360, 2001.
- [38] J. Fan and J. Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20:101–148, 2010.
- [39] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874, 2006.
- [40] G. Forman. A pitfall and solution in multi-class feature selection for text classification. In *Proceedings of the 21st international conference on Machine learning*, volume 69 of *ACM International Conference Proceeding Series*, page 38, 2004.
- [41] G. Fort and S. Lambert-Lacroix. Classification using partial least squares with penalized logistic regression. *Bioinformatics*, 21(7):1104–1111, 2005.
- [42] Y. Freund and R. Schapire. A short introduction to boosting. *J. Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.
-

-
- [43] E. Freyhult, M. Landfors, J. Onskog, and T. Hvidsten *et al.* Challenges in microarray class discovery: a comprehensive examination of normalization, gene selection and clustering. *BMC Bioinformatics*, 11:503, 2010.
- [44] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *J. Statistical Software*, 33(1):1–22, 2010.
- [45] H. Frohlich, N. Speer, A. Poustka, and T. Beißbarth. GOSim - an R-package for computation of information theoretic go similarities between terms and gene products. *BMC Bioinformatics*, 8:166, 2007.
- [46] W. J. Fu. Penalized regression: The bridge versus the lasso. *J. Computational and Graphical Statistics*, 7:397–416, 1998.
- [47] C. Furlanello, M. Serafini, S. Merler, and G. Jurman. Entropy-based gene ranking without selection bias for the predictive classification of microarray data. *BMC Bioinformatics*, 4:54, 2003.
- [48] W. Fury, F. Batliwalla, P. Gregersen, and L. Wentian *et al.* Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion. In *Proceedings of the 28th Annual International conference of the IEEE: IEEE Engineering in medicine and biology*, pages 5531–5534, 2006.
- [49] T. Golub, D. Slonim, P. Tamayo, and C. Huard *et al.* Molecular classification of cancer: Class discovery and class prediction by gene expression. *Science*, 286:531–537, 1999.
- [50] J. Gorodkin. Comparing two K-category assignment by a K-category correlation coefficient. *Computational Biology and Chemistry*, 28:367–374, 2004.
-

- [51] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *J Machine Learning Research*, 3:1157–1182, 2003.
- [52] I. Guyon and A. Elisseeff. An introduction to feature extraction. In I. Guyon, S. Gunn, M. Nikravesh, and L. Zadeh, editors, *Feature Extraction, Foundations and Applications*, Studies in Fuzziness and Soft Computing, pages 1–25, 2006.
- [53] I. Guyon, J. Weston, S. Barhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.
- [54] M. Hall. *Correlation-based feature selection for machine learning*. PhD thesis, Department of Computer Science, University of Waikato, 1999.
- [55] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2001.
- [56] T. Helleputte and P. Dupont. Partially supervised feature selection with regularized linear models. In *Proceedings of the 26th International conference on Machine Learning*, pages 409–416, 2009.
- [57] R. Hu, X. Qiu, and G. Glazko. A new gene selection procedure based on the covariance distance. *Bioinformatics*, 26(3):348–354, 2010.
- [58] J. Hua, W. Tembe, and E. Dougherty. Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition*, 42:409–424, 2009.
- [59] H-L. Huang, C-C Lee, and S-Y Ho. Selecting a minimal number of relevant genes from microarray data to design accurate tissue classifiers. *BioSystems*, 90:78–86, 2007.

- [60] N. Iizuka, M. Oka, H. Yamada-Okabe, and M. Nishida *et al.* Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *The Lancet*, 361:923–929, 2003.
- [61] I. Inza, P. Larranaga, R. Blanco, and A. Cerrolaza. Filter versus wrapper gene selection approaches in dna microarray domains. *Artificial Intelligence Medicine*, 31:91–103, 2004.
- [62] M.T. Jensen. Reducing run-time complexity of multiobjective EAs: The NSGA-II and other algorithms. *IEEE Transactions on Evolutionary Computation*, 7(5):503–515, 2003.
- [63] T. Jirapech-Umpai and S. Aitken. Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes. *BMC Bioinformatics*, 6:148, 2005.
- [64] M. Johannes, J. Brase, H. Frohlich, and S. Gade. Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. *Bioinformatics*, 26(17):2136–2144, 2010.
- [65] G. Jurman, S. Merler, A. Barla, and S. Paoli. Algebraic stability indicators for ranked lists in molecular profiling. *Bioinformatics*, 24(2):258–264, 2008.
- [66] K. Kadota, Y. Nakai, and K. Shimizu. A weighted average difference method for detecting differentially expressed genes from microarray data. *BMC Bioinformatics*, 3:8, 2008.
- [67] D. Kai-Bo, J.C. Rajapakse, H. Wang, and F. Azuaje. Multiple SVM-RFE for gene selection in cancer classification with expression data. *IEEE Transactions on Nanobioscience*, 4:228–234, 2005.

- [68] A. Kalousis, J. Prados, and M. Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems*, 12(1):95–116, 2007.
- [69] S-Y Kim. Effects of sample size on robustness and prediction accuracy of a prognostic gene signature. *BMC Bioinformatics*, 10:147, 2009.
- [70] K. Kira and L. A. Rendell. A feature selection problem: traditional methods and a new algorithm. In *Proc. of the 10th National Conference on Artificial Intelligence*, pages 129–134. AAAI Press, 1992.
- [71] R. Kohavi and G. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.
- [72] D. Koller and M. Sahami. Towards optimal feature selection. In *Proceedings of the 13th International conference on Machine Learning, ICML*, pages 284–292, 1996.
- [73] I. Kononenko. Estimating attributes: Analysis and extension of RELIEF. In F. Bergadano and L. De Raedt, editors, *Proc. of the European Conference on Machine Learning*, pages 171–182. Springer, 1994.
- [74] L. Kuncheva. A stability index for feature selection. In *Proceedings of the 25th IASTED International conference on Artificial intelligence and applications*, pages 390–395, 2007.
- [75] C. Lai, M. Reinders, L. van’t Veer, and L. Wessels. A comparison of univariate and multivariate gene selection techniques for classification of cancer datasets. *BMC Bioinformatics*, 7:235, 2006.
- [76] T. Lal, O. Chapelle, J. Weston, and A. Elisseeff. Embedded methods. In *Feature Extractions Foundations and Applications*. Springer Publications, 2006.

- [77] L. Lancashire, C. Lemetre, and G. Ball. An introduction to artificial neural networks in bioinformatics-application to complex microarray and mass spectrometry datasets in cancer studies. *Briefings in Bioinformatics*, Article in Press, 2009.
- [78] G. V. Lashkia and L. Anthony. Relevant, irredundant feature selection and noisy example elimination. *IEEE Transactions on Systems, Man, And Cybernetics-part B*, 34(2):888–897, 2004.
- [79] Y. LeCun, J. Denker, S. Solla, and R. Howard *et al.* Optimal brain damage. In D. Touretzky, editor, *Advances in Neural Information Processing Systems II*, pages 598–605, 1990.
- [80] J. Lee, J. Lee, M. Park, and S. Song. An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics and Data Analysis*, 48:869–885, 2005.
- [81] M.-L. Lee, F. Kuo, G. Whitmore, and J. Sklar. Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proceedings of National Academy of Sciences*, 97(18):9834–9839, 2000.
- [82] Y. Leung and Y. Hung. A multiple-filter-multiple-wrapper approach to gene selection and microarray data classification. *IEEE Transactions on Computational Biology and Bioinformatics*, 7(1):108–117, 2010.
- [83] F. Li and Y. Yang. Analysis of recursive gene selection approaches from microarray data. *Bioinformatics*, 21(19):3741–3747, 2005.
- [84] L. Li, L. Pedersen, T. Darden, and C. Weinberg. Computational analysis of leukemia microarray expression data using the GA-KNN. In *Proceedings of the*

- 1st Conference on Critical Assessment of Microarray Data Analysis, CAMDA2000*, 2000.
- [85] L. Li, L. Pedersen, T. Darden, and C. Weinberg. Gene selection for sample classification based on gene-expression expression data: study of sensitivity to choice of parameters of the GA-KNN method. *Bioinformatics*, 17(12):1131–1142, 2001.
- [86] S. Li, X. Wu, and M. Tan. Gene selection using hybrid particle swarm optimization and genetic algorithm. *Soft Computing*, 12:1039–1048, 2008.
- [87] T. Li, C. Zhang, and M. Ogihara. A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, 20(15):2429–2437, 2004.
- [88] J. Liang, R. Li, H. Fang, and K-T. Fang. Testing multinormality based on low-dimensional projection. *J. Statistical Planning and Inference*, 86:129–141, 2000.
- [89] D. Lin. An information-theoretic definition of similarity. In *Proc. 15th Int. Conf. Machine Learning*, pages 296–304, 1998.
- [90] T.-H. Lin, N. Kaminski, and Z. Bar-Joseph. Alignment and classification of time series gene expression in clinical studies. *Bioinformatics*, 24:i147–i155, 2008.
- [91] N. Littlestone and C. Mesterharm. An apobayesian relative of WINNOW. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press Cambridge, 1997.
- [92] H. Liu, H. Motoda, and L. Yu. A selective sampling approach to active feature selection. *Artificial Intelligence*, 159:49–74, 2004.

- [93] J. Liu, G. Cutler, W. Li, and Z. Pan *et al.* Multiclass cancer classification and biomarker discovery using GA-based algorithms. *Bioinformatics*, 21(11):2691–2697, 2005.
- [94] J. Liu, S. Kang, C. Tang, and L.B. Ellis *et al.* Meta-prediction of protein subcellular localization with reduced voting. *Nucleic Acid Research*, 35(15):e96, 2007.
- [95] K-H. Liu and C-G. Xu. A genetic programming based approach to the classification of multiclass microarray datasets. *Bioinformatics*, 25(3):331–337, 2009.
- [96] X. Liu, A. Krishnan, and A. Mondry. An entropy-based gene selection method for cancer classification using microarray data. *BMC Bioinformatics*, 6:76, 2005.
- [97] Y. Liu and Y. Zheng. FS-SFS: a novel feature selection method for support vector machines. *Pattern Recognition*, 39:1333–1345, 2006.
- [98] Z. Liu, R. Gartenhaus, M. Tan, and F. Jiang *et al.* Gene and pathway identification with l_p penalized bayesian logistic regression. *BMC Bioinformatics*, 9:412, 2008.
- [99] P. Lord, R. Stevens, A. Brass, and A. Goble. Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19:1275–1283, 2003.
- [100] S. Loscalzo, L. Yu, and C. Ding. Consensus group stable feature selection. In *Proceedings of the 15th ACM SIGKDD International conference on Knowledge discovery and data mining*, pages 567–575, 2009.
- [101] N. Louw and S. Steel. Variable selection in kernel fisher discriminant analysis by means of recursive feature elimination. *Computational Statistics and Data Analysis*, 51:2043–2055, 2006.

- [102] S. Ma and J. Huang. Penalized feature selection and classification in bioinformatics. *Briefings in Bioinformatics*, 9(5):392–403, 2008.
- [103] M. W. Mak and S.Y. Kung. A solution to the curse of dimensionality problem in pairwise scoring techniques. In I. King, J. Wang, L. Chan, and D. Wang, editors, *Proceedings of the 13th International conference on Neural information processing, ICONIP*, volume 4232 of *LNCS*, pages 314–323, 2006.
- [104] E. Marshall. Getting the noise out of gene arrays. *Science*, 306:630–631, 2004.
- [105] G. McLean, N. Carragher, E. Avizienyte, and J. Evans *et al.* The role of focal-adhesion kinase in cancer – a new therapeutic opportunity. *Nature Reviews Cancer*, 5(7):505–515, 2005.
- [106] C. De Mol, S. Mosci, M. Traskine, and A. Verri. A regularized method for selecting nested groups of relevant genes from microarray data. *J. Computational Biology*, 16(5):677–690, 2009.
- [107] D. Nguyen and D. Rocke. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50, 2002.
- [108] K. Nicodemus, J. Malley, C. Strobl, and A. Ziegler. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, 11:110, 2010.
- [109] S. Nijima and S. Kuhara. Recursive gene selection based on maximum margin criterion: a comparison with SVM-RFE. *BMC Bioinformatics*, 7:543, 2006.
- [110] T. Okegawa, R-C. Pong, Y. Li, and J-T. Hsieh. The role of cell adhesion molecule in cancer progression and its application in cancer therapy. *Acta Biochimica Polonica*, 51(2):445–457, 2004.

- [111] C. Ooi, M. Chetty, and S. Teng. Differential prioritization between relevance and redundancy in correlation-based feature selection techniques for multiclass gene expression data. *BMC Bioinformatics*, 7:320–339, 2006.
- [112] C. Ooi, M. Chetty, and S. Teng. Characteristics of predictor sets found using differential prioritization. *Algorithms for Molecular biology*, 2:7, 2007.
- [113] M. Osl, S. Dreiseitl, F. Cerqueira, and M. Netzer. Demoting redundant features to improve the discriminatory ability in cancer data. *J. Biomedical Informatics*, 42:721–725, 2009.
- [114] M. Pechenizkiy, S. Puuronen, and A. Tsymbal. The impact of sample reduction on PCA-based feature extraction for supervised learning. In *Proceedings of the 21st ACM Symposium on Applied Computing*, pages 553–558. ACM Press, 2006.
- [115] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 27:1226–1237, 2005.
- [116] M. Pirooznia, J.Y. Yang, M. Yang, and Y. Deng. A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*, 9(1):S13, 2008.
- [117] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, and L. Sturla *et al.* Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442, 2002.
- [118] J. Qi and J. Tang. Gene ontology driven feature selection from microarray gene expression data. In *Proceedings of the 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB*, pages 428–434, 2006.

- [119] J. Qi and J. Tang. Integrating gene ontology into discriminative powers of genes for feature selection in microarray data. In *Proceedings of the 2007 ACM Symposium on Applied Computing (SAC)*, pages 430–434, 2007.
- [120] J. Quackenbush. Microarray data normalization and transformation. *Nature Genetics*, 32:496–501, 2002.
- [121] J.C. Rajapakse, D. Kai-Bo, and W. Yeo. Proteomic cancer classification with mass spectrometry data. *American J Pharmacogenomics*, 5:281–292, 2005.
- [122] A. Rakotomamonjy. Variable selection using SVM criteria. *J. Machine Learning Research (Special Issue on Variable Selection)*, 3:1357–1370, 2003.
- [123] S. Ramaswamy, P. Tamayo, R. Rifkin, and S. Mukherjee *et al.* Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of National Academy of Sciences*, 98(26):15149–15154, 2001.
- [124] J. Rice. *Mathematical Statistics and Data Analysis*. Thomson Brooks/cole, 2007.
- [125] M. Robin-Sikonja and I. Kononenko. An adaptation of relief for attribute estimation in regression. In *Proceedings of the 14th International conference on Machine Learning, ICML*, pages 296–304, 1997.
- [126] D. T. Ross, U. Scherf, M.B. Eisen, and C. Perou *et al.* Systematic variation in gene expression patterns in human cancer cell. *Nature Genetics*, 24(3):227–235, 2000.
- [127] R. Ruiz, J. Riquelme, and J. Aguilar-Ruiz. Incremental wrapper-based gene selection from microarray data for cancer classification. *Pattern Recognition*, 39:2383–2392, 2006.
- [128] Y. Saeys, I. Inza, and P. Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

-
- [129] J. Sevilla, V. Segura, A. Podhorski, and E. Guruceaga *et al.* Correlation between gene expression and GO semantic similarity. *IEEE Transactions on Computational Biology Bioinformatics*, 2:330–338, 2005.
- [130] L. Shen and E. Tan. Dimension reduction based penalized logistic regression for cancer classification using microarray data. *IEEE Transactions on Computational Biology and Bioinformatics*, 2:166–175, 2005.
- [131] Q. Shen, Z. Mei, and B-X. Ye. Simultaneous genes and training samples selection by modified particle swarm optimization for gene expression data classification. *Computers in Biology and Medicine*, 39:646–649, 2009.
- [132] Q. Shen, W-M. Shi, and W. Kong. New gene selection method for multiclass tumor classification by class centroid. *J. Biomedical Informatics*, 42(1):59–65, 2009.
- [133] L. Shi, W. Tong, H. Fang, and U. Scherf *et al.* Cross-platform comparability of microarray technology: Intra-platform consistency and appropriate data analysis procedures are essential. *BMC Bioinformatics*, 6(Suppl 2):S12, 2005.
- [134] A. Shieh and Y. Hung. Detecting outlier samples in microarray data. *Statistical Applications in Genetics and Molecular Biology*, 8(1):13, 2009.
- [135] J. Shim, I. Sohn, S. Kim, and J. Lee *et al.* Selecting marker genes for cancer classification using supervised weighted kernel clustering and the support vector machine. *Computational Statistics and Data Analysis*, 53(5):1736–1742, 2009.
- [136] C. Sima and E. Dougherty. The peaking phenomenon in the presence of feature-selection. *Pattern Recognition Letters*, 29:1667–1674, 2008.
- [137] D. Singh, P. Febbo, K. Ross, and D. Jackson *et al.* Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209, 2002.
-

-
- [138] A. Statnikov, C. Aliferis, I. Tsamardinos, and D. Hardin *et al.* A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics*, 21(5):631–643, 2005.
- [139] A. Statnikov, L. Wang, and C. Aliferis. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, 9:319, 2008.
- [140] J. Staunton, D. Slonim, H. Collier, and P. Tamayo *et al.* Chemosensitivity prediction by transcriptional profiling. *Proceedings of National Academy of Science*, 98(19):10787–10792, 2001.
- [141] G. Stolovitzky. Gene selection in microarray data: the elephant, the blind men and our algorithms. *Current Opinion in Structural Biology*, 13(3):370–376, 2003.
- [142] A. Su, J. Welsh, L. Sapinoso, and S. Kern *et al.* Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Research*, 61:7388–7393, 2001.
- [143] F. Tai and W. Pan. Incorporating prior knowledge of gene functional groups into regularized discriminant analysis of microarray data. *Bioinformatics*, 23(23):3170–3177, 2007.
- [144] A. Tan, D. Naiman, L. Xu, and R. Winslow *et al.* Simple decision rules for classifying human cancers from gene expression profiles. *Bioinformatics*, 21(20):3896–3904, 2005.
- [145] E. Tang, P. Suganthan, and X. Yao. Gene selection algorithms for microarray data based on least square support vector machine. *BMC Bioinformatics*, 7:95, 2006.
-

- [146] Y. Tang, Y-Q Zhang, and Z. Huang. Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis. *IEEE Transactions on Computational Biology and Bioinformatics*, 4(3):365–381, 2007.
- [147] Y. Tang, Y-Q Zhang, Z. Huang, and X. Hu *et al.* Recursive fuzzy granulation for gene subset extraction and cancer classification. *IEEE Transactions on Information Technology in Biomedicine*, 12(6):723–730, 2008.
- [148] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal Stat Society B*, 58:267–288, 1996.
- [149] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of National Academy of Science*, 99(10):6567–6572, 2002.
- [150] O. Troyanskaya, M. Garber, P. Brown, and D. Botstein *et al.* Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, 18(11):1454–1461, 2002.
- [151] Y-S. Tsai, C-T. Lin, G. Tseng, and I-F. Chung *et al.* Discovery of dominant and dormant genes from expression data using a novel generalization of snr for multi-class problems. *BMC Bioinformatics*, 9:425, 2008.
- [152] I. Tsamardinos and C. Aliferis. Towards principled feature selection: Relevancy, filters and wrappers. In *Proceedings of the 9th International workshop on Artificial Intelligence and Statistics, AISTats*, 2003.
- [153] Y. Tu, G. Stolovitzky, and U. Klein. Quantitative noise analysis for gene expression microarray experiments. *Proceedings of National Academy of Science*, 99(22):14031–14036, 2002.

- [154] E. Tuv, A. Borisov, G. Runger, and K. Torkkola. Feature selection with ensembles, artificial variables, and redundancy elimination. *J. Machine Learning Research*, 10:1341–1366, 2009.
- [155] V. N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience publications, 1998.
- [156] M. Wahde and Z. Szallasi. Improving the prediction of the clinical outcome of the breast cancer using evolutionary algorithms. *Soft Computing*, 10(4):338–345, 2006.
- [157] M. Wahde and Z. Szallasi. A survey of methods for classification of gene-expression data using evolutionary algorithm. *Expert Review Molecular Diagnostics*, 6(1):101–110, 2006.
- [158] H. Wang, F. Azuaze, O. Bodenreider, and J. Dopazo. Gene expression correlation and gene ontology based similarity: An assessment of quantitative relationships. In *Proc. 2004 IEEE Symposium Computational Intelligence in Bioinformatics and Computational Biology*, pages 25–31, 2004.
- [159] L. Wang, J. Zhu, and H. Zou. Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*, 24(3):412–419, 2008.
- [160] Y. Wang, D. Miller, and R. Clarke. Approaches to working in high-dimensional data spaces: gene expression microarrays. *British Journal of Cancer*, 98:1023–1028, 2008.
- [161] Y. Wang, I. Tetko, M. Hall, and E. Frank *et al.* Gene selection from microarray data for cancer classification - a machine learning approach. *Computational Biology and Chemistry*, 29:37–46, 2005.
- [162] P. Wei and W. Pan. Incorporating gene networks into statistical tests for genomic data via a spatially correlated mixture model. *Bioinformatics*, 24(4):404–411, 2008.

- [163] Z. Wei and H. Li. A markov random field model for network-based analysis of genomic data. *Bioinformatics*, 23(12):1537–1544, 2007.
- [164] M. West, C. Blanchette, H. Dressman, and E. Huang *et al.* Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of National Academy of sciences*, 98(20):11462–11467, 2001.
- [165] Z.-X. Xie, Q.-H. Hu, and D.-R. Yu. Improved feature selection algorithm based on SVM and correlation. In J. Wang, Z. Yi, J. Zurada, B.-L. Lu, and H. Yin, editors, *Proceedings of the third international symposium on Neural Networks, ISNN 2006*, volume 3971 of *LNCS*, pages 1373–1380, 2006.
- [166] J. Xuan, Y. Wang, Y. Dong, and Y. Feng *et al.* Gene selection formulticlass prediction by weighted fisher criterion. *EURASIP J Bioinformatics and Systems Biology*, 2007 article id 64628, 2007.
- [167] A-J Yang and X-Y Song. Bayesian variable selection for disease classification using gene expression data. *Bioinformatics*, 26(2):215–222, 2010.
- [168] J. Yang, G-Z. Li, H-H. Meng, and M. Yang *et al.* Improving prediction accuracy of tumor classification by reusing genes discarded during gene selection. *BMC Genomics*, 9(Supp 1):S3, 2008.
- [169] T. Y. Yang. Simple bayesian bianry framework for discovering significant genes and classifying cancer diagnosis. *Computational Statistics and Data Analysis*, 53:1743–1754, 2009.
- [170] X. Yang, S. Bentink, S. Scheid, and R. Spang. Similarities of ordered gene lists. *J. Bioinformatics and Computational Biology*, 4(3):693–708, 2006.

- [171] M. Yousef, S. Jung, L. Showe, and M. Showe. Recursive cluster elimination (RCE) for classification and feature selection from gene expression data. *BMC Bioinformatics*, 8:144, 2007.
- [172] M. Yousef, M. Ketany, L. Manevitz, and L. Showe. Classification and biomarker identification using gene network modules and support vector machines. *BMC Bioinformatics*, 10:337, 2009.
- [173] L. Yu and H. Liu. Efficient feature selection via analysis of relevance and redundancy. *J. Machine Learning Research*, 5:1205–1224, 2004.
- [174] S. Zadeh and M. Moradi. An evaluation of gene ranking methods by gene ontology. In *Proceedings of the 8th International conference on Signal processing, ICSP*, pages 16–20, 2006.
- [175] M. Zervakis, M. Blazadonakis, G. Tsiliki, and V. Danilatu *et al.* Outcome prediction based on microarray analysis: a critical perspective on methods. *BMC Bioinformatics*, 10:53, 2009.
- [176] H. Zhang, J. Ahn, X. Lin, and C. Park. The regularization and variable selection via the elastic net. *Bioinformatics*, 22:88–95, 2006.
- [177] J. Zhang and H. Deng. Gene selection for classification of microarray data based on bayes error. *BMC Bioinformatics*, 8:370, 2007.
- [178] J. Zhang, R. Jin, Y. Yang, and A.G. Hauptmann. Modified logistic regression: An approximation to svm and its applications in large-scale text categorization. In *Proceedings of 20th International Conference on Machine Learning, ICML’03*, 2003.

-
- [179] M. Zhang, L. Zhang, J. Zou, and C. Yao *et al.* Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics*, 25(13):1662–1668, 2009.
- [180] Y. Zhang, C. Ding, and T. Li. Gene selection algorithm by combining reliefF and mRMR. *BMC Genomics*, 9(Suppl 2):S27, 2008.
- [181] X. Zhou and K. Mao. LS bound based gene selection dna microarray data. *Bioinformatics*, 21(8):1559–1564, 2005.
- [182] X. Zhou and T. P. Tuck. MSVM-RFE: Extensions of SVM-RFE for multiclass gene selection on dna microarray data. *Bioinformatics*, 23(9):1106–1114, 2007.
- [183] Z. Zhou, Y-S. Ong, and M. Dash. Markov-blanket embedded genetic algorithm for gene selection. *Pattern Recognition*, 40:3236–3248, 2007.
- [184] Z. Zhou, Y-S. Ong, and M. Dash. Wrapper-filter feature selection algorithm using memetic framework. *IEEE Transactions on Systems, Man and Cybernetics: Part B*, 37(1):70–76, 2007.
- [185] Z. Zhu, Y-S. Ong, and J. Zurada. Identification of full and partial class relevant genes. *IEEE Transactions on Computational Biology and Bioinformatics*, 7(2):263–277, 2010.
- [186] H. Zou. The adaptive lasso and its oracle properties. *J. American Statistical Association*, 101:1418–1429, 2006.
- [187] H. Zou and T. Hastie. The regularization and variable selection via the elastic net. *J. Royal Stat Society B*, 67:301–320, 2005.
-

- [188] M. Zucknick, S. Richardson, and E. Stronach. Comparing the characteristics of gene expression profiles derived by univariate and multivariate classification methods. *Statistical Applications in Genetics and Molecular Biology*, 7(1):7, 2008.