KU LEUVEN

ARENBERG DOCTORAL SCHOOL
Faculty of Engineering Science

Deepening the methodology behind Data integration and Dimensionality reduction

Applications in Life Sciences

Minta Thomas

Supervisor:
Prof. dr. ir. B. De Moor

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

April 2017

Deepening the methodology behind Data integration and Dimensionality reduction

Applications in Life Sciences

Minta THOMAS

Examination committee:
Prof. dr. ir. J. Berlamont, chair
Prof. dr. ir. B. De Moor, supervisor
Prof. dr. ir. J.A.K. Suykens
Prof. dr. ir. L. De Raedt
Prof. dr. ir. J. Aerts
Prof. dr. ir. K. Marchal
(Ghent University)

Dissertation presented in partial fulfillment of the requirements for the degree of Doctor of Engineering Science (PhD): Electrical Engineering

April 2017

© 2017 KU Leuven – Faculty of Engineering Science
Uitgegeven in eigen beheer, Minta Thomas, ESAT - STADIUS, Stadius Centre for Dynamical Systems, Signal Processing and Data Analytics, Kasteelpark Arenberg 10, bus 2446, 3001 Leuven-Belgium (Belgium)

Alle rechten voorbehouden. Niets uit deze uitgave mag worden vermenigvuldigd en/of openbaar gemaakt worden door middel van druk, fotokopie, microfilm, elektronisch of op welke andere wijze ook zonder voorafgaande schriftelijke toestemming van de uitgever.

All rights reserved. No part of the publication may be reproduced in any form by print, photoprint, microfilm, electronic or any other means without written permission from the publisher.

Preface

This thesis summarizes my research work as a PhD student in the bioinformatics group, ESAT - STADIUS, Stadius Centre for Dynamical Systems, Signal Processing and Data Analytics, KU Leuven Belgium. It has been an amazing experience to work in this group and I would like to give acknowledgment to all the people who have helped and supported me during the entire PhD studies.

First and foremost, I would like to thank my supervisor Prof. Bart De Moor for giving me the opportunity to start my PhD research under his supervision. His joint-expertise, constant support, advice and encouragements were a great help for me to find my way out. I appreciate a lot for Bart' fully understanding and continuous funding to support my research work in Leuven. Next, I would like to express my grateful thank to Prof. Johan Suykens for his all guidance and support throughout my PhD. The active discussions with him helped me to get a comprehensive understanding of machine learning, to answer all my questions and dedicated contributions to the manuscript. I also wish to thank the various members of the examination committee of my PhD thesis for their valuable contributions to this thesis. I appreciate them for spending time to read my doctoral thesis and giving constructive comments to improve my thesis. In particular I would like to thank Prof. Luc De Raedt and Prof. Jan Aerts, for their thorough reviewing of the manuscript and many suggestions for improvement that you have applied. I would like to express my grateful thank to Prof. Kathleen Marchal from Ghent University, acts as an external jury for this thesis and valuable suggestions for the improvement of the thesis. I am also grateful to Prof. Jean Berlamont, to be the chairman of jury.

My colleagues in the bioinformatics and machine learning group, also influenced and contributed to my work. Specifically, I would like to thank Anneleen Daemen and Olivier Gevaert for their advice and guidance, in particular during the early stage of my PhD work. Furthermore, I would like to thank Kris De Brabanter for taking time to answer all my machine learning related research questions, for all informative discussions, and dedicated contributions to the manuscripts which

we prepared together. I would like to express my grateful thank to Prof. Yves Moreau who has given us the opportunity to present research work in weekly bioinformatics meetings and for all his valuable suggestions. A part of my work has been done in collaboration with Laboratory for Organic Microwave-Assisted Chemistry (LOMAC), Department of Chemistry and Centre of Microbial and Plant Genetics (CMPG), Department of Microbial and Molecular Systems, KU Leuven. This research work has given me an opportunity to work on real time data sets. I would like to thank Hans P. Steenackers and Prof. Marc de Maeyer for many insights and informative discussions on chemoinformatics research.

I have got the opportunity to work in very pleasant environment with many nice colleagues such as Olivier Gevaert, Griet Laenen, Ryo Sakai, Raf Winand, Sarah Elshal, Pooya Zakeri, Marc Claesen, Tunde Adefioye, Leon-Charles Tranchevent, Nico Verbeeck, Dusan Popovic, Georgios Pavlopoulos, Ernesto Iacucci, Daniela Nitsch, Yu Shi, Sonia Leach, Raf Van de Plas, Lieven Thorrez, Yousef El Aalamat, Joana Goncalves and Jiqiu Cheng, you are all wonderful friends. Thank you so much for all your support, help and friendship. Special thanks to Tunde for his effort on proof reading of the manuscript. I also want to thank Inge for the excellent cooperation during the proposal-writing and external research collaborations. Martin and Elizabeth, thanks for the IT support. Ida, John, Elsy and Mimi thanks for the great help that they offered in each and every steps of the administration process. A very special thanks to Ida to answer all my queries immediately and supported me throughout the entire PhD, especially when I was away from ESAT.

Ultimately, I would like to give thank to my lovely parents and in-laws in India, my friends and families in Belgium and India for their comfort and encouragement during the entire PhD. A simple acknowledgement could never do justice to your support throughout the years. Thank you for everything. Finally, I wish to thank my parents for creating an environment in which I can freely pursue my dreams. My brother and sister for their support through out the entire student life. Perhaps most remarkable is how rarely I have been conscience about this support or how it helped me grow. Finally my husband and daughter who travelled together with me in this journey as a part of my life. Thanks for all your support, understanding and sacrifice during these years.

Minta Thomas
Leuven, Belgium
April 2017

Abstract

The problems of high dimensionality and heterogeneity of data always raise lots of challenges in computational biology and chemistry. As the size of data sets increase, as well their complexity, dimensionality reduction and advanced analytics will gain its importance. The past 10 years or so, data integration has become an active area of research in the field of machine learning, bioinformatics and chemoinformatics.

Several dimensionality reduction and data integration methods are currently available for analyzing and classifying biological data. In the first part of this thesis, we concentrate on dimensionality reduction techniques such as the Generalized Eigenvalue Decomposition (GEVD) and Robust Principal Component Analysis (RPCA). We will investigate the generalized eigenvalue decomposition (GEVD) in a maximum likelihood setting, in which we employ a technique relying on the generalization of the singular value decomposition (SVD). We will elaborate the similarity between maximum likelihood estimation via a generalized eigenvalue decomposition (MLGEVD) and generalized ridge regression. This relationship reveals an important mathematical property of GEVD in which one of the matrices acts as prior information in the model development. Later we present GEVD for the integration of microarray and literature information. Then robust PCA (RPCA) is applied on a weighted matrix for the identification of differentially expressed genes of colon cancer.

In the second part of the thesis, we propose a data-driven bandwidth selection criterion for kernel PCA (KPCA), which is a non-linear dimensionality reduction technique. We center our discussion on feature selection/transformation techniques in medical diagnostics. We show how to build stable, robust and interpretable classifiers on non-linearly separable data.

In the third part of the thesis we investigate a machine learning approach, a weighted LS-SVM classifier to integrate two data sources. This algorithm offers a single mathematical framework for data integration and classification

problems, hence providing solutions for many real bioinformatics applications. Finally, based on PCA, we define new chemical descriptors from the connection-table of chemical compounds. In addition, we develop a new machine learning approach for the identification of biofilm inhibitors of *Salmonella* Typhimurium and *Pseudomonas aeruginosa*. Here, PCA converts the connection-table of each compound into a structural descriptor of two vectors: one corresponding to atoms and the other to bonds. As a supervised classification algorithm, a weighted least squares support vector machine is used in which a table enumerating the atoms is weighted against a table enumerating the bonds. We apply this framework to a given experimental data set on activity of collection of compounds against *Salmonella* and *Pseudomonas* biofilms. This trained model predicts the activity of new compounds on these biofilms.

Beknopte samenvatting

Hoge dimensionaliteit en heterogeniteit van data vergroten de uitdagingen in computationele biologie en chemie. Naarmate de grootte en de complexiteit van de datasets vergroot, zullen dimensie reductie en geavanceerde analytics steeds belangrijker worden. In de laatste 10 jaar is data integratie een actief research onderwerp geworden in machine learning, bioinformatics en chemoinformatics.

Verschillende dimensie reductie en data integratie methoden zijn op dit moment beschikbaar om biologische data te analyseren en classificeren. In het eerste deel van de thesis gaan we ons concentreren op dimensie reductie technieken zoals Generalized Eigenvalue Decomposition (GEVD) and Robust Principal Component Analysis (RPCA). We onderzoeken de generalized eigenvalue decomposition (GEVD) techniek in een maximum likelihood setting waarbij we gebruik maken van een techniek gebaseerd op de generalisatie van de singular value decomposition (SVD). We gaan in op de gelijkenis tussen maximum likelihood estimation via een generalized eigenvalue decomposition (MLGEVD) en generalized ridge regression. Dit verband toont een belangrijke wiskundige eigenschap van GEVD waar een van de matrices zich voordoet als prior information in de ontwikkeling van het model. Later gaan we GEVD voorstellen als integratie techniek van microarray data en literatuur informatie. Daarna wordt Robuste PCA toegepast op een gewogen matrix voor de identificatie van verschillend uitgedrukte genen van darmkanker.

In het tweede deel van de thesis stellen we een data-gedreven selectie criterium voor voor kernel PCA (KPCA). KPCA is een niet-lineaire dimensie reductie techniek. We focussen onze discussie op feature selectie/transformatie technieken in medische diagnostistiek. We tonen hoe een stabiele, robuste en interpreteerbare classifier te maken op niet lineair scheidbare data.

In het derde deel van de thesis onderzoeken we een machine learning aanpak, een gewogen LS-SVM classifier, om twee data bronnen te integreren. Dit algoritme biedt een enkel wiskundig raamwerk voor data integratie en classificatie

problemen, dus het voorziet oplossingen voor vele echte toepassingen in bioinformatics. Tenslotte, gebaseerd op PCA, gaan we nieuwe chemische beschrijvingen definiëren uit de aansluittabel van chemische verbindingen. Daarnaast ontwikkelen we een nieuwe machine learning aanpak voor de identificatie van biofilm remmers van Salmonella Typhimurium en Pseudomonas aeruginosa. Hier zal PCA de aansluittabel van elke verbinding converteren in een gestructureerde beschrijving van twee vectoren: een corresponderend met de atomen en de andere met verbindingen. Als gesuperviseerd classificatie algoritme wordt een gewogen least squares support vector machine gebruikt waarbij een tabel die de atoomaantallen weergeeft gewogen wordt tegen een tabel die de verbindingen weergeeft. We passen dit raamwerk toe op een gegeven experimentele dataset over de activiteit van een verzameling van verbindingen van Salmonella versus Pseudomonas biofilms. Dit getrainde model voorspelt de activiteit van nieuwe verbindingen op deze biofilms.

Glossary

AD	Applicability Domain
ALM	Augmented Lagrange Multipliers
AUC	Area Under the Curve
CacyBP	Calcyclin Binding Protein
cDNA	Complementary DNA
CRC	Colorectal Cancer
CTL	Cytotoxic T cells
DEG	Differentially Expressed Genes
DNA	Deoxyribonucleic acid
EDMD	Early-early Stage Duchenne Muscular Dystrophy
eIF	Eukaryotic initiation factor
ER	Estrogen Receptor
EVD	Eigenvalue Decomposition
FDA	Fisher Discriminant Analysis
FN	False Negatives
FP	False Positives
GEV	Generalized Eigenvectors
GEVD	Generalized Eigenvalue Decomposition
GO	The Gene Ontology
GSVD	Generalized Singular Value Decomposition
HER2	Human Epidermal Growth Factor Receptor 2
HGF	Hepatocyte growth factor
HLA	Human Leukocyte Antigen

IALM	Inexact ALM
ITTACA	Integrated Tumor Transcriptome Array and Clinical data Analysis Database
KPCA	Kernel PCA
Lasso	Least Absolute Shrinkage and Selection Operator
LOO	Leave-One-Out
LOO-CV	Leave-One-Out cross validation
LS-SVM	Least Squares Support Vector Machines
MLGEVD	Maximum Likelihood estimation via Generalized Eigenvalue Decomposition
MLPCA	Maximum Likelihood Principal Component Analysis
mRNA	Messenger RNA
NCI	National Cancer Institute
NSC	Nearest Shrunken Centroids
OAZ1	Ornithine Decarboxylase
PAM	Prediction Analysis of Microarrays
PCA	Principal Component Analysis
PR	Progesterone Receptor
QP	Quadratic Programming
QSAR	Quantitative Structure Activity Relationships
RBF	Radial Basis Function
RNA	Ribonucleic Acid
ROC	Receiver Operating Characteristic
RPCA	Robust PCA
RPL41	Ribosomal Protein L41
RPs	Ribosomal Proteins
rRNAs	Ribosomal RNAs
SRF	Serum Response Factor
SVD	Singular Value Decomposition
SVM	Support Vector Machines

TCTP	Translationally Controlled Tumor Protein
TLS	Total Least Squares
TN	True Negatives
TP	True Positives
Tra1	Tumor Rejection Antigen1
UC	Ulcerative Colitis

Contents

Abstract iii

Contents xi

List of Figures xv

List of Tables xix

1 Introduction 1

1.1 Data Integration and Dimensionality Reduction Techniques . . 1

1.1.1 Data Integration and Dimensionality Reduction in Life Sciences 2

1.2 Motivation and Problem Statement 6

1.3 Thesis overview 12

2 Methodology Overview 17

2.1 Dimensionality Reduction Techniques 18

2.1.1 Principal Component Analysis (PCA) 18

2.1.2 Singular Value Decomposition(SVD) 19

2.1.3 Eigenvalue Decomposition(EVD) 21

2.1.4 Generalized Singular Value Decomposition(GSVD) . . . 21

2.1.5 Generalized EVD(GEVD) 23
2.1.6 Robust Principal Component Analysis (RPCA) 23
2.2 Kernel Methods 25
2.2.1 Support Vector Machines 25
2.2.2 Least Squares Support Vector Machines 28
2.2.3 Kernel PCA 29
2.3 Performance Measures 31

3 A mathematical framework for the incorporation of prior information in medical data analysis 33

3.1 Introduction 34
3.2 Classification Problem 35
3.2.1 Microarray Data 35
3.2.2 Clinical Data 36
3.3 Feature Selection Problem 36
3.3.1 Microarray Data 36
3.3.2 Literature Information 37
3.4 Methods 37
3.4.1 Maximum Likelihood Estimation of Generalized Eigenvalue Decomposition: 37
3.5 Results 40
3.5.1 Classification of Breast Cancer Patients 40
3.5.2 Identification of differentially expressed genes in colon cancer 42
3.6 Discussion 45
3.7 Conclusion 46

4 Robust PCA improves biomarker discovery in colon cancer with incorporation of literature information 47

4.1 Introduction 48
4.2 Method 49
4.2.1 Generalized Eigenvalue Decomposition 49
4.2.2 The RPCA Model of Gene Expression Data 50
4.3 Results 51
4.4 Discussion 55
4.5 Conclusion 56

5 Bandwidth selection criterion of KPCA: Applications in Bioinformatics 57

5.1 Introduction 58
5.2 Data sets 60
5.3 Methods 60
5.3.1 Data-Driven Bandwidth Selection for KPCA 61
5.4 Results 64
5.4.1 Proposed Criterion with PCA 64
5.4.2 Proposed Criterion with Existing Optimization Algorithm for RBF-KPCA 66
5.4.3 Proposed Criterion with Other Classifiers 67
5.5 Discussions 68
5.6 Conclusion 69

6 Predicting breast cancer using an expression values weighted clinical classifier 71

6.1 Background 72
6.2 Data sets 74
6.2.1 Microarray Data 74
6.2.2 Clinical Data 75
6.3 Methods 75

6.3.1 kernel GEVD 76
6.3.2 Weighted LS-SVM classifier 78
6.4 Results 81
6.4.1 Kernel GEVD 81
6.4.2 Weighted LS-SVM classifier 82
6.5 Discussion 85
6.6 Conclusion 86

7 A novel chemoinformatics method for identification of biofilm inhibitors 88

7.1 Introduction 89
7.2 Data Sets 91
7.3 Methods 91
7.3.1 Principal Component Analysis (PCA) 91
7.3.2 Weighted Least Squares-Support Vector Machine (Weighted LS-SVM) Classifier 93
7.4 Result 95
7.5 Discussion 103
7.6 Conclusion 107

8 Conclusion and Future Research 108

8.1 Conclusion 109
8.2 Future Research 112

Bibliography 115

Curriculum vitae 133

List of Figures

1.1 The general picture of the experimental steps involved in microarray analysis 4

1.2 Examples of 1D, 2D, and 3D descriptors: Acetone 6

1.3 QSAR Process 7

1.4 The structure of the thesis and dependence between Chapters . 16

2.1 PCA: (a) Each dot represents a cancer sample plotted against its expression levels for two genes. (In a–c, samples are colored according to type of classes: class 1, red; class 2, black). (b) PCA identifies the two directions (PC1 and PC2) along which the data have the largest spread. (c) Samples plotted in one dimension using their projections onto the first principal component (PC1) for class 1, class 2 and all samples separately. 19

2.2 SVD form of a matrix A 19

2.3 SVD scatter plots on artificial data sets 20

2.4 GSVD decomposition of two expression data sets collected from two experimental conditions. 22

2.5 Robust PCA decomposes the given observation (A) into low rank matrix (R) and sparse matrix (Z), a graphical representation of the equation A=R+Z. 24

2.6 Support Vector Machine For Non-linearly Separable Data: Initially the data is nonlinearly separable. Using feature map, data is transformed from the input space into a high dimensional feature space. After transformation, a linearly assumed classification problem is solved, resulting in a linear hyperplane in feature space. 27

2.7 PCA and KPCA: If the data is not linearly separable, linear projections of the data does not make the data linearly separable, while non-linear projection does. The inverse transform of non-linear projections of the data brings it back to the original space. 29

5.1 Bandwidth selection of KPCA for cervical and colon cancer data on fixed number of components. The plot $J(h)$ vs. h maximizes $J(h)$ at optimal bandwidth h. (a) 5th principal component for cervical cancer data and (b) 15th principal component for colon cancer 62

5.2 Data-Driven Bandwidth Selection for KPCA 63

5.3 The surface plot of Equation (5.3) for various values of h and k. Model selection for KPCA-optimal bandwidth and number of components.(a) Cervical cancer (b) Colon cancer 63

5.4 Slice plot for the Model selection for KPCA for the optimal bandwidth.(a) Cervical cancer (b) Colon cancer 64

5.5 The three dimensional surface plot of LOO-CV performance of optimization algorithm [134] on high-grade glioma data set . . 67

6.1 Overview of the algorithm. The data sets represented as matrices, with rows corresponding to patients and columns corresponding to genes and clinical parameters respectively for first and second data sets. LOO-CV is applied to select the optimal parameters. 80

6.2 Boxplots depict the summary of the classification performances (averaged test AUC on ROC curve in 100 repetitions) of 5 breast cancer cases. CL and MA are the clinical and microarray kernels of RBF kernel functions. CL+LS-SVM and MA + LS-SVM indicates that LS-SVM classifier applied on clinical and microarray kernels. We have used GEVD and kernel GEVD as preprocessing step and then LS-SVM classifier applied on the transformed data set. Finally weighted LS-SVM classifier used as a single mathematical framework for the integration of clinical and microarray kernels and further classification (a) Case I (b) Case II (c) Case III (d) Case IV (e) Case V. 87

7.1 Connection Table of Chemical Compound Melatonin with pubChem ID: 896. 92

7.2 An Example of a Connection-Table 95

7.3 An overview of chemical descriptor formation from the connection-table of compounds. PCA is applied to the connection-table of each compounds to define a new structural descriptor in terms of two vectors. This results into two matrices: atoms vs. compounds and bonds vs. compounds. The weighted LS-SVM framework integrates these two vectors into a single vector named as weighted chemical descriptor and performs further prediction. LOO-CV is applied to select the optimal parameters. 97

7.4 *Salmonella*: Error Bar represents the averaged classification performances in terms of Accuracy, test AUC and F-score over 30 iterations of different descriptors: proposed descriptor, MACCS keys, ECF and Path keys to identify the active and inactive compounds in *Salmonella* biofilm. Proposed descriptor outperformed in terms of test AUC and F-score and MACCS keys outperformed in terms of test AUC. In the error bar x-axis denotes the descriptors and y-axis denotes the test AUC/test accuracy/F-score. 101

7.5 *Pseudomonas*: Error Bar represents the averaged classification performances in terms of Accuracy, test AUC and F-score over 30 iterations of different descriptors: proposed descriptor, MACCS keys, ECF and Path keys to identify the active and inactive compounds in *Pseudomonas* biofilm. Proposed descriptor outperformed in terms of test AUC and F-score and Path keys outperformed in terms of test AUC. In the error bar x-axis denotes the descriptors and y-axis denotes the test AUC/test accuracy/F-score. 102

List of Tables

3.1 Summary of the 3 breast cancer data sets. 35

3.2 MLGEVD and GEVD were used as a pre-processing step which transforms the clinical data onto the direction of generalized eigenvectors. LS-SVM classifier with linear, RBF and polynomial kernel functions applied on these projected clinical data for the patient classification. The performance are measured in terms of average test AUC (std) and average F-score (std) over 30 iterations 41

3.3 Summary of the Data Sets - Identification of differentially expressed genes in colon cancer. 42

3.4 The 50 top ranked genes for relevance in colon cancer diagnosis identified by MLGEVD and GEVD, with the literature references. 43

3.5 LS-SVM model for prediction of tumor and non-tumor samples of colon cancer on whole sets of genes and subsets of genes selected by MLGEVD and GEVD. Average classification performances test AUC (std) are given in terms of test AUC. 44

4.1 Averaged LS-SVM classifier performance, for classification of colon cancer patients, on whole genes, subset of differentially expressed genes, and subset of both differentially expressed and co-expressed genes, over 30 iterations. 52

4.2 23 Differentially expressed genes of colon cancer identified by the proposed method 53

5.1 Summary of the 11 binary disease data sets. 61

5.2 Comparison of classifiers: Mean AUC(std) of 30 iterations . . 65

5.3 Summary of averaged execution time of each classifiers over 30 iterations in seconds. 66

5.4 KPCA + LS-SVM Classifier: Comparison of performance of proposed bandwidth selection criterion for KPCA with the method proposed by Pochet *et al.* [134]: Averaged test AUC(std) over 30 iterations and execution time in minutes 67

5.5 Summary of the range (minimum to maximum) of features selected by t-test over 30 iterations. 69

6.1 Summary of the 5 breast cancer data sets. 74

6.2 The summary of the classification performances (averaged test AUC (std) on ROC curve in 100 repetitions) of 5 breast cancer cases. CL+LS-SVM and MA + LS-SVM indicates that LS-SVM classifier applied on clinical and microarray kernels. Then we have used GEVD and kernel GEVD as preprocessing step and then applied LS-SVM classifier on the transformed data set. Finally the weighted LS-SVM classifier used for data integration and classification on clinical and microarray kernels. The AUC values obtained with different techniques were compared using a paired test, Wilcoxon signed rank test. 83

6.3 Comparisons of RBF with clinical kernel functions: On weighted LS-SVM framework, we evaluated the LOO-CV performances of, clinical kernel function in [39] and RBF microarray kernels, with RBF clinical and microarray kernels. In the weighted LS-SVM classifier framework, RBF kernel functions of clinical parameters performs better than clinical kernel functions on three case studies. 84

7.1 Table enumerating atoms: data source I 96

7.2 Table enumerating bonds: data source II 96

7.3 Comparison of averaged classification performances of different descriptors: proposed descriptor, MACCS keys, ECF, Path keys and BCUT descriptors to identify the active and inactive compounds in *Salmonella* and *Pseudomonas* biofilm. On both case studies, the proposed descriptor outperformed other descriptors in terms of averaged accuracy and F-score. 99

7.4 Comparison of averaged prediction performances of different descriptors - *Thrombin*, *Trypsin* and *FactorXa*. 100

7.5 Compound Id, Structure and IC50 for biofilm prevention of 10 novel compounds used for validation 103

7.6 The 10 validation compounds identified as active by each descriptor (proposed, MACCS, ECF and Path) with different cut-off values (10,20,30,40,50 and 100) for IC50 - *Salmonella* and *Pseudomonas*. The numbers in the bold represent the compound which are correctly identified as active. 104

7.7 Comparison of prediction performances of different descriptors (proposed, MACCS, ECF and Path) to identify active compounds in *Salmonella* biofilm formation. The results are given in terms of test AUC, F-score and accuracy which illustrating the ability of the proposed descriptor to identify compounds with very high activity against the biofilms. The other descriptors missed the ability to identify active compounds at the lower IC50 cut-off values, but with increasing cut-off values, MACCS on *Pseudomonas* outperformed all the other descriptors. 105

7.8 Comparison of prediction performances of different descriptors (proposed, MACCS, ECF and Path) to identify active compounds in *Pseudomonas* biofilm formation. The results are given in terms of test AUC, F-score and accuracy which illustrating the ability of the proposed descriptor to identify compounds with very high activity against the biofilms. The other descriptors missed the ability to identify active compounds at the lower IC50 cut-off values. 106

Chapter 1

Introduction

Bioinformatics and chemoinformatics have evolved at the interface between chemistry, biology and information technology. In many areas of biological and chemical science, the huge amount of data and information produced by high-throughput technologies such as microarray can only be processed and analyzed using computational techniques. Predictive informatics methods employ statistical techniques to mine this data for hidden correlations and to retrieve molecules, genes or patterns with specific properties or desirable biological activities from large datasets. Furthermore, many of these problems are so complex that they require informatics methods for solving them. Informatics methods have been especially valuable in drug design/development and personalized medicine, but are also increasingly employed in several other disciplines governed by complex systems. This calls for an interdisciplinary approach to studying these problems.

1.1 Data Integration and Dimensionality Reduction Techniques

Data integration is the process of integrating data from multiple sources into meaningful and valuable information. The process has an important role in several situations, including business and scientific domains. The data integration aims at combining selected data sources which form one single comprehensive view of data sources. There are varieties of applications that benefit from data integration. In the area of business intelligence, integrated data sources can be used for querying and reporting on business activities,

for statistical analysis, data mining, and machine learning in order to enable forecasting, decision making, and predictions. Data integration in the life sciences becomes a more complex challenge considering the current "data explosion" [69]. A lot of theoretical work dealing with data integration and numerous open problems remain unsolved, mainly on resolving semantic conflicts between heterogeneous data sources [67].

In the last few years, machine learning techniques have been successfully applied to many application areas such as information retrieval [151], image processing [114], computational biology, and chemistry. To understand and explore the real datasets, we often apply machine learning techniques such as clustering or classification in a high dimensional space. However, developing these machine learning models on large data sets can be very time-consuming because of its high dimensionality. Dimensionality reduction is the most important technique in unsupervised learning [43], to get a meaningful structure or previously unknown patterns in the multivariate data.

1.1.1 Data Integration and Dimensionality Reduction in Life Sciences

Data integration in the life sciences is an important area of research, but it is a difficult task. The current advancement in science and technology gives rise to an increasingly wide range of data sources, which in turn must be combined to get an integrated single view of the biological systems. Indeed, the increasing importance of computational biology, the science of using biological data to develop algorithms and relations among various biological systems, emphasizes the importance of data integration in the life sciences.

In the post-genomic era [116], modern biology has generated tremendous information by biological high-throughput technologies, such as proteomics and transcriptomics. This omics data can be very useful, but the real challenge is to analyze all this data, as a whole, after integrating it [153]. Data integration helps us get a comprehensive view of different, heterogeneous and distributed biomedical data sources. Data integration solutions can be very useful in biomedical information retrieval, clinical diagnosis, system biology, drug design etc [153].

The data integration plays [199] an important role in the growth of data-driven knowledge discovery in life sciences. The integration of diverse data helps the biologists to get the interesting patterns or behavior from the data or perform comparative analyses among different biological systems. The effective integration of information from different biological data sources is considered as

the pre-requisite for many computational biology/bioinformatics research and has advantages in a wide range of use cases such as analysis and understanding of omics data, biomarker discovery, and analysis of pathways for drug discovery.

The problem of high dimensionality can be approached with the use of dimensionality reduction methods. Principal component analysis (PCA), singular value decomposition (SVD), eigenvalue decomposition (EVD), generalized singular value decomposition (GSVD) and generalized eigenvalue decomposition (GEVD) are commonly used, dimensionality reduction techniques, for the analysis of high-dimensional data such as genomics, proteomics data etc.

Application Areas

The following sections provide a brief description of the application areas of life sciences which we discuss throughout the thesis: microarray data analysis [29], an introduction to quantitative structure–activity relationship (QSAR) models [75] and clinical data analysis.

Microarray Data Analysis

Microarray technology has become one of the crucial tools that many biologists use to monitor genome-wide expression levels of genes in a given organism. Microarrays may be used to measure gene expression in many ways, but one of the most popular applications is to compare the expression of a set of genes from a cell maintained in a particular condition (Test Cells) to the same set of genes from a reference cell maintained under normal conditions (Control Cells) [173]. Figure 1.1 gives a general picture of the experimental steps involved in microarray data.

The following procedures, we quote from [29]. First, RNA is extracted from the cells, which reverse transcribed into cDNA by using an enzyme reverse transcriptase and nucleotides labeled with different fluorescent dyes (red and green). Once the samples have been differently labeled, they are allowed to hybridize onto the same glass slide. Following the hybridization step, the spots in the hybridized microarray are excited by a laser and scanned to detect the red and green dyes. The amount of fluorescence emitted upon excitation corresponds to the amount of bound nucleic acid. If the gene was expressed to the same extent in both conditions, one would find the spot to be yellow, and if the gene was not expressed in both conditions, the spot would be black. At the end of the experimental stage, we get an image of the microarray, in which

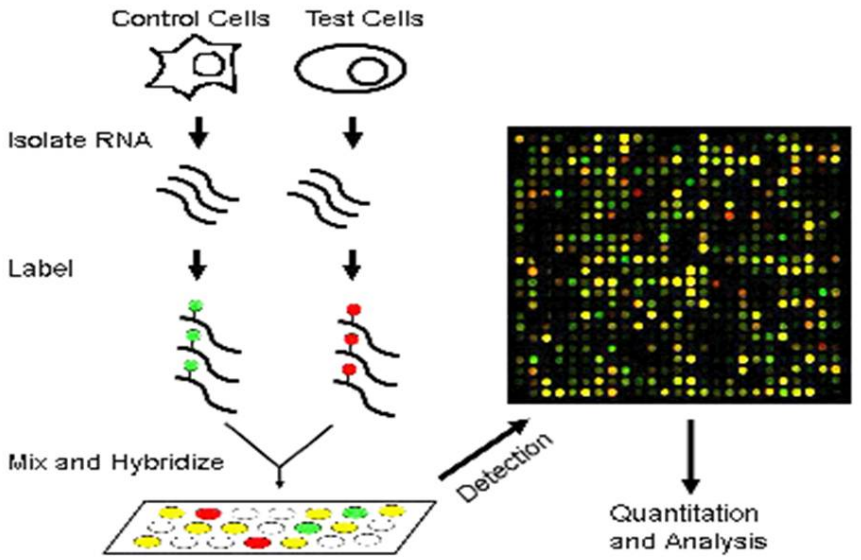


Figure 1.1: The general picture of the experimental steps involved in microarray analysis

each spot representing a gene with fluorescence value indicating the relative expression level of that gene.

The microarray is scanned following hybridization and an image file is normally generated. Then the image is analyzed to identify spots. In the case of microarrays, the spots are arranged in an orderly manner into sub-arrays or pen groups, which makes spot identification straightforward. After identifying areas corresponding to sub-arrays, an area within the sub-array must be selected to get a measure on the spot and background intensity. The relative expression level of a gene can be measured as the amount of red or green light emitted after excitation [9]. The most common metric used to relate this information is called expression ratio. It is denoted here as T_k and defined as
$T_k = R_k/G_k$
For each gene k on the array, where R_k represents the spot intensity metric for the test sample and G_k represents the spot intensity metric for the reference sample. The processed data, after the normalization procedure (total intensity normalization or Mean log centering), can then be represented in the form of a

matrix, often called gene expression matrix. Each row in the matrix corresponds to a particular gene and each column could either correspond to an experimental condition or a time at which expression of the genes has been measured. The expression levels of a gene across different experimental conditions are named as the gene expression profile, and the expression levels of all genes under an experimental condition are named as the sample expression profile.

Quantitative structure–activity relationship (QSAR) models

Molecular descriptors are numerical values that characterize the properties of chemical compounds. Descriptors are frequently divided into 1D, 2D, or 3D descriptors, depending on the dimensionality of the molecular representation from which they can be calculated [20]. Specifically, 1D descriptors are based exclusively on the type of atoms which make up the molecule. In addition to the types of atoms, 2D descriptors also incorporate the bonding pattern of the molecule. 3D descriptors consider the spatial arrangement of the atoms in the molecule. Figure 1.2 shows examples of 1D, 2D, and 3D descriptors of Acetone (PubChem CID:180). The QSAR models [75], are focused on estimating the biological activity of a molecule. QSAR modeling generates predictive models derived from the application of statistical/machine learning techniques correlating biological activity (including desirable therapeutic effect and undesirable side effects) or physio-chemical properties with descriptors representatives of molecular structure or properties. A good quality QSAR model depends on many factors, such as the quality of input data, the choice of descriptors and statistical methods for modeling and for validation. The Figure 1.3 shows the main steps involved in a QSAR process. Any QSAR modeling [196] should ultimately lead to statistically robust and predictive models, capable of making accurate and reliable predictions about the activity of new compounds.

Clinical Data Analysis

Clinical decision support systems are essential to improve the cost-effectiveness of overall health care systems, especially since the cost of health care for age-related diseases such as cancer is significantly increased due to aging populations [12]. In medical applications, clinical decision support system helps clinicians to improve their clinical decision making. Clinicians are also overwhelmed by a tsunami of data for each single patient. Clinical data based on patient history, tumor characteristics, laboratory analysis, ultrasound parameters, or environmental factors are more easily accessible [63]. Information obtained from genome, proteome, transcriptome, contributes to personalize medicine equally

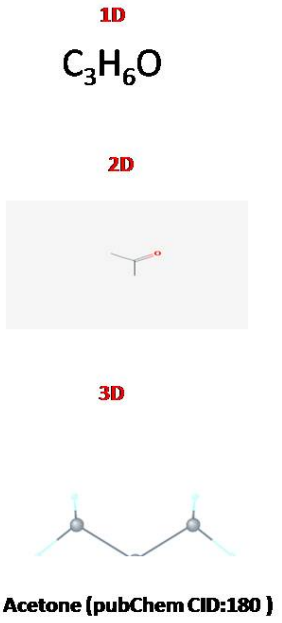


Figure 1.2: Examples of 1D, 2D, and 3D descriptors: Acetone

well. This enables clinicians to make decisions regarding diagnosis, prognosis and response to therapy based on real biological insights on why tumors behave differently from patient to patient.

1.2 Motivation and Problem Statement

Genomics, proteomics, and clinical data being generated from clinical tumors have the potential to transform cancer management. The rapid developments of high-throughput technologies of molecular profiling at the genome, transcriptome, epigenome, protein, and pharmacological levels demand efficient computational approaches for combining or analyzing the results of those technologies.

Recently, data-driven computational cancer modeling has become an active field of cancer research [47]. In particular, the development of cancer models that encompass different biological scales in time and space (i.e. multiscale cancer

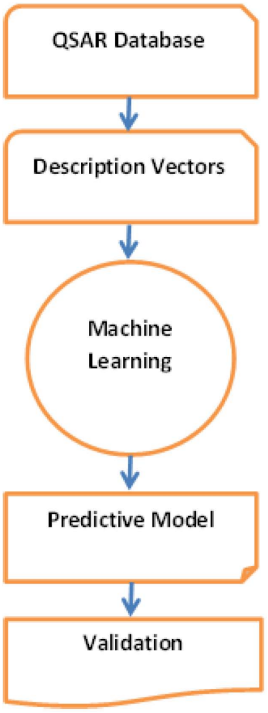


Figure 1.3: QSAR Process

models), has gained attention in view of the potential to integrate disparate kinds of patient data and to enable patient-specific prediction and assist in treatment planning [89]. Several computational tools that assist biologists and human geneticists have been developed, including tools to organize and query scientific literature (such as e.g. Pubmed, GOPubmed [49]) or tools to analyze and interpret high-throughput data such as expression data (GeneSpring, Bioconductor [61]).

Recent technological advances in high-throughput biology have generated vast amounts of disparate biological data describing different aspects of cellular functioning also known as omics layers[66]. It has largely been accepted that a comprehensive understanding of a biological system (cell) can come only from a joint analysis of all omics layers [69, 91]. One of the main aims is to define efficient algorithms that combine existing knowledge with raw data in order to create novel hypotheses to be experimentally assayed and eventually enrich our knowledge. The main goal of any data integration methodology is to extract additional biological information from multiple data sets that cannot be

gained from any single data set alone[66]. To reach this goal, data integration methodologies have to meet many computational challenges. These challenges arise owing to different sizes, formats, and dimensionalities of the data being integrated, also due to their complexity, noisiness, information content etc [66].

Alter *et al.* [3] proposed a comparative mathematical framework, based on generalized singular value decomposition (GSVD), for two genome-scale expression data sets. This framework formulates expression as the superposition of the effects of regulatory programs, biological processes, and experimental artifacts common to both data sets, as well as those that are exclusive to one data set or the other, by using GSVD. This framework enables comparative reconstruction and classification of the genes and arrays of both data sets. This GSVD framework can be used for comparison of two genomic data sets from the same repeated experiments or two different types of genomic information such as DNA copy number, mRNA expression, or protein abundance, collected from the same set of samples to explain the molecular composition of the overall biological signal in these samples, to illustrate the relation between chromosomes of the same organism [3]. GSVD has been successfully used in many bioinformatics applications. As the performance of the model completely depends on choosing parameter estimation criterion, it is necessary to obtain the best estimate of the parameters associated with the GSVD (in terms of generalized eigenvectors). In GSVD, the generalized eigenvectors are generally obtained using matrix decomposition techniques [68]. Of course, there are a variety of optimization criteria to be considered when evaluating parameter estimation methods (e.g. robustness, bias and variance of the estimators) and none is universally the best [36]. One widely used approach is to employ a maximum likelihood criterion. Maximum likelihood estimation, a method of estimating the parameters of a statistical model, help us to obtain the most appropriate estimators of the generalized eigenvectors.

In this part, we focus on the following three research questions:

Q1 Develop a mathematical framework for the maximum likelihood estimation via generalized eigenvectors?

Q2 Does the generalized eigenvectors obtained by maximum likelihood estimation techniques, perform better than generalized eigenvectors obtained by matrix decomposition techniques?.

Q3 Does the data integration using GEVD framework improve the classification/clustering performance in decision-making?

Machine learning techniques have been commonly applied to many scientific domains such as engineering, finance, biology, remote sensing, and economics. The dimensionality of this data could be more than thousands, such as digital

images and videos, gene expressions and DNA copy numbers, documents, and financial time series. The data analysis of such data sets always suffers from the curse of dimensionality. To solve this problem, the dimensionality reduction algorithms have been proposed to project the original high-dimensional feature space to a low-dimensional space, keeping as much information as possible from the original space. The unsupervised dimensionality reduction methods are more useful in the practical applications, especially if we do not have any prior knowledge to new scientific problems, for example, the labeled data are not available.

In microarray analysis, it has been observed that although there are thousands of genes for each observation, a few underlying gene components may account for much of the data variability. Principal component analysis (PCA) provides an efficient way to find these underlying gene components and reduce the input dimensions [15]. This linear transformation has been widely used in gene expression data analysis and compression [200]. If the data are concentrated in a linear subspace, PCA provides a way to compress data and simplify the representation without losing much information. However, if the data are concentrated in a nonlinear subspace, PCA will fail to work well. In this case, one may need to consider kernel principal component analysis (KPCA). KPCA is a nonlinear version of PCA. It has been studied intensively in the last several years in the field of machine learning and has claimed success in many applications [126]. As a kernel method, KPCA suffers from the problem of choosing hyperparameters for kernel functions. No well-founded methods, however, have been established for this based on unsupervised learning. Most of the existing approaches for parameter estimation of KPCA were coupled with the final classifier. In this case, the performance of KPCA obviously depends on the choice of the classifier. This shows the importance of a mathematical technique which selecting the hyperparameters of kernel function based on unsupervised learning.

In this part, we focus on the following three research questions:

Q4 Does the kernel PCA perform better than PCA as a pre-processing step in classification/clustering tasks?

Q5 Design a parameter optimization criterion for RBF-KPCA based on an unsupervised learning?

Q6 What are the advantages and disadvantages of using feature selection and feature transformation techniques?

Principal components analysis (PCA) is a very popular dimension reduction technique which is widely used as a first step in the analysis of high-dimensional microarray data. However, the classical approach which is based on the mean

and the sample covariance matrix of the data is very sensitive to outliers [85]. Also, classification methods based on this covariance matrix give bad results in the presence of outliers in the data. In real-world applications, the data outliers often largely appear in the datasets, thus PCA may not get the optimal performance. Moreover, PCA and KPCA transform the data into a new space, losing the original feature space and hence, further feature selections are made impossible.

In this part, we focus on the following research question:

Q7 What are the alternative analysis for PCA and KPCA, if the data are highly corrupted with noise/outliers?

With the recent rapid developments in high-throughput technologies, such as next-generation sequencing, array comparative hybridization, and mass spectrometry, databases are increasing in both the amount and the complexity of the data they contain. In bioinformatics, it is a challenge to integrate these data to improve the available biological information. Some of the most powerful methods for integrating heterogeneous data types are kernel-based methods [37, 103]. The heterogeneous genomics data, such as expression data, amino acid sequence information, protein-protein interaction data, have been integrated to solve single classification problem: classification of transmembrane and non-transmembrane proteins [103]. Each data was transformed into a kernel matrix and integration occurred on this kernel level without referring back to the data. However, it was important to generate a framework to understand the importance of each data by assigning weights to the kernel matrix. In [37], Daemen and colleagues investigated whether clinical and microarray data can be efficiently combined using a kernel-based approach. They opted for intermediate integration in which the data sets are treated as separate entities and then combined at the kernel level–possibly weighted as before, building one final model. A kernel based network prediction method with automatic data selection for gene expression, phylogenetic profiles and amino acid sequences was proposed in [94]. The heterogeneous data, such as protein characteristics and sequence alignment features, were integrated into a kernel framework for fold recognition [42]. A hierarchical multi-label prediction of gene function was done by combining predictions of multiple SVM classifiers in a Bayesian framework [11]. In all these cases, high accuracy was obtained than when based on any of the data sets individually. However, the final performance of the model depends on the hyperparameters associated with kernel functions, the chosen weights of kernel matrix in the data integration framework and the classifier applied on the integrated data sets. This requires a simple and efficient algorithm to solve both data fusion and non-linear classification problems in a single mathematical framework.

In this part, we focus on the following three research questions:

Q8 Do the non-linear data integration techniques perform better than linear data integration?

Q9 Develop a simple mathematical framework for non-linear data integration?

Q10 What are the benefits of using this framework in data integration tasks?

The chemoinformatics generally intends simply to produce useful computational models that can predict chemical and biological properties of compounds given the chemical structure of a molecule [122]. Traditionally the identification of active molecules is usually performed by screening a libraries of molecules in a wet lab experiment. These procedures remain costly and time-consuming as the number of molecules increases in this library. Virtual screening [140] is based on a computational model which identify the activity of molecules in a specific biological condition of their structure. In this case, an initial set of molecules with known activity information is used to build a model that relates the structure of molecules to its activity. This model development usually involves statistical and machine learning procedures. To build the models first, we need to convert the structural description of the compounds into a numerical representation. The way of representing structural information of molecules into numerical form usually referred to as descriptors in chemoinformatics [71]. More specifically, they are quantitative representations of physical, chemical or topological characteristics of molecules that summarize the molecular structure and activity from different aspects. Deriving a chemical descriptor which completely describes the chemical structure of the compound is one of the most challenging tasks in chemoinformatics. Examples of few chemical descriptors are given below:

- **Extended Connectivity Fingerprints** are generated by first assigning some initial label to each atom and then applying a Morgan type algorithm [123] to generate the fingerprints. Morgan's algorithm consists of l iterations. In each iteration, a new label is generated and assigned to each atom by combining the current labels of the neighboring atoms (i.e, connected via a bond) [105]. The union of the labels assigned to all the atoms over all the l iterations is used as the descriptors to represent each compound.
- **Maccs Keys** [10] are the sets of descriptors based on structural fragments, that have been identified a priori by a domain expert [51]. Each such structural fragment becomes a key and occupies a fixed position in the descriptor space.

- **BCUT descriptor** [21] are based on an earlier descriptor developed by Burden [21] that is calculated from a matrix representation of a molecule's connection table [105]. These descriptors were designed to encode atomic properties relevant to atomic molecular interactions.

Several research scientists have already designed and developed various chemical descriptor and fingerprints for chemical compound representation. Besides their extensive usage in QSAR modeling, based on machine learning techniques [26, 124, 113, 98], these descriptors have a significant potential for the identification of bimolecular targets and network analysis of protein-ligand interactions. By combining the chemical similarity and side-effect similarity, certain potential targets has been identified in [23]. The relationships between proteins function similarity and the ligand structure similarity has been investigated in [95] to predict new high-potential drug targets. Furthermore, several studies used chemical descriptors to predict the drug target interaction [77, 184, 167, 41, 130] and to characterize the structural information of amino acids for developing more effective proteins [25, 70, 48].

The most common representation of the structure of chemical compounds is a connection-table, i.e., a table enumerating the atoms and another enumerating the bonds. Studies already used connection table partially to represent the chemical compounds based on the subdivision and classification of the molecular surface area according to atomic properties [102]. As the connection-table completely represents the chemical structure of chemical compounds, it is important to define a chemical descriptor based on this table, which completely representing the structural properties of chemical compounds.

In this part, we focus on the following three research questions:

Q11 Do the 3D chemical descriptors perform better than 2D descriptors in QSAR process?

Q12 What are the advantages of using chemical descriptors based on connection table of compounds?

Q13 What are the applications of QSAR modelling in biological sciences?

1.3 Thesis overview

This thesis starts with a general introduction to data integration and dimensionality reduction techniques, then continue with motivation and problem statement, and finally the outline of the thesis. The general methodology introduction is given in Chapter 2. Subsequently, Chapter 3 introduces the

maximum likelihood interpretation of the generalized eigenvalue decomposition (GEVD) in which one of the data set acts as prior information in the model development. In real data examples, we show how to incorporate external knowledge extracted from microarray data/literature information into medical diagnosis. In Chapter 4, we use a robust PCA (RPCA) approach as a framework for identifying differentially expressed genes from microarray and literature information. Chapter 5 introduces a new data-driven bandwidth selection criterion for kernel PCA (KPCA) which is related to least squares cross-validation for kernel density estimation. In Chapter 6 we propose a machine learning approach, a weighted LS-SVM classifier to integrate two data sources: microarray and clinical parameters. The proposed model has been shown to be a promising mathematical framework in both data fusion and non-linear classification problems. Chapter 7 proposes a new chemical descriptor for chemical compounds and shows how to predict the activity of chemical compounds in a biological condition. Finally, Chapter 8 presents the conclusion of our work. By summarizing the proposed computational techniques for biological data analysis, this Chapter also elaborates on the opportunities and directions for future research.

Unless stated otherwise, the chemical data analyzed in this thesis (Chapter 7) have been prepared and kindly made available by our collaborative partners:

- Chemical synthesis of the compounds and contribution of compounds: Prof. Marc de Maeyer and Xiaoyu Qing: Laboratory for Organic Microwave-Assisted Chemistry (LOMAC), Department of Chemistry, KU Leuven, Celestijnenlaan 200F, B-3001 Leuven, Belgium.
- Anti-biofilm testing: Prof. Jos Vanderleyden and Dr. Hans P. Steenackers: Centre of Microbial and Plant Genetics (CMPG), Department of Microbial and Molecular Systems, KU Leuven, Kasteelpark Arenberg 20, box 2460, B-3001 Leuven, Belgium.

The main results of this thesis overall are the following:

- The main aim of the work is to show the equivalence between maximum likelihood estimation via a generalized eigenvalue decomposition (MLGEVD) and generalized ridge regression. This relationship reveals an important mathematical property of the generalized eigenvalue decomposition (GEVD), in which the second data matrix acts as prior information in the model. We illustrate the importance of prior knowledge in clinical decision making/identifying differentially expressed genes with case studies for which microarray data sets with corresponding clinical/literature information are available. In our analysis, we have

shown that MLGEVD can be used as an alternative of GEVD for better classification/prediction.
This work was published in the following journal:
Thomas M., Daemen, A., De Moor B. Maximum Likelihood Estimation of GEVD: Applications in Bioinformatics. Computational Biology and Bioinformatics, IEEE/ACM Transactions on. Volume: 11, Issue: 4, 673 - 680: 2014

- In our work a data integration approach is used, in which microarray expressions values are weighted with literature information. Initially, we present the GEVD in terms of the ordinary eigenvalue decomposition (EVD) for the integration of microarray and literature information. Then robust PCA (RPCA) is used for the identification of differentially expressed genes in colon cancer. Initially, we apply RPCA on colon cancer data and then on weighted colon cancer data with literature information. The results suggest that the incorporation of external knowledge into microarray analysis improves the identification of disease-specific genes. Further to identify the co-expressed genes of the obtained disease-specific genes, we make the network analysis of the selected genes using the GeneMania tool. We obtained sets of co-expressed genes which are parts of colon cancer data, but not identified by our approach. Thus, searching for co-expressed genes helped us to identify disease related genes which are really missing in our analysis.

- The ultimate goal of our work is to design a powerful preprocessing step, decoupled from the classification method, for large dimensional data sets. By following the idea of least squares cross-validation in kernel density estimation, we propose a new data-driven bandwidth selection criterion to tune the LS-SVM formulation of RBF-KPCA. The tuned LS-SVM formulation to KPCA is applied to several data sets and serves as a dimensionality reduction technique for a final classification task.
This work was published in the following journal:
Thomas M., De Brabanter K., De Moor B.: New bandwidth selection criterion for Kernel PCA: Approach to Dimensionality Reduction and Classification Problems. BMC Bioinformatics 2014, 15:137 (2014)

- In this work, while bringing up the benefits of LS-SVM classifiers and generalized eigenvalue/singular value decompositions, we propose a machine learning approach, a single mathematical framework for data integration and classification: weighted LS-SVM classifier. The advantages of this new classifier will be demonstrated in five breast cancer case studies, for which expression data and an extensive collection of clinical data are publicly available.
This work was published in the following journal:

Thomas M., De Brabanter K., Suykens J.A.K., De Moor B.: Predicting breast cancer using an expression values weighted clinical classifier. BMC Bioinformatics 2014, 15:6603 (2014).

- Finally, we intend to derive a new chemical descriptor from the connection-table of chemical compounds, allowing a better distinction between biologically active and inactive compounds. Our method is applied to the identification of inhibitors of *Salmonella* and *Pseudomonas* biofilm formation. Development of this type of anti-microbial is urgently needed as biofilms, surface-associated bacterial communities embedded in a self-produced polymeric matrix, provide strong protection against the activity of antibiotics, disinfectants, and the immune system. The results of the validation set illustrate the ability of the proposed descriptor to identify compounds with very high activity against the biofilms. To the best of our knowledge, this is the first time in chemoinformatics that a prediction model is developed on a connection-table of chemical descriptors and a weighted LS-SVM classifier.
  Minta Thomas, Hans P Steenackers, Marc De Maeyer, Johan AK Suykens, Inge Thijs, Xiaoyu Qing, Tran Thi Thu Tran, Erik Van der Eycken, Jos Vanderleyden and Bart De Moor : Chemoinformatics approach to identify new compounds which inhibit bio films formed by either Salmonella or Pseudomonas. Internal Report 16-169, ESAT, KU Leuven (Leuven, Belgium), 2016

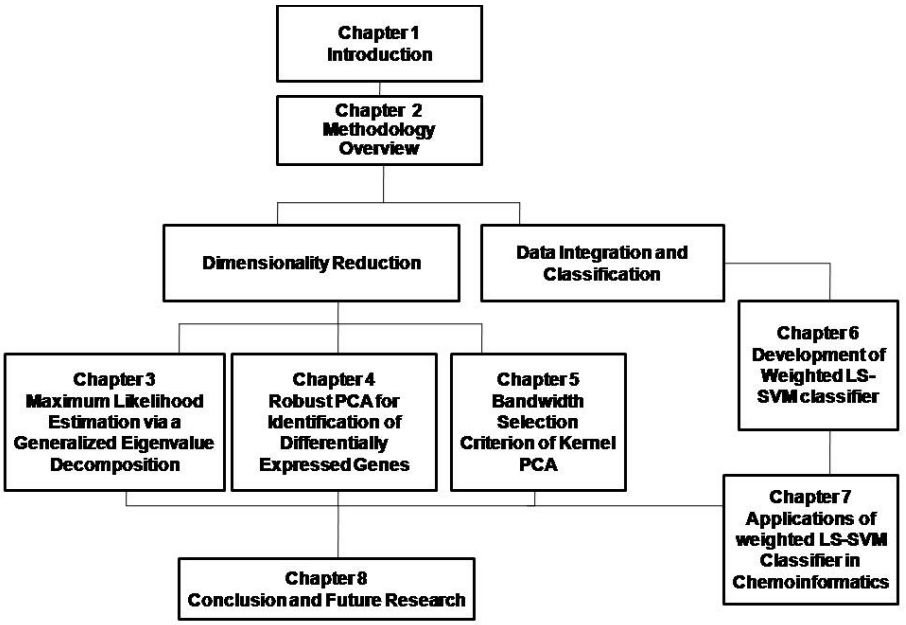


Figure 1.4: The structure of the thesis and dependence between Chapters

Chapter 2

Methodology Overview

High-dimensional datasets present many mathematical challenges, and are bound to give rise to new theoretical developments [50]. One of the problems with high-dimensional datasets is that, in many cases, not all the measured variables are important for understanding the underlying phenomena of interest. While certain computationally expensive novel methods can construct predictive models with a high accuracy of high-dimensional data [8], it is still of interest in many applications to reduce the dimension of the original data prior to any modeling of the data.

Computational techniques for analyzing high-throughput data in genomics, proteomics, and visualization have been extensively studied and have played vital roles in understanding biological mechanisms. Machine learning [99] and related techniques such as support vector machines [185], decision trees [137], and neural networks [191] have been increasingly used to solve problems in genomics and systems biology. Machine learning is a branch of artificial intelligence that induces pattern from past experience or large, complex data by optimizing a performance criterion. It has been shown that machine learning methods substantially improve performances compared to traditional statistical techniques [35]. We have chosen machine learning approaches such as the LS-SVM classifier [170] in connection with dimensionality reduction and data integration for the analysis of biological data sets.

2.1 Dimensionality Reduction Techniques

Dimensionality reduction [133] can also be seen as the process of deriving a set of degrees of freedom which can be used to reproduce most of the variability of a data set. Due to the increase in large data sets, the use of dimensionality reduction techniques has become a necessity in many biological data sets. Many algorithms for dimensionality reduction have been developed to accomplish these tasks. While all of these methods have a similar goal, approaches to the problem are different. Dimensionality reduction can be accomplished in two ways: feature selection [72] and feature transformation [129]. With the feature transformation techniques, the information contained in the original data set to transform into a reduced set of new variables. Principal component analysis (PCA)[128] is a linear-transformation that project data into a lower dimensional space. Although these transformed features may provide a better discriminative ability, they lack a clear physical interpretation. Feature selection techniques, on the other hand, select important features from the given data set, without changing the original representation of the variables. Since both of these techniques are well known pre-processing steps in the field of bioinformatics, we have considered both of them in this dissertation.

2.1.1 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) [128] is mathematically defined as an orthogonal transformation that converts the data into a new coordinate system, such that the largest variance by any projection of the data lies on the first coordinate (first principal component), the second largest variance in the second coordinate, and so on. The full principal components decomposition of $m \times n$ matrix A can therefore, be given as follows:

$$Score = A * Coef$$

where each column of the $n \times n$ matrix $Coef$ are the eigenvectors of $A^T A$ and $Score$, $m \times n$ matrix, is the representation of A in the principal component space.

The basics of PCA can be explained with simple geometrical interpretations of the data as shown in Figure 2.1. To allow for such interpretations, imagine that the microarray in our example, measured the expression levels of only two genes, gene 1 and gene 2. This simplifies plotting the cancer samples according to their expression profiles, which in this case consist of two numbers (Fig. 2.1a). Patient samples are classified as being either positive (class 1) or negative (class 2).

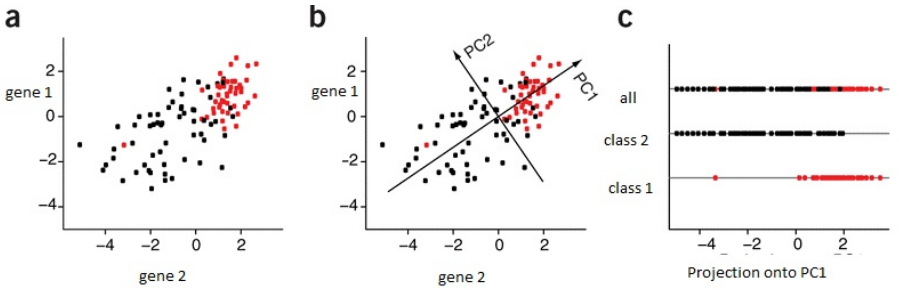


Figure 2.1: PCA: (a) Each dot represents a cancer sample plotted against its expression levels for two genes. (In a–c, samples are colored according to type of classes: class 1, red; class 2, black). (b) PCA identifies the two directions (PC1 and PC2) along which the data have the largest spread. (c) Samples plotted in one dimension using their projections onto the first principal component (PC1) for class 1, class 2 and all samples separately.

2.1.2 Singular Value Decomposition(SVD)

Any real $m \times n$ matrix can be factored as [68]

$$A = U\Sigma V^T$$

where U is an$m \times m$ orthogonal matrix whose columns are the eigenvectors of AA^T, V is an $n \times n$ orthogonal matrix whose columns are the eigenvectors of A^TA, and Σ is an $m \times n$ diagonal matrix. There are several facts about SVD

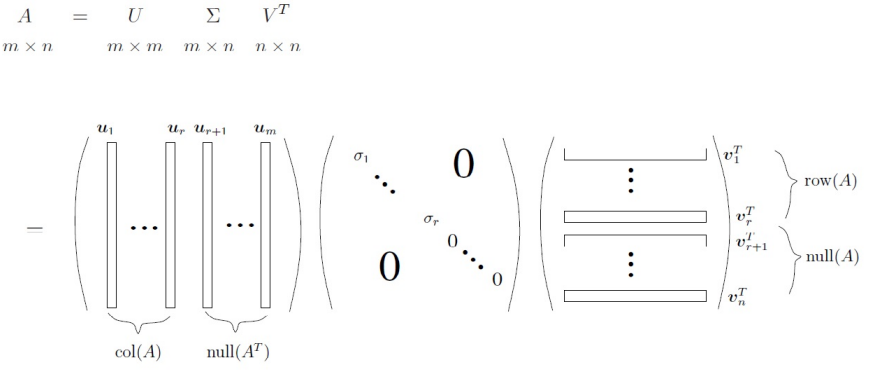


Figure 2.2: SVD form of a matrix A

[88]:
(a) rank(A) = rank(Σ) = r
(b) The column space of A is spanned by the first r columns of U.
(c) The null space of A is spanned by the last n - r columns of V.
(d) The row space of A is spanned by the first r columns of V.
(e) The null space of A^T is spanned by the last m - r columns of U.

In gene expression analysis generally, we want to classify samples in a diagnostic study, or classify genes in system biology. Projection of data into SVD subspaces and visualization with scatter plots can reveal structures in the data that may be used for classification. If we have a $m \times n$ gene expression data A, the SVD decomposes it into $A = U\Sigma V^T$. The projection of data into the direction of V, i.e, AV is used to classify genes and the projection of data into the direction of U, i.e, $U^T A$ is used to classify samples. For projection, instead of using all columns of orthogonal matrices U and V, we will choose only the first r columns as shown in Figure 2.2, which capturing the majority of the variance in the data sets. The Figure 2.3 illustrates how the SVD can be useful for

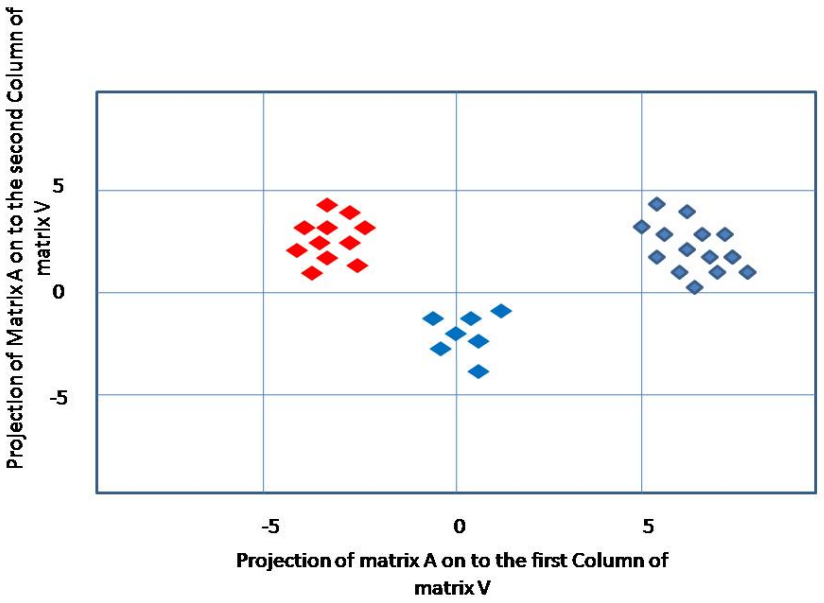


Figure 2.3: SVD scatter plots on artificial data sets

classification/clustering tasks. Assume that the gene expression data A contains three different classes of genes. Initially data is projected onto the subspace of SVD (i.e, the first two columns of the matrix V in Figure 2.3) and then simple scatter plot is used for grouping the genes into different clusters.

2.1.3 Eigenvalue Decomposition(EVD)

The eigenvalue decomposition of matrix A^TA and AA^T [68] is given as follows:

$$A^TA = VDV^T,$$

$$AA^T = UD'U^T,$$

where V is an $n \times n$ orthogonal matrix consisting of the eigenvectors of A^TA, D is an $n \times n$ diagonal matrix with the eigenvalues of A^TA on the diagonal, U an $m \times m$ orthogonal matrix consisting of the eigenvectors of AA^T, and D' is an $m \times m$ diagonal matrix with the eigenvalues of AA^T on the diagonal. It turns out that D and D' have the same non-zero diagonal entries except that the order might be different.

The SVD of a matrix A gives the complete eigensystems of AA^T and A^TA without forming these products explicitly. From the standpoint of accuracy, this is the right way to compute these eigensystems, since information about the smaller eigenvalues are lost when AA^T and A^TA are computed in floating-point arithmetic [192]. Hence we worked with SVD, instead of EVD for which explicit commands are provided in Matlab. The numerically optimal way to calculate GEVD is via 3 SVDs.

2.1.4 Generalized Singular Value Decomposition(GSVD)

The generalized singular value decomposition (GSVD) of $m \times n$ matrix A and $p \times n$ matrix B is [68] is obtained from the singular value decomposition (SVD) of matrix $[A; B]$ as follows:

$$A = U\Sigma_A X^T \tag{2.1}$$

$$B = V\Sigma_B X^T \tag{2.2}$$

where U, V are orthogonal matrices and columns of X are generalized singular vectors.

Figure 2.4 illustrates how GSVD perform simultaneous linear transformation of the two expression data sets from the two m-genes $\times$ n-arrays and p-genes $\times$ n-arrays spaces to the two reduced n-genelets $\times$ n-arraylets spaces [3]. The genelets and arraylets are data-driven decoupled superpositions of genes and

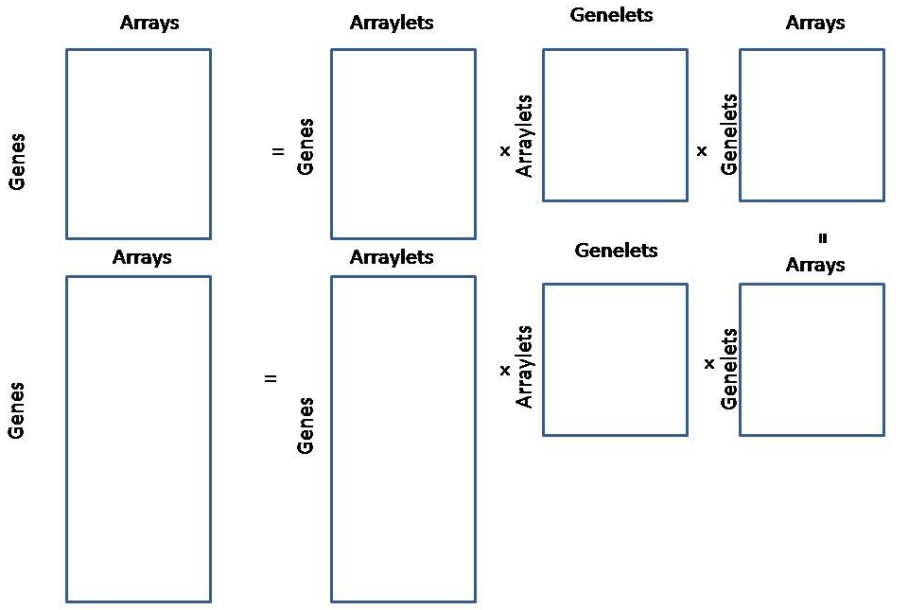


Figure 2.4: GSVD decomposition of two expression data sets collected from two experimental conditions.

arrays. The arraylets are orthogonal projections of the datasets onto the space spanned by the non-orthogonal genelets, that are shared by both datasets [3]. In the GSVD expression $A = U\Sigma_A X^T$, the values in the i^{th} row of U express this projection for the i^{th} gene. In that row, the column with the largest absolute value corresponds to the genelet which explains the greatest proportion of variance. The m^{th} row vector in X^T lists the expression signal of the m^{th} genelet across the different arrays in both data sets simultaneously. The GSVD framework provides a technique to combine two gene expression datasets with only partial overlap of both gene sets and experimental conditions. The framework help us to identify the genes which present in only one dataset co-expressed with a target gene present exclusively in the other dataset, even when experimental conditions for the two datasets are not identical [148].

2.1.5 Generalized EVD(GEVD)

If B^TB is invertible, then the GEVD [68] of A^TA and B^TB can be obtained from Equations 2.1 and 2.2 as follows:

$$A^TA(X^T)^{-1} = B^TB(X^T)^{-1}\Lambda. \tag{2.3}$$

where $(X^T)^{-1}$ is generalized eigenvectors, Λ is a diagonal matrix with diagonal entries $\Lambda_{ii} = (\frac{\Sigma_{A_{ii}}}{\Sigma_{B_{ii}}})^2$, $i = 1, \ldots, N$. $A(X^T)^{-1}$ indicates projection of data A into the direction of generalized eigenvectors. Similar to EVD problem, these projected data can be used further for classification of samples or clustering of features.

2.1.6 Robust Principal Component Analysis (RPCA)

Finding a low-rank decomposition of a matrix is an essential tool in data mining and information retrieval [7]. Prominent applications include in summarizing adjacency matrices for social network analysis or term-document matrices for text classification [78]. In addition to missing values, microarray data are often corrupted with extreme values (outliers). Since the ordinary SVD is not robust enough to outliers, RPCA is an alternate decomposition technique to obtain the low-rank approximation of the matrix.

RPCA, a new method for matrix recovery has been introduced recently in the field of signal processing [24]. The problem of matrix recovery can be described as follows, assume that all the data points are stacked as column vectors of a matrix A, and the matrix (approximately) have low rank:

$$A = R + Z$$

where R has low-rank and Z is a perturbation matrix. The robust PCA proposed by Candes *et al.* can recover a low-rank matrix R from highly corrupted measurements A [24]. Here, the entries in Z can have an arbitrarily large magnitude, and their support is assumed to be sparse but unknown. Figure 2.5 demonstrates how the RPCA decomposing the given observation into a low-rank matrix and an error matrix.

RPCA was proposed by Candes and colleagues in [24]. Let $\|A\|_* = \Sigma_i \sigma_i(A)$ denote the nuclear norm of the matrix A, that is, the sum of its singular values, and $\|R\|_1 = \Sigma_{ij}|R_{ij}|$ denote the l_1-norm of R, which is efficient and robust to outliers. For a given data matrix A, RPCA solves the following optimization problem:

$$\min_{R,Z} \|R\|_* + \lambda\|Z\|_1, \tag{2.4}$$

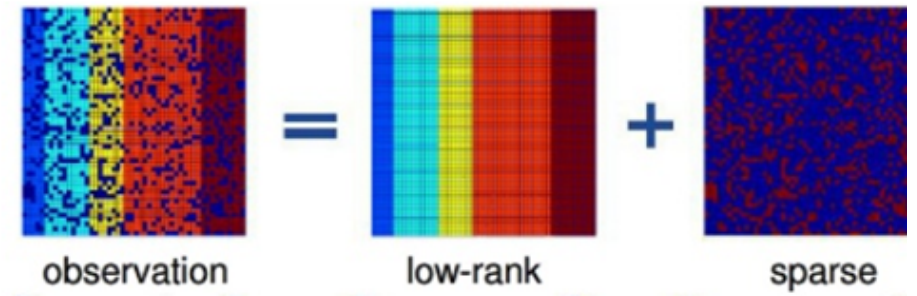


Figure 2.5: Robust PCA decomposes the given observation (A) into low rank matrix (R) and sparse matrix (Z), a graphical representation of the equation A=R+Z.

subject to $A = R + Z$, where λ is a positive regulation parameter. According to [110], the Augmented Lagrange multiplier method on the Lagrangian function can be applied: Thus the two components R and Z can be exactly recovered through solving the following convex problem,

$$L(A, R, Z, \mu) = \|A\|_* + \lambda\|R\|_1 + < Y, A - R - Z > +(\mu/2)\|A - R - Z\|_F^2$$

where Y can now be interpreted as an estimate of a dual variable, μ is a positive scalar and $\|.\|_F^2$ denotes the Frobenius norm. Several mathematical techniques are available to solve the optimization problem in Equation 2.4. Here we have chosen the inexact ALM (IALM) algorithm proposed in [110] to solve the RPCA problem due to its accuracy, stability, and fast convergence. The IALM algorithm for solving the RPCA problem can be designed by adapting the algorithm of Augmented Lagrange Multipliers. The exact convergence rate of the IALM is difficult to obtain in theory, that is why it is known as inexact, but extensive numerical experiments have shown that it still converges Q-linearly [154]. An Augmented Lagrangian algorithm [13] consists of a sequence of outer iterations. At each outer iteration, a minimization problem with simple constraints is approximately solved whereas Lagrange multipliers Y and penalty parameter μ are updated in the master routine.

2.2 Kernel Methods

Kernel methods are a class of algorithms for pattern analysis, to find and study general types of relations in general types of data such as sequences, text documents, images, etc [147]. These methods work by mapping data x into a high dimensional feature space with a nonlinear feature map $\phi(x)$. This kernel function forms an inner product $k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ between all pairs of data items x_i and x_j in the feature space. Any symmetric, positive semidefinite function is a valid kernel function, resulting in many possible kernels [82]. The functions that are most frequently employed in classification problems are the linear kernel ${x_i}^T x_j$, the polynomial kernel$({x_i}^T x_j + b)^d$ with - as kernel parameters - the intercept constant $b \in \mathbb{R}^+$ and degree $d \in \mathbb{N}$, the radial basis function (RBF) $\exp(-||x_i - x_j||_2^2/\sigma^2)$ with $\sigma \in \mathbb{R}^+$ representing the bandwidth [147].

2.2.1 Support Vector Machines

Support vector machines (SVM) [185] are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. In addition to performing linear classification, SVMs can efficiently perform non-linear classification using the so-called kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.

More formally, a support vector machine constructs a hyperplane or set of hyperplanes in a high - or infinite - dimensional space, which can be used for classification, regression, or other tasks. Intuitively, a good separation is achieved by the hyperplane that has the largest distance to the nearest training data point of any class, since in general the larger the margin the lower the generalization error of the classifier.

Consider a training set for classification $\{x_i, y_i\}_{i=1}^N$ of N samples with feature vectors $x_i \in \mathbb{R}^p$ and binary output labels $y_i \in \{-1, +1\}$. The aim is to design a function $f(x) = y$ that correctly classifies unseen samples $\{x, y\}$. Data points x_i with $f(x_i) \geq 0$, are assigned to the label +1, data points with $f(x_i) \leq 0$ to the label -1. Vapnik [185] considered a set of hyperplanes $\{w^T x + b = 0\}$, corresponding to a linear function of the form $f(x) = w^T x + b$. Variable b represents the bias term and w is the normal vector to hyperplane. Vapnik

further introduced a principled way to choose the best possible hyperplane among all hyperplanes that separates two classes of samples. Simply for linear separable data, the corresponding classifier was defined as $y(x) = sign[w^T x + b]$, with the closet data points x_i satisfying the equality constraint $[w^T x_i + b] = 1$ or $[w^T x_i + b] = -1$. This optimal discriminant boundary is known as the linear SVM. The above constraint forces the margin (that is,the distance between the nearest points of both classes) to be equal to $\frac{2}{\|w\|_2}$. The SVM classifier is thus obtained from the solution to a convex optimization problem in which the margin is maximized or equivalently $\|w\|_2^2 = w^T w$ minimized, subject to the constraint of correctly classifying the training data points with strong confidence.

The optimization problem in primal weight space, is defined as [185]:

$$\min_{w,b}(\frac{1}{2}w^T w)$$

subject to

$$y_i[w^T x_i + b] \geq 1, i = 1, \ldots, N$$

where N is the number of data points.

As the dimension of w is determined by the number of elements in the feature vector x_i, the calculation of w can be avoided for problems with high dimensional data by solving the problem as a quadratic programming (QP) problem in dual space. The unknown variables in dual space are the Lagrange multipliers α_i, referred to as support values. Vector α is of size N and thus typically N small for high-throughput data sets. The resulting classifier equals:

$y(x) = sign[\Sigma_{i=1}^N \alpha_i y_i {x_i}^T x + b]$.

In most binary classification problems, classes are characterized by overlapping distributions in such a way that a separating hyperplane does not exist. To obtain an algorithm that can cope with misclassifications and a certain fraction of outliers, the problem is extended with slack variables η. The non-negative slack variables, η_i, which measure the degree of misclassification of the data x_i, are introduced as:
$y_i[w^T x_i + b] \geq 1 - \eta_i$, $i = 1, \ldots, N$.

For data that are not linearly separable, the decision boundary can directly be constructed in the input space and the problem formulation in primal space changes into [185]:

$$\min_{w,b}(\frac{1}{2}w^T w + C\Sigma_{i=1}^N \eta_i)$$

subject to

$$y_i[w^T x_i + b] \geq 1 - \eta_i,$$

$\eta_i \geq 0,\ i = 1, \ldots, N$

The positive hyper parameter C represents the trade-off between the requirement of large margin and that of few misclassifications (that is, $\eta_i > 1$) or classifications with little confidence ($0 < \eta_i \leq 1$). Increasing C forces the function $f(x) = w^T x + b$ to have a smaller margin, but correctly classifying more data points with strong confidence. In the case of nonlinearly separable or

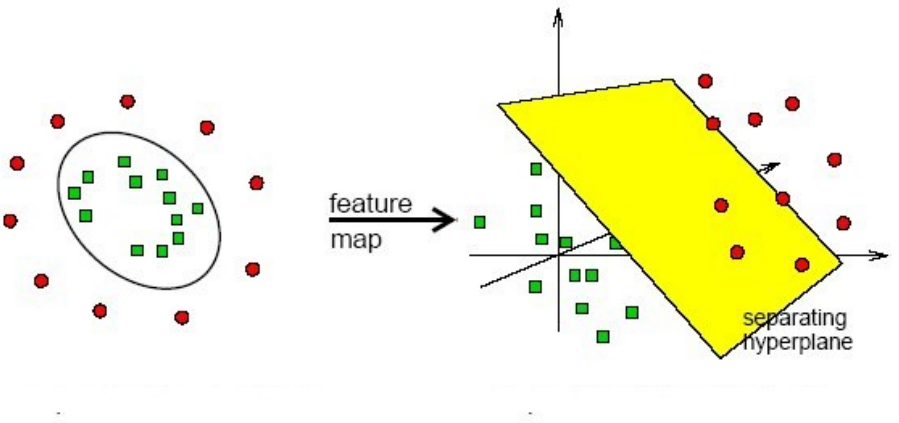


Figure 2.6: Support Vector Machine For Non-linearly Separable Data: Initially the data is nonlinearly separable. Using feature map, data is transformed from the input space into a high dimensional feature space. After transformation, a linearly assumed classification problem is solved, resulting in a linear hyperplane in feature space.

non-separable data, a transformation of the data from the input space into a high dimensional feature space is required as a $\phi(x)$. The constructed hyperplane corresponds to the nonlinear discriminant boundary $y(x) = sign[w^T \phi(x) + b]$ in the original input space. Figure 2.6 illustrates the principle of the SVM for non-linearly seperable data. The linear SVM is thus easily extendable to the nonlinear case by replacing x_i by $\phi(x_i)$ in the problem formulation and applying the kernel trick to avoid explicit knowledge of $\phi(x)$, which is often infinite dimensional as

well as w. The nonlinear SVM therefore needs to be solved in dual space with as discriminant boundary $y(x) = sign[\Sigma_{i=1}^{N} \alpha_i y_i k(x_i, x) + b]$. The kernel function $k(x_i, x)$ interprets inner products between all pairs of data items x_i and x in the feature space. Support vector machines are also suited for regression problems. In case of a linear problem, the function $f(x) = w^T x + b$ is estimated using the training data with continuous out variable $y \in R$. Similarly to classification, this strategy can be extended to nonlinear regression problem by replacing x by $\phi(x)$. The resulting SVM model becomes $y(x) = \Sigma_{i=1}^{N}[\alpha_i k(x_i, x) + b]$.

2.2.2 Least Squares Support Vector Machines

A modified version of the SVM, the LS-SVM, was developed by Suykens *et al* [170, 172]. The main difference between standard SVM and the LS-SVM is that the difference in loss function and the inequality constraints in the formulation of SVM are replaced by equality constraints, thereby considering the value 1 at the right-hand side as a target value rather than a threshold value. In this way, the interpretation of slack variables η_i slightly changes into the offset with respect to target 1, and the variables are referred to as error variable e_i. Moreover, the second term in the objective function is replaced by the squared error contributions. These modifications transform the QP problem, which requires expensive computations, into a much simpler set of linear equations. The LS-SVM is therefore much faster on high-dimensional data sets with low computational cost. The formula for this constrained convex optimization problem equals:

$$\min_{w,b,e} \frac{1}{2} w^T w + \gamma \frac{1}{2} \Sigma_{i=1}^{N} e_i^2$$

subject to:

$$y_i[w^T \phi(x_i) + b] = 1 - e_i, \qquad i = 1, \ldots, N$$

with γ the regularization parameter, representing-in analogy to the hyperparameter C the trade-off between margin maximization and minimization of the squared error contribution. In dual space the equivalent problem is obtained by solving the Lagrangian, resulting in a system of linear equations in functions of the number of data points N:

$$\begin{bmatrix} 0 & y^T \\ y & \Omega + \frac{I}{\gamma} \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 1_N \end{bmatrix}$$

with α and b are known variables and $\Omega_{ij} = y_i y_j k(x_i, x_j)$, $i, j = 1, \ldots, N$.

2.2.3 Kernel PCA

KPCA, which is a generalization of PCA, a nonlinear dimensionality reduction technique that has proven to be a powerful pre-processing step for classification algorithms. It has been studied intensively in the last several years in the field of machine learning and has claimed success in many applications [126]. An algorithm for classification using KPCA was developed by Liu *et al.* [111]. KPCA was proposed by Schölkopf and Smola [146], by considering a mapping to a high-dimensional feature space (possible infinite) and applying Mercer's theorem. Figure 2.7 illustrates how does KPCA find a projection of the data which makes data linearly separable.

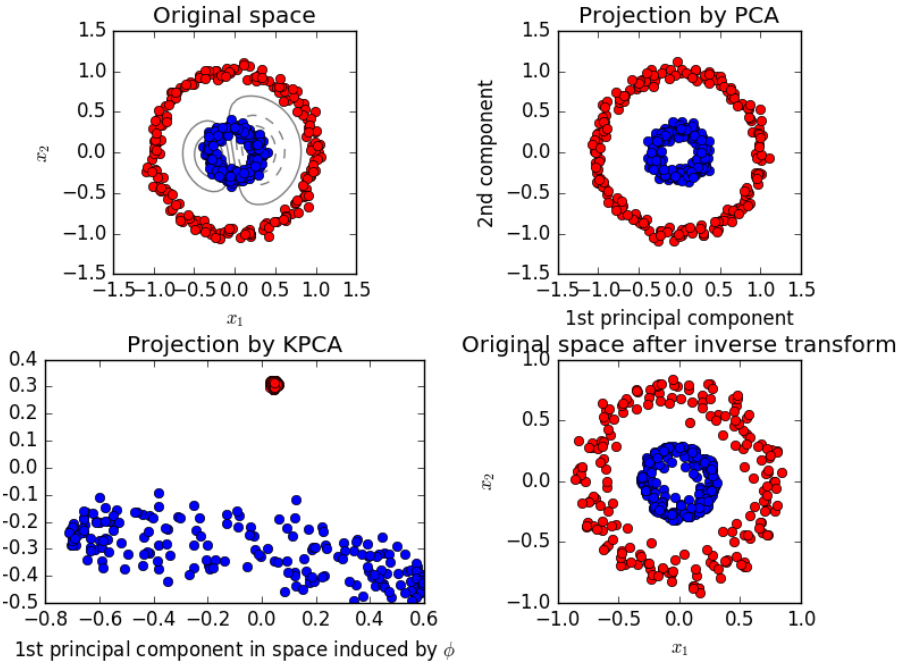


Figure 2.7: PCA and KPCA: If the data is not linearly separable, linear projections of the data does not make the data linearly separable, while non-linear projection does. The inverse transform of non-linear projections of the data brings it back to the original space.

Suykens *et al.* [170, 171] proposed a simple and straightforward primal-dual support vector machine formulation to the PCA problem as follows: The PCA analysis problem is interpreted as a one-class modeling problem with a target

value equal to zero around which the variance is maximized. This results into a sum of squared error cost function with regularization. The score variables are taken as additional error variables. We now follow the usual SVM methodology of mapping the data from the input space to a high-dimensional feature space $\phi : \mathbb{R}^d \to \mathbb{R}^{n_h}$, where n_h can be infinite, and apply Mercer's theorem [118].

Our objective is the following

$$\max_v \sum_{k=1}^{N} [0 - v^T(\phi(x_k) - \hat{\mu}_\phi))]^2$$

with $\hat{\mu}_\phi = (1/N) \sum_{k=1}^{N} \phi(x_k)$ and v is the eigenvector in the primal space with maximum variance. This formulation states that one considers the difference between $v^T(\phi(x_k) - \hat{\mu}_\phi)$ (the projected data points to the target space) and the value 0 as error variables. The projected variables correspond to what is called *score* variables. These error variables are maximized for the given N data points. Next, by adding a regularization term we also want to keep the norm of v small. The following optimization problem is formulated now in the primal weight space

$$\max_{v,e} J_P(v, e) = \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2} v^T v$$

such that

$$e_k = v^T(\phi(x_k) - \mu_\phi), k = 1, ..., N.$$

The Lagrangian yields

$$\mathcal{L}(w, e; \alpha) = \gamma \frac{1}{2} \sum_{k=1}^{N} e_k^2 - \frac{1}{2} v^T v - \sum_{k=1}^{N} \alpha_k (e_k - v^T(\phi(x_k) - \hat{\mu}_\phi))$$

with conditions for optimality

$$\frac{\partial \mathcal{L}}{\partial v} = 0 \to v = \sum_{k=1}^{N} \alpha_k (\phi(x_k) - \hat{\mu}_\phi)$$

$$\frac{\partial \mathcal{L}}{\partial e_k} = 0 \to \alpha_k = \gamma e_k$$

$$\frac{\partial \mathcal{L}}{\partial \alpha_k} = 0 \to e_k - v^T(\phi(x_k) - \hat{\mu}_\phi) = 0$$

By elimination of variables e and w, one obtains

$$\frac{1}{\gamma} \alpha_k - \sum_{l=1}^{N} \alpha_l (\phi(x_l) - \hat{\mu}_\phi)^T (\phi(x_k) - \hat{\mu}_\phi) = 0 \quad k = 1, \ldots, N.$$

Defining $\lambda = \frac{1}{\gamma}$, one obtains the following dual problem

$$\Omega_c \alpha = \lambda \alpha$$

where Ω_c denotes the centered kernel matrix with ijth entry: $\Omega_{c,i,j} = K(x_i, x_j)$ - $\frac{1}{N}\sum_{r=1}^{N} K(x_i, x_r)$ - $\frac{1}{N}\sum_{r=1}^{N} K(x_j, x_r)$ + $\frac{1}{N^2}\sum_{r=1}^{N}\sum_{s=1}^{N} K(x_r, x_s)$.

2.3 Performance Measures

In machine learning and statistics, two classes of samples that need to be distinguished are often referred to as the positive and negative class. Throughout this dissertation, the positive class is often annotated with disease samples, while the negative class contains the normal samples.

For the evaluation of a binary classifier, the predicted labels are compared with the given output labels, for all samples. The measures are given by, true positives (TP) - the number of correctly classified positive samples, false positives (FP) - the number of incorrectly classified negative samples, true negatives (TN) - the number of correctly classified negative samples and false negatives (FN) - the number of incorrectly classified positive samples. The accuracy represents the proportions of N samples that are correctly classified. The sensitivity or true positive rate equals the proportion of positive samples that are correctly classified as such, while specificity, also known as true negative rate, measures the proportion of negative samples identified as negative.

- $Accuracy = \frac{(TP+TN)}{N}$
- $Sensitivity = \frac{TP}{(TP+FN)}$
- $Specificity = \frac{TN}{(TN+FP)}$
- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$
- $F - Score = 2.\frac{Precision*Recall}{Precision+Recall}$

For all of the above performance criteria, samples are assigned to the positive or negative class by a binary classifier. The receiver operating characteristics (ROC) curve [14, 119] summarizes the performance of a classifier by showing the true positive rate (sensitivity) versus the false positive rate (1-specificity) as the discrimination threshold is varied. Each threshold corresponding to a particular operating point on the ROC curve. Increasing the threshold simultaneously enlarges the number of true positives and false alarms, whilst reducing the number of false positives automatically reduces the number of hits. The area under the ROC curve (AUC) is an evaluation measure that is independent of the operating point. An AUC of 0.5 corresponds to a random classifier lacking discriminative power, while a perfect classifier is characterized by an AUC of 1. Thus, the larger the AUC, the better the classifier, with a high sensitivity at a small positive rate (high specificity). The F-Score are indicative of precision and recall, with scores ranging from 0 to 1.

Chapter 3

A mathematical framework for the incorporation of prior information in medical data analysis

The work was published as Thomas, M., Daemen, A., De Moor, B., Maximum Likelihood Estimation of GEVD: Applications in Bioinformatics, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Volume: 11, Issue: 4, 673 - 680: 2014

Abstract: We propose a method, maximum likelihood estimation via a generalized eigenvalue decomposition (MLGEVD), that employs a well known technique relying on a generalization of singular value decomposition (SVD). The main aim of the work is to show the equivalence between MLGEVD and generalized ridge regression. This relationship reveals an important mathematical property of the generalized eigenvalue decomposition (GEVD), in which the second argument acts as prior information in the model. Thus we show that MLGEVD allows the incorporation of external knowledge on the quantities of interest into the estimation problem. We illustrate the importance of prior knowledge in clinical decision making/in identifying differentially expressed genes with case studies for which microarray data sets with corresponding clinical/literature information are available. MLGEVD results in significantly improved diagnosis, prognosis and prediction of therapy response.

3.1 Introduction

Microarray technology is a significant tool in gene expression analysis and cancer diagnosis. These technologies are typically used for class discovery [136, 144] and prediction [33, 157]. For most diseases and examinations, clinical data such as age, gender and medical history guide clinicians in diagnosis. The effective management of these data always leads to better clinical prognosis. Microarray data is in general much more difficult and expensive to collect while clinical parameters are routinely measured by clinicians. A vital study on the prediction of breast cancer outcome has suggested that despite the emergence of these high-throughput technologies, clinical markers and profiles have similar power for prognosis [52]. Studies show that clinical and microarray data sets improve the prediction accuracy [39] in clinical decision making.

Biomarker discovery and prognosis prediction are essential for improved personalized treatment of cancer. Principal component analysis (PCA) and PCA-based approaches for example, were used for the identification of differentially expressed genes (DEG) in pulmonary adenocarcinoma [76] and E. coli [90]. Troyanskaya and colleagues developed nonparametric methods to identify DEG in microarray data [180]. Besides well-known statistical tests such as the chi-square test [155], Chun and colleagues proposed a new test, the 'half Student's t-test', specifically for detecting DEG in heterogeneous diseases [83]. The singular value decomposition (SVD) and the generalized SVD (GSVD) have been shown to have great potential within bioinformatics for extracting common information from data sets such as genomics and proteomics data [152, 3]. Maximum likelihood principal component analysis (MLPCA) is an error-in-variables modeling method in that it accounts for measurement errors in the estimation of model parameters. Wentzell *et al.* [194] generalized the PCA method to MLPCA [193, 195]. The tight equivalence between MLPCA and total least squares (TLS) is explored in [149]. Finally, several studies have developed methods for integrating literature and microarray data sets for identifying disease related genes [56, 64].
In this Chapter we propose a method which incorporates external knowledge of interest in the analysis of microarray and clinical data sets. The main aim of the work is to show the equivalence of maximum likelihood estimation via the generalized eigenvalue decomposition (MLGEVD) with generalized ridge regression. This reveals an important mathematical property of GEVD in which the second matrix acts as the prior information in the model. We incorporate microarray/literature information as prior information in the model to improve the accuracy in clinical decision making.

3.2 Classification Problem

Breast cancer is one of the most extensively studied cancer types for which many microarray data sets are publicly available. Among them, we selected three cases for which also clinical information was available [31, 79, 158]. All the three data sets are available in the Integrated Tumor Transcriptome Array and Clinical data Analysis database (ITTACA): `http://bioinfo-out.curie.fr/ittaca/`. Overview of all the data sets are given in Table 3.1.

Table 3.1: Summary of the 3 breast cancer data sets.

Case Study	#Samples		#Genes	#Clinical variables
	Class 1	Class2		
Case I	85	25	4997	Age, Ethnicity, ER status, PR status, Radiation treatment, Chemotherapy, Hormonal therapy, Nodal status, Metastasis, Tumor stage, Tumor size, Tumor grade.
Case II	33	96	5997	Age, Ethnicity, pretreatment tumor stage, nodal status, nuclear grade, ER status, PR status, HER2 status.
Case III	112	65	12633	Age, Tumor size, Nodal status, ER status, Tamoxifen treatment.

3.2.1 Microarray Data

The microarray data were obtained with the Affymetrix technology and preprocessed with MAS5.0, the GeneChip Microarray Analysis Suite 5.0 software (Affymetrix)[112]. However, as probe selection for the Affymetrix gene chips relied on earlier genome and transcriptome annotation that are significantly different from current knowledge, an updated array annotation was used for the conversion of probes to Entrez Gene IDs, lowering the number of false positives [40]. Finally, the low signal-to-noise [121] ratio of microarray data was taken into account by unsupervised exclusion of genes with low variation (variance less than the 20th percentile), retaining the 4997, 5997 and 12633 most varying genes for the first, second and third microarray data respectively [31, 79, 158]. All the data set that we have used are available in the Integrated Tumor Transcriptome Array and Clinical data Analysis database (ITTACA):`http://bioinfo-out.curie.fr/ittaca/`. This resulted in a $p \times n$ matrix B with p =number of genes and n =number of samples.

3.2.2 Clinical Data

The first case study of 129 patients contained information on 17 available clinical variables. Five variables were excluded [31]: two redundant variables that were least informative based on univariate analysis in those variable pairs with a correlation coefficient exceeding 0.7, and three variables with too many missing values. After exclusion of patients with missing clinical information, this data set consisted of 110 patients, of which 85 were disease free whilst in 25 patients the disease occurred [38].

The second case study, in which response to treatment was studied, entailed 12 variables for 133 patients [79]. Patient and variable exclusion performed as described above, resulted in 129 patients and 8 variables. Of the 129 remaining patients, 33 showed complete response to treatment while, 96 patients were characterized by residual disease.

In the last case study, relapse was studied in 187 patients [158]. After preprocessing, this data set retained information on 5 variables for 177 patients. In 112 patients, no relapse occurred while 65 patients had a relapse.

The clinical data contain three different types of variables: continuous (C), ordinal (O) and nominal (N). Normalization is required to make these variables comparable to each other. Rank order, min-max and square root transformations were applied to the ordinal, continuous and nominal variables, respectively. This resulted in a $m \times n$ matrix A with m =number of clinical parameters and n =number of samples.

3.3 Feature Selection Problem

Colon cancer is a common malignancy affecting both women and men. Including the literature information in the analysis of gene expression data offers an opportunity to incorporate functional information about the genes when identifying disease associated genes [139]. In this study, we use gene expression data and the corresponding literature information of colon cancer to identify differentially expressed genes.

3.3.1 Microarray Data

The colon cancer data set investigated here was taken from the Bioinformatics Research Group Repository: `http://www.upo.es/eps/bigs/datasets.html`. It contains 62 samples, among them 40 colon tumor samples and 22 normal

colon samples, with 1,988 genes and 12 controls. Data were standardized to a mean of 0 and standard deviation of 1. This resulted in a $m \times n$ matrix A with $m = 62$ and $n = 2000$.

3.3.2 Literature Information

We used a well defined cancer vocabulary with 2406 terms from the NCI Dictionary of Cancer Terms: `http://www.cancer.gov/dictionary`. Pubmed abstracts with the terms in the vocabulary were extracted using Perl, version 5.10.1 for windows. We defined literature information as a matrix with row corresponding to cancer related terms and the columns with the same genes as in Section 3.3.1. Each entry in the matrix corresponding to the number of Pubmed abstracts in which the gene and term co-occur. We chose to retrieve entries containing the official gene name, abbreviations or aliases in the corresponding field, following the same strategy as used by Gevaert et al [64]. Finally, the cosine similarity measure was used to obtain gene-to-gene distances between 0 and 1, derived from the literature information. This resulted in a $p \times n$ matrix B with $p = 2406$ and $n = 2000$.

3.4 Methods

In this section, we formulate a mathematical framework, the maximum likelihood estimation via a generalized eigenvalue decomposition (MLGEVD) and show the similarity between MLGEVD and generalized ridge regression.

3.4.1 Maximum Likelihood Estimation of Generalized Eigenvalue Decomposition:

The GEVD in Equation 2.3 can now be rewritten as EVD problem:

$$(B^TB)^{-1/2}A^TA(B^TB)^{-1/2}[(B^TB)^{1/2}(X^T)^{-1}] = [(B^TB)^{1/2}(X^T)^{-1}]\Lambda. \quad (3.1)$$

If we call $(B^TB)^{1/2}(X^T)^{-1} = W$, then Equation 3.1 can be rewritten as:

$$(B^TB)^{-1/2}A^TA(B^TB)^{-1/2}W = W\Lambda.$$

where $(B^TB)^{-1/2}A^TA(B^TB)^{-1/2}$, is a symmetric matrix.

Let SVD of $A(B^TB)^{-1/2}$ be

$$A(B^TB)^{-1/2} = PSQ^T \tag{3.2}$$

Then

$$(B^TB)^{-1/2}A^TA(B^TB)^{-1/2}Q = QS^TS$$

where $S^TS = \Lambda$ and $W = Q$ with $Q^TQ = I_n$.

Let define,

$$D = A(B^TB)^{-1/2}Q = A(X^T)^{-1} \tag{3.3}$$

with $(X)^{-1}(B^TB)(X^T)^{-1} = I_n$ and columns of $(X^T)^{-1}$ are GEVs.

Equation 3.3 shows that projection of weighted matrix $A(B^TB)^{-1/2}$ onto eigenvectors Q are equivalent to the projection of matrix A onto GEVs $(X^T)^{-1}$.

MLGEVD is an analog to GEVD that incorporates information about projection errors to develop GEVD models. The theoretical foundations of MLGEVD are initially established using GEVD and extended to the framework of ridge regression. An efficient iterative algorithm based on the ridge regression method is described.

MLGEVD problem which estimates the optimal GEVs are formulated as follows: minimize $||D - A(X^T)^{-1}||^2$ subject to $[(X)^{-1}(B^TB)(X^T)^{-1}] = I_n$.

Maximum Likelihood Estimation of GEVD can be formulated with the assumptions that B^TB is invertible, let SVD of $A(B^TB)^{-1/2} = PSQ^T$ and $D = A(B^TB)^{-1/2}Q$ are known. $D^{(0)} = \tilde{A}(B^TB)^{-1/2}Q$, where $\tilde{A}$ is the rank k truncated SVD approximation of A. D_i and $D_i^{(0)}$ are the i^{th} columns of the matrices D and $D^{(0)}$ respectively. Each column of the matrix D can be considered to represent a point in the m-dimensional row space, with the true measurements corrupted by normally distributed errors. Here D_i is a column vector of D, $D_i^{(0)}$ represents the error-free column vector and $\eta_i = D_i - D_i^{(0)}$ is the vector of measurement errors, which has an error covariance matrix $F_i = cov(\eta_i)$ and $\tilde{e_i} = A\tilde{r}_i$ where r_i is unknown (i^{th} column of $(X^T)^{-1}$).

One defines the Lagrangian of MLGEVD problem as follows:

$$\mathcal{L} = \Sigma_{i=1}^n(D_i - \tilde{e_i})^TF_i^{-1}(D_i - \tilde{e_i}) - \Sigma_{i=1}^n(1 - r_i^TB^TBr_i) - \Sigma_{i=1}^n\alpha_i(A\tilde{r}_i - \tilde{e_i}) \ .$$

with the optimality conditions,

$$\frac{\partial\mathcal{L}}{\partial\tilde{e_i}} = -2F_i^{-1}(D_i - \tilde{e_i}) + \alpha_i^T = 0.$$

$$\frac{\partial\mathcal{L}}{\partial r_i} = -A^T\alpha^T + 2(B^TB)r_i = 0.$$

$\frac{\partial \mathcal{L}}{\partial \alpha_i} = \tilde{e_i} - A\tilde{r_i} = 0,\ i = 1 \ldots,\ n.$

Eliminations of $\tilde{e_i}$ and α_i yields an equation in the form of

$$\tilde{r_i} = (A^T F_i^{-1} A + B^T B)^{-1} A^T F_i^{-1} D_i, i = 1 \ldots, n.$$

Thus, the Maximum Likelihood estimation of GEVD is

$$\tilde{r}_{iMLGEVD} = (A^T F_i^{-1} A + B^T B)^{-1} A^T F_i^{-1} D_i, i = 1 \ldots, n. \tag{3.4}$$

which is in the form of generalized ridge regression. Thus $B^T B$ is interpreted as the prior information to obtain the optimal GEVs.

To obtain MLGEVD via GEVD, an iterative algorithm based on generalized regression has been proposed as follows:
An Algorithm for MLGEVD

1. Input data matrices $A \in \mathbb{R}^{m \times n}$, $B \in \mathbb{R}^{p \times n}$.
2. Initial approximation: $A(B^T B)^{-1/2} = PSQ^T$, $D = A(B^T B)^{-1/2} Q$, $D^{(0)} = \tilde{A}(B^T B)^{-1/2} Q$ where $\tilde{A}$ is the rank k truncated SVD approximation of A, D_i and $D_i^{(0)}$ are the i^{th} column of matrices D and $D^{(0)}$ respectively. F_i is the error covariance of $(D_i - D_i^{(0)})$.
3. j=0;
4. repeat
5. Compute the solution of ML GEVD: $r_i = (\tilde{A}^T F_i^{-1} \tilde{A} + B^T B)^{-1} \tilde{A}^T F_i^{-1} D_i$, $i = 1 \ldots$,n, are the GEVs (columns of $R^{(j)}$).
6. Compute $D^{(j+1)} = \tilde{A} R^{(j)}$ using Equation 3.3.
7. j=j+1
8. Until$||D^{(j)} - D^{(j-1)}||_F / ||D^{(j)}||_F \leq \varepsilon$, where ε is the convergence parameter.
9. Output: $\tilde{D} = D^{(j)}, \tilde{R} = R^{(j)}$

The MLGEVD algorithm, an iterative algorithm monotonically decreases the cost function value and the convergence of the algorithm is quite reliable. The convergence rate, however, is linear and depends on the distribution of the singular values of $A(B^T B)^{-1/2}$.

3.5 Results

The proposed MLGEVD algorithm obtains an optimal GEVs for the GEVD. GEVD and MLGEVD capture common information from two data sets in terms of GEVs by matrix decomposition [68] and ML estimation framework described in Section 3.4.1, respectively. In the model development, initially we used matrix B^TB as the prior information to obtain the GEVs. Then we applied these technique as a pre-processing step which projected the matrix A onto the directions of GEVs, i.e, $A(X^T)^{-1}$. We illustrated the applications of MLGEVD/GEVD in bioinformatics with two problems: one for the classification of breast cancer patients and the second for the identification of differentially expressed genes in colon cancer.

3.5.1 Classification of Breast Cancer Patients

The summary of the data sets are given in Table 3.1. In all the three case studies, clinical data A contain measurements on m clinical parameters, for n samples and microarray data B contain expression level of p genes over these n samples. For all case studies, 2/3rd of clinical and microarray data were randomly assigned to the training set and the remaining to the test set. The split was performed with the relative proportion of outcomes sampled in both training and test set was similar to the original proportion in the full data set, which resulted in four matrices: A_{train}, A_{test}, B_{train} and B_{test}. In the first step, GEVs were obtained from both clinical and microarray - training data sets by MLGEVD and GEVD, respectively. Projected variables corresponding to clinical training data set was obtained by projecting these data onto the direction of GEVs, $A_{train}(X^T)^{-1}$. Next, the LS-SVM classifier [170] was trained using $A^T_{train}A_{train}(X^T)^{-1}$, followed by classification of the $A^T_{test}A_{train}(X^T)^{-1}$. Classification performance was given in terms of test set Area Under the ROC Curve (AUC) and F-score [156]. In this section all the steps were implemented using Matlab R2012b and LS-SVMlab v1.8 toolbox [46] with the default parameter settings.

Table 3.2 summarizes the average test AUC of LS-SVM classifier on 30 random iterations of trained and test data sets. For the sake of comparison, LS-SVM classifiers with the linear, RBF and polynomial kernel function were applied. The LS-SVM classifier with the linear kernel function resulted in the best test AUC for both GEVD and MLGEVD. Probably the data distribution is linearly separable in the projected space hence, with the kernel function we obtained a good prediction performances in all cases. The F-scores for these classifiers shown in Table 3.2 are indicative of precision and recall, with scores ranging

Table 3.2: MLGEVD and GEVD were used as a pre-processing step which transforms the clinical data onto the direction of generalized eigenvectors. LS-SVM classifier with linear, RBF and polynomial kernel functions applied on these projected clinical data for the patient classification. The performance are measured in terms of average test AUC (std) and average F-score (std) over 30 iterations

	kernel function	linear	RBF	polynomial
Case I				
test AUC	MLGEVD	**0.80**(0.09)	**0.79**(0.01)	**0.77**(0.07)
	GEVD	0.77(0.08)	0.74(0.09)	0.63(0.02)
	p-value	0.03	0.01	2.72E-10
test F-score	MLGEVD	**0.64**(0.01)	**0.57**(0.02)	**0.51** (0.13)
	GEVD	0.55(0.01)	0.49 (0.01)	0.26(0.03)
	p-value	0.17	0.23	0.03
Case II				
test AUC	MLGEVD	**0.80**(0.05)	0.75(0.09)	**0.70**(0.10)
	GEVD	0.79(0.06)	**0.78**(0.08)	0.61(0.06)
	p-value	0.01	0.04	0.01
test F-score	MLGEVD	**0.60**(0.03)	**0.46** (0.06)	0.47(0.01)
	GEVD	0.56(0.02)	0.51(0.07)	**0.50**(0.05)
	p-value	0.02	0.06	0.11
Case III				
test AUC	MLGEVD	**0.67**(0.07)	**0.60**(0.08)	**0.60**(0.04)
	GEVD	0.66(0.08)	0.57(0.05)	0.54(0.06)
	p-value	0.02	0.26	0.86
test F-score	MLGEVD	**0.44**(0.04)	**0.29**(0.46)	**0.28**(0.03)
	GEVD	0.32(0.08)	0.23 (0.12)	0.24 (0.08)
	p-value	0.06	0.18	0.04

p-value: two-sided sign test ; RBF: radial basis function

from 0 to 1. While comparing the case III with other cases in Table 3.2, it is observed that case III perform much worse than other case studies. As the number of genes increase, the linear projections based on GEVD cannot capture the non-linear pattern in the data well and in addition, there is a high chance of overfitting. The solution for this problem, especially for the large data set, is the projection based on kernel GEVD.

High-throughput data, such as microarray data are in general much more difficult and expensive to collect while clinical parameters are routinely measured by clinicians. We have used the MLGEVD/GEVD framework as a pre-processing step in which B^TB were used as a prior information to obtain the GEVs. Later the matrix A is projected onto the directions of GEVs to transform the data onto the common space. The final classification is performed on clinical data in the transformed space, that is, $A_{test}^T A_{train}(X^T)^{-1}$.

3.5.2 Identification of differentially expressed genes in colon cancer

The summary of the Data Sets are given in Table 3.3. In this problem,

Table 3.3: Summary of the Data Sets - Identification of differentially expressed genes in colon cancer.

Matrix	Rows	Columns
A	62 Patients	2000 Genes
B	2406 Terms	2000 Genes

matrix A was microarray data (62 samples $\times$ 2000 genes) and matrix B was literature information (2406 terms $\times$ 2000 genes). The GEVs were obtained from microarray data and literature information using MLGEVD and GEVD framework. Finally microarray data was divided into two groups: normal and cancerous samples. Each set of these samples were projected onto the direction of GEVs, resulted in two sets of scores Z^1 and Z^2. Let $g_i = Z_i^1 - Z_i^2$ be the difference in score for gene i between normal and cancerous samples.

Each gene was represented graphically as a point in the n-dimensional space (with n the number of GEVs selected for projection). The gene i with similar expression levels has approximately the same scores $Z_i^1 \approx Z_i^2$ and form a cloud of points around the origin. Differentially expressed genes have significantly different scores and are located away from the origin. To identify the outliers in this n-dimensional space, the Mahalanobis distance is calculated for each gene $MD_i^2 = (g_i - c)\Sigma^{-1}(g_i - c)^T$, with c the multivariate arithmetic mean and

Table 3.4: The 50 top ranked genes for relevance in colon cancer diagnosis identified by MLGEVD and GEVD, with the literature references.

MLGEVD Gene Symbol	Ref	GEVD Gene Symbol	Ref
IGLC1	[84]	EIF4A1	[87]
RPLP1	[87]	N2b5HR	[87]
TMSB4X	[87]	ITIH1	[87]
FTL	[87]	IGHG3	[59]
IGKC	[186]	IGLC1	[84]
TCTP	[87]	RPLP2	[87]
EIF4A1	[87]	MYL6	[87]
S100A6	[87]	TCTP	[87]
RPS	[87]	RPL41	[87]
SELENBP1	–	ACTB	[87]
RPS29	[87]	RPS9	[87]
RPSA	[87]	HSP90B1	[87]
RPL30	[87]	RPL37A	[87]
RPL37A	[87]	BBC1	[87]
RPL32	[87]	RPLP1	[87]
IGHG3	[59]	YBX1	[87]
YBX1	[87]	UBB	[87]
CPSF1	–	RPSA	[87]
RPL37A	–	GAPDH	[87]
LGALS3	–	SRF	[87]
RPS18	[87]	IGF2	–
UBB	[87]	RPL37A	–
PFN1	[87]	RPL1	–
RPS6	[87]	HSPB1	–
GAPDH	[87]	RPS29	[87]
HSP90B1	[87]	RPS18	[87]
RPS24	[87]	FTL	[87]
BBC1	[87]	IGKC	[186]
RPS28	[87]	RPL30	[87]
RPL38	[2]	RPS	[87]
MUC2	[84]	RPS28	[87]
IGHG3	[58]	HLA-B	[87]
ITIH1	[87]	S100A6	[87]
RPLP2	[87]	EEF1A2	[87]
RPL41	[87]	IFI27	–
ALDOA	[87]	EEF1B2	–
ACTB	[87]	JUND	–
RPS9	[87]	MT1G	–
OAZ	[87]	SELENBP1	–
HSP90AB1	[55]	RPS8	–
RPS24	[2]	ARNT	–
B2M	[87]	TSPAN8	–
MAMDC2	[87]	OAZ	[87]
SRF	[87]	RPS11	[87]
DESMIN	[202]	RPS24	[87]
LYZ	[198]	MUC2	[84]
N2b5HR	[87]	TPM2	[202]
MYL6	[87]	RPS19	[2]
FCGRT	–	RPL32	[87]
RPL37	–	LYZ	[198]

Σ^{-1} the inverse of the covariance matrix of the differences in scores [90]. Genes with the largest Mahalanobis distances are defined as the most differentially expressed genes.

Table 3.4 shows the top 50 differentially expressed genes obtained with MLGEVD and GEVD. Among these genes, relevance for colon cancer has been shown for 44 and 38 genes respectively with MLGEVD and GEVD. An LS-SVM model with RBF kernel was built for the prediction of tumor vs. non-tumor samples. The microarray data was split into training and test data which followed the same strategy of breast cancer case studies. In Table 3.5, we compared the prediction performances (test AUC) of LS-SVM classifier on full

Table 3.5: LS-SVM model for prediction of tumor and non-tumor samples of colon cancer on whole sets of genes and subsets of genes selected by MLGEVD and GEVD. Average classification performances test AUC (std) are given in terms of test AUC.

Genes selected by	kernel function	test AUC	p-value[a]
full data set	RBF	0.821(0.147)	0.019
GEVD	RBF	0.841(0.087)	0.072
MLGEVD	RBF	**0.895(0.060)**	

[a] two-sided sign test for the comparison of full data sets and GEVD with MLGEVD.

sets of genes, genes obtained from GEVD and MLGEVD. LS-SVM classifier offered the best prediction performances (see Table 3.5) on the 50 genes obtained from MLGEVD. Results shows that MLGEVD based gene selection obtained the disease specific genes better than GEVD, which significantly improved the classification performance.

Several genes selected by this approach are known to be involved in, and important for, colon cancer. The ribosome, the essential cellular organelle for protein synthesis in all cells, consists of ribosomal RNAs (rRNAs) and ribosomal proteins (RPs). Ribosomal protein L41 (RPL41) is a microtubule-associated protein essential for functional spindles and for the integrity of centrosome. Abnormal mitosis and a disrupted centrosome associated with RPL41 down-regulation may be related to malignant transformation [188]. In our analysis, RPL41 ranked as one of the differentially expressed gene. Studies in [87] and [109] already reported on the importance of RPL41 in colon cancer.

Ectopic expression of tumor rejection antigen 1 (Tra1) was detected in the ulcerative colitis (UC) affected colonic mucosa [101]. Tra1 is reported as a differentially expressed gene in colon cancer [87]. Calcyclin binding protein (CacyBP) was a promising candidate biomarker for colorectal cancer (CRC) metastasis and also sheds light on the underlying molecular mechanism by which CacyBP promotes CRC metastasis [65]. 60S acidic ribosomal protein P1 was reported as a top-ranked gene in colon cancer [87, 190].

Translationally controlled tumor protein (TCTP) is a highly conserved and ubiquitously expressed protein in all eukaryotes highlighting its important functions in the cell. Previous studies revealed that TCTP is implicated in many biological processes, including cell growth, tumor reversion, and induction of pluripotent stem cells. In human colon cancer, the level of TCTP mRNA was detected in three human colon carcinoma cell lines (SNU-C2A, SNU-C4, and SNU-C5) [27]. Ornithine decarboxylase (OAZ1) catalyzes the conversion of ornithine to putrescine in the first and apparently rate-limiting step in

polyamine biosynthesis. The ornithine decarboxylase antizymes play a role in the regulation of polyamine synthesis by binding to and inhibiting ornithine decarboxylase. OAZ was reported as a top ranked gene in colon cancer [87, 2].

Alterations in the distribution and/or adhesiveness of laminin receptors in colon cancer cell lines were suggested to be associated with increased tumorigenicity [96]. A study of cultured colon cancer cells suggests that laminin may play an important role in hematogeneous metastasis by mediating, tethering and spreading of colon cancer cells under blood flow [97]. In general, the markers are involved in cell signaling, adhesion and communication, immune response, heat shock, and DNA repair [87].

In short, out of 50 genes identified as differentially expressed by MLGEVD, majority of these genes is reported as top ranked genes in colon cancer in various studies. In addition, LS-SVM classifier on selected genes by MLGEVD offered the best prediction performances than LS-SVM classifier on whole gene sets and subsets of genes selected by the GEVD framework.

3.6 Discussion

Several studies have already used GSVD as a comparative mathematical framework for two data sets [3, 107]. In this study, we showed that one of the data matrix in GSVD/GEVD framework acts as a prior information in the model development to obtain the GEVs. In addition, we showed tight equivalence between MLGEVD and regularized regression. MLGEVD was applied to four case studies for which gene expression with corresponding clinical/literature information were available. Microarray and clinical parameters were gathered from patients with breast cancer. Literature information from Pubmed were collected for colon cancer. The main aim of the work was to interpret the GEVD problem in the framework of maximum likelihood estimation. To validate the merit of MLGEVD over GEVD/GSVD, models were built for classifying patients. In this study we performed MLGEVD, on clinical data sets with microarray as prior information and on microarray data with literature as prior information. In both cases, initially GEVs obtained from MLGEVD and GEVD respectively. Subsequently, clinical parameters/microarray were projected onto the generalized eigenvectors, referred to as the projected clinical space/gene space.

In the breast cancer case study, MLGEVD performed better than GEVD in classification while in colon cancer the incorporation of external knowledge into the analysis of microarray improves the identification of disease related genes.

The proposed model is very cost effective. In breast cancer case studies, the high throughput technologies which are difficult and expensive to collect are used only for the model development. The clinical parameters which are routinely measured by clinicians are used for prediction. In real data examples, we have shown how to incorporate external knowledge, extracted from microarray data/literature information into medical diagnosis. The proposed method provides a general way to incorporate such ever-increasing amounts of prior knowledge in the analysis and further improve the predictive performance.

3.7 Conclusion

In this work, we developed a mathematical framework MLGEVD which relied on a generalization of SVD and then compared its performance with GEVD. We showed that the proposed approach could be used as an alternative of GEVD which significantly improved diagnosis, prognosis and prediction of therapy response. Both GEVD and MLGEVD used high-throughput data, which were difficult and expensive to collect only for model development. In the near future, we will investigate the applicability of MLGEVD to more than two matrices and interpret these matrix results in a Bayesian context.

Chapter 4

Robust PCA improves biomarker discovery in colon cancer with incorporation of literature information

Microarray technology handles thousands of genes of several hundreds of patients at a time. It is hard, however, to extract relevant information about genes and diseases from these data. Bioinformatics and statistical methods collect the appropriate information from these data sets in light of the specific application. To date, many research groups identified differentially expressed genes based on microarray data and literature information. Still, the lack of efficient methods for assessing the biological implications of gene expression data remains an important difficulty in exploiting this information.

In this study a data integration approach is used, in which microarray expressions values are weighted with literature information. First, we present the generalized eigenvalue decomposition (GEVD) in terms of the ordinary eigenvalue decomposition (EVD) for the integration of microarray and literature information. Then robust PCA (RPCA) is used for the identification of differentially expressed genes of colon cancer. Initially, we apply RPCA on colon cancer data only and then on both colon cancer and literature information. Finally, we search for, the co-expressed genes of differentially expressed ones, which are part of the colon cancer data.

To evaluate the obtained genes which are really associated with colon cancer, we perform classification of patients using an LS-SVM classifier on this subset of genes. Then we compare the prediction performances on the whole set of genes, subsets of differentially expressed genes and subsets of both differentially expressed and co-expressed genes. In colon cancer, the highest leave-one-out (LOO) areas under the receiver operating characteristic curves (AUC) are obtained ranging from 0.868 to 0.989 on the subsets of differentially expressed and co-expressed genes.

The results suggest that the incorporation of external knowledge into microarray analysis improves the identification of disease specific genes. This emphasizes the importance of data integration in gene expression analysis.

4.1 Introduction

Biomarker discovery and prognosis prediction are essential for improved, personalized treatment of cancer. Microarray technology is a significant tool in gene expression analysis and cancer diagnosis. It can simultaneously handle thousands of genes. Microarray data are typically used for class discovery [136, 144] and prediction [33, 157]. The challenge in dealing with microarray data lies in the fact that there is a difference in orders of magnitude between the number of samples (typically less than hundred) and the number of genes (typically tens of thousands). The measurements also contain both measurement and systematic noise with an impact on classification accuracy.

An ever increasing number of techniques has been available for the discovery of clusters of samples with similar gene expressions in microarray data. Troyanskaya and colleagues developed nonparametric methods to identify differentially expressed genes (DEG) in microarray data [180]. Principal component analysis (PCA) and PCA-based approaches were used as well for the identification of DEG in pulmonary adenocarcinoma [76] and E coli [168]. Besides well-known statistical tests such as the chi-square test [155], Chun and colleagues proposed a new test, the 'half Student's t-test', specifically for detecting DEG in heterogeneous diseases [83]. Generalized Eigen Value Decomposition has been shown to have great potential within bioinformatics for extracting common information from datasets such as genomics and proteomics data [3, 152]. Recently the robust PCA (RPCA) based method for discovering differentially expressed genes was proposed in [111]. Several studies used microarray and literature information together to obtain differentially expressed genes with elements for biological understanding, generating and validating new biological hypotheses [56, 30, 28].

In this Chapter, we propose a data integration strategy in which gene expression values are weighted with literature information. We represent the generalized eigenvalue decomposition (GEVD) in terms of the eigenvalue decomposition (EVD) for the integration of two data sets. Then we apply robust PCA (RPCA) to discover the differentially expressed genes in colon cancer. Finally we assume that there are disease-specific genes that are not identified by the analysis. Thus we search GeneMANIA tool: `www.genemania.org` for finding co-expressed genes related to the obtained disease specific genes of colon cancer data.

4.2 Method

The methods used for this study can be subdivided into two categories: representation of GEVD in terms of EVD which offers a framework in which the microarray data set is weighted with literature information and then the robust PCA approach is applied on these data sets to recover a low rank and a sparse matrix.

4.2.1 Generalized Eigenvalue Decomposition

The GEVD in Equation 2.3 can now be rewritten as EVD problem:

$$(B^TB)^{-1/2}A^TA(B^TB)^{-1/2}W = W\Lambda.$$

where $W = (B^TB)^{1/2}(X^T)^{-1}$, and $(B^TB)^{1/2}$ is a symmetric square root the matrix B^TB. Thus, if B^TB is invertible, the GEVD can be represented as a PCA estimation of weighted matrix $(B^TB)^{-1/2}A^TA(B^TB)^{-1/2}$.

Let $D = (B^TB)^{-1/2}A^T$, and its SVD be

$$D = PSQ^T. \tag{4.1}$$

The matrix $(B^TB)^{-1/2}$ is defined [80] as follows: Let the EVD of $B^TB = T\Sigma T^T$, where the columns of T are the eigenvectors and Σ is a diagonal matrix. $(B^TB)^{1/2} = T\Sigma^{1/2}T^T$ and $(B^TB)^{-1/2} = T\Sigma^{-1/2}T^T$.

We can estimate the low rank decomposition of matrix D by SVD as shown in Equation 4.1. But standard SVD is highly susceptible to outliers. This problem can be addressed by using a robust analysis.

4.2.2 The RPCA Model of Gene Expression Data

We followed the same RPCA model for gene expression data, which is proposed in [111], but here we considered both gene expression data and literature information. In this section, we used the same data sets which we described in Sections 3.3.1 and 3.3.2. We used Equation 4.1 in which a $m \times n$ matrix A is microarray ($m = 62$ samples and $n = 2000$ genes) and $p \times n$ matrix B is literature information ($p = 2406$ terms and $n = 2000$ genes). We obtain a $n \times m$ matrix $D = (B^T B)^{-1/2} A^T$, a gene expression data weighted with literature information. Each row of D represents the transcriptional responses of a gene weighted with literature information in all the m samples, and each column of D represents the weighted expression levels of all the n genes in one sample.

Our goal of using RPCA model on weighted microarray data is to identify disease associated genes. Assume that the matrix decomposition $D = R + Z$ has been done by using RPCA. By choosing the appropriate parameter λ, the positive regularization parameter in Equation 2.4, the sparsity of the perturbation matrix Z can be influenced, i.e. most of the entries in Z are zero. Studies show that in RPCA estimation, the low-rank component R corresponds to the stationary patterns and the sparse component Z captures the differentially expressed pattern [24]. Hence the differentially expressed genes can be treated as sparse perturbation signals Z.

The sparse matrix Z, in which most of the entries are zero, can be denoted as:

$$Z = \begin{bmatrix} z_{11} & z_{12} & .. & z_{1m} \\ z_{21} & z_{22} & .. & z_{2m} \\ ... & ... & ... & ... \\ z_{n1} & z_{n2} & .. & z_{nm} \end{bmatrix}$$

In RPCA estimation, the low-rank component R corresponding to the stationary pattern and the sparse component Z containing the differentially expressed patterns. In microarray analysis, the normal genes obtain similar profiles over all samples and differentially expressed ones behave differently on diseased and normal samples. Thus we assume that the non-zero entries in the matrix Z corresponds to differentially expressed patterns. The matrix Z contains both positive and negative values. Hence to obtain the differentially expressed features, the absolute value of the entries in the each row of the matrix Z need to be considered. For that the following two steps are executed: firstly, the absolute values of the entries in the sparse matrix Z are found out; secondly, to get the evaluating vector $\tilde{Z}$, the matrix is summed by row. This resulted into an $n \times 1$ vector in which each row corresponding to the gene and each entry in the vector represents the sum of the absolute values of entries in the corresponding row of the sparse matrix Z.

$\tilde{Z} = [\Sigma_{i=1}^{m} \|s_{1i}\| \dots \Sigma_{i=1}^{m} \|s_{ni}\|]^T$.

Consequently, the evaluating vector, the sum of the different expression profile $\hat{Z}$ is sorted in descending order, to arrange genes from most differentially expressed ones. Without loss of generality, suppose that the first c_1 entries in $\hat{Z}$ are non-zero, that is,

$\hat{Z} = [\hat{z_1}, \dots, \hat{z_{c1}}, 0, \dots, 0]^T$.

Thus, the larger the element in $\hat{Z}$, is the more different gene in weighted expression data. The indices of $\hat{Z}$ correspond to differentially expressed genes. Thus we identified differentially expressed genes of colon cancer data set.

To evaluate the relevance of identified genes in colon cancer, we applied LS-SVM classifier on these subsets of genes to classify cancerous and non-cancerous samples. In addition literature references helped us to understand the biological importance of these genes in colon cancer.

4.3 Results

In our analysis, we collected a publicly available binary class colon data set (disease vs. normal) of 1,988 genes and 62 samples (See Section 3.3.1). Then we generated a gene-term matrix as described in Section 3.3.2 resulting in a matrix with 1,988 genes and 2,406 terms. In order to integrate these two data sets, we have applied the Equation 4.1, on an 62×1,988 matrix A and a 2,406×1988 matrix B. Thus the colon cancer data set is weighted with the literature information resulting in a single matrix $D = (B^T B)^{-1/2} A^T$.

Then we applied RPCA on the integrated data set by solving the Equation 2.4 using the inexact ALM (IALM) algorithm proposed in [110]. In RPCA model, the matrix has been decomposed as $D = R + Z$. By choosing the appropriate parameter λ, the positive regularization parameter in Equation 2.4, the sparse perturbation matrix S can be obtained, i.e., most of entries in S are zero or near-zero. We adjusted the λ to obtain approximately 20 non-zero entries in the matrix S. The IALM algorithm is simple to implement, each iteration involves computing a partial SVD of a matrix D, and converges to the true solution in a small number of iterations.

Liu *et al* [111] already has used RPCA to identify the differentially expressed genes from colon cancer data. Our main objective is to compare the performance of the RPCA model to identify disease related genes from gene expression data with both gene expression and literature information. To validate the relevance of obtaining genes in colon cancer, we performed LS-SVM classifier [172] on the

subsets of genes selected by RPCA. Classification of patients was performed on subsets of genes using an LS-SVM classifier, in which 2/3rd samples are split into training and the remaining as test sets. In all cases, a linear kernel function was used with the LS-SVM classifier. To further identify the co-expressed genes of the obtained disease specific genes, we make the network analysis of the selected genes using the GeneMania tool `www.genemania.org`. Table 4.1 shows the averaged classification performance of colon cancer data set on whole data sets, subsets of genes identified by RPCA in [111] and proposed approach, over 30 iterations. The results show that better prediction performance in terms of test areas under curve (AUC) is obtained with the proposed approach. We obtained sets of co-expressed genes which are parts of colon cancer data, but not identified by our approach. Thus searching for co-expressed genes helped us to identify disease related genes which are really missing in our analysis. Then we estimated prediction performance to classify patients on sets of genes, including both co-expressed and differentially expressed ones. The predictive performance has improved while considering both co-expressed genes and differentially expressed ones. Thus the incorporation of literature information into microarray analysis improves the identification of disease associated genes and hence offer better diagnosis, progonosis, and patient therapy.

Table 4.1: Averaged LS-SVM classifier performance, for classification of colon cancer patients, on whole genes, subset of differentially expressed genes, and subset of both differentially expressed and co-expressed genes, over 30 iterations.

Data Sources	No. of genes	Methods	test AUC
Microarray and literature information	23	proposed approach	0.834(0.087)
Microarray	30	RPCA in [111]	0.834(0.085)
Microarray and literature information	33	proposed approach and GeneMania	**0.870**(0.078)
Microarray	46	RPCA in [111]and GeneMania	0.795(0.095)
Microarray	2000	whole data set	0.834(0.097)

The colon cancer related genes were selected by the proposed approach are shown in Table 4.2. Several genes selected by this approach are known to be involved in, and important for, colon cancer. Ribosome, the essential cellular organelle for protein synthesis in all cells, consists of ribosomal RNAs (rRNAs) and ribosomal proteins (RPs). Ribosomal protein L41 (RPL41) is a microtubule-associated protein essential for functional spindles and for the integrity of centrosome and that the abnormal mitosis and disrupted centrosome associated with the RPL41 down-regulation may be related to malignant transformation [188]. In our analysis, RPL41 (H55933) ranked as the most differentially expressed gene.

Table 4.2: 23 Differentially expressed genes of colon cancer identified by the proposed method

Gene Id		Gene Annotation
Hsa.3004	H55933	Homo sapiens mRNA for homologue to yeast ribosomal protein L41
Hsa.2357	T52342	Human tra1 mRNA for human homologue of murine tumor rejection antigen gp96
Hsa.474	L28809	Homo sapiens dbpB-like protein mRNA
Hsa.6080	J02763	Human calcyclin gene
Hsa.20836	R02593	60S acidic ribosomal protein P1
Hsa.45293	H86060	Negative factor (Simian immunodeficiency virus)
Hsa.20836	R02593	60S acidic ribosomal protein P1 (Polyorchis penicillatus)
Hsa.13491	R39465	Eukaryotic Initiation factor 4A (Oryctolagus cuniculus)
Hsa.3835	H79852	60S acidic ribosomal protein P2 (Babesia bovis)
Hsa.37254	R85482	Serum response factor (Homo sapiens)
Hsa.909	M11799	Human MHC class I HLA-Bw58 gene
Hsa.749	T63508	Ferritin in heavy chain (HUMAN)
Hsa.2597	T49423	Breast basic conserved protein 1 (HUMAN)
Hsa.750	T72863	Ferritin light chain (HUMAN)
Hsa.2800	X55715	Human Hums3 mRNA for 40S ribosomal protein s3
Hsa.8068	T57619	40S ribosomal protein S6 (Nicotiana tabacum)
Hsa.467	H20709	MYOSIN LIGHT CHAIN ALKALI, SMOOTH-MUSCLE ISOFORM (HUMAN)
Hsa.3087	T65938	Translationally controlled tumor protein (HUMAN)
Hsa.5710	T63484	Human ornithine decarboxylase antizyme (Oaz) mRNA
Hsa.3061	X63469	Transcription initiation factor iie-beta chain (HUMAN)
Hsa.878	T61609	Laminin receptor (HUMAN)
Hsa.9994	T51539	Hepatocyte growth factor-like protein precursor (Homo sapiens)
Hsa.4689	T95018	40s ribosomal protein s18 (Homo sapiens)

Studies in [87] and [109] already reported that RPL41 (H55933) gene has high rank in colon cancer.

Ectopic expression of tumor rejection antigen 1 (Tra1) were detected in the UC affected colonic mucosa [101]. Tra1 is reported as a differentially expressed gene in colon cancer [87]. Calcyclin binding protein (CacyBP) as a promising candidate biomarker for colorectal cancer (CRC) metastasis and also sheds light on the underlying molecular mechanism by which CacyBP promotes CRC metastasis [65]. 60S acidic ribosomal protein P1 (R02593) were reported as a top-ranked gene in colon cancer [87, 190].

Eukaryotic initiation factor (elF) functions as a subunit of the initiation factor complex elF4F, which mediates the binding of mRNA to the ribosome. Serum response factor (SRF)regulates transcription of many serum-inducible and muscle-specific genes. It binds to the serum response element, a DNA sequence required for the transcription of a number of genes in response to growth factor or mitogen stimulation. These genes might provide an indication of the migratory capacity of the cells in the specimens and hence their propensity for metastasis [87, 159].

Human Leukocyte Antigen (HLA) class I molecules are of major importance for cell-mediated anti-tumor immune responses. Expression of HLA class I/2-microglobulin (2-m) complexes carrying tumor-specific peptides is a prerequisite for adaptively matured cytotoxic T cells (CTLs) to be able to recognize tumor cells [179]. Loss of expression of HLA class I molecules has been frequently reported for colorectal tumors [92, 22].

Low serum ferritin levels are associated with patients having serious gastrointestinal pathologies such as neoplasia and acid peptic disease [182]. Previous work has shown that the majority of colorectal adenocarcinomas exhibit ferritin expression [34], but the clinical significance remains unknown. Myosin light chain alkali, smooth-muscle isoform (H20709) also reported as top ranked genes in colon cancer on several studies [87].

Translationally controlled tumor protein (TCTP) is a highly conserved and ubiquitously expressed protein in all eukaryotes—highlighting its important functions in the cell. Previous studies revealed that TCTP is implicated in many biological processes, including cell growth, tumor reversion, and induction of pluripotent stem cell. In human colon cancer, the level of TCTP mRNA was detected in three human colon carcinoma cell lines (SNU-C2A, SNU-C4, and SNU-C5) [27]. Ornithine decarboxylase (OAZ1) catalyzes the conversion of ornithine to putrescine in the first and apparently rate-limiting step in polyamine biosynthesis. The ornithine decarboxylase antizymes play a role in the regulation of polyamine synthesis by binding to and inhibiting ornithine

decarboxylase. OAZ(T63484) were reported [87, 2] as a top ranked gene in colon cancer.

Alterations in the distribution and/or adhesiveness of laminin receptors in colon cancer cell lines may be associated with increased tumorigenicity [96]. A study of cultured colon cancer cells suggests that laminin may play an important role in hematogeneous metastasis by mediating tethering and spreading of colon cancer cells under blood flow [97]. In general, the markers are involved in cell signaling, adhesion and communication, immune response, heat shock, and DNA repair [87]. Hepatocyte growth factor (HGF) signaling has been implicated in a broad spectrum of human cancers. Studies show that HGF-induced colon tumor cell proliferation, invasion as well as tumor growth and metastasis in xenograft models [165].

In short, out of 23 genes identified as differentially expressed genes of colon cancer, majority of these genes are reported on several studies as top ranked genes in colon cancer. In addition, the selected genes clearly distinguish cancerous and non-cancerous samples indicating that these genes are part of colon cancer.

4.4 Discussion

The proposed method has been applied to colon cancer data of which the corresponding literature information is available. Literature information was gathered from Pubmed by searching gene and term co-occurrence, in which columns corresponded to genes, and rows to cancer related terms. We chose to retrieve column entries containing the official gene name, abbreviations or aliases and row entries corresponding to cancer related terms such as colon cancer, crohn's disease, abdominal pain etc. To verify the merit of our approach over the use of RPCA on the single expression data source, models were built for classifying colon cancer patients. In many studies, single data sources are explored for the identification of differentially expressed genes. In our opinion, a single data source is inadequate to explain complex networks of genes underlying a disease. In this study, microarray data are weighted with literature information. These types of external knowledge in the analysis of microarray improves the identification of cancer related genes accurately.

In RPCA, firstly, the matrix microarray data weighted with literature information D is decomposed into a low rank matrix R and perturbation matrix Z by using Equation 2.4; secondly, the differentially expressed genes were discovered as explained in Section 4.2.2. The proposed approach offer an efficient approach to incorporate prior information into microarray analysis for gene selection. In addition, compared to other statistical approaches for

gene selection, it does not widely vary on the number of genes selected on each iterations. Finally the biological relevance and the prediction performance of selected genes emphasize the relevance of our approach for the identification of differentially expressed genes.

4.5 Conclusion

The results suggest that the proposed approach to colon cancer data, improve the identification of differentially expressed genes and performance of decision support in cancer. With real data examples, we have shown how to incorporate external knowledge, extracted from literature information on medical diagnosis. The proposed method provides a general way to incorporate such ever-increasing amounts of prior knowledge into the analysis and to further improve the predictive performance. These results emphasize the need for comprehensive prior knowledge gathered with microarray data, but it is unknown which type and levels of external knowledge are most relevant for the biomarker discovery and prognostic prediction.

Chapter 5

Bandwidth selection criterion of KPCA: Applications in Bioinformatics

This paper was published in BMC Bioinformatics: Thomas M., De Brabanter K., De Moor B.: New bandwidth selection criterion for Kernel PCA: Approach to Dimensionality Reduction and Classification Problems. BMC Bioinformatics 2014, 15:137 (2014).

Background: DNA microarrays are potentially powerful technology for improving diagnostic classification, treatment selection, and prognostic assessment. The use of this technology to predict cancer outcome already has a history of almost a decade. Disease class predictors can be designed for known disease cases and provide diagnostic confirmation or clarify abnormal cases. The main input of these class predictors is high dimensional data with many variables and only a few observations. Reducing the dimensionality of the feature set significantly speeds up the prediction task. Feature selection (t-test) and feature transformation (Principal Component Analysis (PCA)) methods are well known preprocessing steps in the field of bioinformatics. Several prediction tools are available based on these techniques.

Results: Studies show that a well tuned Kernel PCA (KPCA) is a valuable preprocessing step for dimensionality reduction, but the available bandwidth selection method for KPCA was computationally expensive. In this paper, we propose a new data-driven bandwidth selection criterion for radial basis function (RBF) KPCA which is related to least squares cross-validation for kernel density

estimation. We propose a new prediction model with a well tuned KPCA and Least Squares Support Vector Machine (LS-SVM). We estimate the accuracy of the newly proposed model on 9 case studies. Then, we compare its performance (in terms of test set Area Under the ROC Curve (AUC) and computation time) with other well known techniques such as LS-SVM on the whole data set, PCA + LS-SVM, t-test + LS-SVM, Prediction Analysis of Microarrays (PAM) and Least Absolute Shrinkage and Selection Operator (Lasso). Finally, we assess the performance of the proposed strategy with an existing KPCA parameter tuning algorithm by means of two extra case studies.

Conclusion: We propose, evaluate, and compare several mathematical/statistical techniques that apply feature transformation/selection for subsequent classification, and consider their application to medical diagnostics. Both feature selection and feature transformation perform well on classification tasks. Due to the dynamic selection property of feature selection, it is hard to define significant features for the classifier which predicts classes of future samples. Moreover, the proposed strategy enjoys a distinctive advantage with its relatively low time complexity.

5.1 Introduction

Biomarker discovery and prognosis prediction are essential for improved, personalized treatment of cancer. Microarray technology is a significant tool in gene expression analysis and cancer diagnosis. It can simultaneously handle thousands of genes. Microarray data are typically used for class discovery [144, 136] and prediction [157, 33]. The high dimensionality of the input feature space in comparison with the relatively small number of subjects (curse of dimensionality) is a widespread concern, so some form of dimensionality reduction is often applied. Feature selection and feature transformation are two commonly used dimensionality reduction techniques. The key difference between feature selection and feature transformation is that in the former only a subset of original features is selected while the latter is based on the generation of completely new features.

In this genomic era, several classification and dimensionality reduction methods are available for analyzing and classifying microarray data. Prediction Analysis of Microarrays (PAM) [177] is a statistical technique for class prediction from gene expression data using Nearest Shrunken Centroids (NSC). PAM identifies subsets of genes that best characterize each class. LS-SVM [170, 172] is a promising method for classification because of its solid mathematical foundations which convey several salient properties that other methods hardly provide. A

commonly used technique for feature selection, the t-test, assumes that the feature values from two different classes follow normal distributions. Several studies, especially microarray analysis, have used t-test and LS-SVM together to improve the prediction performance by selecting important features [32, 83]. The Least Absolute Shrinkage and Selection Operator (Lasso) [176] is often used for gene selection and parameter estimation in high-dimensional microarray data [93]. The Lasso shrinks some of the coefficients to zero, and the amount of shrinkage is determined by the tuning parameter, often determined by cross validation.

Inductive learning systems were successfully applied in a number of medical domains, e.g. in localization of a primary tumor, prognostic of recurrence of breast cancer, diagnosis of thyroid diseases, and rheumatology [57]. An induction algorithm is used to learn a classifier, which maps the space of feature values to the set of class values. The classifier is later used to classify new instances with unknown classifications (class labels). Researchers and practitioners realize that the effective use of these inductive learning systems requires data preprocessing before a learning algorithm can be applied [129]. Due to the large variability of feature selection techniques, it may be difficult or even impossible to remove irrelevant and/or redundant features from a data set. Feature transformation techniques, such as KPCA, discover a new feature space having fewer dimensions through a functional mapping while keeping as much information in the data as possible.

KPCA, which is a generalization of PCA, is a nonlinear dimensionality reduction technique that has proven to be a powerful pre-processing step for classification algorithms. It has been studied intensively in the last several years in the field of machine learning and has claimed success in many applications [126]. An algorithm for classification using KPCA was developed by Liu *et al.* [111]. KPCA was proposed by Schölkopf and Smola [146], by considering a mapping to a high-dimensional feature space (possible infinite) and applying Mercer's theorem. Suykens *et al.* [170, 171] proposed a simple and straightforward primal-dual support vector machine formulation to the PCA problem.

Pochet *et al.*[134] proposed an optimization algorithm for KPCA with RBF kernel followed by Fisher Discriminant Analysis (FDA) is to find the parameters of KPCA. In the latter case, parameter selection is coupled with the corresponding classifier. This means that the performance of the final procedure depends on the chosen classifier. Such a procedure could produce very bad results in case of weak classifiers. In addition, this appears to be time consuming when tuning the parameters of KPCA.

Most classification methods have problem with high dimensionality of microarray data and require dimensionality reduction first. The ultimate goal of our work is

to design a powerful preprocessing step, decoupled from the classification method, for large dimensional data sets. In this Chapter, we explain an LS-SVM approach to KPCA. Next, by following the idea of least squares cross-validation in kernel density estimation, we propose a new data-driven bandwidth selection criterion to tune the LS-SVM formulation of KPCA. The tuned LS-SVM formulation to KPCA is applied to several data sets and serves as a dimensionality reduction technique for a final classification task. In addition, we compared the proposed strategy with an existing optimization algorithm for KPCA as well as with other preprocessing steps. Finally, for the sake of comparison, we applied LS-SVM without dimensionality reduction, PCA+LS-SVM, t-test + LS-SVM, PAM and Lasso. Randomization on all data sets are carried out in order to get a more reliable idea of the expected performance.

5.2 Data sets

In our analysis, we collected 11 publically available binary class data sets (disease vs. normal). The data sets are: colon cancer data [2], breast cancer data [79], pancreatic cancer premalignant data [81], cervical cancer data [197], acute myeloid leukemia data[166], ovarian cancer data[`http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp`], head & neck squamous cell carcinoma data [100], early-early stage duchenne muscular dystrophy(EDMD) data [131], HIV encephalitis data [115], high grade glioma data [127], and breast cancer data [181]. Missing values of these data sets have been imputed based on the nearest neighbor method. An overview of the characteristics of all the data sets can be found in Table 5.1. In breast cancer II and high grade glioma data sets, all data samples have already been assigned to a training set or test set. In all other cases, 2/3rd of the data samples of each class are assigned randomly to the training and the rest to the test set. These randomization are the same for all numerical experiments on all data sets. This split was performed stratified to ensure that the relative proportion of outcomes sampled in both training and test set was similar to the original proportion in the full data set. In all these cases, the data were standardized to zero mean and unit variance.

5.3 Methods

Link with an LS-SVM approach to KPCA in Section 2.2.3 and the idea of least squares cross-validation in kernel density estimation, we propose a new data-driven bandwidth selection criterion to tune the LS-SVM formulation of KPCA.

Table 5.1: Summary of the 11 binary disease data sets.

Data set	#Samples Class 1	Class2	#Genes
1: Colon	40	22	2000
2: Breast cancer I	34	99	5970
3: Pancreatic	50	50	15154
4: Cervical	8	24	10692
5: Leukemia	26	38	22283
6: Ovarian	91	162	15154
7: Head & neck squamous cell carcinoma	22	22	12625
8: Duchenne muscular dystrophy	23	14	22283
9: HIV encephalitis	16	12	12625
10: High grade glioma	28	22	12625
11: Breast cancer II	46	51	24188

5.3.1 Data-Driven Bandwidth Selection for KPCA

Model selection is a central issue in all learning tasks, especially in KPCA. Since KPCA is an unsupervised technique, formulating a data-driven bandwidth selection criterion is not trivial. Analogue to least squares cross validation [19, 145] in kernel density estimation, we propose a new data driven selection criterion for KPCA. Let

$$z_n(x) = \Sigma_{i=1}^{N} \alpha_i^{(n)} K(x_i, x)$$

where $K(x_i, x_j) = \exp(-\frac{||x_i - x_j||^2}{2h^2})$ (RBF kernel with bandwidth h) and set the target equal to 0 and denote by $z_n(x)$ the score variable of sample x on n^{th} eigenvector $\alpha^{(n)}$. Here, the score variables are expressed in terms of kernel expressions in which every training point contributes. These expansions are typically dense (nonsparse). Therefore we have chosen the L_1 loss function to induce sparsness in kernel PCA. We propose the following tuning criterion for the bandwidth h:

$$J(h) =_{h \in \mathbb{R}_0^+} \mathrm{E} \int |z_n(x)| dx, \tag{5.1}$$

where E denotes the expectation operator. Maximizing Equation (5.1) would lead to overfitting since we used all the training data in the criterion. Instead, we work with Leave One Out (LOO) cross validation estimation of $z_n(x)$ to obtain the optimum bandwidth h of KPCA, which gives projected variables

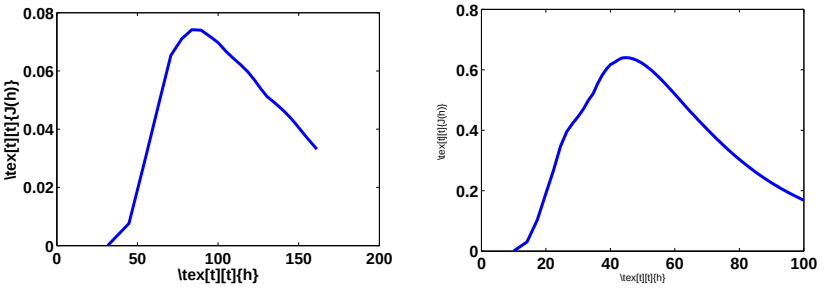


Figure 5.1: Bandwidth selection of KPCA for cervical and colon cancer data on fixed number of components. The plot $J(h)$ vs. h maximizes $J(h)$ at optimal bandwidth h. (a) 5th principal component for cervical cancer data and (b) 15th principal component for colon cancer .

with maximal variance. A finite approximation to Equation (5.1) is given by

$$J(h) =_{h\in\mathbb{R}_0^+} \frac{1}{N}\sum_{j=1}^{N}\int |z_n^{(-j)}(x)|dx \tag{5.2}$$

where N is the number of samples and $z_n^{(-j)}$ denotes the score variable with the jth observation is left out. In case the leave one out approach is computationally expensive, one could replace it with a leave v group out strategy (v- fold cross-validation). Integration can be performed by means of any numerical technique. The final model with optimum bandwidth is constructed as follows:

$$\Omega_{c,\hat{h}_{max}}\alpha = \lambda\alpha,$$

where $\hat{h}_{max} = \max_{h\in\mathbb{R}_0^+} \frac{1}{N}\sum_{j=1}^{N}\int |z_n^{(-j)}(x)|dx$. Figure 5.1 shows the bandwidth selection for cervical and colon cancer data sets for fixed number of components. To also retain the optimum number of components k of KPCA, we modify Equation (5.2) as follows: Figure 5.3 shows the surface plot of Equation (5.3) for various values of h and k.

$$J(h,k) =_{h\in\mathbb{R}_0^+,k\in\mathbb{N}_0} \frac{1}{N}\sum_{n=1}^{k}\sum_{j=1}^{N}\int |z_n^{(-j)}(x)|dx \tag{5.3}$$

where $k = 1,\ldots,N$. Figure 5.2 illustrate the proposed model. Thus, the proposed data-driven model can obtain the optimal bandwidth for KPCA, while retaining minimum number of eigenvectors which capture the majority of the variance of the data. Figure 5.4 shows a slice of the surface plots. The values of

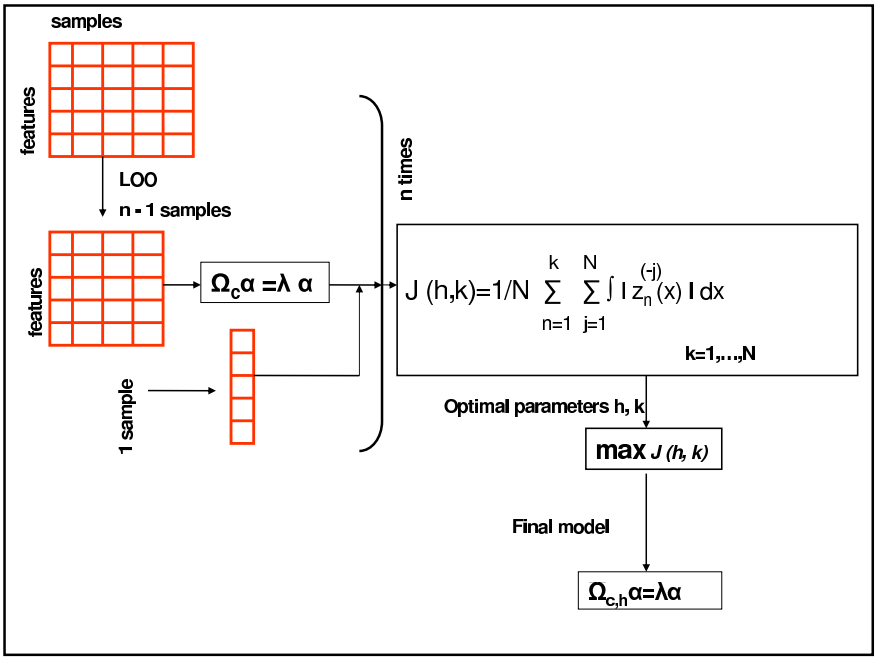


Figure 5.2: Data-Driven Bandwidth Selection for KPCA

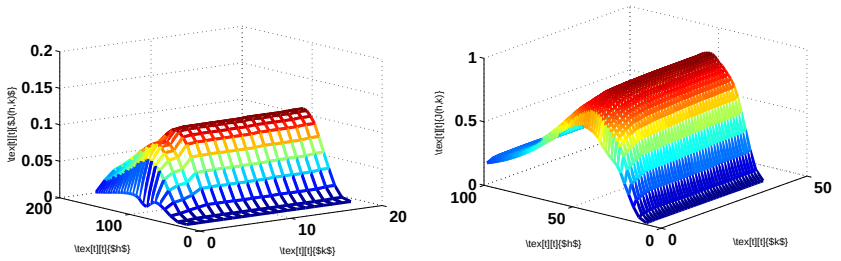


Figure 5.3: The surface plot of Equation (5.3) for various values of h and k. Model selection for KPCA-optimal bandwidth and number of components.(a) Cervical cancer (b) Colon cancer .

the proposed criterion were rescaled to be maximum 1. The parameters that maximize Equation (5.3) are $h = 70.71$ and $k = 5$ for cervical cancer data and $h = 43.59$ and $k = 15$ for colon cancer data.

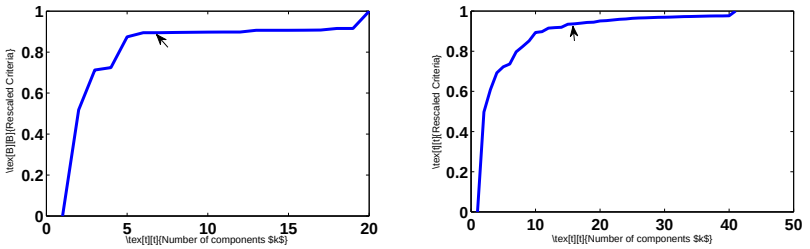


Figure 5.4: Slice plot for the Model selection for KPCA for the optimal bandwidth.(a) Cervical cancer (b) Colon cancer .

5.4 Results

First we considered nine data sets described in Table 5.1. We have chosen the RBF kernel $K(x_i, x_j) = \exp(-\frac{||x_i - x_j||^2}{2h^2})$ for KPCA. In this section all the steps are implemented using Matlab R2012b and LS-SVMlab v1.8 toolbox [46]. Next, we compared the performance of the proposed method with classical PCA and an existing tuning algorithm for RBF-KPCA developed by Pochet *et al.* [134]. Later, with the intention to comprehensively compare PCA+LS-SVM and KPCA+LS-SVM with other classification methods, we applied four widely used classifiers to the microarray data, being LS-SVM on whole data sets, t-test + LS-SVM, PAM and Lasso. To fairly compare kernel functions of the LS-SVM classifier; linear, RBF and polynomial kernel functions are used (in Table 5.2 referred to as linear/poly/RBF). The average test accuracies and execution time for all these methods when applied to the 9 case studies are shown in Table 5.2 and Table 5.3 respectively. For all these methods, training on 2/3rd of the samples and testing on 1/3rd of the samples was repeated 30 times.

5.4.1 Proposed Criterion with PCA

For each data set, the proposed methodology is applied. This methodology consists of two steps. First, Equation (5.3) is maximized in order to obtain an optimal bandwidth h and corresponding number of components k. Second, the reduced data set is used to perform a classification task with LS-SVM. We retained 5 and 15 components respectively for cervical and colon cancer data sets. For PCA, the first k components capturing most of the variability of the original dataset were used. We retained 13 components (83.68% of variance explained) and 15 components (83.20% of variance explained) for cervical and colon cancer respectively for PCA. Similarly, we obtained number of components of PCA

Table 5.2: Comparison of classifiers: Mean AUC(std) of 30 iterations

Data set	Kernel function	preprocessing + LS-SVM classifier				PAM	Lasso
		whole data	PCA	KPCA	t-test($p<0.05$)		
	RBF	0.769(0.127)	0.793(0.081)	0.822(0.088)	0.816(0.094)		
I	lin	0.822(0.068)	0.837(0.088)	**0.864**(0.078)	**0.858**(0.077)	0.787(0.097)	**0.837**(0.116)
	poly	0.818(0.071)	0.732(0.072)	0.825(0.125)	0.829(0.071)		
	RBF	0.637(0.146)	0.749(0.093)	**0.780**(0.076)	0.760(0.080)		
II	lin	**0.803**(0.059)	0.772(0.094)	**0.790**(0.075)	0.764(0.067)	0.659(0.084)	0.766(0.074)
	poly	0.701(086)	0.752(0.063)	0.753(0.072)	0.766(0.064)		
	RBF	0.832(0.143)	0.762(0.066)	0.879(0.058)	0.913(0.047)		
III	lin	**0.915**(0.043)	0.785(0.063)	0.878(0.066)	**0.913**(0.047)	0.707(0.067)	**0.9359**(0.0374)
	poly	0.775(0.080)	0.685(0.105)	0.8380(0.068)	0.913(0.047)		
	RBF	0.615(0.197)	0.853(0.112)	0.867(0.098)	0.853(0.187)		
IV	lin	**0.953**(0.070)	0.917(0.083)	**0.929**(0.077)	**0.924**(0.070)	0.759(0.152)	0.707(0.194)
	poly	0.762(0.118)	0.811(0.140)	0.840(0.131)	0.733(0.253)		
	RBF	0.807(0.238)	0.790(0.140)	0.976(0.035)	0.950(0.150)		
V	lin	**0.997**(0.005)	0.528(0.134)	0.982(0.022)	**0.999**(0.001)	0.923(0.062)	0.934(0.084)
	poly	0.942(0.051)	0.804(0.121)	0.975(0.028)	**0.999**(0.002)		
	RBF	**0.998**(0.001)	0.982(0.002)	0.984(0.012)	**0.998**(0.004)		
VI	lin	0.990(0.005)	0.973(0.002)	0.978(0.013)	0.993(0.013)	0.960(0.016)	0.951(0.045)
	poly	**0.998**(0.006)	0.985(0.016)	0.973(0.018)	0.955(0.042)		
	RBF	0.946(0.098)	0.941(0.057)	0.932(0.071)	0.940(0.098)		
VII	lin	**0.983**(0.025)	0.947(0.047)	**0.954**(0.051)	**0.983**(0.031)	0.931(0.058)	0.952(0.030)
	poly	0.785(0.143)	0.903(0.078)	0.915(0.080)	0.920(0.025)		
	RBF	0.823(0.159)	0.923(0.096)	0.858(0.113)	0.950(0.150)		
VIII	lin	0.840(0.164)	0.969(0.044)	0.800(0.019)	**0.999**(0.005)	**0.982**(0.050)	0.890(0.081)
	poly	0.781(0.186)	0.870(0.117)	0.785(0.121)	**0.998**(0.007)		
	RBF	0.638(0.210)	0.823(0.159)	0.852(0.180)	**0.873**(0.166)		
IX	lin	**0.931**(0.126)	0.840(0.164)	0.846(0.143)	**0.875**(0.136)	0.703(0.175)	0.705(0.174)
	poly	0.841(0.176)	0.781(0.186)	0.798(0.193)	0.798(0.193)		

Table 5.3: Summary of averaged execution time of each classifiers over 30 iterations in seconds.

Dataset	whole data	PCA	KPCA	t-test ($p < 0.05$)	PAM	Lasso
1: Colon	17	10	18	13	8	72
2: Breast	56	38	54	42	12	258
3: Pancreatic	17	12	26	19	20	453
4: Cervical	43	28	29	33	43	106
5: Leukemia	225	185	184	195	28	680
6: Ovarian	51	25	39	44	19	865
7: Head & neck squamous cell carcinoma	59	39	45	47	30	238
8: Duchenne muscular dystrophy	146	115	113	110	80	20100
9: HIV encephalitis	45	27	27	28	88	118

and the number of components with corresponding bandwidth for KPCA for the remaining data sets.

The score variables (the new coordinate of samples projected onto the KPCA or PCA components having maximum variance) are used to develop an LS-SVM classification model. The test AUC values averaged over the 30 random repetitions are reported.

While comparing the performance (test AUC and execution time) KPCA outperformed PCA on a majority of the cases in terms of test AUC with only very small difference in execution time.

5.4.2 Proposed Criterion with Existing Optimization Algorithm for RBF-KPCA

We selected two experiments from Pochet *et al.* [134] (last two data sets in Table 5.1), being high-grade glioma and breast cancer II data sets. We repeated the same experiments as reported in Pochet *et al.* [134] and compared with the proposed strategy. The results are shown in Table 5.4. The three dimensional surface plot of LOO-CV performance of Pochet *et al.* method [134] for the high-grade glioma data set is shown in Figure 5.5, with the optimal $h = 114.018$ and $k = 12$. The optimum parameters are $h = 94.868$ and $k = 10$ obtained by the proposed strategy (see Equation (5.3)) for the same data set. When looking at test AUC in Table 5.4, both case studies applying the proposed strategy, perform better than the method proposed by Pochet *et al.*[134] with

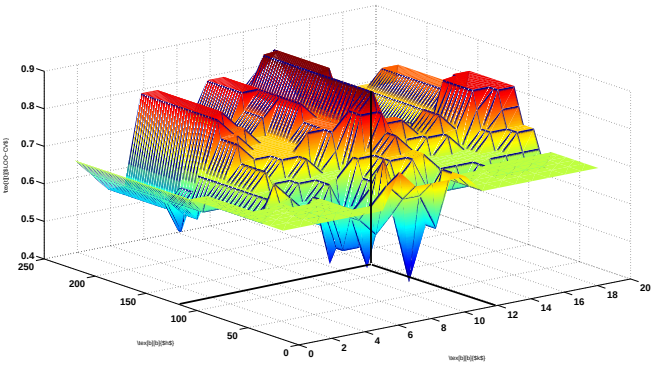


Figure 5.5: The three dimensional surface plot of LOO-CV performance of optimization algorithm [134] on high-grade glioma data set

Table 5.4: KPCA + LS-SVM Classifier: Comparison of performance of proposed bandwidth selection criterion for KPCA with the method proposed by Pochet *et al.* [134]: Averaged test AUC(std) over 30 iterations and execution time in minutes

Data set	proposed strategy		Pochet *et al.* [134]	
	Test AUC	time	Test AUC	time
high-grade glioma data	0.746(0.071)	2	0.704(0.104)	38
breast cancer II	0.6747(0.1057)	4	0.603(0.157)	459

less variability. In addition, the tuning method Pochet *et al.* [134] appears to be quite time consuming, whereas the proposed model enjoys a distinctive advantage with its low time complexity to carry out the same process.

5.4.3 Proposed Criterion with Other Classifiers

When looking specifically at all these methods in term of test AUC, we note that LS-SVM performance was slightly low on PCA. On breast cancer I, cervical cancer and HIV encephalitis data sets LS-SVM with linear kernel performs significantly better in terms of test AUC. The t-test + LS-SVM classifier shows the best test AUC for Leukemia and EDMD data sets. LS-SVM with linear kernel and t-test + LS-SVM classifiers have approximately the same test AUC on ovarian cancer and head & neck squamous cell carcinoma data sets. The

proposed strategy with LS-SVM (RBF) classifiers offer better test AUC for colon cancer, breast cancer I, cervical cancer and head & neck squamous cell carcinoma data sets. Only on pancreatic data set, Lasso outperformed all other case studies. The test AUC of PAM was significantly worse on all data sets except DMD data set.

5.5 Discussions

While analyzing the test AUC of different classifiers on nine data sets does not direct to a common conclusion that one method outperforms the other. Instead, it shows that each of these methods has their own advantage in classification tasks. When considering classification problems without dimensionality reduction, the regularized LS-SVM classifier on half of data sets shows a good performance. Up till now, most microarray data sets are quite small, but it can be expected that these data sets will become larger or perhaps represent more complex classification problems in the future. In this case, dimensionality reduction processes (feature selection and feature transformation) become important.

The selected features on feature selection methods such as t-test, PAM and Lasso widely vary for each random iteration. Further, the classification performance of these methods on each iteration depends on the number of features selected. Table 5.5 shows the range, i.e. minimum and maximum number of features selected on 30 iterations. PAM and Lasso only outperformed in two case studies. However, PAM is a user friendly toolbox for gene selection and classification tasks, its performance depends really on the selected features. In addition, it is interesting that the Lasso selected only very small subsets of the actual data sets. But, in the lasso, the amount of shrinkage varies, depending on the value of the tuning parameter, which is often determined by cross validation [187]. The number of genes selected as the outcome-predictive genes generally decrease as the value of the tuning parameter increases. The optimal value of the tuning parameter that maximizes the prediction accuracy is determined; however, the set of genes identified using the optimal value contains the non-outcome-predictive genes (ie, false positive genes) in many cases [93].

The test AUC on all nine case studies show that KPCA performs better than classical PCA. But the parameters of KPCA need to be optimized. We note that an already existing optimization algorithm for KPCA proposed by Pochet *et al.* [134] is completely coupled with the subsequent classifier. In addition, it appears to be very time-consuming. By the data-driven parameter selection of KPCA, the proposed strategy enhances KPCA as a real preprocessing step.

Table 5.5: Summary of the range (minimum to maximum) of features selected by t-test over 30 iterations.

Dataset	t-test ($p < 0.05$)	PAM	Lasso
1: Colon	173-522	15-373	8-36
2: Breast	746-1124	13-4718	7-87
3: Pancreatic	2564-4855	3-1514	12-112
4: Cervical	954-2108	2-10692	5-67
5: Leukemia	742-3468	137-11453	2-69
6: Ovarian	321-950	34-278	62-132
7: Head and neck squamous cell carcinoma	1761-2828	1-12625	3-35
8: Duchenne muscular dystrophy	3066-4536	129-22283	8-24
9: HIV encephalitis	620-2013	1-12625	1-20

In combination with classification methods, microarray data analysis can be useful to guide clinical management in cancer studies. In this study, several mathematical and statistical techniques evaluated and compared in order to optimize the performance of clinical predictions based on microarray data. Considering the possibility of increasing size and complexity of microarray data sets in the future, dimensionality reduction and nonlinear techniques have its own significance. In many cases, in a specific application context the best feature set is still important (e.g. drug discovery). While considering the stability and performance (both accuracy and execution time) of classifiers, the proposed methodology has its own importance to predict classes of future samples of known disease cases.

5.6 Conclusion

The objective in class prediction with microarray data is an accurate classification of cancerous samples, allowing for directed and more successful therapies. In this paper, we proposed a new data-driven bandwidth selection criterion for KPCA (which is a well defined preprocessing technique). In particular, we optimize the bandwidth and the number of components to maximize the projected variance of KPCA. In addition, we compared several data preprocessing techniques prior to classification. In all case studies, most of these data preprocessing steps performed well on classification with approximately similar performance. We observed that feature selection based methods selected features widely vary on each iteration. Hence it is difficult, even impossible to design a stable class

predictor for future samples with these methods. Experiments on nine data sets show that the proposed strategy provides a stable preprocessing algorithm for classification of high dimensional data with good performance on test data.

The advantages of the proposed KPCA+LS-SVM classifier were presented in four aspects. First, we propose a data-driven bandwidth selection criterion for KPCA by tuning the optimum bandwidth and the number of principal components. Second, we illustrate that the performance of the proposed strategy is significantly better than an existing optimization algorithm for KPCA. Third, its classification performance is not sensitive to any number of selected genes, so the proposed method is more stable than others proposed in literature. Fourth, it reduces the dimensionality of the data while keeping as much information as possible of the original data. This leads to computationally less expensive and more stable results for massive microarray classification.

Chapter 6

Predicting breast cancer using an expression values weighted clinical classifier

This paper was published in BMC Bioinformatics: Thomas M., De Brabanter K., Suykens J.A.K., De Moor B.: Predicting breast cancer using an expression values weighted clinical classifier. BMC Bioinformatics 2014, 15:411 (2014).

Background: Clinical data, such as patient history, laboratory analysis, ultrasound parameters-which are the basis of day-to-day clinical decision support-are often used to guide the clinical management of cancer in the presence of microarray data. Several data fusion techniques are available to integrate genomics or proteomics data, but only a few studies have created a single prediction model using both gene expression and clinical data. These studies often remain inconclusive regarding an obtained improvement in prediction performance. To improve clinical management, these data should be fully exploited. This requires efficient algorithms to integrate these data sets and design a final classifier.

LS-SVM classifiers and generalized eigenvalue/singular value decompositions are successfully used in many bioinformatics applications for prediction tasks. While bringing up the benefits of these two techniques, we propose a machine learning approach, a weighted LS-SVM classifier to integrate two data sources: microarray and clinical parameters.

Results: We compared and evaluated the proposed methods on five breast

cancer case studies. Compared to LS-SVM classifier on individual data sets, the generalized eigenvalue decomposition (GEVD) and the kernel GEVD, the proposed weighted LS-SVM classifier offers good prediction performance, in terms of test area under the ROC Curve (AUC), on all breast cancer case studies.

Conclusions: A clinical classifier weighted with microarray data set results in significantly improved diagnosis, prognosis and prediction responses to therapy. The proposed model has been shown to be a promising mathematical framework in both data fusion and non-linear classification problems.

6.1 Background

Microarray technology, which can handle thousands of genes of several hundreds of patients at a time, makes it hard for scientists to manually extract relevant information about genes and diseases, especially cancer. Moreover this technique suffers from a low signal-to-noise ratio. Despite the rise of high-throughput technologies, clinical data such as age, gender and medical history, guide clinical management of most diseases and examinations. A recent study [183] shows that the importance of the integration of microarray and clinical data have a synergetic effect on predicting breast cancer outcome. Gevaert et al. [64] have used a Bayesian framework to combine expression and clinical data. They found that decision integration, and partial integration leads to a better performance, whereas full data integration showed no improvement. These results were obtained by using a cross validation approach on the 78 samples in the van't Veer et al.[181] data set. On the same data set, Boulesteix et al. [18] employed random forests and partial least squares approaches to combine expression and clinical data. In contrast, they reported that microarray data do not noticeably improve the prediction accuracy yielded by clinical parameters alone.

The representation of any data set with a real-valued kernel matrix, independent of the nature or complexity of data to be analyzed, makes kernel methods ideally positioned for heterogeneous data integrations. Integration of data using kernel fusion is featured by several advantages. Biological data has diverse structures, for example, high dimensional expression data, the sequence data, the annotation data, the text mining data and heterogeneous nature of clinical data and so on. The main advantage is that the data heterogeneity is rescued by the use of kernel trick, where data which has diverse data structures are all transformed into kernel matrices of the same size. To integrate them, one could follow the classical additive expansion strategy of machine learning to combine them

linearly. These nonlinear integration methods of kernels have attracted great interests in recent machine learning research.

Daemen et al.[39] proposed kernel functions for clinical parameters and pursued an integration approach based on combining kernels (kernel inner product matrices derived from the separate data types) for application in a Least Squares Support Vector Machine (LS-SVM)[170]. They explained that the newly proposed kernel functions for clinical parameter do not suffer from the ambiguity of data preprocessing by equally considering all variables. That means, a distinction is made between continuous variables, ordinal variables with an intrinsic ordering, but often lacking equal distance between two consecutive categories and nominal variables without any ordering. They concluded that the clinical kernel functions represent similarities between patients more accurately than linear or polynomial kernel function for modeling clinical data. Pittman et al. [132] combined clinical and expression data for predicting breast cancer outcome by means of a tree classifier. This tree classifier was trained using meta-genes and/or clinical data as inputs. They explained that key metagenes can up to a degree, replace traditional risk factors in terms of individual association with recurrences. But the combination of metagenes and clinical factors currently defines models most relevant in terms of statistical fit and also, more practically, in terms of cross-validation predictive accuracy. The resulting tree models provide an integrated clinico-genomic analysis that generate substantially accurate and cross-validated predictions at the individual patient level.

Singular Value Decomposition (SVD) and generalized SVD (GSVD) have been shown to have great potential within bioinformatics for extracting common information from data sets such as genomics and proteomics data [152, 3]. Several studies have used LS-SVM as a prediction tool, especially in microarray analysis [32, 83].

In this Chapter, we propose a machine learning approach for data integration: a weighted LS-SVM classifier. Initially we will explain generalized eigenvalue decomposition (GEVD) and kernel GEVD. Later we will explore the relationships of kernel GEVD with weighted LS-SVM classifier. Finally, the advantages of this new classifier will be demonstrated on five breast cancer case studies, for which expression data and an extensive collection of clinical data are publicly available.

6.2 Data sets

Breast cancer is one of the most extensively studied cancer types for which many microarray data sets are publicly available. Among them, we selected five cases for which a sufficient number of clinical parameters were available [31, 79, 158, 181, 120]. All the data sets that we have used are available in the Integrated Tumor Transcriptome Array and Clinical data Analysis database (ITTACA): `http://bioinfo-out.curie.fr/ittaca/`. Overview of all the data sets are given in Table 6.1.

Table 6.1: Summary of the 5 breast cancer data sets.

Case Study	#Samples Class 1	Class2	#Genes	#Clinical variables
Case I	85	25	5000	Age, Ethnicity, ER status, PR status, Radiation treatment, Chemotherapy, Hormonal therapy, Nodal status, Metastasis, Tumor stage, Tumor size, Tumor grade.
Case II	33	96	6000	Age, Ethnicity, pretreatment tumor stage, nodal status, nuclear grade, ER status, PR status, HER2 status.
Case III	112	65	5000	Age, Tumor size, Nodal status, ER status, Tamoxifen treatment.
Case IV	46	51	12192	Age, Tumor size, Grade, Erp, Angioinvasion, Lymphocytic Infiltrate, PRp. Lymphocytic Infiltrate, PRp.
Case V	58	193	20055	Age, Tumor size, Grade, ER, Prp, Lymph node.

6.2.1 Microarray Data

For the first three data sets, the microarray data were obtained with the Affymetrix technology and preprocessed with MAS5.0, the GeneChip Microarray Analysis Suite 5.0 software (Affymetrix). However, as probe selection for the Affymetrix gene chips relied on earlier genome and transcriptome annotation that are significantly different from current knowledge, an updated array annotation was used for the conversion of probes to Entrez Gene IDs, lowering the number of false positives [40].

A fourth data set consists of two groups of patients[181]. The first group of patients, the training set, consists of 78 patients of, which 34 patients belonged to the poor prognosis group and 44 patients belonged to the good prognosis group. The second group of patients, the test set, consists of 19 patients, of which 12 patients belonged to the poor prognosis group and 7 patients belonged

to the good prognosis group. The microarray data was already background corrected, normalized and log-transformed. Preprocessing step removes genes with small profile variance, less than the 10th percentile.

The last data set consists of transcript profiles of 251 primary breast tumors were assessed by using Affymetrix U133 oligonucleotide microarrays. cDNA sequence analysis revealed that 58 of these tumors had p53 mutations resulting in protein-level changes, whereas the remaining 193 tumors were p53 wt [120].

6.2.2 Clinical Data

The first data of 129 patients contained information on 17 available clinical variables, 5 were excluded [31]: two redundant variables that were least informative based on univariate analysis in those variable pairs with a correlation coefficient exceeding 0.7, and three variables with too many missing values. After exclusion of patients with missing clinical information, this data set consisted of 110 patients remained in 85 of whom disease did not recur whilst in 25 patients disease recurred.

The second data in which response to treatment was studied, consisted of 12 variables for 133 patients [79]. Patient and variable exclusion as described above resulted in this data set. Of the 129 remaining patients, 33 showed complete response to treatment, while 96 patients were characterized as having residual disease.

In the third data, relapse was studied in 187 patients [158]. After preprocessing, this data set retained information on 5 variables for 177 patients. In 112 patients, no relapse occurred while 65 patients were characterized as having a relapse.

The fourth data set [181] consisted of predefined training and test sets same as that of corresponding microarray data. The last data set consisted of 251 patients with 6 available clinical variables [120]. After exclusion of patients with missing clinical information, this data set consisted of 237 patients, of which 55 patients with p53 mutant breast tumor and the remaining patients without p53 mutant breast tumor.

6.3 Methods

In the first section, we will discuss the GEVD and represent it in terms of an ordinary EVD. Next, we formulate an optimization problem for kernel GEVD in primal space and solution in dual space. Finally, by generalizing

this optimization problem in terms of LS-SVM classifier, we propose a new machine learning approach for data fusion and classifications, a weighted LS-SVM classifier.

6.3.1 kernel GEVD

The GEVD in Equation 2.3 can now be rewritten as EVD problem :

$$(B^TB)^{-1/2}A^TA(B^TB)^{-1/2}W = W\Lambda.$$

where $W = (B^TB)^{1/2}(X^T)^{-1}$. The SVD of matrix $A(B^TB)^{-1/2}$ is given below:

$$D = A(B^TB)^{-1/2} = PSQ^T. \tag{6.1}$$

The matrix $(B^TB)^{-1/2}$ is defined [80] as follows: Let EVD of $B^TB = T\Sigma T^T$, where the columns of T are eigenvectors and Σ is a diagonal matrix. $(B^TB)^{1/2} = T\Sigma^{1/2}T^T$ and $(B^TB)^{-1/2} = T\Sigma^{-1/2}T^T$.

In this section we discuss LS-SVM formulations to kernel GEVD, which is a non-linear GEVD of $m \times n$ matrix A, and $p \times n$ matrix B, and a weighted LS-SVM classifier. LS-SVM formulations to different problems were discussed in [170]. This class of kernel machines emphasizes primal-dual interpretations in the context of constrained optimization problems.

Given a training data set of n points $\mathcal{D} = \{x_i^{(1)}, x_i^{(2)}, y_i\}_{i=1}^n$ with output data $y_i \in \mathbb{R}$ and input data sets $x_i^{(1)} \in \mathbb{R}^m$, $x_i^{(2)} \in \mathbb{R}^p$ ($x_i^{(1)}$ and $x_i^{(2)}$ are the i^{th} column of matrices A and B respectively).

Consider the feature maps $\varphi^{(1)}(.) : \mathbb{R}^m \to \mathbb{R}^{n_1}$ and $\varphi^{(2)}(.) : \mathbb{R}^p \to \mathbb{R}^{n_2}$ to a high dimensional feature space $\mathcal{F}$, which is possibly infinite dimensional. The centered feature matrices $\Phi_c^{(1)} \in \mathbb{R}^{n_1 \times N}$, $\Phi_c^{(2)} \in \mathbb{R}^{n_2 \times N}$ become

$$\Phi_c^{(1)} = [\varphi^{(1)}(x_1^{(1)})^T - \hat{\mu}_{(\varphi_1)}^T; \ldots; \varphi^{(1)}(x_N^{(1)})^T - \hat{\mu}_{(\varphi_1)}^T]^T$$

$$\Phi_c^{(2)} = [\varphi^{(2)}(x_1^{(2)})^T - \hat{\mu}_{(\varphi_2)}^T; \ldots; \varphi^{(2)}(x_N^{(2)})^T - \hat{\mu}_{(\varphi_2)}^T]^T,$$

where $\hat{\mu}_{\varphi l} = \frac{1}{N}\Sigma_{i=1}^N \varphi^{(l)}(x_i^{(l)})$, $l = 1, 2$

LS-SVM approach to Kernel GEVD

Kernel GEVD is a nonlinear extension of GEVD, in which the data are first embedded into a high dimensional feature space introduced by the kernel and

then linear GEVD is applied. While considering the matrix $A(B^TB)^{-1/2}$ in Equation 6.1 and the feature maps $\varphi^{(1)}(.) : \mathbb{R}^m \rightarrow \mathbb{R}^{n_1}$ and $\varphi^{(2)}(.) : \mathbb{R}^p \rightarrow \mathbb{R}^{n_2}$ described in the previous section, the approximation of covariance matrix for $A(B^TB)^{-1/2}$ by explicit feature map in the feature space becomes $C \approx \Phi_c^{(1)}(\Phi_c^{(2)^T}\Phi_c^{(2)})^{-1}\Phi_c^{(1)^T}$ with eigendecomposition $Cv = \lambda v$.

While considering the kernel PCA formulation based on the LS-SVM framework in [4] and EVD of $Cv = \lambda v$ in primal space, our objective is to find the directions in which projected variables have maximal variance.

The LS-SVM approach to kernel GEVD is formulated as follows:

$$\boxed{\begin{array}{c} \min_{v,e} \; J(v,e) \; = \; \gamma\tfrac{1}{2}e^T(\Phi_c^{(2)^T}\Phi_c^{(2)})^{-1}e - \tfrac{1}{2}v^Tv \\ \\ \textit{such that} \quad e = \Phi_c^{(1)^T}v \; , \end{array}} \tag{6.2}$$

where v is the eigenvector in the primal space, $\gamma \in \mathbb{R}^+$ is a regularization constant and e are the projected data points to the target space.

Defining the Lagrangian

$$\mathcal{L}(v,e;\alpha) = \tfrac{\gamma}{2}e^T(\Phi_c^{(2)^T}\Phi_c^{(2)})^{-1}e - \tfrac{1}{2}v^Tv - \alpha^T\{(e - \Phi_c^{(1)^T}v)\},$$

with optimality conditions,

$$\tfrac{\partial\mathcal{L}}{\partial v} = 0 \rightarrow v = \Phi_c^{(1)} \; \alpha,$$

$$\tfrac{\partial\mathcal{L}}{\partial e} = 0 \rightarrow \alpha = \gamma(\Phi_c^{(2)^T}\Phi_c^{(2)})^{-1}e,$$

$$\tfrac{\partial\mathcal{L}}{\partial \alpha} = 0 \rightarrow e = \Phi_c^{(1)^T}v,$$

elimination of v and e will yield an equation in the form of a GEVD:

$$\boxed{\Omega_c^{(1)}\alpha = \lambda\Omega_c^{(2)}\alpha,}$$

where $\lambda = \frac{1}{\gamma}$ eigenvalue, $\Omega_c^{(1)}$, $\Omega_c^{(2)}$ are centered kernel matrices and α are generalized eigenvectors. The symmetric kernel matrices $\Omega_c^{(1)}$ and $\Omega_c^{(2)}$ resolves the heterogeneities of clinical and microarray data by the use of kernel trick, where data which have diverse data structures are transformed into kernel matrices with same size.

In a special case of GEVD, if one of the data matrix is the identity matrix, it will be equivalent to an ordinary EVD. If $(\Phi_c^{(2)^T}\Phi_c^{(2)})^{-1} = I$, then the optimization

problem proposed for kernel GEVD (See Equation 6.2)) will be equivalent to optimization problem in [171] for the LS-SVM approach to kernel PCA.

6.3.2 Weighted LS-SVM classifier

Our objective is to represent kernel GEVD in the form of weighted LS-SVM classifier. Given the link between LS-SVM approach to kernel GEVD in Equation 6.2 and the weighted LS-SVM classifier (see [169] in a different type of weighting to achieve robustness), one considers the following optimization problem in primal weight space:

$$\boxed{\begin{array}{c} \min_{v,e,b} \; J(v,e) \;=\; \gamma \tfrac{1}{2} e^T (\Phi_c^{(2)^T} \Phi_c^{(2)})^{-1} e + \tfrac{1}{2} v^T v \\ \\ \textit{such that} \quad y = \Phi_c^{(1)^T} v + b 1_N + e, \end{array}}$$

with $e = [e_1, \ldots, e_N]^T$ a vector of variables to tolerate misclassifications, weight vector v in primal weight space, bias term b and regularization parameter $\gamma \in \mathbb{R}^+$. Compared to the constrained optimization problem for least squares support vector machine (LS-SVM) [170, 172], in this case, the error variables are weighted with a matrix $(\Phi_c^{(2)^T} \Phi_c^{(2)})^{-1/2}$.

The weight vector v can be infinite dimensional, which makes the calculation of v impossible in general. One defines the Lagrangian
$\mathcal{L}(v, e, b; \alpha) = \frac{1}{2} v^T v + \frac{\gamma}{2} e^T (\Phi_c^{(2)^T} \Phi_c^{(2)})^{-1} e - \alpha^T \{((\Phi_c^{(1)^T} v) + b 1_N) + e - y\}$,

with Lagrange multipliers $\alpha \in \mathbb{R}^N$.

$\frac{\partial \mathcal{L}}{\partial v} = 0 \rightarrow v = \Phi_c^{(1)} \; \alpha$

$\frac{\partial \mathcal{L}}{\partial b} = 0 \rightarrow 1_N^T \alpha = 0$

$\frac{\partial \mathcal{L}}{\partial e} = 0 \rightarrow \alpha = \gamma (\Phi_c^{(2)^T} \Phi_c^{(2)})^{-1} e$

$\frac{\partial \mathcal{L}}{\partial \alpha_i} = 0 \rightarrow e + \Phi_c^{(1)^T} v + b = y$

Elimination of v and e yields a linear system

$$\begin{bmatrix} 0 & 1_N^T \\ 1_N & \Omega_c^{(1)} + \gamma^{-1} \Omega_c^{(2)} \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{6.3}$$

with $y = [y_1, ..., y_N]^T$, $1_N = [1, ..., 1]^T$, $\alpha = [\alpha_1, ..., \alpha_N]^T$, $\Omega_c^{(1)} = {\Phi_c^{(1)}}^T \Phi_c^{(1)}$ and $\Omega_c^{(2)} = {\Phi_c^{(2)}}^T \Phi_c^{(2)}$.

The resulting classifier in the dual space is given by

$$y(x) = \sum_{i=1}^{N} \alpha_i([K^{(1)}(x, x_i) + \frac{1}{\gamma} K^{(2)}(x, x_i)] + b) \tag{6.4}$$

with α_i are the Lagrange multipliers, γ is a regularization parameter chosen by the user, $K^{(1)}(x, z) = {\varphi^{(1)}(x)}^T \varphi^{(1)}(z)$, $K^{(2)}(x, z) = {\varphi^{(2)}(x)}^T \varphi^{(2)}(z)$ and $y(x)$ is the output corresponding to validation point x. The LS-SVM for nonlinear function estimation in [169] is similar to the proposed weighted LS-SVM classifier.

The symmetric, kernel matrices $K^{(1)}$ and $K^{(2)}$ resolve the heterogeneities of clinical and microarray data sources such that they can be merged additively as a single kernel. The optimization algorithm for the weighted LS-SVM classifier is given below:

Algorithm: Optimization algorithm for the weighted LS-SVM classifier

1. Given a training data set of N points $\mathcal{D} = \{x_i^{(1)}, x_i^{(2)}, y_i\}_{i=1}^N$ with output data $y_i \in \mathbb{R}$ and input data sets $x_i^{(1)} \in \mathbb{R}^m$, $x_i^{(2)} \in \mathbb{R}^p$.
2. Calculate Leave-One-Out cross validation (LOO-CV) performances of the training set with different combinations of γ and σ_1, σ_2 (bandwidths of kernel functions $K^{(1)}$, $K^{(2)}$) by solving linear system Equation 6.3 and Equation 6.4. In case the Leave-One-Out (LOO) approach is computationally expensive, one could replace it with a leave p group out strategy (p-fold cross-validation)
3. Obtain the optimal parameters combinations (γ, σ_1, σ_2) which have the highest LOO-CV performance.

The proposed optimization problem is similar to the the weighted LS-SVM formulation in [4] which replaced $({\Phi_c^{(2)}}^T \Phi_c^{(2)})^{-1}$ with a diagonal matrix to achieve sparseness and robustness.

The proposed method is a new machine learning approach in data fusion and subsequent classifications. In this study, the advantages of a weighted LS-SVM classifier were explored, by designing a clinical classifier. This clinical classifier combined kernels by weighting kernel inner product from one data set with that from the other data set. Here we considered microarray kernels as weighting

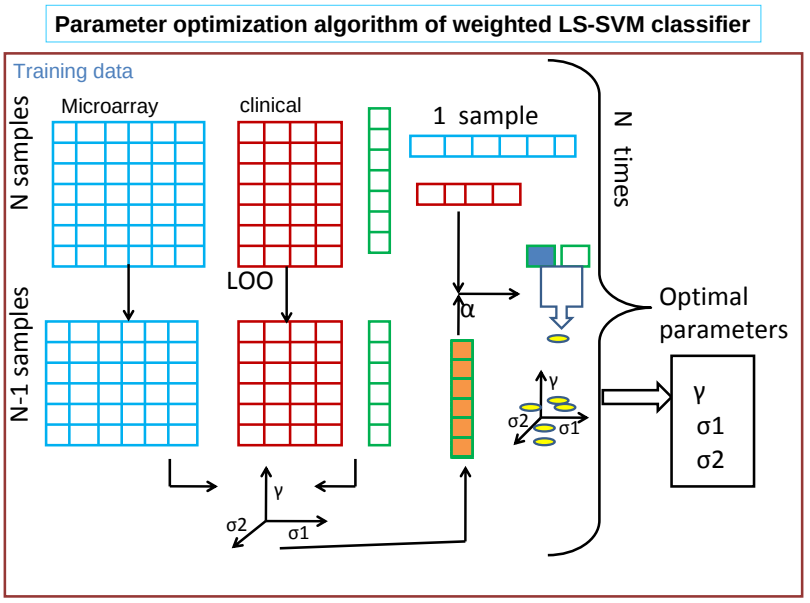


Figure 6.1: Overview of the algorithm. The data sets represented as matrices, with rows corresponding to patients and columns corresponding to genes and clinical parameters respectively for first and second data sets. LOO-CV is applied to select the optimal parameters.

matrix for clinical kernels. In each of these case studies, we compared the prediction performance of individual data sets with GEVD, kernel GEVD and weighted LS-SVM classifier. In kernel GEVD, σ_1 and σ_2 are the bandwidth of RBF-kernel function $K(x, z) = \exp(-\frac{||x-z||^2}{2\sigma^2})$ of clinical and microarray data sets respectively. The parameters σ_1, σ_2 were chosen as to obtain the highest LOO-CV performance. The parameter selection (see Algorithm) for the weighted LS-SVM classifier is illustrated in Figure 6.1. For several possible values (10^{-2} to 10^2, with grid size 10) of the kernel parameters σ_1 and σ_2, the LOO cross validation performance is computed for each possible combination (10^{-2} to 10^2, with grid size 10) of γ. The optimal parameters are the combinations (σ_1, σ_2, γ) with best LOO-CV performance. Remark the complexity of this optimization procedure because both the kernel parameters (σ_1 and σ_2) and γ need to be optimized in the sense of the LOO-CV performance.

6.4 Results

The summary of data sets for the five case studies is given in Table 6.1. In all the case studies, clinical data A contain measurements of m clinical parameters, for n samples and microarray data B contain expression levels of p genes over these n samples of breast cancer data. In all cases except the fourth, 2/3rd of the data samples of each class are assigned randomly to the training and the rest to the test set. This randomization is the same for all numerical experiments on all data sets. This split was performed to ensure that the relative proportion of outcomes sampled in both training and test set was similar to the original proportion in the full data set. In all these cases, the microarray data were standardized to zero mean and unit variance. Normalization of training sets as well as test sets is done by using the mean and standard deviation of each gene expression profile of the training sets. In the fourth data set [181], all data samples have already been assigned to a training set or test set.

Initially, LS-SVM classifier applied to RBF kernel functions of clinical and microarray data, individually for cancer patient classification. Then GEVD and kernel GEVD have used as a pre-processing step and LS-SVM classifier applied on the reduced data sets. Thus we compared the classification performances on single data source (microarray or clinical data) with two data sources (microarray and clinical data). In addition, we compared the performance of pre-processing with GEVD and kernel GEVD on classification task. Finally, we integrated the two data sets, microarray and clinical, and perform further classification using the weighted LS-SVM classifier. In addition, we compared the proposed weighted LS-SVM classifiers with other existing data fusion techniques.

6.4.1 Kernel GEVD

In kernel GEVD, initially we formulated two symmetric kernel matrices $\Phi_c^{(1)}$ and $\Phi_c^{(2)}$ for clinical and microarray data respectively. The optimal parameters of these kernel matrices (σ_1 and σ_2, the bandwidths of clinical and microarray kernels) in GEVD framework were selected using LOO-CV performance. $N-1$ samples were used to create microarray and clinical kernels. Kernel GEVs α were obtained by applying GEVD on these matrices. Then the clinical kernel matrix projected on to the direction of kernel GEV α resulting in the training scores. Then the validation score on left-out sample obtained by projecting the validation clinical kernel onto the direction of kernel GEV α. Similar to GEVD, LS-SVM model was trained and tested on the training and test scores respectively. This procedure repeated for each sample in the given data sets with different combinations of σ_1 and σ_2. The combinations of σ_1 and σ_2 values

which offered the best CV-result will be chosen as the optimum ones. Then LS-SVM classifier is applied on the training and test clinical data projected onto the direction of kernel GEV α, to train the model and prediction. High-throughput data such as microarray have used only for the model development, that is, for obtaining generalized eigenvector α. Further the classifications are performed only on clinical kernels. Initially we applied LS-SVM classifier on microarray and clinical kernels individually. Then we used the GEVD and kernel GEVD as pre-processing steps and the LS-SVM classifier applied on the transformed clinical data. The results show that considerations of two data sets in a single framework improve the prediction performance than individual data sets. In addition, pre-processing with the kernel GEVD significantly improved the prediction performance than GEVD. The results of the five case studies are shown in Table 6.2 and Figure 6.2. We represented expression and clinical data with kernel matrix, based on RBF kernel function. The RBF kernel functions makes each of the these data which had diverse structures, transformed into kernel matrices with same size.

6.4.2 Weighted LS-SVM classifier

We proposed a weighted LS-SVM classifier, a useful technique in data fusion as well as in supervised learning. The parameters (γ in Equation 6.3 and σ_1, σ_2 the bandwidths of microarray and clinical kernel functions) associated with this method are selected by Algorithm. For several possible values (10^{-2} to 10^2 with grid size 10) of the kernel parameters σ_1 and σ_2, the LOO cross validation performance is computed for each possible combination (10^{-2} to 10^2 with grid size 10) of γ. In each LOO-CV, 1 samples is left out and models are built for all possible combinations of parameters on the remaining $N-1$ samples. The optimization problem is not sensitive to small changes of bandwidths of microarray and clinical kernel functions. Careful tuning of γ allows tackling the problem of overfitting and tolerating misclassifications. Model parameters are chosen corresponding to the model with highest LOO AUC. The LOO-CV approach takes less than a minute for a single iteration of the first three case studies and 1-2 minutes for the rest of case studies on Windows 7, Intel core i3 processor. Statistical significance test are performed in order to allow a correct interpretations of the results. A non-parametric paired test, the Wilcoxon signed rank test (signrank in Matlab) [45], has been used in order to make general conclusions. A threshold of 0.05 is chosen, which means that the two results are significantly different if the value of the Wilcoxon signed rank test applied to both of them is lower than 0.05. On all case studies, weighted LS-SVM classifier outperformed all other discussed methods (LS-SVM on RBF clinical kernel and microarray kernel individually, GEVD and kernel GEVD as pre-processing step,

Table 6.2: The summary of the classification performances (averaged test AUC (std) on ROC curve in 100 repetitions) of 5 breast cancer cases. CL+LS-SVM and MA + LS-SVM indicates that LS-SVM classifier applied on clinical and microarray kernels. Then we have used GEVD and kernel GEVD as preprocessing step and then applied LS-SVM classifier on the transformed data set. Finally the weighted LS-SVM classifier used for data integration and classification on clinical and microarray kernels. The AUC values obtained with different techniques were compared using a paired test, Wilcoxon signed rank test.

Classifiers	Case I	Case II	Case III	Case IV	Case V
CL +LS-SVM					
test AUC	0.7795(0.0687)	0.7772(0.0554)	0.6152(0.0565)	0.6622(0.0628)	0.7740(0.0833)
p-value	0.0039	1.48E-04	0.0086	5.21E-06	0.1602
MA+LS-SVM					
test AUC	0.7001(0.0559)	0.8065(0.0730)	0.6217(0.0349)	0.7357(0.0085)	0.6166(0.0508)
p-value	0.0059	0.0140	0.0254	2.41E-04	0.0020
GEVD+LS-SVM					
test AUC	0.7801(0.0717)	0.7673(0.0548)	0.6196(0.0829)	0.7730(0.1011)	0.8001(0.0648)
p-value	0.0137	3.41E-05	0.0040	0.1558	0.0840
KGEVD+LS-SVM					
test AUC	0.7982(0.0927)	0.8210(0.0670)	0.6437(0.0313)	0.7901(0.0917)	0.8031(0.0624)
p-value	0.0195	0.1144	0.0020	0.6162	0.0720
weighted LS-SVM					
test AUC	**0.8177**(0.0666)	**0.8465**(0.0480)	**0.6985**(0.0443)	**0.8119**(0.0893)	**0.8210**(0.0477)

p-value: a paired test, Wilcoxon signed rank test.
CL and MA are the clinical and microarray kernels of RBF kernel functions.

followed by LS- SVM on reduced data), in terms of test AUC, as shown in Table 6.2 and Figure 6.2. The weighted LS-SVM performance is slightly better on the second and fourth cases, but not significantly, than the kernel GEVD.

Then we compare the proposed weighted LS-SVM classifiers with other data fusion techniques which integrate microarray and clinical data sets. Daemen *et al* [39] investigated the effect of data integration on performance with three case studies [31, 79, 158]. They reported that a better performance was obtained when considering both clinical and microarray data with the weights (μ) assigned

Table 6.3: Comparisons of RBF with clinical kernel functions: On weighted LS-SVM framework, we evaluated the LOO-CV performances of, clinical kernel function in [39] and RBF microarray kernels, with RBF clinical and microarray kernels. In the weighted LS-SVM classifier framework, RBF kernel functions of clinical parameters performs better than clinical kernel functions on three case studies.

Kernel functions	Case I	Case II	Case III	Case IV	Case V
Clinical kernel	0.8108(0.0351)	**0.8315** (0.0351)	**0.7479**(0.0111)	0.7385(0.1100)	0.7673(0.0213)
RBF	**0.8243**(0.0210)	0.8202(0.0100)	0.7143(0.0217)	**0.7846**(0.0699)	**0.7862**(0.0221)

to them optimized (μClinical+(1-μMicroarray)). In addition, they concluded from their 10-fold AUC measurements that the clinical kernel variant, led to a significant increase in performance, in the kernel based integration approach of clinical and microarray. The first three case studies, we have taken from the work of Daemen *et al* [39]. They have considered the 200 most differential genes selected from the training data with the Wilcoxon rank sum test, for the kernel matrix obtained from microarray. The fourth case study, we have taken from the paper of Gevaert *et al* [64] in which they investigated different types of integration strategies, with Bayesian network classifier. They concluded that partial integration performs better in terms of test AUC. Our results also confirm that consideration of microarray and clinical data sets together, improves prediction performances than individual data sets. In addition, non-linear projections of data using kernel GEVD significantly improved the prediction performance than the linear GEVD because these projections captured the non-linear relationship in the data. In case III, feature selection of microarray has reduced the overfitting in the trained model, but the number of features of clinical data are not sufficient to generate a good model, hence the model did not perform well during the validation/testing step.

In our analysis, microarray-based kernel matrix are calculated on full data set without preselecting genes and thus avoiding potential selection bias [5]. In addition, we compared RBF kernel with the clinical kernel function [39] in terms of LOO-CV performance. Results are given in Table 6.3. We followed the same strategy which was explained for weighted LS-SVM classifier, except the clinical kernel function have been used for the clinical parameters. On three out of five case studies, RBF kernel functions performed better than clinical kernel function. Clinical kernel function outperformed RBF kernel function in cases II and III. But further study is required to make a conclusion that why did the clinical kernel function perform well in few cases, in which the RBF kernel function failed to perform, especially on case III.

6.5 Discussion

Integrative analysis has been primarily used to prioritize disease genes or chromosomal regions for experimental testing, to discover disease subtypes or to predict patient survival or other clinical variables. The ultimate goal of this work is to propose a machine learning approach which is functional in both data fusion and supervised learning. We further analyzed the potential benefits of merging microarray and clinical data set for prognostic application in breast cancer diagnosis.

We integrate microarray and clinical data into one mathematical model, for the development of highly homogeneous classifiers in clinical decision support. For this purpose, we present a kernel based integration framework in which each data set is transformed into a kernel matrix. Integration occurs on this kernel level without referring back to the data. Some studies [39, 183] already reported that intermediate integration of clinical and microarray data set, improves prediction performance on breast cancer outcome. In primal space, the clinical classifier is weighted with expression values. The solution in dual space is given on Equations 6.3 and 6.4 which provides a way to integrate two kernel functions explicitly and perform further classifications.

To verify the merit of the proposed approach over the single data sources such as clinical and microarray data, the LS-SVM were built on all data sets individually for classifying cancer patients. Next, GEVD and kernel GEVD were used as pre-processing step. Then the data in the projected space (scores) have used to build the LS-SVM classifier. Results show that these types of integration information helped us to achieve better prediction performances than considering single data sets. Integration of different data sources is relevant in cancer studies for better diagnosis, prognosis and personal therapy. In addition, the results suggest that kernel based data integration increases the predictive performance of clinical decision support models. This indicates that there might be non-linear pattern in the data that effectively modelled with kernel based techniques. Finally weighted LS-SVM approach was used for the integration of both microarray and clinical kernel functions and performed subsequent classifications. The weighted LS-SVM classifier proposes a new optimization framework to solve the problem of classification using features of different types such as clinical and microarray data. In this framework, kernel functions were applied to each data sets separately and a decision boundary was made in high dimensional feature space. Due to its unique property, that is, data fusion and classification in a single framework with kernel methods, it became a competitive non-linear data fusion and classification technique.

We should note that the models proposed in this paper are expensive, but

less than the other kernel-based data fusion techniques. Since the proposed weighted LS-SVM classifier simplified both data fusion and classification in a single framework, it does not have an additional cost for tuning parameters for kernel-based classifiers. And it is given that, the weighting matrix should be invertible in the optimization problem of kernel GEVD and the weighted LS-SVM classifier.

In life science research, there is an increasing need for heterogeneous data integration such as proteomics, genomics, mass spectral imaging and so on. Such studies are required to determine, which data sets are most significant to be considered as weighting matrix. The proposed weighted LS-SVM classifier integrates heterogeneous data sets to achieve best performing and affordable classifiers.

6.6 Conclusion

The results suggest that the use of our data integration approach on gene expression and clinical data can improve the performance of decision making in cancer. We proposed a weighted LS-SVM classifier for the integration of two data sources and further prediction task. Each data set is represented by a kernel matrix, based on the RBF kernel function. The proposed clinical classifier gives a step towards improving predictions for individual patients about prognosis, metastatic phenotype and therapy responses.

Because the parameters (bandwidth for kernel matrices and regularization term γ of weighted LS-SVM) had to be optimized, all possible combinations of these parameter were investigated with a LOO-CV. Since these parameters optimization strategy is time consuming, one can further investigate a parameter optimization criterion for kernel GEVD and weighted LS-SVM.

The applications of the proposed method are not limited to, clinical and expression data sets. Possible additional applications of weighted LS-SVM include integration of genomic information collected from different sources and biological processes. In short, the proposed machine learning approach is a promising mathematical framework in both data fusion and non-linear classification problems.

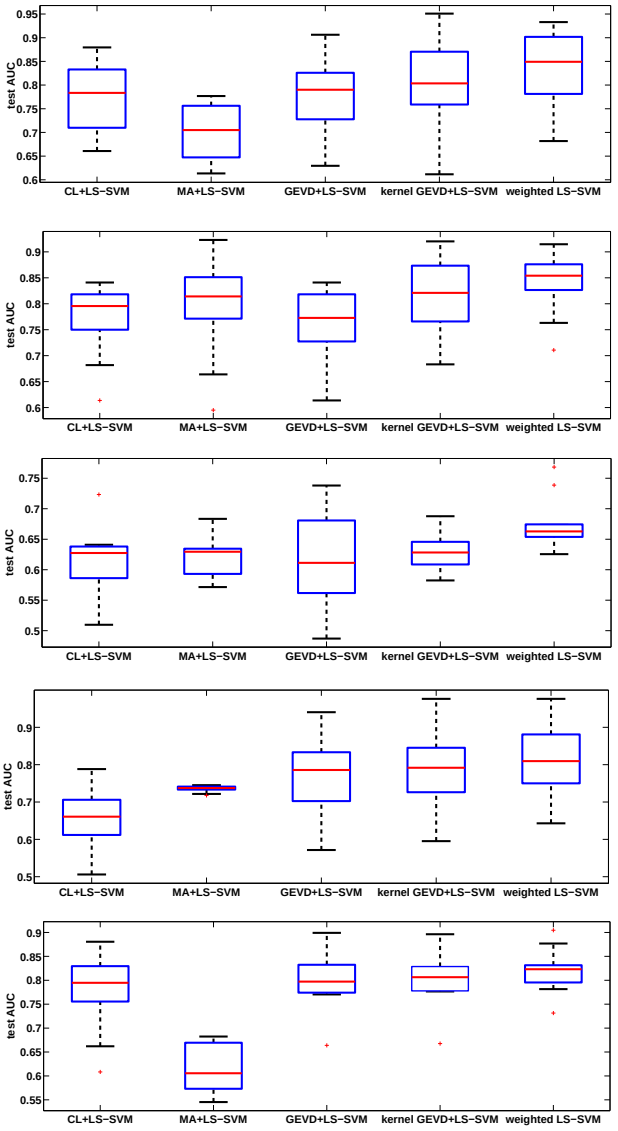


Figure 6.2: Boxplots depict the summary of the classification performances (averaged test AUC on ROC curve in 100 repetitions) of 5 breast cancer cases. CL and MA are the clinical and microarray kernels of RBF kernel functions. CL+LS-SVM and MA + LS-SVM indicates that LS-SVM classifier applied on clinical and microarray kernels. We have used GEVD and kernel GEVD as preprocessing step and then LS-SVM classifier applied on the transformed data set. Finally weighted LS-SVM classifier used as a single mathematical framework for the integration of clinical and microarray kernels and further classification (a) Case I (b) Case II (c) Case III (d) Case IV (e) Case V.

Chapter 7

A novel chemoinformatics method for identification of biofilm inhibitors

Background: The most common representation of compounds in chemoinformatics is a connection-table, i.e., a table enumerating the atoms and another enumerating the bonds. In this study, we intend to derive a new chemical descriptor from this table, allowing a better distinction between biologically active and inactive compounds. Our method was applied to the identification of inhibitors of *Salmonella*, *Pseudomonas* biofilm formation and inhibitors of thrombin, trypsin, and factor Xa. Development of this type of anti-microbial is urgently needed as biofilms, surface-associated bacterial communities embedded in a self-produced polymeric matrix, provide strong protection against the activity of antibiotics, disinfectants and the immune system.

Method: We propose a new machine learning approach for the identification of biologically active compounds. Principal Component Analysis (PCA) converts the connection-table of each compound into a structural descriptor of two vectors: one corresponding to atoms and the other to bonds. As a supervised classification algorithm, a weighted least squares support vector machine (LS-SVM) is used in which the table enumerating the atoms is weighted against the table enumerating the bonds. We apply this framework to, a given experimental data set of preventive activity of 308 2-aminoimidazole-based compounds against *Salmonella* and *Pseudomonas* biofilms and its 10 newly synthesized validation compounds, and 72 inhibitors of the benzamidine type with respect

to their binding affinities toward thrombin, trypsin, and factor Xa. Prediction performances are evaluated and compared with the LS-SVM classifier on other well-known chemical descriptors such as Extended-Connectivity Fingerprints, Path-Length Fingerprints, MACCS Keys and Burden Eigenvalues descriptors.

Result: To evaluate the performance of the proposed machine learning approach, initially we randomly split the given compounds into two groups: training (2/3rd of compounds) and test data sets. This split was performed with the relative proportion of active and inactive compounds in both training and test sets. The classification model was built on the training data and the prediction performance of the method was evaluated on the test data. In the trained model, the averaged performance measurements on 30 iterations of the test data were calculated. The proposed machine learning approach obtained the best averaged test accuracy with F-score compared to the LS-SVM classifier on the other discussed chemical descriptors. In addition, it is experimentally observed that the proposed approach is able to discriminate the few compounds with the highest activity of the 10 validation compounds of *Salmonella* and *Pseudomonas*.

Conclusion: The results suggest that the newly proposed approach, the weighted chemical descriptors of molecular structure, identified accurately the inhibitors on *Salmonella* and *Pseudomonas* biofilms formation, than other discussed descriptors. The proposed machine learning approach could be applicable to any problem in which the property or activity of interest is dependent upon the molecular structure.

7.1 Introduction

Computational methods involving machine learning techniques are among the most popular tools used in chemoinformatics tasks [62, 117, 150]. Most of these techniques in chemoinformatics are used for the classification of chemical compounds based on a descriptor representation of the compounds. Historically, the most widely used descriptors have been based on fingerprints, such as the Extended-Connectivity Descriptor in which an iterative process assigns numeric identifiers to each atom, independent of the original numbering of the atoms [142]. Path-Length Fingerprints are generated with respect to path lengths of chemical compounds [178]. MACCS Keys [10] are the sets of descriptors based on structural fragments, that have been identified a priori by domain experts [51]. Burden eigenvalue descriptors (BCUT) rely on an eigenvalue factorization of molecular connection table [21].

Biofilms have been found to be involved in a wide variety of microbial infections in the body and it was observed that 80% of all human bacterial infections are biofilm-associated [44, 73]. Biofilms of *Pseudomonas aeruginosa* can cause chronic infections, which are a serious problem for medical care in industrialized societies, especially for immunocompromised patients and the elderly [143]. They often cannot be treated effectively with traditional antibiotic therapy. Biofilm formation is also an important survival strategy of *Salmonella* Typhimurium, both within and outside the host [163]. Given the extent of problems caused by biofilms, there has been a significant effort to develop new anti-biofilm strategies [16, 104]. Traditionally biologists identify biofilm inhibitors by wet-laboratory screening of thousands of compounds. This strategy is time consuming and expensive. Chemoinformatics approaches employ a range of methods which are primarily used for the rapid evaluation and prioritization of compounds prior to wet-laboratory testing. Although molecular docking, a well established computational technique, was applied for the identification of potential histidine kinase inhibitors to combat *Staphylococcus epidermidis* [135] and for the selection of potential biofilm inhibitors of *Pseudomonas aeruginosa* [201], chemoinformatics approaches have not been fully exploited yet in the area of bacterial biofilm inhibition.

The main purpose of this work is to introduce a machine learning approach, a weighted chemical descriptor based on the connection-table of compounds, for designing a prediction model for identification of the activity of chemical compounds in a specific biological condition. This method was applied for the identification of inhibitors of *Salmonella* and *Pseudomonas* biofilm formation and inhibitors of Thrombin, Trypsin and Factor Xa. Initially, molecular structures were converted into structural descriptors in terms of two vector spaces: atoms and bonds by Principal Component Analysis (PCA), which transforms the two dimensional connection-table into a single dimensional space. Next, a weighted least squares support vector machine (weighted LS-SVM) classifier [175] was applied on these two vector spaces to integrate them into a single vector, named as a weighted chemical descriptor (which is completely representing the connection-table) and further predict the biological activity of the chemical compounds. Finally, we compared the prediction performance of the weighted LS-SVM classifier on the newly proposed descriptor with the LS-SVM classifier on other well known descriptors such as Extended-Connectivity Fingerprints, Path-Length Fingerprints, MACCS Keys and BCUT descriptors.

7.2 Data Sets

In our analysis, we have a connection-table of 308 2-aminoimidazole-based compounds [162, 164, 54, 161, 160] that have the capability to prevent biofilm formation of *S.* Typhimurium or *P. aeruginosa*, for model development and an independent set of 10 novel compounds for validation. In addition, we have IC50 values of these 308 compounds. All compounds were synthesized according to the previously reported protocols in [162, 164, 54, 161, 160]. The IC50 value is defined as the concentration at which biofilm formation is inhibited by 50% and thus predicts the effectiveness of these compounds against biofilm formation. The lower the IC50, the more effective a compound is against the biofilm formation. Later we used Bohm Serin Protease Inhibitor Data Set to identify the inhibitors of Thrombin, Trypsin and Factor Xa [17]. As an example, connection-table of chemical compound *Melatonin* is given in Figure 7.1.

7.3 Methods

The methods used for this study can be subdivided into two categories: feature selection from the connection-table (table enumerating atoms and table enumerating bonds) using PCA and prediction with either weighted LS-SVM classifier or LS-SVM classifier.

7.3.1 Principal Component Analysis (PCA)

PCA was applied to convert the connection-table of each compound in the data sets, i.e. inhibitors of *S.* Typhimurium and *P. aeruginosa* biofilm formation, into a novel chemical descriptor of two vectors: one corresponding to atoms and the other to bonds.

PCA [128] is mathematically defined as an orthogonal transformation that converts the data into a new coordinate system, such that the largest variance by any projection of the data lies on the first coordinate (first principal component), the second largest variance on the second coordinate, and so on. The full principal components decomposition of $m \times n$ matrix A can therefore be given as follows:

$$Score = A * Coef$$

where each column of the $n \times n$ matrix $Coef$ are the eigenvectors of $A^T A$ and $Score$, $m \times n$ matrix, is the representation of A in the principal component space.

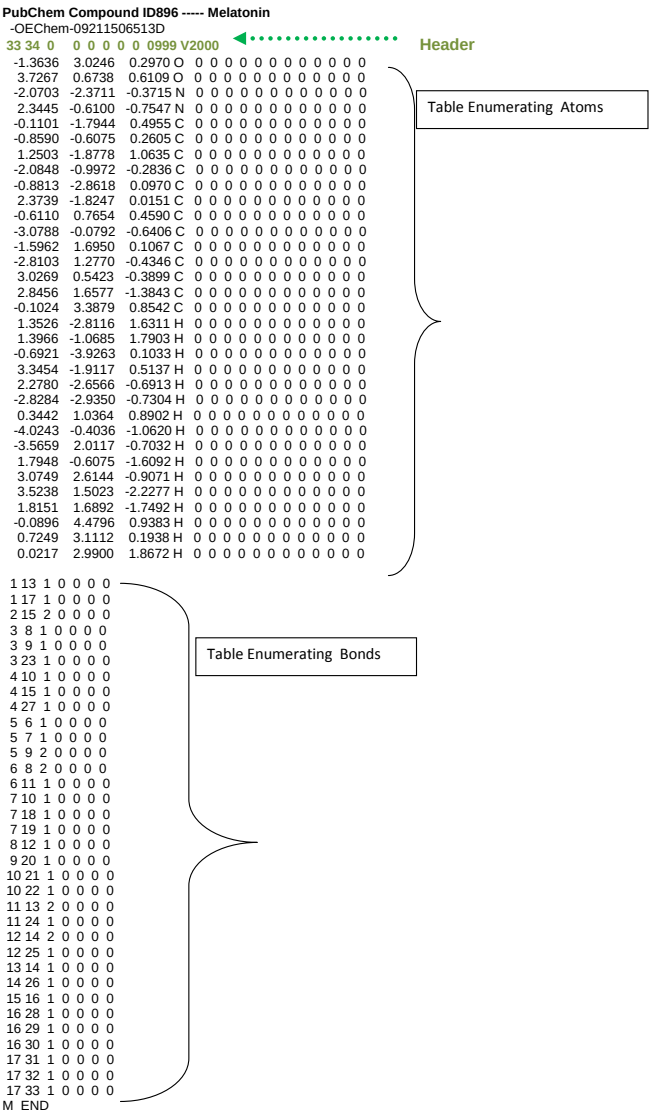


Figure 7.1: Connection Table of Chemical Compound Melatonin with pubChem ID: 896.

7.3.2 Weighted Least Squares-Support Vector Machine (Weighted LS-SVM) Classifier

Initially PCA transforms the atom and bond table of each compound into two descriptor vectors, one in atom space and the other in bond space. These two vectors should be considered together for the complete representation of a chemical compound. These descriptors are effectively integrated back into a single space by using the weighted LS-SVM classifier proposed in [175] and performs further prediction.

Given a training data set of N points $\mathcal{D} = \{x_i^{(1)}, x_i^{(2)}, y_i\}_{i=1}^N$ with output data $y_i \in \mathbb{R}$ and input data sets $x_i^{(1)} \in \mathbb{R}^m$, $x_i^{(2)} \in \mathbb{R}^p$. Consider the feature map $\varphi^{(1)}(.) : \mathbb{R}^m \to \mathbb{R}^{n_1}$ and $\varphi^{(2)}(.) : \mathbb{R}^p \to \mathbb{R}^{n_2}$ to high dimensional feature space, which is possibly infinite dimensional. The centered feature matrices $\Phi_c^{(1)} \in \mathbb{R}^{n_1 \times N}$, $\Phi_c^{(2)} \in \mathbb{R}^{n_2 \times N}$ become

$$\Phi_c^{(1)} = [\varphi^{(1)}(x_1^{(1)})^T - \hat{\mu}_{(\varphi_1)}^T; \ldots; \varphi^{(1)}(x_N^{(1)})^T - \hat{\mu}_{(\varphi_1)}^T]^T$$

$$\Phi_c^{(2)} = [\varphi^{(2)}(x_1^{(2)})^T - \hat{\mu}_{(\varphi_2)}^T; \ldots; \varphi^{(2)}(x_N^{(2)})^T - \hat{\mu}_{(\varphi_2)}^T]^T,$$

where $\hat{\mu}_{\varphi l} = \frac{1}{N}\Sigma_{i=1}^N \varphi^{(l)}(x_i^{(l)})$, $l = 1, 2$.

The weighted LS-SVM classifier proposed in [175] optimizes the following problem

$$\boxed{\text{P}}: \quad \min_{v,e,b} \; J(v,e) \; = \; \gamma \tfrac{1}{2} e^T (\Phi_c^{(2)^T} \Phi_c^{(2)})^{-1} e + \tfrac{1}{2} v^T v$$

$$\textit{such that} \quad y = \Phi_c^{(1)^T} v + b 1_N + e,$$

with $e = [e_1, \ldots, e_N]^T$ a vector of variables to tolerate misclassifications, weight vector v in primal weight space, bias term b and regularization term $\gamma \in \mathbb{R}^+$. Solution to the above problem in dual space,

$$\begin{bmatrix} 0 & 1_N^T \\ 1_N & \Omega_c^{(1)} + \gamma^{-1}\Omega_c^{(2)} \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{7.1}$$

with $y = [y_1, ..., y_N]^T$, $1_N = [1, ..., 1]^T$, $\alpha = [\alpha_1, ..., \alpha_N]^T$, $\Omega_c^{(1)} = \Phi_c^{(1)^T} \Phi_c^{(1)}$ and $\Omega_c^{(2)} = \Phi_c^{(2)^T} \Phi_c^{(2)}$.

The resulting classifier in the dual space is given by

$$y(x) = \text{sign}[\sum_{i=1}^{N} \alpha_i([K^{(1)}(x, x_i) + \frac{1}{\gamma}K^{(2)}(x, x_i)] + b)]. \quad (7.2)$$

with α_i the Lagrange multipliers, $K^{(1)}(x, z) = \varphi^{(1)}(x)^T\varphi^{(1)}(z)$, $K^{(2)}(x, z) = \varphi^{(2)}(x)^T\varphi^{(2)}(z)$ and $y(x)$ is the output corresponding to validation point x. Throughout the Chapter we use radial basis function (RBF) kernel: $K(x, x_i) = \exp(-\frac{||x-x_i||^2}{2\sigma^2})$ (RBF kernel with bandwidth σ).

Here the advantages of a weighted LS-SVM classifier are thus explored by designing two kernel matrices: a measure of similarity between two compounds based on their atoms and bonds description. In the primal space of the weighted LS-SVM framework, each atom description is weighted with the bond description and the solution in dual space, provides a way to integrate these two kernel matrices efficiently (see Equation 7.2). The resultant chemical descriptors are completely representing the connection-table of chemical compounds.

The parameter selection for the weighted LS-SVM classifier is illustrated on the Algorithm box. For several possible values of the kernel parameters σ_1 and σ_2, the leave-one-out cross validation (LOO-CV) performance is computed for each possible combination of γ . The optimal parameters are the combinations (σ_1, σ_2, γ) with best LOO-CV performance. Note the complexity of this optimization procedure because both the kernel parameters (σ_1 and σ_2) and γ need to be optimized in the sense of the LOO-CV performance.

Box: Algorithm: weighted LS-SVM classifier for prediction of chemical compounds

1. Given a training data set of N points $\mathcal{D} = \{x_i^{(1)}, x_i^{(2)}, y_i\}_{i=1}^N$ with output data $y_i \in \mathbb{R}$ and input data sets $x_i^{(1)} \in \mathbb{R}^m$, $x_i^{(2)} \in \mathbb{R}^p$.

2. Find optimal γ and σ_1, σ_2 bandwidths of kernel functions $K^{(1)}$, $K^{(2)}$ by solving linear system Equation 7.1 using leave-one-out cross validation (LOO-CV).

3. Predict the performance of the model by plugging the optimal parameters from previous steps into Equation 7.2.

7.4 Result

The proposed method defines new chemical descriptor of compounds from their connection-table, a representation of chemical molecules. An example of the connection-table, a portion of a structure-data file (SDF file) is given in Figure 7.2:

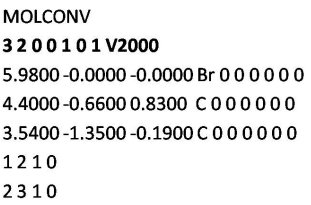


Figure 7.2: An Example of a Connection-Table

The first row of the table is the header line. The remaining rows are actual part of the connection-table in which the second row indicates: 3 atoms, 2 bonds, ..., V2000 standard. Rows 3-5 represent the atom block (1 line for each atom) and rows 6-7 represent the bond block (1 line for each bond). The columns of the atoms block specify the X,Y,Z-coordinates, atom symbol, isotope, charge, stereo code and the columns of the bond block specify the row numbers of atoms, and codes for bond type, bond stereochemistry etc. The table contains non-numerical data, i.e. atom symbol, which is converted to a numerical representation as follows while the other columns remain the same:

```
C-0  N-1  H-2  F-3  Br-4  O-5  S-6  CL-7  Si-8  K-9 ...
```

The molecule in the given example, has three atoms (3 rows with 10 columns) and two bonds (2 rows with 4 columns). A chemical compound can be denoted with a variety of different connection tables describing one and the same compound but with different numbering of atoms. Canonicalization is the task which taking one of the numberings as the standard one and derive a unique code from it. It is accomplished by numbering the atoms of a molecule so that it is represented by only one connection table. Morgan algorithm is used for deriving a canonical code for the molecule [60, 123].

Table 7.1: Table enumerating atoms: data source I

5.9800	-0.0000	-0.0000	Br	0	0	0	0	0	0
4.4000	-0.6600	0.8300	C	0	0	0	0	0	0
3.5400	-1.3500	-0.1900	C	0	0	0	0	0	0

Table 7.2: Table enumerating bonds: data source II

1	2	1	0
2	3	1	0

We form two tables (referred to as Table 7.1 and Table 7.2) corresponding to the atoms and the bonds description of the chemical compounds, respectively. The size of the atom table will be number of atoms $\times$ 10 and bond table be number of bond $\times$ 4. Our aim is to generate two vectors one representing the atom and other to bond table. These vectors should able to capture as much information as possible from the original tables. Hence, PCA is applied to each of these tables individually to define chemical descriptors based on both atom and bond tables. The first component which captures the majority of the variance of the data (for each compound, the first component explaining minimally 70% variance) is selected for each table. After transforming the data onto direction of the first component, we obtain two vectors, one corresponding to a table enumerating the atoms and another to a table enumerating bonds for each compound. Thus, we defined two chemical descriptors for each compound. Based on this, we construct two tables: compounds vs. atoms and compounds vs. bonds. The number of atoms and bonds are different for each compound. The entries in the descriptor vectors are filled with zeros, if the compound does not have an atom or bond corresponding to an entry. An overview of the chemical descriptor formation from the connection-table of compounds is given in Figure 7.3 and Box Algorithm II.

Box: Algorithm II: Chemical descriptor and weighted LS-SVM classifier for prediction of chemical compounds

1. Split the connection table of each compound into atom and bond table as shown in Figures 7.1 and 7.3.

2. To generate chemical descriptors based on the atom and bond properties of chemical compounds, we applied PCA on these two tables individually and obtain the first eigenvector explaining minimally 70% variance of the data. Then projecting these tables onto the direction of the corresponding

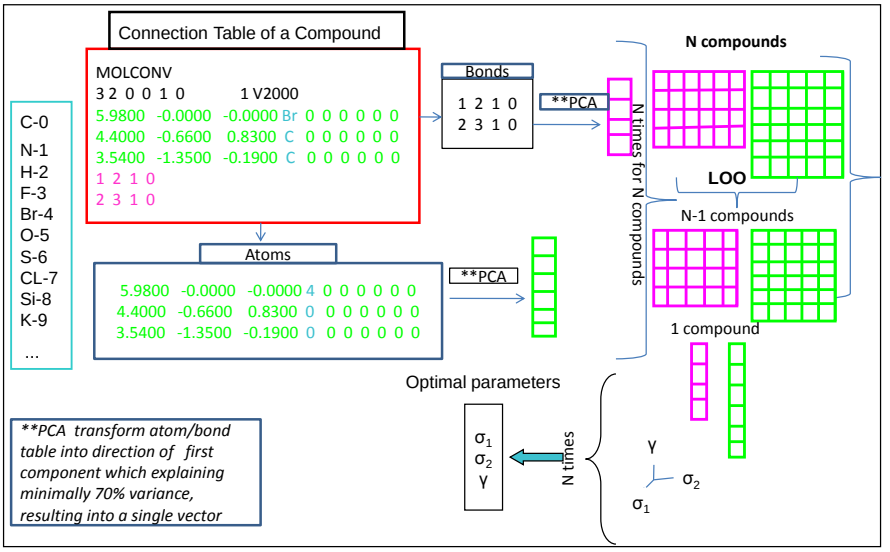


Figure 7.3: An overview of chemical descriptor formation from the connection-table of compounds. PCA is applied to the connection-table of each compounds to define a new structural descriptor in terms of two vectors. This results into two matrices: atoms vs. compounds and bonds vs. compounds. The weighted LS-SVM framework integrates these two vectors into a single vector named as weighted chemical descriptor and performs further prediction. LOO-CV is applied to select the optimal parameters.

eigenvector resulting in two descriptor vectors, one corresponding to atom and other to bond.

3. Repeat the previous steps for each compound. Finally we get two tables, one corresponding to atom and other to bonds. For 308 compounds, we generate 308 $\times$ atoms matrix for atom and 308 $\times$ bonds matrix for bond properties.

4. Connection table is the structural representation of the compounds. For the complete representation of chemical compounds, we need to integrate these two descriptors properly. We used the weighted LS-SVM classifier for data integration and further classification.

5. Find optimal γ and σ_1, σ_2 bandwidths of kernel functions $K^{(1)}$, $K^{(2)}$ by solving linear system Equation 7.1 using leave-one-out cross validation (LOO-CV) on training data sets.

6. Predict the performance of the model by plugging the optimal parameters from previous steps into Equation 7.2 on validation set.

In the weighted LS-SVM classifier, the RBF kernel function provides a measure of similarity between two compounds based on their atoms and bonds description, resulting in two kernel matrices. In the primal space of the weighted LS-SVM framework, each atom description is weighted with the bond description and the solution in dual space, provides a way to integrate these two kernel matrices efficiently (see Equation 7.2). The weighting scheme in the weighted LS-SVM classifier with RBF kernel function provides a way to represent chemical descriptors and perform further classification. In order to obtain the optimized values on Equation 7.2 for σ_1 and σ_2, LOO-CV performance is computed with each possible combination of γ, with values ranging from 10^{-3} to 10^3, with grid size 10. The combinations of the parameters which offer the highest LOO-CV performance is selected as the optimal parameters.

To evaluate the performance of the proposed machine learning approach, initially we randomly split the given compounds into two groups: training (2/3rd of compounds) and test data sets. This split was performed with relative proportion of active and inactive compounds in both training and test sets. The classification model was built on the training data and the prediction performance of the method was evaluated on the test data. In the case of *Pseudomonas*, we have eliminated 78 compounds for which the corresponding IC50 values were not available. Compounds were categorized into two sets: active or less (not) active based on the IC50 values. Selecting a cut-off value of IC50 to classify a compound as active or inactive is arbitrary. In the trained model of *Salmonella* and *Pseudomonas*, the averaged performance measurements on 30 iterations of the test data were calculated for a range of IC50 values (e.g. IC50 = 10 μM,..., IC50 = 100 μM). The prediction performance of the model was represented in terms of test areas under the receiver operating characteristic curves (AUC), accuracy and F-score. The F-score indicates the balance between precision and recall. The optimal cut-off value, below which a compound was considered active, was calculated based on the best prediction performances. Models with the optimal cut-off value could offer best performance for the future predictions. We have determined the optimal cut-off value of IC50 as 10 μM for both the *Salmonella* and *Pseudomonas* and 5 μM for Thrombin, Trypsin and Factor Xa case studies. At optimal cut-off value, the *Salmonella* data set contains 47 active and 159 inactive compounds as training set, and 23 active and 79 inactive compounds as test set. While the *Pseudomonas* data set contains 32 active and 122 inactive compounds as training set and 15 active and 61 inactive compounds as test set.

Table 7.3 and Figures 7.4, 7.5 compare the prediction performances (with optimal

Table 7.3: Comparison of averaged classification performances of different descriptors: proposed descriptor, MACCS keys, ECF, Path keys and BCUT descriptors to identify the active and inactive compounds in *Salmonella* and *Pseudomonas* biofilm. On both case studies, the proposed descriptor outperformed other descriptors in terms of averaged accuracy and F-score.

	Proposed Method	MACCS Keys	ECF	Path keys	BCUT
Salmonella					
Accuracy(std)	**0.8071(0.0581)**	0.7706(0.0500)	0.7686(0.0438)	0.7735(0.0454)	0.766(0.04)
p-value		0.0100	0.0100	0.0005	0.012
Test AUC(std)	0.6453(0.0409)	**0.6988(0.0294)**	0.6525(0.0263)	0.6425(0.0328)	0.665 (0.077)
p-value		6.73E-05	0.3528	0.2800	0.031
F-score(std)	**0.3058(0.0375)**	0.0212(0.0333)	0.0315(0.0241)	0.036(0.0142)	0.484(0.0823)
p-value		1.29E-16	5.52E-17	1.00E-17	0.064
Pseudomonas					
Accuracy(std)	**0.8179(0.0492)**	0.7908(0.0441)	0.7829(0.0360)	0.7882(0.0313)	0.65(0.055)
p-value		0.0001	0.0019	0.0004	0.012
Test AUC(std)	0.6549(0.0267)	0.7041(0.0258)	0.6559(0.0397)	**0.7118(0.0395))**	0.588(0.066)
p-value		5.06E-06	0.9104	3.30E-05	0.212
F-score(std)	**0.3277(0.0346)**	0.0211(0.0330)	0.1401(0.0854)	0.0874(0.0469)	0.743(0.059)
p-value		2.34E-17	1.38E-07	8.71E-14	0.024

cut-off) of the proposed weighted LS-SVM descriptor, with LS-SVM on 2D chemical descriptors; Extended-Connectivity Fingerprints, MACCS keys, Path-Length Fingerprints and descriptor based on connection table; BCUT. Table 7.4 illustrates the averaged prediction performances of different descriptors on Thrombin, Trypsin and Factor Xa data sets. Each traditional chemical descriptor is a complete representation of a compound, thus for final prediction, LS-SVM classifier is applied directly. While the PCA based descriptor spaces should be considered together to obtain a complete representation of chemical compounds, which necessitates the use of the weighted LS-SVM classifier to integrate these two data sets and further perform predictions. Statistical significance tests were performed in order to allow a correct interpretation of the results. A non-parametric paired test, the signed rank test has been used in order to make general conclusions. Two results are significantly different if the value of the

Table 7.4: Comparison of averaged prediction performances of different descriptors - *Thrombin*, *Trypsin* and *FactorXa*.

	BCUT	ECF	MACCS Keys	Path keys	Proposed Method
Thrombin					
Test AUC (std)	0.625(0.101)	0.673(0.012)	0.643(0.013)	0.5(0.0)	0.666(0.133)
Test Accuracy (std)	0.855(0.058)	0.784(0.21)	0.801(0.124)	0.610(0.070)	0.822(0.083)
F-Score(std)	NIL	NIL	NIL	NIL	0.600(0.050)
Trypsin					
Test AUC (std)	0.53(0.085)	0.625(0.091)	0.695(0.0)	0.500(0.0)	0.653(0.125)
Test Accuracy (std)	0.817(0.027)	0.764(0.042)	0.768(0.0)	0.166(0.125)	0.777(0.058)
F-Score (std)	0.2(0.03)	NIL	NIL	NIL	0.58(0.021)
FactorXa					
Test AUC (std)	0.594(0.078)	0.493(0.015)	0.571(0.022)	0.654(0.033)	0.582(0.099)
Test Accuracy(std)	0.727(0.0924)	0.766(0.023)	0.733(0.035)	0.694(0.047)	0.790(0.078)
F-Score (std)	0.835(0.070)	NIL	NIL	NIL	0.802(0.053)

signed rank test (p-value) applied to both of them is lower than 0.05. In both case studies, the proposed novel machine learning approach performed well with test accuracy and F-score, while the best test AUC returned by MACCS keys for *Salmonella* and Path keys for *Pseudomonas*. In Tables 7.3 and 7.4, BCUT descriptor performed better than proposed method in terms of F-score and test accuracy, respectively. The results show the proposed chemical descriptor could represent the complete chemical structure of the compounds very well in a non-linear mathematical framework hence, there is a significant improvement in the performances for all case studies.

Finally, we used the designed model to predict the activity of 10 novel synthesized compounds (see Table 7.5) against biofilms. These compounds were synthesized and *in vitro* tested for preventive activity against *Salmonella* and *Pseudomonas* biofilm formation according to previously published procedures [162].

First, we trained the model based on 308 compounds for different cut-off values of the IC50 (e.g. IC50 = 10 μM, IC50 = 20μM, IC50 = 30 μM, IC50 = 50 μM and IC50 = 100 μM) to define activity and non-activity of the compounds. Then we used these trained models to predict the activity of the novel compounds based on different cut-off values. Table 7.8 lists the compounds identified by each descriptor for different IC50 cut-off values. Finally, with the *in vitro* measured IC50 values of the 10 experimentally synthesized compounds, the prediction performances of all the discussed descriptors were obtained and represented in Tables 7.7 and 7.8 in terms of test AUC, accuracy and F-score. The results

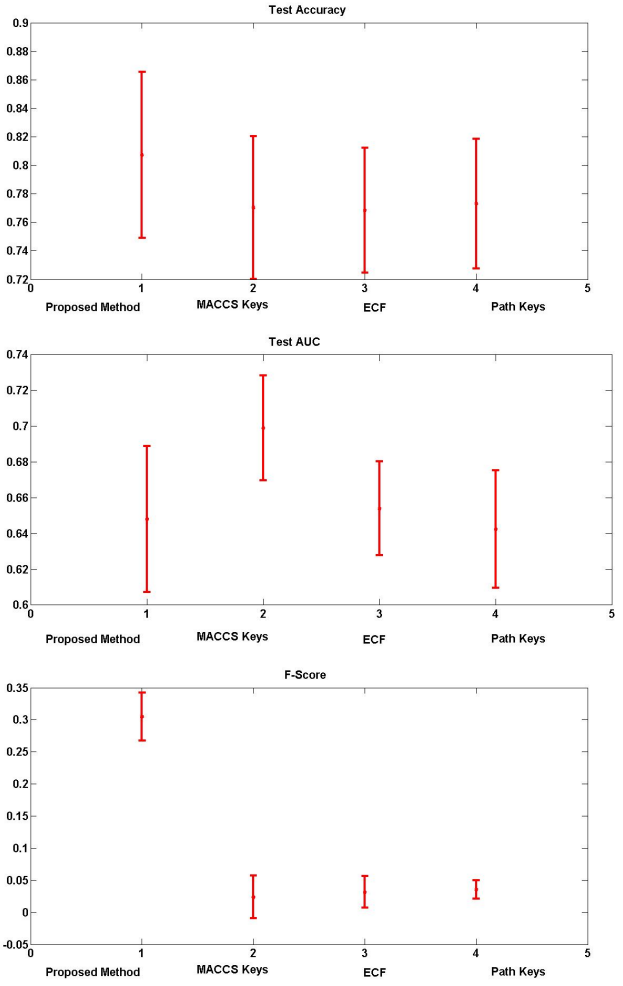


Figure 7.4: *Salmonella*: Error Bar represents the averaged classification performances in terms of Accuracy, test AUC and F-score over 30 iterations of different descriptors: proposed descriptor, MACCS keys, ECF and Path keys to identify the active and inactive compounds in *Salmonella* biofilm. Proposed descriptor outperformed in terms of test AUC and F-score and MACCS keys outperformed in terms of test AUC. In the error bar x-axis denotes the descriptors and y-axis denotes the test AUC/test accuracy/F-score.

illustrated the ability of the proposed descriptor to identify compounds with

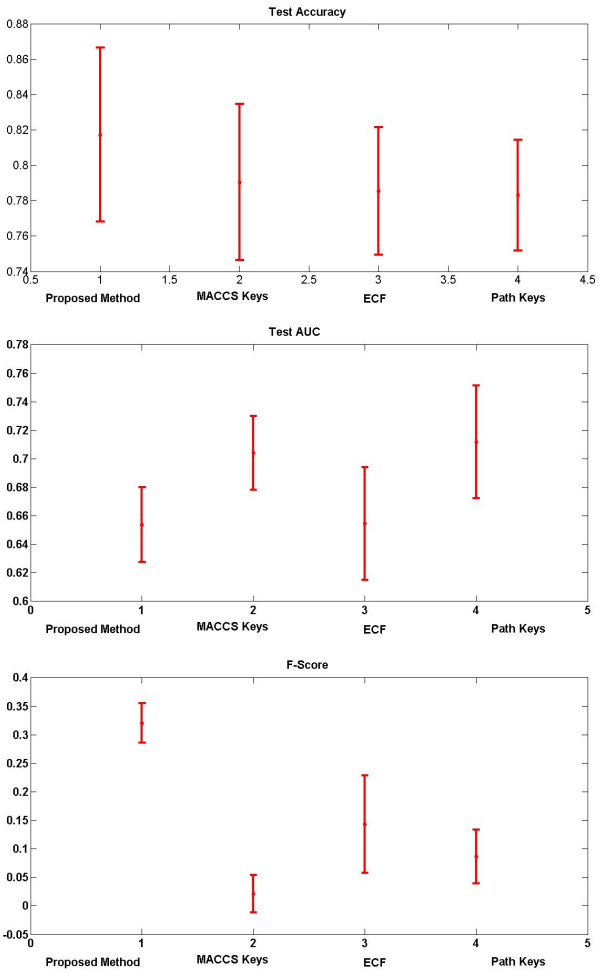


Figure 7.5: *Pseudomonas*: Error Bar represents the averaged classification performances in terms of Accuracy, test AUC and F-score over 30 iterations of different descriptors: proposed descriptor, MACCS keys, ECF and Path keys to identify the active and inactive compounds in *Pseudomonas* biofilm. Proposed descriptor outperformed in terms of test AUC and F-score and Path keys outperformed in terms of test AUC. In the error bar x-axis denotes the descriptors and y-axis denotes the test AUC/test accuracy/F-score.

very high activity against the biofilms.

Table 7.5: Compound Id, Structure and IC50 for biofilm prevention of 10 novel compounds used for validation

N^o	Structure	*Salmonella* IC50 in μM	*Pseudomonas* IC50 in μM
1		69.12(50.43 to 94.74)	211.90(142.2 to 315.9)
2		42.86(19.02 to 96.56)	60.21(27.25 to 133.0)
3		44.46(22.41 to 88.20)	greater than 400
4		115.1(65.28 to 203.0)	358.20(223.5 to 574.1)
5		48.62(36.67 to 64.45)	54.53(24.69 to 120.4)
6		53.79(21.01 to 137.7)	greater than 400
7		27.47(18.96 to 39.82)	74.74(43.41 to 128.7)
8		10.78(6.488 to 17.92)	13.92(7.851 to 24.69)
9		25.06(15.38 to 40.83)	28.70(16.04 to 51.37)
10		119.00(80.56 to 175.8)	15.92(2.630 to 96.32)

The values between brackets are the 95% confidence interval

7.5 Discussion

In recent years, the development of computational techniques that build models to accurately classify chemical compounds, have been an active area of research. Most of the best performing techniques in this area use chemical descriptors of compounds. Recently, connection-tables have become the most complete structural representation of chemical compounds [106]. In this Chapter we proposed a new chemical descriptor derived from the connection-table of compounds by PCA analysis, which transforms the two dimensional connection-table into a single dimensional subspace in terms of two vectors: one corresponding to atoms and the other to bonds for each compound. The

Table 7.6: The 10 validation compounds identified as active by each descriptor (proposed, MACCS, ECF and Path) with different cut-off values (10,20,30,40,50 and 100) for IC50 - *Salmonella* and *Pseudomonas*. The numbers in the bold represent the compound which are correctly identified as active.

SALMONELLA						
IC50 cut-off (μM)	10	20	30	40	50	100
Proposed Method	NIL	1,2,**9**,10	4,**7**,**9**,10	**3**,4,6,**7**,**9**,10	**3**,4,**6**,**7**,**9**,10	**3**,4,**5**,**6**,**7**,**8**,**9**
MACCS	NIL	NIL	NIL	**3**,**9**	**3**,**9**	**3**,**5**,**7**,**8**,**9**,10
ECF	NIL	NIL	**3**,**7**,**8**	**3**,**7**,**8**	**7**,**8**	**3**,**5**,**6**,**7**,**8**,10
Path	NIL	**9**	**9**	**9**	**9**	**1**,**3**,**5**,**7**,**8**,**9**
PSEUDOMONAS						
IC50 cut-off (μM)	10	20	30	40	50	100
Proposed Method	**8**,**10**	6,7,**8**	3,6,7,**8**	3,6,7,**8**	3,4,6,7,**8**,**9**	1,**2**,3,4,**5**,6,**7**,**8**,**9**
MACCS	NIL	**8**,9	**8**,**10**	3,**8**,**9**,**10**	3,8,9,10	1,**2**,3,4,**5**,**7**,**8**,**9**,**10**
ECF	NIL	**9**	NIL	3,**10**	3,7,**10**	1,**2**,3,4,**5**,6,**7**,**8**,**9**,**10**
Path	NIL	**9**	**9**	**8**,**9**	3,**9**	1,**2**,3,4,**5**,6,**7**,**8**,**9**,**10**

RBF kernel function provides a measure of similarity between two compounds based on their atoms and bonds description, resulting in two kernel matrices. In the primal space of the weighted LS-SVM framework, each atom description is weighted with the bond description and the solution in dual space provides a way to integrate these two kernel matrices efficiently.

In order to select the best prediction model, we considered three different performance measures: accuracy, F-score and test AUC. Generally an AUC score between 0.6 and 0.7 indicates that the model has sufficient ability to discriminate positive and negative classes [86]. All discussed descriptors obtained AUC scores ranging between 0.6 and 0.7, whereas MACCS keys outperformed in the *Salmonella* and Path Keys in the *Pseudomonas* case study. While considering the prediction performance in terms of accuracy, the proposed approach significantly outperformed all other descriptors. But for highly class imbalanced data set with more negative samples, a model can predict the value of the majority class for all predictions and achieve a high classification accuracy. In this context, additional measures such as F-score are required to evaluate a classifier. F-score is often used as an aggregated performance score for the evaluation of algorithms, which considers both precision and recall. The F-score will be equal to 1 if the algorithm has perfect precision and recall. F-score with values between 0 and 1 usually gives a reasonable rank ordering of different

Table 7.7: Comparison of prediction performances of different descriptors (proposed, MACCS, ECF and Path) to identify active compounds in *Salmonella* biofilm formation. The results are given in terms of test AUC, F-score and accuracy which illustrating the ability of the proposed descriptor to identify compounds with very high activity against the biofilms. The other descriptors missed the ability to identify active compounds at the lower IC50 cut-off values, but with increasing cut-off values, MACCS on *Pseudomonas* outperformed all the other descriptors.

	IC50	10	20	30	40	50	100
SALMONELLA							
Proposed Approach	AUC	0.500	0.722	0.691	0.548	0.548	0.625
	F-score	NIL	0.286	0.571	0.444	0.546	0.800
	accuracy	0.900	0.600	0.700	0.500	0.500	0.700
MACCS	AUC	0.500	0.500	0.500	0.700	0.643	0.625
	F-score	NIL	NIL	NIL	0.714	0.444	0.714
	accuracy	0.900	0.800	0.700	0.700	0.500	0.700
ECF	AUC	0.500	0.500	0.762	0.762	0.643	0.562
	F-score	NIL	NIL	0.667	0.667	0.444	0.769
	accuracy	0.900	0.800	0.800	0.800	0.500	0.600
Path	AUC	0.500	0.750	0.667	0.667	0.571	0.875
	F-score	NIL	0.200	0.500	0.500	0.250	0.857
	accuracy	0.900	0.900	0.800	0.600	0.400	0.800

NIL: If all 10 compounds are identified as inactive by the specific descriptor with corresponding IC50

Table 7.8: Comparison of prediction performances of different descriptors (proposed, MACCS, ECF and Path) to identify active compounds in *Pseudomonas* biofilm formation. The results are given in terms of test AUC, F-score and accuracy which illustrating the ability of the proposed descriptor to identify compounds with very high activity against the biofilms. The other descriptors missed the ability to identify active compounds at the lower IC50 cut-off values.

	IC50	10	20	30	40	50	100
PSEUDOMONAS							
Proposed Approach	AUC	1.000	0.625	0.4524	0.4524	0.4167	0.4167
	F-score	1.000	0.400	0.444	0.444	0.500	0.286
	accuracy	1.000	0.700	0.500	0.500	0.400	0.500
MACCS	AUC	0.500	0.687	0.833	0.929	0.792	0.625
	F-score	NIL	0.500	0.800	0.857	0.750	0.800
	accuracy	0.800	0.800	0.900	0.900	0.800	0.700
ECF	AUC	0.500	0.563	0.500	0.595	0.542	0.500
	F-score	NIL	0.363	NIL	0.400	0.545	0.750
	accuracy	0.800	0.700	0.700	0.700	0.500	0.600
Path	AUC	0.500	0.563	0.667	0.833	0.595	0.500
	F-score	NIL	0.364	0.500	0.800	0.400	0.571
	accuracy	0.800	0.700	0.800	0.900	0.600	0.600

NIL: If all 10 compounds are identified as inactive by the specific descriptor with corresponding IC50

algorithms/classifiers. Compared to other descriptors, the proposed weighted chemical descriptor significantly outperformed in terms of F-score in both case studies.

The results of the validation set illustrated the ability of the proposed descriptor to identify compounds with very high activity against the biofilms. It is noted that the other descriptors have missed the ability to identify active compounds at the lower IC50 cut-off values. However, with increasing cut-off values, MACCS on *Pseudomonas* outperformed all the other descriptors. The proposed model with validation data obtained the best performance at IC50 = 10 μM, similar to the test data. It is noted that for the lower IC50 cut-off values, the number of active compounds in the trained model was very small. Even with this small number of active compounds in the trained model, the weighted

chemical descriptor captured enough information to distinguish active and inactive compounds. This indicates that an increase in the ratio of positive to negative training instances greatly influences the performance of the classifiers.

To the best of our knowledge, this is the first time in chemoinformatics that a prediction model is developed on a connection-table based chemical descriptors and a weighted LS-SVM classifier. We admit that the time needed for building the predictive model for the weighted LS-SVM classifier was expensive in terms of parameter optimization. In general, the use of connection-table based descriptor in combination with the weighted LS-SVM approach could be applied successfully in the area of of chemoinformatics to identify very active compounds in a given biological condition.

7.6 Conclusion

It is a challenge for the researcher in chemoinformatics to extract knowledge from the chemical structure. In our analysis, PCA was used to decompose the connection-table of each compound effectively to a low dimensional representation, as such defining a new structural descriptor of chemical compounds. Then a weighted LS-SVM approach was used to design a weighted chemical descriptor and to predict the biological activity of chemical compounds. The results illustrate that the obtained descriptor offers an improved model to identify very active compounds in a specific biological condition. The proposed machine learning technique could be applied to any classification/prediction problem which is based on the molecular structure of compounds.

Chapter 8

Conclusion and Future Research

High dimensional and heterogeneous biological data sets always raises challenges in computational biology and chemistry. The aim of this thesis is therefore, to develop and use mathematical techniques for dimensionality reduction and data integration. Microarray data analysis offers the improved health care, but to get there, genomic data must be integrated with clinical information. A huge amount of important biomedical data is hidden in the bulk of research articles in the biomedical fields. The integration of information from the literature, gathered by text mining, with microarray data are essential for better biological understanding and validation of new biological hypothesis. We therefore, focused not only on the integration of clinical and microarray data sets to improve clinical decision support but also including the incorporation of literature information into microarray data analysis for better prognosis of diseases. In addition to the applications of these techniques in bioinformatics, we explained the importance of data integration in chemoinformatics with case studies, especially for identification of the activity of chemical compounds for a specific biological condition.

Presently, due to the availability of massive biomedical data on each individual, both health care and life science is becoming data-driven. The input features are structured/un-structured data with many challenges, including sparse-binary features, non-unique distributed structure, and high dimensional data, which reducing accurate clinical decision in clinical practice [53]. In recent decades, considerable effort has been made toward overcoming most of these challenges, but still there is an essential need for significant improvements in

this field for future personalized medicine. We suggest future work be conducted to investigate and use big data analytics and large-scale machine learning frameworks to tackle most of the challenges and provide high-quality health care with reduced cost.

8.1 Conclusion

Data integration plays an important role in combining clinical, and environmental data with high-throughput genomic data to identify functions of genes, proteins, and other aspects of the genome [74]. Integrating data from different sources is, therefore, an important part of current research in genomics and proteomics. The identification of disease-associated genes and classification of patients require not only an understanding of the genetic basis of the disease, but also the correlation of this data with knowledge normally processed in the clinical setting [6]. Although several research groups have already proposed GSVD [3] to obtain common information between two data sets, we considered MLGEVD/GEVD as a pre-processing step in which both clinical and microarray data were used together to obtain the generalized eigenvectors, a process common to both methods. We integrated two data sources with a common process in a maximum likelihood framework [174] to obtain the generalized eigenvectors in which one data source acts as the prior information. This data framework will be useful if the availability of one type of biological information is limited and considered them as the prior information in the model development. Microarray data, which was difficult and expensive to collect were incorporated as prior information into clinical decision-making, improving the classification performance and offering better diagnosis and prognosis. Incorporation of literature information into microarray analysis improved the possibility for obtaining stable disease associated genes. In cancer studies, generally combining these types of heterogeneous data improved the performance of decision support. In the near future, one can investigate the applicability of MLGEVD to more than two matrices and interpret these matrix results in a Bayesian context. If the prior data set is very large with many features, overfitting may occur in the model development which reducing the performance of the model on unseen/validation data. The linear projections based on GEVD will not perform very well in the model development, if the primary data source contains only limited number of features. To tackle these problems either we have to perform the projection based on kernel based GEVD or remove irrelevant features from the prior data using feature selection techniques.

Considering the statistically noisy nature of microarray data (much more variables than observations) and large collection of existing biological knowledge,

it is essential to exploit that knowledge for analysis and understanding of microarray data [125]. Text mining techniques constitute a promising technology for automating the incorporation of scientific knowledge in the microarray data mining process [125]. In recent years, there has been a growing interest in identifying/designing data integration and visualization techniques which simultaneously discover patterns occurring in multiple data sources. Incorporation of literature information into gene expression data analysis is an example of such a scenario, which is concerned with the analysis of the actual expression data in conjunction with existing textual information on genes, proteins, diseases, and so on.

The data integration approach in Chapter 3 is built on the SVD framework, but one major disadvantage of classical SVD is its brittleness with respect to grossly corrupted or outlying observations. Random errors are unavoidable in modern applications in imaging and bioinformatics, where some measurements may be arbitrarily corrupted or simply irrelevant to the structure we are trying to identify. We, therefore, developed an RPCA approach in Chapter 4 for integrating data resources and finding and exploiting low-dimensional structure in high-dimensional data where data sets now routinely lie in thousand- or even million-dimensional observation spaces. RPCA is a modification of the widely used statistical procedure PCA, which works well with respect to corrupted observations. Microarray data are commonly perceived as being extremely noisy because of many imperfections inherent in the current technology. Hence, we focused on RPCA to identify differentially expressed genes from highly corrupted microarray data set. Moreover, our studies show that RPCA on the weighted microarray data sets with literature information, identify disease-associated genes much accurately than RPCA on microarray data. The Augmented Lagrange multiplier based RPCA algorithms are simpler to analyze and easier to implement [110], compared to accelerated proximal gradient based methods. Moreover, they are also of much higher accuracy as the iterations are proven to converge to the exact solution of the problem. Incorporation of biological literature information into microarray data analysis improved the identification of stable genes associated with disease and offer better diagnosis and prognosis in cancer clinical decision making.

In Chapters 3 and 4, we discussed MLGEVD and RPCA, which are linear dimensionality reduction techniques, which fail to capture the nonlinear relationship in the data. Kernel methods provide a vectorial representation of any kind of data by mapping them into high dimensional feature space. Kernel based machine learning algorithms become an important research area and has wide applications in image, signal processing, and pattern recognition [108]. To discover the non-linear relationship in the data, kernel methods are widely used in dimensionality reduction techniques, namely KPCA, kernel discriminant

analysis (KDA) etc. Studies show that KPCA performs better than PCA in many applications [1]. But the performance of KPCA completely depends on the parameter selection of kernel functions. In addition, the unsupervised dimensionality reduction methods are most useful in the practical applications in which the labeled data are usually expensive to collect. Chapter 5 provides a solution to these two problems, that is, it offers a data driven bandwidth selection criterion for KPCA which is executing in an unsupervised mode.

Although several researchers have already proposed several non-linear data integration models, they are coupled with the selected classifiers. We proposed a kernel-based mathematical framework for data integration and classification: a weighted LS-SVM classifier. Compared with the existing approaches, the proposed approach will be a simple mathematical framework for kernel based data integration. This framework can be applied to any two complex data sources which have a common space and the final goal is to make a prediction or classification based on this common space. In this framework, the weight assigned to the second kernel matrix can be obtained from the data, using cross-validation approaches. In our integration framework, two data sets should not be completely redundant. This mathematical framework developed based on kernel GEVD approach have used the second matrix as the weighted matrix for the model development. The noisy/corrupted data source generally considered as the second matrix which adding additional information to the primary resource (first matrix) to perform better classification/prediction. This approach could be considered as a standard mathematical problem to produce better classification performance based on heterogeneous data integration.

Machine learning techniques have been widely used in drug discovery and development especially in the areas of chemoinformatics. In chemoinformatics, machine learning has been widely used in QSAR studies. In a generalized machine learning algorithm, modern QSAR is characterized by the use of chemical descriptors based on the structure of chemical compounds. Molecular descriptors are generally used for representing structural and physiochemical properties of compounds, ranging molecular weight to complex 2D and 3D descriptors. As the descriptors based on molecule's connection table, represent the complete representation of the structure of the compounds, usually the machine learning models based on these descriptors offers good prediction performances. In Chapter 7, we proposed a new chemical descriptor from the connection-table of compounds in terms of two vectors: one corresponding to the atoms and the other to the bonds of each compound. To the best of our knowledge, this is the first time in chemoinformatics a prediction model is developed on the connection-table based chemical descriptors and weighted LS-SVM classifier. This work further can be extended to integrate any two 2D or 3D descriptors and perform further prediction/classification. In addition to their

application in traditional virtual screening, this framework can be combined with structural information of chemical compounds in molecular docking to predict potential off-target properties of any given compound.

8.2 Future Research

Integrating data from different sources creates many challenges in bioinformatics [74]. In dealing with heterogeneous data, for example, one needs to convert data from different sources into a single view or common dimension. It is important to identify the methods which captures as much information as possible from each individual data sources while integrating them. Data from different sources might have different quality and informativity depending on their source such as text, image, sequences and the experimental conditions that generated the data. Data from different sources might also have different informativity even if their quality is good and reliable; thus one source of data might give us more information than the other in answering the biological question of interest [74]. Extensive research is, therefore, needed for developing quality and informativity scores for various genomic, genetic, and proteomic data. The data exploitation aspect of data integration requires more attention in the usage of prior knowledge, development of statistical methods and visualization tools to analyze heterogeneous data sets [69].

The term bigdata intuitively describes a situation present in many research fields: the amount of data generated by instruments is exploding, and in many cases doubling over short periods of time [69]. Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include different types such as structured/unstructured and streaming/batch, and different sizes from terabytes to zettabytes [138]. The real challenges in big data analytics are gathering, storing large collections of data and extracting useful information from these data set. Using advanced analytics techniques such as text analytics, machine learning, predictive analytics, data mining, and statistics, researchers and businesses can analyze previously unused data sources independently or together with their existing enterprise data to gain new insights resulting in significantly better and faster decisions. The traditional analytics techniques are not scalable enough to support big data platform, only few machine learning algorithms are available to process large data sets in a reasonable amount of time. The following points are the challenges with the techniques proposed for handling big data sets in this dissertation and to be tackled in future.

- The dimensionality reduction techniques such as PCA, SVD, EVD, GSVD and GEVD, fail on large data sets due to the lack of computing resources

such as memory space or, processing time. The big data technologies, process large quantities of data within tolerable elapsed time. Therefore in the future one can propose the implementation of these dimensionality reduction techniques for these platform and overcome the technical challenges with big data. One drawback of the RPCA approach is their batch-processing nature. If the data set grows or changes over time, the algorithm needs to run from scratch. This raises the question of how the existing models can be extended to include the scalable data sets.

- In addition to class discovery, microarray experiments often aim to identify individual genes that are differentially expressed under distinct conditions, such as between two or more phenotypes, cell lines, under different treatment types or diseased and healthy subjects [141]. As it is currently largely unclear how molecular variants and their interactions determine cancer pathogenesis and propensity, marker identification is valuable for improving understanding of the molecular mechanisms of cancers and for suggesting novel drug targets [189]. Kernel techniques help us to construct different nonlinear versions of any algorithm which can be expressed in terms of dot products, known as the kernel trick [147]. Thus, kernel algorithms avoid the explicit usage of the input variables in the statistical learning task. Kernel machines can be used to implement several learning algorithms but they usually act as a black-box with respect to the input variables[141]. Since biomarker signature discovery is an important area in cancer studies, classification based on KPCA could be extended further to uncover small sets of interpretable features/disease associated genes from microarray data sets. For large data sets, the computation time needed for matrix decomposition using the eigenvalue decomposition of the kernel matrix may be excessive. Advanced techniques are required to formulate the multivariate statistics, such as kernel PCA and kernel regression, in matrix form over big data platforms.

- Besides the LS-SVM classifier that was used in this dissertation to improve clinical decision support in cancer, the weighted LS-SVM classifier has been investigated for data integration as well. In the future, our kernel based integration method should be compared with other existing data integration framework. Moreover, as these frameworks are complementary, an ensemble approach can be applied. More accurate classifiers can be obtained by not only combining different data types but also an outcome of multiple classifiers. Furthermore the parameter optimization strategy, especially for the bandwidth for kernel matrices and regularization term can be improved. The work on novel chemoinformatics method for identification of biofilm inhibitors can be further extended to screen a set of compounds from PubChem, which are active in specific biological

conditions prior to wet-laboratory testing. In addition, the current method cannot address the questions such as the parameters which are important for the activity of compound and how to modify molecules to improve the activity in a specific biological condition etc. One can further extend this method to answer such questions which helps medicinal chemist for the discovery and development of new therapeutic agents.

Bibliography

[1] Ahmadinejad, M. M., and Sherly, E. A comparative study on pca and kpca methods for face recognition. *IJSR 5* (2016), 2589–2594.

[2] Alon, U., Barkai, N., A Notterman, D., Gish, K., Ybarra, S., Mack, D., and J Levine, A. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS 96* (1999), 6745–6750.

[3] Alter, O., Brown, P. O., and Botstein, D. Generalized singular value decomposition for comparative analysis of genomescale expression data sets of two different organisms. *PNAS 100* (2003), 3351–3356.

[4] Alzate, C., and Suykens, J. A. K. Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA. *IEEE Transactions on Pattern Analysis and Machine Intelligence 32*, 2 (2010), 335–347.

[5] Ambroise, C., and McLachlan, G. J. Selection bias in gene extraction on the basis of microarray gene-expression data. *PNAS 99* (2002), 6562–6566.

[6] Analyti, A., Kondylakis, H., Manakanatas, D., Kalaitzakis, M., Plexousakis, D., and Potamias, G. *Integrating Clinical and Genomic Information Through the PrognoChip Mediator*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006, pp. 250–261.

[7] Azar, Y., Fiat, A., Karlin, A. R., McSherry, F., and Saia, J. Spectral analysis of data. In *33rd Symposium on Theory of Computing (STOC)* (2001), p. 619–626.

[8] AzhaguSundari, B., and Thanamani, A. S. Feature selection based on fuzzy entropy. *IJETTCS 2*, 2 (2013), 30–34.

[9] Babu, M. M. An introduction to microarray data analysis. *Computational Genomics: Theory and Application. Edited by: Richard P.* (2004), 225–249.

[10] Barnard, J. M., and Downs, G. M. Chemical fragment generation and clustering software. *J Chem Inf Comput Sci 37* (1997), 141–142.

[11] Barutcuoglu, Z., Schapire, R. E., and Troyanskaya, O. G. Hierarchical multi-label prediction of gene function. *Bioinformatics 22*, 7 (2006), 830–836.

[12] Berger, N. A., Savvides, P., Koroukian, S. M., Kahana, E. F., Deimling, G. T., Rose, J. H., Bowman, K. F., and Miller, R. H. Cancer in the elderly. *Trans Am Clin Climatol Assoc.*, 117 (2006), 147–156.

[13] Bertsekas, D. P. *Constrained optimization and Lagrange multiplier methods.* Academic Press, New York, 1982.

[14] Bewick, V., Cheek, L., and Ball, J. Statistics review 13: receiver opearting characterstics curves. *Critical Care 6*, 8 (2004), 508–512.

[15] Bicciato, S., Luchini, A., and Di Bello, C. PCA disjoint models for multiclass cancer analysis using gene expression data. *Bioinformatics 19*, 5 (2003), 571–578.

[16] Bjarnsholt, T., Tolker-Nielsen, T., Hoiby, N., and Givskov, M. Interference of pseudomonas aeruginosa signalling and biofilm formation for infection control. *Expert Reviews in Molecular Medicine 12* (2010), e11.

[17] Bohm, M., Sturzebecher, J., Klebe, G., and Teschendorff, A. Three-dimensional quantitative structureactivity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor xa. *J. Med. Chem. 42* (1999), 458 – 477.

[18] Boulesteix, A. L., Porzelius, C., and Daumer, M. Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics 24* (2008), 1698–1706.

[19] Bowman, A. W. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika 71* (1984), 353–360.

[20] Bunin, B. A., Bajorath, J., Siesel, B., and Morales, G. *Chemoinformatics: Theory, Practice and Products.* Springer, 2007.

[21] Burden, F. R. Molecular identification number for substructures search. *J. Chem. Inf. Comput. Sci. 29* (1989), 225–227.

[22] Cabrera, T., Collado, A., Fernandez, M. A., Ferron, A., Sancho, J., Ruiz-Cabello, F., and Garrido, F. High frequency of altered hla class i phenotypes in invasive colorectal carcinomas. *Tissue Antigens*, 52 (1998), 114–123.

[23] Campillos, M., Kuhn, M., Gavin, A. C., Jensen, L. J., and Bork, P. Drug target identification using side-effect similarity. *Science 321*, 5886 (2008), 263–266.

[24] Candes, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis. *Journal of the ACM 58*, 11 (2011).

[25] Cao, D., Xiao, N., Xu, Q., and Chen, A. F. Rcpi: R/bioconductor package to generate various descriptors of proteins, compounds and their interactions. *Bioinformatics 31*, 2 (2015), 279–281.

[26] Cao, D., Yang, Y., Zhao, J., Yan, J., Liu, S., Hu, Q., Xu, Q., and Liang, Y. Computeraided prediction of toxicity with substructure pattern and random forest. *J Chemometr 26*, 1 (2012), 7–15.

[27] Chan, T. H. M., Chen, L., and Guan, X.-Y. Role of translationally controlled tumor protein in cancer progression. *Biochem Res Int. 7* (2012), 1–5.

[28] Chaussabel, D., and Sher, A. Mining microarray expression data by literature profiling. *Genome Biology*, 3 (2002).

[29] Chen, G. *Exploring Topologies of Genetic Networks*. PhD thesis, University of Illinois Chicago, Computer Science Department, 2008.

[30] Cheung, W. A., Ouellette, B. F., and Wasserman, W. W. Inferring novel gene-disease associations using medical subject heading over-representation profiles. *Genome Medicine*, 4 (2012), 75.

[31] Chin, K., De Vries, S., Fridlyand, J., Spellman, P. T., Roydasgupta, R., Kuo, W.-L., Lapuk, A., Neve, R. M., Qian, Z., Ryder, T., Chen, F., Feiler, H., Tokuyasu, T., Kingsley, C., Dairkee, S., Meng, Z., Chew, K., Pinkel, D., Jain, A., Ljung, B. M., Esserman, L., Albertson, D. G., Waldman, F. M., and Gray, J. W. Genomic and transcriptional aberrations linked to breast cancer pathophysiologies. *Cancer Cell. 10* (2006), 529–541.

[32] Chu, F., and Wang, L. Application of support vector machine to cancer classification with microarray data. *International Journal of Neural systems, World Scientific 5* (2005), 475–484.

[33] Conde, L., Mateos, A., Herrero, J., and J, D. Improved class prediction in dna microarray gene expression data by unsupervised reduction of the dimensionality followed by supervised learning with a perceptron. *Journal of VLSI Signal Processing 35* (2003), 245–253.

[34] Corver, W. E., Koopman, L. A., Mulder, A., Cornelisse, C. J., and Fleuren, G. J. Distinction between hla class i-positive and -negative cervical tumor subpopulations by multiparameter dna flow cytometry. *Cytometry*, 41 (2000), 73–80.

[35] Cruz, J. A., and Wishart, D. S. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform. 2* (2006), 59–78.

[36] D Wentzell, P., T Andrews, D., C Hamilton, D., Faber, K., and R Kowalski, B. Maximum likelihood principal component analysis. *J of Chemometrics 11* (1997), 339–366.

[37] Daemen, A., Gevaert, O., and De Moor, B. Integration of clinical and microarray data with kernel methods. *Conf Proc IEEE Eng Med Biol Soc.* (2007), 5411–5.

[38] Daemen, A., Gevaert, O., Ojeda, F., Debucquoy, A., Suykens, J. A. K., Sempoux, C., Machiels, J.-P., Haustermans, K., and De Moor, B. Kernel-based integration of genome-wide data for clinical decision support. *Genome Medicine I* (2009), 39.1–39.17.

[39] Daemen, A., Timmerman, D., Van den Bosch, T., Bottomley, C., Kirk, E., Van Holsbeke, C., Valentin, L., Bourne, T., and De Moor, B. Improved modeling of clinical data with kernel methods. *Artificial Intelligence in Medicine 54* (2012), 103–114.

[40] Dai, M., Wang, Pand Boyd, A. D., Kostov, G., Athey, B., Jones, E. G., Bunney, W. E., Myers, R. M., Speed, T. P., Akil, H., and Watson, S Jand Meng, F. Evolving gene/transcript definitions significantly alter the interpretation of genechip data. *Nucleic Acids Res. 33* (2005), e175.

[41] Dakshanamurthy, S., Issa, N. T., Assefnia, S., Seshasayee, A., Peters, O. J., Madhavan, S., Uren, A., Brown, M. L., and Byers, S. W. Predicting new indications for approved drugs using a proteochemometric method. *J Med Chem 55*, 15 (2012), 6832–6848.

[42] Damoulos, T., and Girolami, M. A. Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. *Bioinformatics 10* (2008), 1264–1270.

[43] Dash, M., Liu, H., and Yao, J. Dimensionality reduction of unsupervised data. *In Tools with Artificial Intelligence, Ninth IEEE International Conference on IEEE* (1997), 532–539.

[44] Davies, D. Understanding biofilm resistance to antibacterial agents. nature reviews. *Drug Discovery 2*, 2 (2003), 114–122.

[45] Dawson-Saunders, B., and Trapp, R. G. *Basic Clinical Biostatistics*. Prentice-Hall International Inc., London, 1994.

[46] De Brabanter, K., Karsmakers, P., Ojeda, F., Alzate, C., De Brabanter, J., Pelckmans, K., De Moor, B., Vandewalle, J., and Suykens, J. A. K. Ls-svmlab toolbox user's guide version 1.8.

[47] Deisboeck, T. S. Personalizing medicine: a systems biology perspective. *Mol Syst Biol. 5* (2009), 249.

[48] Dimitrov, I., Naneva, L., Doytchinova, I., and Bangov, I. Allergenfp: allergenicity prediction by descriptor fingerprints. *Bioinformatics 30*, 6 (2014), 846–851.

[49] Doms, A., and Schroeder, M. Gopubmed: exploring pubmed with gene ontology. *Nucleic Acids Research 33*, 783-786 (2005).

[50] Donoho, D. L. High-dimensional data analysis: The curses and blessings of dimen- sionality. In *Mathematical Challenges of the 21st Century, The American Math. Society, Los Angeles* (2000).

[51] Durant, J. L., Leland, B. A., Henry, D. R., and Nourse, J. G. Reoptimization of mdl keys for use in drug discovery. *J Chem Inf Model 6*, 42 (2002), 1273–1280.

[52] Eden, P., Ritz, C., Rose, C., Ferno, M., and Peterson, C. 'good old' clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *Eur J Can. 40* (2004), 1837–41.

[53] Elsebakhi, E., , Asparouhov, O., and Al-Ali, R. Novel incremental ranking framework for biomedical data analytics and dimensionality reduction: Big data challenges and opportunities. *Journal of Computer Science and Systems Biology* (2015).

[54] Ermolat'ev, D. S., Bariwal, J. B., Steenackers, H. P., de Keersmaecker, S. C., and van der Eycken, E. V. Concise and diversity-oriented route toward polysubstituted 2-aminoimidazole alkaloids and their analogues. *Angew. Chem. Int. Ed. Engl.*, 49 (2010), 9465–9468.

[55] Fang, O. H., Mustapha, N., and N, S. M. Integrative gene selection for classification of microarray data. *Computer and Information Science 4* (2011), 55–63.

[56] Faro, A., Giordano, D., and Spampinato, C. Combining literature text mining with microarray data: advances for system biology modeling. *Brief Bioinform 1*, 13 (2012), 61–82.

[57] Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R. *Advances in Knowledge Discovery and Data Mining*. AAAI/ MIT Press, 1997.

[58] Filippone, M., Masulli, F., and Rovetta, S. Simulated annealing for supervised gene selection. *Springer-Verlag* (2010).

[59] Fogel, G, B., and Corne, D, W. *Computational intelligence in bioinformatics*. IEEE Press Series on Computstional Intelligence.

[60] Gasteiger, J., and Engel, T. *Chemoinformatics: A textbook*. WILEY-VCH GMbH and Co. KGaA, 2003.

[61] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., and J, Z. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol. 5*, 10 (2004), R80.

[62] Geppert, H., Vogt, M., and Bajorath, J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J Chem Inf Model 50* (2010), 205–216.

[63] Gevaert, O., Smet, F., Timmerman, D., Moreau, Y., and De Moor, B. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics 22* (2006), e184–e190.

[64] Gevaert, O., Smet, F., Timmerman, D., Moreau, Y., and De Moor, B. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics 22* (2006), e184–e190.

[65] Ghosh, D., Li, Z., Tan, X. F., Lim, T. K., Mao, Y., and Lin, Q. itraq based quantitative proteomics approach validated the role of calcyclin binding protein (cacybp) in promoting colorectal cancer metastasis. mol cell proteomics. *Mol Cell Proteomics 7* (2013), 1865–80.

[66] Gligorijević, V., and Pržulj, N. Methods for biological data integration: perspectives and challenges. *J R Soc Interface 12*, 112 (2015).

[67] Goble, C., and stevens, R. State of the nation in data integration for bioinformatics. *J Biom Inf. 41* (2008), 687–693.

[68] Golub, G. H., and Van Loan, C. F. *Matrix Computations*. 2nd ed. (Baltimore: Johns Hopkins University Press), 1989.

[69] Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A., and Tegnér, J. Data integration in the era of omics: current and future challenges. *BMC Systems Biology 8* (2014), 1 –10.

[70] Gonzalez-Diaz, H., Vilar, S., Santana, L., and Uriarte, E. Medicinal chemistry and bioinformatics-current trends in drugs discovery with networks topological indices. *Curr Top Med Chem 7*, 10 (2007), 1015–1029.

[71] Guha, R., and Willighagen, E. A survey of quantitative descriptions of molecular structure. *Curr Top Med Chem. 12*, 18 (2012), 1946–1956.

[72] Guyon, I. An introduction to variable and feature selection. *JMLR 3* (2003), 1157–1182.

[73] Hall-Stoodley, L., and Stoodley, P. Evolving concepts in biofilm infections. *Cellular Microbiology 11*, 7 (2009), 1034–1043.

[74] Hamid, J. S., Hu, P., Roslin, N. M., Ling, V., Greenwood, C. M. T., and Beyene, J. Data integration in genetics and genomics: Methods and challenges. *Hum Genomics Proteomics* (2015).

[75] Hansch, C., Maloney, P., Fujita, T., and Muir, R. Correlation of biological activity of phenoxyacetic acids with hammett substituent constants and partition coefficients. *Nature*, 194 (1962), 178–180.

[76] Harriet, W., Eeva, K., Jouni, K. S., Antti, K., Jaakko, H., Sisko, A., and Sakari, K. Identification of differentially expressed genes in pulmonary adenocarcinoma by using cdna array. *Oncogene 21* (2002), 5804–5813.

[77] He, Z., Zhang, J., Shi, X., Hu, L., Kong, X., Cai, Y., and Chou, K. Predicting drugtarget interaction networks based on functional groups and biological features. *Plos One 5*, 3 (2010), e9603.

[78] Heged, I., Jelasity, M., Kocsis, L., and Benczúr, A. A. Fully distributed robust singular value decomposition. In *14-th IEEE International Conference on Peer-to-Peer Computing* (2014), p. 1–9.

[79] Hess, K. R., Anderson, K., Symmans, W. F., Valero, V., Ibrahim, N., Mejia, J. A., Booser, D., Theriault, R. L., Buzdar, A. U., Dempsey, P. J., Rouzier, R., Sneige, N., Ross, J. S., Vidaurre, T., Gómez, H. L., Hortobagyi, G. N., and Pusztai, L. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *J Clin Oncol 24* (2006), 4236–4244.

[80] Higham, N. Newton's method for the matrix square root. *Mathematics of Computation 46*, 174 (April 1986), 537–549.

[81] Hingorani, S. R., Petricoin, E. F., Maitra, A., Rajapakse, V., King, C., Jacobetz, M. A., Ross, S., Conrads, T. P., Veenstra, T. D., Hitt, B. A., Kawaguchi, Y., Johann, D., Liotta, L. A., Crawford, H Cand Putt, M. E., Jacks, T., Wright, C. V., Hruban, R. H., Lowy, A. M., and Tuveson, D. A. Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse. *Cancer Cell 6*, 4 (2003), 437–50.

[82] Hofmann, T., Schölkopf, B., and Smola, A. J. Kernel methods in machine learning. *Ann. Statist. 36*, 3 (2008), 1171–1220.

[83] Hsu, C.-L., and Lee, W.-C. Detecting differentially expressed genes in heterogeneous disease using half student's t-test. *Int. I. Epidemiol. 10* (2010), 1–8.

[84] Hu, S., and Sunil, R. J. Statistical redundancy testing for improved gene selection in cancer classification using microarray data. *Cancer Inform 3* (2007), 29–41.

[85] Hubert, M., and Engelen, S. Robust PCA and classification in biosciences. *Bioinformatics 20*, 11 (2004), 1728–1736.

[86] Šimundić, A. Measures of diagnostic accuracy: basic definitions. *Med Biol Sci. 4*, 22 (2008), 61–5.

[87] J Moler, E., L Chow, M., and S Mian, I. Analysis of molecular profile data using generative and discriminative methods. *Physiol Genomics 4* (2000), 109–126.

[88] Jia, Y. B. Singular value decomposition. *University of Iowa, Unpublished paper* (2014).

[89] Johnson, D., McKeever, S., Stamatakos, G., Dionysiou, D., Graf, N., Sakkalis, V., Marias, K., Wang, Z., and

Deisboeck, T. S. Dealing with diversity in computational cancer modeling. *Cancer Informatics 12* (2013), 115–124.

[90] Jonnalagadda, S., and Srinivasan, R. A PCA based approach for gene target selection to improve industrial strains. *Computer Aided Chemical engineering 24* (2007), 1013–1018.

[91] Joyce, A. R., and Palsson, B. . The model organism as a system: integrating omics data sets. *Nat. Rev. Mol. Cell Biol. 7* (2006), 198–210.

[92] Kaklamanis, L., Gatter, K. C., Hil, l. A. B., Mortensen, N., Harris, A. L., Krausa, P., McMichael, A., Bodmer, J. G., and Bodmer, W. F. Loss of hla class-i alleles, heavy chains and beta 2-microglobulin in colorectal cancer. *Int J Cancer*, 51 (1992), 379–385.

[93] Kaneko, S., Hirakawa, A., and Hamada, C. Gene selection using a high-dimensional regression model with microarrays in cancer prognostic studies. *Cancer Inform 11* (2012), 29–39.

[94] Kato, T., Tsuda, K., and Asai, K. Selective integration of multiple biological data for supervised network inference. *Systems biology 21* (2005), 2488–2495.

[95] Keiser, M. J., Roth, B. L., Armbruster, B. N., Ernsberger, P., Irwin, J. J., and Shoichet, B. K. Relating protein pharmacology by ligand chemistry. *Nat Biotech-nol 25*, 2 (2007), 197–206.

[96] Kim, W. H., Lee, B. L., Jun, S. H., and Song, S Y and, K. H. K. Expression of 32/67-kda laminin receptor in laminin adhesionselected human colon cancer cell lines. *British J Cancer 77* (1998), 15–20.

[97] Kitayama, J., Nagawa, H., Tsuno, N., Osada, T., Hatano, K., Sunami, E., Saito, H., and Muto, T. Laminin mediates tethering and spreading of colon cancer cells in physiological shear flow. *British J Cancer 80* (1999), 1927–1934.

[98] Kombo, D. C., Tallapragada, K., Jain, R., Chewning, J., Mazurov, A. A., Speake, J. D., Hauser, T. A., and Toler, S. 3d molecular descriptors important for clinical success. *J Chem Inf ModelJ Chem Inf Model 53*, 2 (2013), 327–342.

[99] Kovahi, R., and Provost, F. Glossary of terms. *Machine Learning 30* (1998), 271–274.

[100] Kuriakose, M. A., Chen, W. T., He, Z. M., and Sikora, A. G. Selection and validation of differentially expressed genes in head and neck cancer. *Cell Mol Life Sci 61(11)* (2004), 1372–83.

[101] Kurokawa, M. S., Hatsugai, M., Noguchi, Y., Yoshioka, T., Mitsui, H., Yasuda, H., and Kato, T. *Proteomic Approaches for Biomarker Discovery in Ulcerative Colitis.* 2011.

[102] Labute, P. Derivation and applications of molecular descriptors based on approximate surface area. *Methods Mol Biol. 275* (2004), 261–78.

[103] Lanckriet, G. R. G., De Bie, T., Cristianini, N., Jordan, M. I., and Noble, S. A statistical framework for genomic data fusion. *Bioinformatics 20*, 16 (2004), 2626–2635.

[104] Landini, P., Antoniani, D., Burgess, J. G., and Nijland, R. Molecular mechanisms of compounds affecting bacterial biofilm formation and dispersal. *Applied Microbiology and Biotechnology 3*, 86 (2010), 813–823.

[105] Leach, A. R., and Gillet, V. J. *An Introduction to Chemoinformatics.* Springer, 2007.

[106] Leach, A. R., and Gillet, V. J. *Representation and Manipulation of 2D Molecular Structures.* Springer, 2007.

[107] Lee, C. H., Alpert, B. O., Sankaranarayanan, P., and Alter, O. Gsvd comparison of patient-matched normal and tumor acgh profiles reveals global copy-number alterations predicting glioblastoma multiforme survival. *PLoS One 1*, 7 (2012).

[108] Li, J.-B., Wang, Y.-H., Chu, S.-C., and Roddick, J. F. Kernel self-optimization learning for kernel-based feature extraction and recognition. *Inf. Sci. 257* (Feb. 2014), 70–80.

[109] Li, X., Rao, S., Wang, Y., and Gong, B. Gene mining: a novel and powerful ensemble decision approach to hunting for disease genes using microarray expression profiling. *Nucleic Acids Res. 9* (2004), 2685–2694.

[110] Lin, Z., Chen, M., Wu, L., and Ma, Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *http://Arxiv.org/abs/1009.5055v2* (2010).

[111] Liu, Z., Chen, D., and Bensmail, H. Gene expression data classification with kernel principal component analysis. *J Biomed biotechnol 2* (2005), 155–159.

[112] Lu, X., and Zhang, X. The effect of genechip gene definitions on the microarray study of cancers. *BioEssays 28* (2006), 739–746.

[113] Lv, W., and Xue, Y. Prediction of acetylcholinesterase inhibitors and characterization of correlative molecular descriptors by machine learning methods. *Eur J Med Chem 45*, 3 (2010), 1167–1172.

[114] Lézoray, O., Charrier, C., Cardot, H., and Lefèvre, S. Machine learning in image processing. *EURASIP Journal on Advances in Signal Processing 2008*, 1 (2008), 1–2.

[115] Masliah, E., Roberts, E. S., Langford, D., and Everall, I. Patterns of gene dysregulation in the frontal cortex of patients with hiv encephalitis. *J Neuroimmunol 157(1-2)* (2004), 163–75.

[116] Medini, D., Serruto, D., Parkhill, J., Relman, D. A., Donati, C., Moxon, R., Falkow, S., and Rappuoli, R. Microbiology in the post-genomic era. *Nat Rev Microbiol. 6*, 6 (2008), 419–430.

[117] Melville, J. L., Burke, E. K., and Hirst, J. D. Machine learning in virtual screening. *Comb Chem High Throughput Screen 12* (2009), 332–343.

[118] Mercer, J. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A 209* (1909), 415–446.

[119] Metz, C. E. Basic principles of ROC analysis. *Semin Nucl Med*, 8 (1978), 283–298.

[120] Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., Ploner, A., Pawitan, Y., Hall, P., Klaar, S., Liu, E. T., and Bergh, J. An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *PNAS 102*, 38 (2005), 13550–13555.

[121] Mishra, D., and Sahu, B. Feature selection for cancer classification: A signal-to-noise ratio approach. *International Journal of Scientific Engineering Research 2* (2011), 1–7.

[122] Mitchel, J. B. O. Machine learning methods in chemoinformatics. *Comput Mol Sci 4* (2014), 468–481.

[123] Morgan, H. J. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *J.Chem. Doc. 5*, 2 (1965), 107–113.

[124] Nantasenamat, C., Isarankura-Na-Ayudhya, C., and Prachayasittikul, V. Advances in computational methods to predict the biological activity of compounds. *Expert Opin Drug Dis 5*, 7 (2010), 633–654.

[125] Natarajan, J. Text mining perspectives in microarray data mining. *ISRN Computational Biology 2013* (2013), 1–5.

[126] Ng, A., Jordan, M., and Weiss, Y. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems 14, Proceedings of the 2001* (2001), pp. 849–856.

[127] Nutt, C. L., Mani, D. R., Betensky, R. A., Tamayo, P., Cairncross, J. G., Ladd, Cand Pohl, U., Hartmann, C., McLaughlin, M. E., Batchelor, T. T., Black, P. M., von Deimling, A., Pomeroy, S. L., Golub, T. R., and Louis, D. N. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Res. 63* (2003), 1602–1607.

[128] Pearson, K. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine 2 11* (1901), 559–572.

[129] Pechenizkiy, M., Tsymbal, A., and Puuronen, S. PCA-based feature transformation for classification:issues in medical diagnostics. In *Proceedings of the 17th IEEE Symposium on Computer-Based Medical Systems* (2004).

[130] Perot, S., Sperandio, O., Miteva, M. A., Camproux, A., and Villoutreix, B. O. Druggable pockets and binding site centric chemical space: a paradigm shift in drug discovery. *Drug Discov Today 15*, 15-16 (2010), 656–667.

[131] Pescatori, M., Broccolini, A., Minetti, C., and Bertini, E. Gene expression profiling in the early phases of dmd: a constant molecular signature characterizes dmd muscle from early postnatal life throughout disease progression. *FASEB J 21(4)* (2007), 1210–26.

[132] Pittman, J., Huang, E., Dressman, H., Horng, C., Cheng, S., Tsou, M., Chen, C., Bild, A., Iversen, E., Huang, A., Nevins, J., and West, M. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. *PNAS 101* (2004), 8431–8436.

[133] Plastria, F., De Bruyne, S., and Carrizosa, E. Dimensionality reduction for classification. *Lecture Notes in Computer Science 5139* (2008), 411–418.

[134] Pochet, N., De Smet, F., Suykens, J. A. K., and De Moor, B. Systematic benchmarking of microarray data classification: assessing the role of nonlinearity and dimensionality reduction. *Bioinformatics 20* (2004), 3185–3195.

[135] Qin, Z., Zhang, J., Xu, B., Chen, L., Wu, Y., Yang, X., Shen, X., Molin, S., Danchin, A., Jiang, H., and Qu, D. Structure-based discovery of inhibitors of the yycg histidine kinase: new chemical leads to combat staphylococcus epidermidis infections. *BMC Microbiol 6* (2006), 96.

[136] Qiu, P., and Plevritis, S. K. Simultaneous class discovery and classification of microarray data using spectral analysis. *Journal of Computational Biology 16* (2009), 935–944.

[137] Quinlan, J. R. Induction of decision trees. *Machine Learning 1* (1986), 81–106.

[138] Raj, P., and Deka, G. C. *Handbook of Research on Cloud Infrastructures for Big Data Analytics*. IGI Global: Advances in Data Mining and Database Management Sytems, 2014.

[139] Raychaudhuri, S., Chang, J. T., Imam, F., and Altman, R. B. The computational analysis of scientific literature to define and recognize gene expression clusters. *Nucleic Acids Res 15*, 31 (2003), 4553–60.

[140] Rester, U. From virtuality to reality - virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective. *Current Opinion in Drug Discovery Development. 11*, 4 (2008), 559–68.

[141] Reverter, F., , Vegas, E., and Sánchez, P. Mining gene expression profiles: An integrated implementation of kernel principal component analysis and singular value decomposition. *Genomics Proteomics Bioinformatics 8*, 3 (2010), 200–210.

[142] Rogers, D., Brown, R., and Hahn, M. Using extended-connectivity finger prints with laplacian-modified bayesian analysis in high-throughput screening. *J.Biomol.Screen 7*, 10 (2005), 682–6.

[143] Romling, U., and Balsalobre, C. Biofilm infections, their resilience to therapy and innovative treatment strategies. *Journal of internal medicine*, 272 (2012), 541–561.

[144] Roth, V., and Lange, T. Bayesian class discovery in microarray data. *IEEE Transactions on Biomedical Engineering 51* (2004).

[145] Rudemo, M. Empirical choice of histograms and kernel density estimators. *Scand. J. Statist. 9* (1982), 65–78.

[146] Scholkopf, B., Smola, A. J., and Muller, K.-R. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation. 10* (1998b), 1299–1319.

[147] Schölkopf, B., Tsuda, K., and Vert, J.-P. *Kernel Methods in Computational Methods*. MIT Press, Cambridge, MA, 2004.

[148] Schreiber, A. W., Shirley, N. J., Burton, R. A., and Fincher, G. B. Combining transcriptional datasets using the generalized singular value decomposition. *BMC Bioinformatics 9*, 335 (2008), 1 –15.

[149] Schuermans, M., Markovsky, I., Wentzell, P. D., and Huffel, S. V. On the equivalence between total least squares and maximum likelihood PCA. *Analytica Chimica Acta 544* (2005), 254–267.

[150] Schwaighofer, A., Schroeter, T., Mika, S., and Blanchard, G. How wrong can we get?a review of machine learning approaches and error bars. *Comb Chem High Throughput Screen 12* (2009), 453—-468.

[151] Sebastiani, F. Machine learning in automated text categorization. *ACM computing surveys (CSUR) 34*, 1 (2002), 1–47.

[152] Sedeh, R, S., Bathe, M., and K-J, B. The subspace iteration method in protein normal mode analysis. *J Comput Chem 31* (2010), 66–74.

[153] Seoane, J. A., Aguiar-Pulido, V., Munteanu, C. R., Rivero, D., Rabunal, J. R., Dorado, J., and Pazos, A. Biomedical data integration in computational drug design and bioinformatics. *Curr Comput Aided Drug Des. 1*, 9 (2013), 108–17.

[154] Shannon, B. J., and Paliwal, K. K. Role of phase estimation in speech enhancement. In *INTERSPEECH 2006-ICSLP* (2006).

[155] Shao, L., Huang, Q., Luo, M., and Lai, M. Detection of the differentially expressed gene igf-binding protein-related protein-1 and analysis of its relationship to fasting glucose in chinese colorectal cancer patients. *Endocrine-Related Cancer 11* (2004), 141–148.

[156] Sokolova, M., Japkowicz, N., and Szpakowicz, S. Beyond accuracy, f-score and ROC: a family of discriminnat measures for performance evaluation. *American Association for Artificial Intelligence* (2006).

[157] Somorjai, R. L., Dolenko, B., and Baumgartner, R. Class prediction and discovery using gene microarray and proteomics mass spectroscopy data:curses,caveats, cautions. *Bioinformatics 19* (2003), 1484–1491.

[158] Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., van de Vijver, M. J., Bergh, J., Piccart, M., and Delorenzi, M. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst 98* (2006), 262–272.

[159] Sotiropoulos, A., Gineitis, D., Copeland, J., and Treisman, R. Signal-regulated activation of serum response factor is mediated by changes in actin dynamics. *Cell*, 98 (1999), 159–169.

[160] Steenackers, H., Dubey, A., Robijns, S., Ermolat'ev, D., Delattin, N., Dovgan, B., Girandon, L., Fröhlich, M., De Brucker, K., Cammue, B. P. A., Thevissen, K.,

Balzarini, J., Van der Eycken, E. V., and J, V. Evaluation of the toxicity of 5-aryl-2-aminoimidazole-based biofilm inhibitors against eukaryotic cell lines. *Bone Cells and the Nematode Caenorhabditis elegans Molecules*, 19 (2014), 16707–16723.

[161] Steenackers, H., Ermolat'ev, D., Trang, T., Savalia, B., de Weerdt, A., Shah, A., Vanderleyden, J., and van der Eycken, E. Microwave-assisted one-pot synthesis and anti-biofilm activity of 2-amino-1h-imidazole/triazole conjugates. *Org. Biomol. Chem.*, 12 (2014), 3671–3678.

[162] Steenackers, H., Ermolat'ev, D. S., Savaliya, B., de Weerdt, A., de Coster, D., van der Eycken, E., de Vos, D., Vanderleyden, J., and de Keersmaecker, S. C. Structure activity relationship of 4(5)-phenyl-2-amino-1h-imidazoles, n1-substituted 2-aminoimidazoles and imidazo[1,2-a]pyrimidinium salts as inhibitors of the biofilm formation by salmonella typhimurium and pseudomonas aeruginosa. *J. Med. Chem.*, 54 (2010), 472–482.

[163] Steenackers, H., Hermans, K., Vanderleyden, J., and De Keersmaecker, S. Salmonella biofilms: an overview on occurrence, regulation and eradication. *Food Research International 2*, 45 (2012).

[164] Steenackers, H. P., Ermolat'ev, D. S., Savaliya, B., Weerdt, A. D., Coster, D. D., Shah, A., van der Eycken, E. V., de Vos, D. E., Vanderleyden, J., and de Keersmaecker, S. C. Structure-activity relationship of 2-hydroxy-2-aryl-2,3-dihydro-imidazo[1,2-a]pyrimidinium salts and 2n-substituted 4(5)-aryl-2-amino-1h-imidazoles as inhibitors of biofilm formation by salmonella typhimurium and pseudomonas aeruginosa. *Bioorg. Med. Chem.*, 19 (2011), 3462–3473.

[165] Stein, U., Walther, W., Arlt, F., Schwabe, H., Smith, J., Fichtner, I., Birchmeier, W., and Schlag, P. M. Macc1, a newly identified key regulator of hgf-met signaling, predicts colon cancer metastasis. *Nat Med 1*, 85 (2009), 25–39.

[166] Stirewalt, D. L., Meshinchi, S., Kopecky, K. J., and Fan, W. Identification of genes with abnormal expression changes in acute myeloid leukemia. *Genes Chromosomes Cancer 47(1)* (2008), 8–20.

[167] Strombergsson, H and, L. M., and Kleywegt, G J and, W. J. Towards proteome-wide interaction models using the proteochemometrics approach. *Mol Inform 29*, 6-7 (2010), 499–508.

[168] Sudhakar, J., and Rajagopalan, S. A PCA-based approach for gene target selection to improve industrial strains. *Computer Aided Chemical engineering 24* (2007), 1013–1018.

[169] Suykens, J. A. K., De Brabanter, J., Lukas, L., and Vandewalle, J. Weighted least squares support vector machines: robustness and sparse aprroximation. *Neurocomputing 48* (2002), 85–105.

[170] Suykens, J. A. K., Gestel, T. V., De Brabanter, J., De Moor, B., and Vandewalle, J. *Least Squares Support Vector Machines*. World Scientific, 2002.

[171] Suykens, J. A. K., Van Gestel, T., Vandewalle, J., and De Moor, B. A support vector machine formulation to PCA analysis and its kernel version. *IEEE Transactions on Neural Networks 14*, 2 (Mar 2003), 447–450.

[172] Suykens, J. A. K., and Vandewalle, J. Least squares support vector machine classifiers. *Neural Processing Letters 9* (1999), 293–300.

[173] Tarca, A. L., Romero, R., and Draghici, S. Analysis of microarray experiments of gene expression profiling. *Am J Obstet Gynecol. 2*, 195 (2006), 373–388.

[174] Thomas, M., Daemen, A., and De Moor, B. Maximum likelihood estimation of gevd: Applications in bioinformatics. *IEEE/ACM Trans Comput Biol Bioinform 11*, 4 (2014), 673–80.

[175] Thomas, M., De Brabanter, K., Suykens, J. A. K., and De Moor, B. Predicting breast cancer using an expression values weighted clinical classifier. *BMC Bionformatics 15* (2014), 1–17.

[176] Tibshirani, R. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. B 58* (1996), 267–288.

[177] Tibshirani, R. J., Hastie, T. J., Narasimhan, B., and G, C. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS 99* (2002), 6567–6572.

[178] Todeschini, R., and Consonni, V. *Molecular descriptors for chemoinformatics*. Wiley-VCH Verlag GmbH & Co. KGaA, 1995.

[179] Townsend, A., and Bodmer, H. Antigen recognition by class i-restricted t lymphocytes. *Annu Rev Immunol*, 7 (1989), 601–624.

[180] Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D., and Altman, R. B. Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics 18* (2002), 1454–1461.

[181] van ' t Veer, L. J., Dai, H., Van De Vijver, M. J., HeY, D., Hart, A. A. M., Mao, M., Peterse, H. L., Van Der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernard, R., and Friend, S. H. Gene expression profiling predicts clinical outcome of breast cancer. *Nature 415* (2002), 530–536.

[182] van den Elsen, P. J., Holling, T. M., Kuipers, H. F., and van der Stoep, N. Transcriptional regulation of antigen presentation. *Curr Opin Immunol*, 16 (2004), 67–75.

[183] van Vliet, M. H., Horlings, H. M., van de Vijver, M., and Reinders, M. J. T. Integration of clinical and gene expression data has a synergetic effect on predicting breast cancer outcome. *PLoS ONE 7* (2012), e40385–e40358.

[184] van Westen, G. J. P., Wegner, J. K., IJzerman, A. P., van Vlijmen, H. W. T., and Bender, A. Proteochemometric modeling as a tool to design selective compounds and for extrapolating to novel targets. *Medchemcomm 2*, 1 (2011), 16–30.

[185] Vapnik, V. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[186] Venet, D., Maenhaut, C., and Bersini, H. Separation of samples into their constituents using gene expression data. *Bioinformatics 17* (2001), S279–S287.

[187] Verweij, P. J., and Houwelingen, H. C. Cross-validation in survival analysis. *Stat Med 12* (1993), 2305–14.

[188] Wang, S., Huang, J., He, J., Wang, A., Xu, S., Huang, S. F., and Xiao, S. Rpl41, a small ribosomal peptide deregulated in tumors, is essential for mitosis and centrosome integrity. *Neoplasia 3* (2010), 284–93.

[189] Wang, Y., Miller, D. J., and Clarke, R. Approaches to working in high-dimensional data spaces: gene expression microarrays. *British Journal of Cancer 98*, 6 (2008), 1023–1028.

[190] Wang, Z., and Palade, V. *Fuzzy gene mining: A fuzzy-based framework for cancer microarray data analysis*. A john wiley Sons, 2008.

[191] Warren, M. A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics 4*, 5 (1943), 115–133.

[192] Watkins, D. S. Product eigenvalue problems. *Society for Industrial and Applied Mathematics 47* (2005), 3–40.

[193] Wentzell, P. D., Andrews, D. T., Hamilton, D. C., Faber, K., and Kowalski, B. R. Maximum likelihood principal component analysis. *J. Chemomet 11* (1997), 339–366.

[194] Wentzell, P. D., Andrews, D. T., and Kowalski, B. R. Maximum likelihood multivariate calibration. *Anal. Chem. 69* (1997), 2299–2311.

[195] Wentzell, P. D., and Lohnes, M. T. Maximum likelihood principal componnet analysis with correlated measuremnet errors: theoretical and practical considerations. *Chemomet. Intell. Lab. Syst 45* (1999), 65–85.

[196] Wold, S., and Eriksson, L. Statistical validation of QSAR results. *Waterbeemd, Han van de. Chemometric methods in molecular design. Weinheim: VCH*, ISBN 3-527-30044-9 (1995), 309–318.

[197] Wong, Y. F., Selvanayagam, Z. E., Wei, N., and Porter, J. Expression genomics of cervical cancer: molecular classification and prediction of radiotherapy response by dna microarray. *Clin Cancer Res 9(15)* (2003), 5486–92.

[198] Yana, X., Denga, M., K Fungb, W., and Qiana, M. Detecting differentially expressed genes by relative entropy. *Journal of Theoretical Biology 234* (2005), 395–402.

[199] Yang, J. Y., Yang, M. Q., Zhu, M., Arabnia, H. R., and Deng, Y. Promoting synergistic research and education in genomics and bioinformatics. *BMC Genomics supple I*, 9 (2008), 1–5.

[200] Yeung, K. Y., and Ruzzo, W. L. Principal component analysis for clustering gene expression data. *Bioinformatics 17*, 9 (2001), 763–774.

[201] Zeng, Z., Qian, L., Cao, L., Tan, H., Huang, Y., Xue, X., Shen, Y., and Zhou, S. Virtual screening for novel quorum sensing inhibitors to eradicate biofilm formation of pseudomonas aeruginosa. *Appl Microbiol Biotechnol 79* (2008), 119–126.

[202] Zhang, H., Song, X., and Zhang, X. Miclique: An algorithm to identify differentially coexpressed disease gene subset from microarray data. *Journal of Biomedicine and Biotechnology* (2009).

Curriculum vitae

Minta Thomas was born on May 20, 1983 in Vazhoor, Kottayam, India. In 2003, she obtained bachelor of science for computer applications from the M G University, Kottayam, India and in 2005 she obtained master of science from the M G University, Kottayam, India. In 2007, she completed the master of Philosophy for Bioinformatics from the Kerala University, Trivandrum, India. From 2007-2009, she worked as a faculty in Bioinformatics at Christ College Rajkot and SNGIST, Cochin, India. In 2009, she started her PhD (initial 34- months Funded by Erasmus Mundus Scholarship for PhD student from European Commission) in bioinformatics group at Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, K U Leuven Belgium under the supervision of Prof. Dr. Ir. Bart De Moor. Her research interests include the development of machine learning algorithms, specifically, for dimensionality reduction and data integration and its applications in prediction/classification problems; deep learning, bigdata and text analytics.

List of publications

Published Papers

- Thomas Minta, Daemen Anneleen and De Moor Bart. Maximum likelihood estimation of GEVD: Applications in Bioinformatics. IEEE/ACM Transactions on Computational Biology and Bioinformatics 2014, Volume: 11, Issue: 4, 673:680 (2014).
- Thomas M., De Brabanter K., De Moor B.: New bandwidth selection criterion for Kernel PCA: Approach to Dimensionality Reduction and Classification Problems. BMC Bioinformatics 2014, 15:137 (2014).
- Thomas M., De Brabanter K., Suykens J.A.K., De Moor B.: Predicting breast cancer using an expression values weighted clinical classifier. BMC Bioinformatics 2014, 15:411 (2014).

Internal Reports

- Minta Thomas, Hans P Steenackers, Marc De Maeyer, Johan AK Suykens, Inge Thijs, Xiaoyu Qing, Tran Thi Thu Tran, Erik Van der Eycken, Jos Vanderleyden and Bart De Moor : Chemoinformatics approach to identify new compounds which inhibit bio films formed by either Salmonella or Pseudomonas. Internal Report 16-169, ESAT-SISTA, KU Leuven (Leuven, Belgium), 2016.
- Thomas Minta and De Moor Bart. Robust PCA improves biomarker discovery in colon cancer with incorporation of literature information. Internal Report 14-30, ESAT-SISTA, KU Leuven (Leuven, Belgium), 2014.

Papers presented in international conferences

– Thomas Minta, Daemen Anneleen and De Moor Bart. Maximum likelihood estimation of GEVD: Applications in Bioinformatics. Asia Pacific Bioinformatics Conference (APBC) 2014, Shanghai, China.

Posters presented in international conferences

– Improved classification by integrating multiple patient data sets with literature information using co-inertia analysis. Thomas Minta, Daemen Anneleen and De Moor Bart. European Conference on Computational Biology 2010 (ECCB10).

– A GSVD framework to identify differentially expressed genes for colon cancer by incorporating literature information. Thomas Minta and De Moor Bart. International Conference on Stem Cells and Cancer 2010 (ICSCC10).

FACULTY OF ENGINEERING SCIENCE
DEPARTMENT OF ELECTRICAL ENGINEERING
BIOINFORMATICS
ESAT - STADIUS, Stadius Centre for Dynamical Systems, Signal Processing and Data Analytics
Kasteelpark Arenberg 10, bus 2446, 3001 Leuven-Belgium