MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA

Cristina Botella Pérez

ISBN: 978-84-693-5427-8
Dipòsit Legal: T-1418-2010

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA

Cristina Botella Pérez

DOCTORAL THESIS

UNIVERSITAT ROVIRA I VIRGILI

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA

Cristina Botella Pérez

DOCTORAL THESIS

Supervised by
Dr. Joan Ferré Baldrich and Dr. Ricard Boqué Martí

Department of Analytical Chemistry and Organic Chemistry
Universitat Rovira i Virgili
Tarragona 2010

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

ROVIRA I VIRGILI UNIVERSITY
Department of Analytical Chemistry
and Organic Chemistry

Dr. JOAN FERRÉ BALDRICH and Dr. RICARD BOQUÉ MARTÍ, associate professors of the Department of Analytical Chemistry and Organic Chemistry at Rovira i Virgili University

CERTIFY:

The Doctoral Thesis entitled: 'MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA', presented by CRISTINA BOTELLA PÉREZ to receive the degree of Doctor of the Rovira i Virgili University, has been carried out under our supervision, in the Department of Analytical Chemistry and Organic Chemistry at Rovira i Virgili University, and all the results presented in this thesis were obtained in experiments conducted by the above mentioned student.

Tarragona, March 2010

Dr. Joan Ferré Baldrich

Dr. Ricard Boqué Martí

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

In the middle of difficulty
lies the opportunity

Albert Einstein

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Arriben els últims dies de gairebé cinc anys de camí... d'un camí que no ha estat fàcil, ple de sensacions, experiències i moments compartits amb molta gent. Gent que ha estat amb mi durant part o la totalitat d'aquesta tesi i de la que no em voldria oblidar ara que sembla que arribem al final.

Gràcies al Dr. Joan Ferré i al Dr. Ricard Boqué per confiar en mi.

Joan, Ricard, gràcies pels consells i per donar-me l'oportunitat d'aprendre al vostre costat.

Gràcies a tots els membres del grup de Quimiometria, Qualimetria i Nanosensors per aquests anys. Gràcies als companys amb els que he compartit els inicis, la totalitat o el final del doctorat. Així i sense voler oblidar-me de ningú, gràcies a tot el grup per acollir-me com ho heu fet. Gràcies a Vero i Giselle pel seu suport i ànims sobretot als inicis. Gràcies a Idoia, Vane, Santi, Jordi, Jaume, Carol i Kris per tots els bons moments i els riures de les millors hores de cafè.

Joe, aiix al final mi compi de despacho, cuántas horas compartidas y cuántos buenos momentos, me quedo con ellos, gracias. Igualment, gràcies a Marta S, pels ànims, per preocupar-te i posar-li somriures a aquesta tesi.

Montse, encara que sigui des de la distància, gràcies. Gràcies pels teus correus i els teus ànims. També des de la distància, gràcies a Sílvia, Laia, David i Lluis; des de les nostres terres m'heu acompanyat dia a dia. Les vostres paraules han estat sempre importants.

Laura, Antonio, Rafa ... aquest camí ha tingut sentit gràcies a vosaltres, GRÀCIES per ser com sou, no canvieu mai.

Laura, GRÀCIES. Gracias por preocuparte por mí, por nuestras charlas y por tener siempre una palabra de apoyo y de ánimo preparada, gracias por compartir conmigo estos años.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Antonio, què t'he de dir...tants anys junts, GRÀCIES. *Gràcies pels* breaks, *pels riures que hem compartit i has aconseguit treure'm en els mals dies. Gracias por preocuparte y estar siempre a mi lado.*

Rafa, com tants cops, ara tampoc tinc paraules, simplement GRÀCIES. *Gràcies pel teu suport, els teus ànims en els mals moments i les teves paraules sempre ben escollides. Hem quedo amb les nostres llargues xerrades. Gràcies per escoltar-me i ser-hi sempre.*

Tomàs, la persona que ha compartit amb mi aquest camí, que m'ha apoiat en els moments més durs i ha fet possible que arribés a la fi, GRÀCIES. *Gràcies per no deixar-me defallir i ajudar-me a mirar endavant en tot moment. Sé que no sempre ha estat fàcil.*

I que puc dir d'aquells qui gairebé han fet la tesi per mi i amb mi... els meus pares, GRÀCIES. *Gràcies per estar sempre al meu costat i apoiar-me en qualsevol de les meves decisions, creient amb les meves possibilitats més que ningú.*

A tots, només una paraula més, GRÀCIES.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Table of contents

ISBN:978-84-693-5427-8/DL:T-1418-2010

Structure

This thesis is structured in eight chapters.

Chapter 1. Introduction. This chapter gives an overview of DNA microarrays, their origin, types and applications. The steps involved in the generation of the microarray data, from hybridization to image acquisition and data pre-processing, are described. The need of multivariate data analysis is justified. Finally, the multivariate methods used for analyzing microarray data are cited, focusing on classification methods.

Chapter 2. Thesis Objectives. In this chapter are described the aims of this thesis. These objectives are developed in the publications included in the next chapters.

Chapter 3. Discussion of the implementation of the reject option in *p*-DPLS. This chapter discusses the implementation of the reject option in *p*-DPLS. Firstly, the calculation of the *p*-DPLS model and the class prediction process based on the Bayes rule are detailed. Then, the limitations of the classification based on the Bayes rule are discussed. Two approximations to introduce a reject option that overcome the cited limitations discussed in previous section are presented. Finally, the implications of the reject option in the evaluation of the classifiers are commented.

Chapter 4. Classification from microarray data using *p*-DPLS with reject option. This paper (C. Botella, J. Ferré, R. Boqué, *Talanta, 80* (*2009*) *321-328*) describes the implementation of a reject option in *p*-DPLS models in order to improve the classification of microarray data. The reject option allows a *p*-DPLS model to not classify outliers and ambiguous samples. This ensures that only the samples whose classification is reliable enough are indeed classified. As a consequence, the number of misclassifications decreases and the accuracy of the classifier improves.

Chapter 5. Outlier detection and ambiguity detection for microarray data in *p*-DPLS regression. Outlier detection is often overlooked in microarray data analysis with factor-

based classification methods. However, outlier diagnostics are required when implementing any classification method in real practice. In this paper (C. Botella, J. Ferré, R. Boqué, *Journal of Chemometrics* (*2010*) Accepted*)* two procedures, typically used in chemometrics, are combined with the reject option (chapter 4) to detect outliers and ambiguous samples in *p*-DPLS. The application of these diagnostics increases the accuracy of the *p*-DPLS models and avoids classifying samples from classes that were not modelled.

Chapter 6. Gene selection based on selectivity ratio for *probabilistic* discriminant partial least squares. Gene selection is a fundamental step in microarray data analysis. It allows both identifying the genes that characterize a certain disease and also simplifying and improving classification models by discarding irrelevant genes. In this paper (C. Botella, J. Ferré, R. Boqué, (*2010*) *submitted*) a gene selection procedure that is specific for PLS is used to find the best subset of genes that discriminate between different subtypes of tumours and also between healthy and tumour samples. The procedure is based on selecting the genes that maximize the selectivity ratio (*SR*) index. The paper also shows that the calculated accuracy of a classifier can be largely influenced by how the dataset is splitted into a training set and a test set. Certain splits can lead to a wrong assessment of the validity of the gene selection algorithm. A repetitive procedure consisting of data split, gene selection, training and validation is proposed in order to test the goodness of the genes selected whit the *SR* index.

Chapter 7. Multi-class classification of microarray gene expression data. In most cases, samples to be classified from microarray data may belong to more than two subtypes of a disease. The *p*-DPLS approach used so far only allows discriminating between two subtypes. This chapter (C. Botella, J. Ferré, R. Boqué, (*2010*) *submitted*) describes a classification strategy to be used when there are more than two candidate classes. The method combines the predictions from one-versus-one *p*-DPLS models with the Linear Discriminant Analysis (LDA) classifier.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Chapter 8. Conclusions. This chapter sums up the improvements achieved by the methods presented in this thesis.

The **Appendix** contains a description of the datasets used in this thesis, the list of the abbreviations used, and the list of papers and presentations performed during this period.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

CHAPTER 1 | Introduction

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

1.1 Genetic expression

Deoxyribonucleic acid (DNA) molecules are the genetic material of most living organisms [1]. They are chains of nucleotides (Figure 1). A nucleotide consists of a phosphate group, a deoxyribose sugar molecule and a nitrogenous base (guanine, cytosine, adenine or thymine) [1]. Genes are sequences of hundreds or thousands of these nucleotides that encode the genetic information to make specific proteins [2].

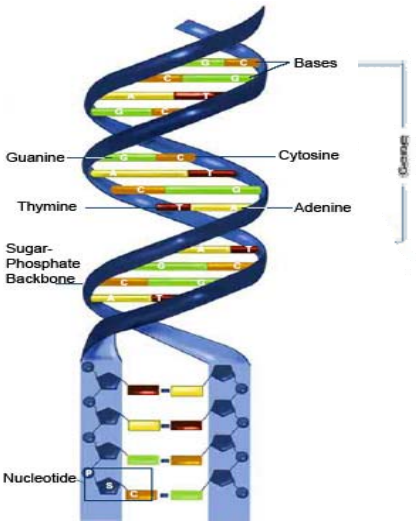
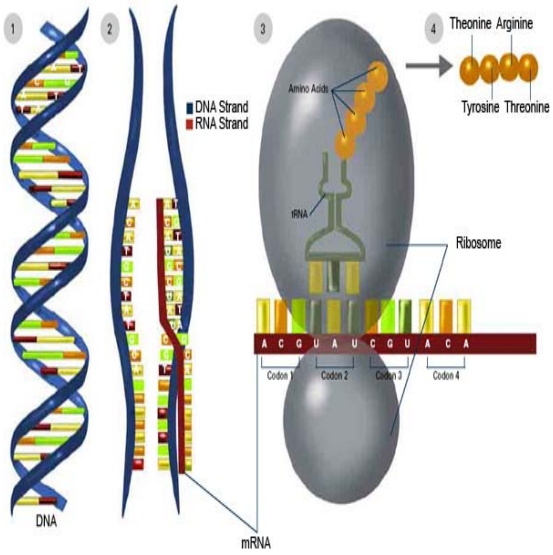


Figure 1. DNA chain. Source: [3].



The protein formation involves a transcription process, in which the genes are mapped into messenger RNA (mRNA) by the RNA polymerase enzyme [1, 4] followed by a translation process, in which the aminoacids encoded by the mRNA codons are joined in the presence of transfer RNA (tRNA) and ribosomal RNA (rRNA) (Figure 2).

Figure 2. Transcription and translation processes in the making of a protein. Source: [3].

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

The genes regulate the protein expressions and consequently the metabolic processes of the living organisms. Some genes are only expressed in particular cell types or in certain development stages [5], so these genes (or their expressed intermediate, mRNA) can be seen as markers to define particular cellular states, such as healthy or tumour [6].

1.2 Microarrays

Microarray technology is a powerful tool for simultaneously evaluating the expression level of thousands of genes in a cell [2] and, hence, the information that is encoded in the DNA [6].

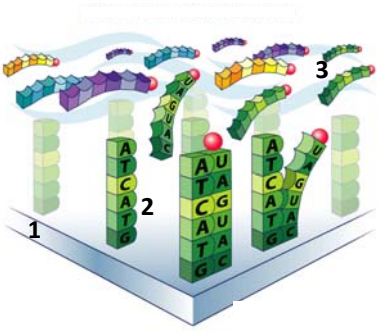


Figure 3. Parts of a microarray. **1.** Slide, **2.** Probe DNA, **3**. Target DNA. Source: *Affymetrix*.

A ***microarray*** is a microscopic slide that contains an ordered series of DNA, RNA proteins or tissues. The DNA microarrays are the most common [7]. A DNA microarray is generally a glass slide or a silicon chip in which thousands of gene sequences are printed (Figure 3). On every spot many copies of a specified DNA sequence are chemically bonded to the surface of the slide [2]. The genes immobilized onto the slide are called the DNA probe. Over this DNA probe, the target DNA or the target RNA (depending on the microarray platform) obtained from the cell under study is hybridized (hydrogen bonded). The amount of hybridization is measured and related to the presence and expression of certain genes in the cell.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Figure 4 shows the workflow process in a microarray experiment. The experimental process varies depending on the microarray platform that is used (see below). After data have been measured and pre-processed, multivariate analysis is needed to deal with the large amount of data that every microarray experiment generates [7-10].

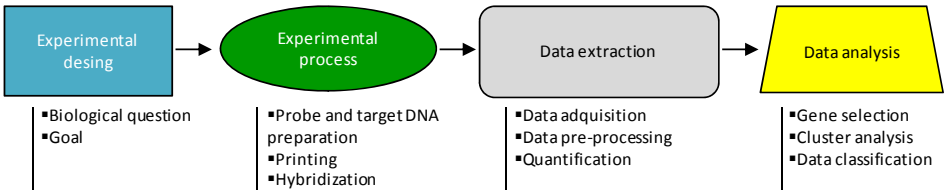


Figure 4. Microarray workflow process.

1.2.1 Microarray platforms and experimentation

The first DNA array was developed by Ed Southern in 1975 [10]. Southern noticed that labelled acid nucleic molecules could be used to evaluate other molecules linked on a solid support. He used the array to verify the presence or the absence of a specific sequence of DNA from the different sources and to identify the size of the restriction fragment.

In 1995, an *in-situ* probe synthesis method for photolitographically manufacturing DNA arrays was developed by Fodor *et al*. [11] and commercialized by Affymetrix Inc. At the same time, pre-synthesized DNA microarrays were popularized by Patrick O. Brown's laboratory at Stanford University [12]. They published step-by-step plans for building a robotic DNA arrayer [13]. This was, together with the development of the Southern blot, one of the milestones in the microarray development because the Brown's method made microarrays affordable for research laboratories, while the early methods for manufacturing miniaturized DNA arrays using *in-situ* probe synthesis required sophisticated and expensive robotic equipment.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Nowadays, there are two main microarray platforms, namely cDNA arrays (where c means complementary) and oligonucleotide arrays. They differ in the preparation and content of the probe, and also on the sample preparation [2, 5] (Table 1). Figure 5 shows the experimental procedure in a cDNA microarray experiment and in a *in-situ* oligonucleotide microarray experiment.

In cDNA microarrays, the probes are cDNAs typically 100-300 bases long. A cDNA strand is a DNA strand synthetized using a reverse transcriptase enzyme, which makes a DNA sequence complementary to the mRNA present in cells [2]. Note that the commonly called DNA microarrays are actually cDNAmicorarrays. The target sample consists of chains of cDNA of the test samples Cy5 labeled and chains of cDNA of refernece sample Cy3 labeled [2, 5]. After the sample has been hybridized, microarrays are washed for several minutes in decreasing salt buffers and finally dried either by centrigugation of the slide or a rinse in isopropanol followed by quick drying with nitrogen gas or filtered air [7]. The raw microarray data are obtained by exciting the fluorescent dyes at each spot and scanning the microarray. One intensity value is generated by the emission from the Cyanine 3 (Cy3) fluorophore and another from Cyanine 5 (Cy5). The total fluorescence emitted by the spot at each wavelength is proportional to the total amount of the dye in the spot. Hence, it is proportional to the total amount of reference or test sample hybridized. When images of both dyes (colour channels) are mixed, the typical microarray picture is obtained [1, 7, 14]. The colours on the microarray image respond to the four respective situations of microarray hybridization (Figure 6): no hybridization (black spot), reference sample hybridization (green spot), target sample hybridization (red spot) and test and reference sample hybridization (yellow spot). Different intensities of the colours indicate different levels of hybridization.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

DNA microarray images from different samples are then transformed onto gene expression data matrices. Each row of the matrix corresponds to a sample and each column corresponds to a gene. Each value characterizes the expression level of the particular gene in that particular sample. The gene expression is given by the ratio between the intensities in the red and the green channels, which are directly related to the level of expression of the transcript [1].

The *in-situ* oligonucleotide arrays, produced by Affymetrix, each gene is represented as a probe set of 10-25 oligonucleotide pairs[1] instead of one full length or partial cDNA clone. These probes are synthesized directly on the surface of the support. The target sample is a cDNA biotin labeled [2, 7]. In contrast on the spotted cDNA arrays, in this case the test and the reference sample are hybridized separately on different chips; then, data adquisition is done by scanning the probe array. It creates a 8×8 pixels (on average) for any probe cell. A single intensity value for every probe cell, representative of the hybridization level of its target, is derived. Finally, the gene expression is given by the differences of PM and MM [1]. The gene expressions of all genes analysed for a sample are given in a row of the gene expression matrix.

Table 1. Types of microarrays. Source: [1].

Probe	Arraying technique	Microarray platform
cDNA	Robotic spotting	Spotted cDNA microarrays
Oligonucleotides	Robotic spotting	Spotted oligonucleotide microarrays
	In-situ synthesis	*In-situ* oligonucleotide microarrays

[1] The oligonucleotide pair (probe pair) comprises one oligonucleotide that perfectly matches the gene sequence (Perfect Match, PM) and a second oligonucleotide having one nucleotide mismatch in the middle of it (Mismatch, MM).

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

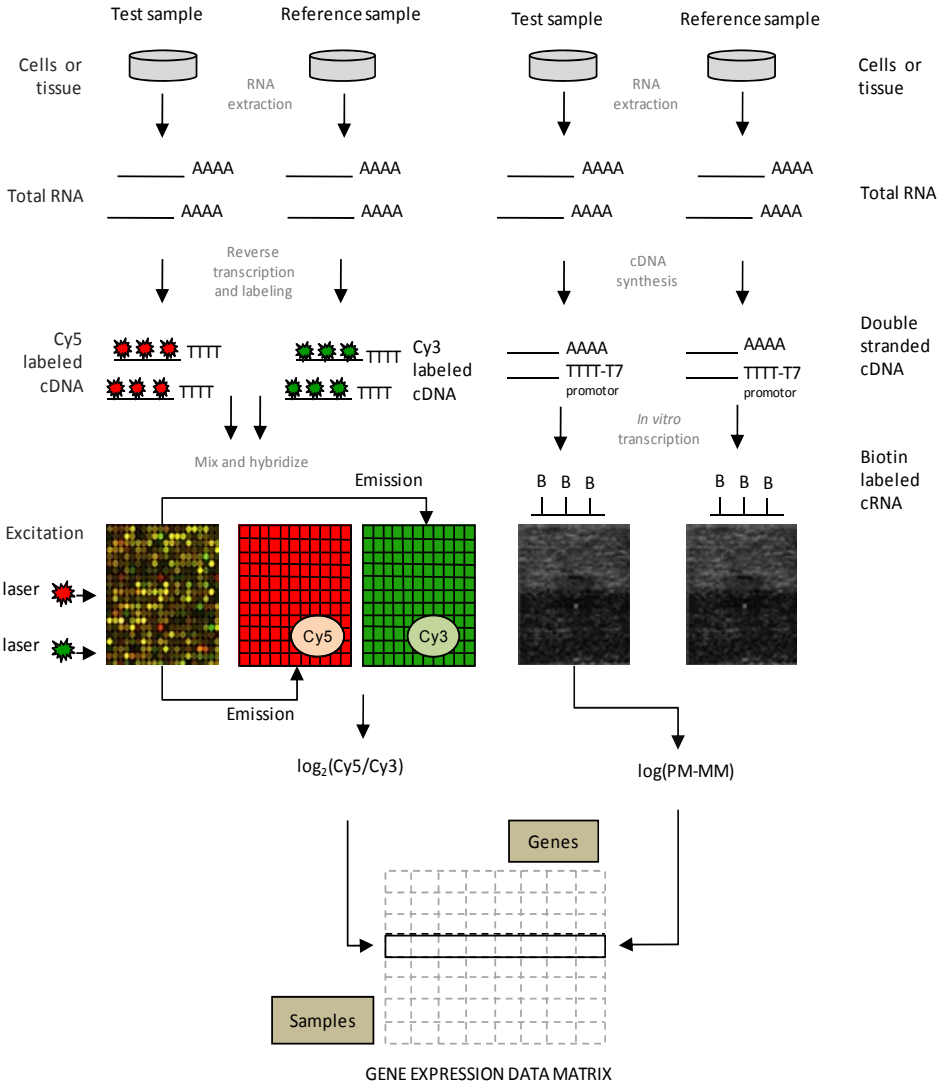


Figure 5. cDNA and *in-situ* oligonucleotide microarray sample preparation, hybridization and data measurement.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

1.2.2. Microarray data

The experimental steps involved in a microarray workflow, from microarray manufacture to microarray data extraction (Figure 4 and 5), may introduce noise and variability in the data. Common sources of variability in microarray experiments are variations related to microarray manufacturing and variations related to microarray scanning [7]. Variability related to microarray manufacturing is due to dye effects, slide effects or print-tip effects. The variability of microarray scanning is due to scanner manufacturing and to a non specific background. The most common origins of both [15] are summarized in Table 2. To minimize the effect of the sources of variation that may affect microarray data a proper pre-processing data is fundamental in microarray data analysis. This pre-processing transforms the data to make them suitable for analysis [1]. Pre-processing of microarray data is done in the steps described next [16].

Table 2. Sources of variations of microarray data.

Microarray manufacturing	**Dye effects**	▪ Different incorporation of dyes ▪ Dye instability ▪ Gene label interaction
	Slide effects	▪ Printing variability ▪ Different pin efficiency over time ▪ Array coating ▪ Slide inhomogeneities ▪ Efficiency of the hybridization reaction ▪ Background noise on the slide
	Spatial, Print-tip or Plate effects	▪ Different amounts of RNA of probes and DNA target sample ▪ Temperature and humidity ▪ PCR amplification ▪ Sample preparation protocols
Microarray scanning	▪ Scanner manufacture for example due: laser wrongly adjusted or laser misaligned. ▪ Non specific background and over shining, non specific radiations and signals from neighbouring. ▪ Image analysis, non linear transmission characteristics, saturation effects and variations in spot shape.	

Abbreviations. PCR: Polymerase chain reaction.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Background subtraction

Signal intensities of a gene include contributions from non specific hybridizations and other fluorescences from the glass. This background fluorescence is estimated from the pixels that are near the feature but are not a part of a spot [17]. Local background for each channel and spot is evaluated focusing on small regions surrounding the spot mask (region 2 in Figure 6). Then, the median or the mean of pixel values in this region is calculated for each channel and subtracted from the spot intensity [14].

A less used alternative calculates a global background for each slide: an average of negative control spot intensities is used as background value, being the empty spots the negative control spots.

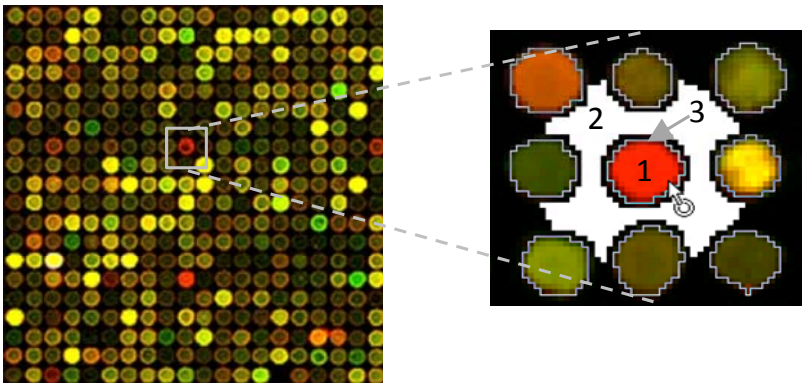


Figure 6. Scanned Microarray image. **1.** Feature pixels **2.** Background pixels **3.** Two-pixel exclusion region. Source: GENEPIX PRO [17].

In *in-situ* oligonucleotide arrays a local background is calculated for each probe and then a weighted combination of these backgrounds is substracted from all the probes of the microarray.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Treatment of missing values

Microarray datasets frequently contain missing values, either because the spot is empty (intensity=0), or because the background intensity is higher than the spot intensity (intensity with background corrected<0). These values need to be deleted or estimated and replaced, in a process called imputation, for subsequent data mining [18].

In the imputation, the missing values may be replaced by a 1 (i.e. since log(1)=0, what means no gene expression) or replaced by the mean of the intensities of the gene among all the samples.

Particularly, in affymetrix datasets, when the intensity of the Mismatch probe cell is higher than the Perfect match intensity, this probe has not physiological sense, in such a case a value called Change Threshold is used instead of the Mismatch intensity [7].

Filtering bad data

Filtering excludes from the data the observations that do not fulfil a pre-formulated presumption [4]. For example, too low intensity values that cannot be trusted due to instrumental limitations of the scanner. Typically, the lowest intensity value of the reliable microarray data, referred as “floor”, is 10. Values below “floor” are usually removed (filtered) from the data because they are not reliable enough. Similarly, the array elements at the high end of the fluorescence intensities may saturate the detector. The threshold referred to as “ceiling” value is set at 16.000 and values over “ceiling” are removed too [4, 19].

Fold change, log_2 (two fold)

In cDNA microarrays the expression of a gene in a sample is the ratio of the intensities in both channels for that gene. Although these ratios provide an intuitive measure of expressions changes, they have the disadvantage of treating up- and down-regulated genes differently. Genes up-regulated by a factor of 2 have an expression ratio of 2, whereas those down-regulated by the same factor have an expression ratio of 0.5. The most widely used transformation of the ratio is the logarithm base 2, which treats up-regulated and down-regulated genes symmetrically, so that a gene up-regulated by a factor of 2 has a $\log_2$(ratio) = 1, a gene down-regulated by a factor of 2 has a $\log_2$(ratio) = −1, and a gene expressed at a constant level (with a ratio of 1) has a $\log_2$(ratio) = 0. So, $\log_2$(ratio) will be used to represent expression levels [19].

In some cases the log transformation may be too “strong” and have the effect of increasing the importance of the low intensities. Then, a weaker transformation like a cube root is used [6].

Normalization

Normalization consists of removing arbitrary variations in the measured gene expression levels of hybridized samples so that biological differences (different gene expressions) can be more easily distinguished. Table 3 summarises the main normalization criteria used and the systematic variation they remove.

The most used method is the LOcally WEighted Scatterplot Smoothing (LOWESS) correction [20] for non linear data, and total intensity normalization or median subtraction otherwise. In *in-situ* microarrays analysis a separate probe array experiment is performed, which is used by scaling techniques to minimize differences

in overall signal intensities between the two arrays, allowing for a more reliable detection of biologically relevant changes in the sampples [1, 7].

Table 3. Strategies for microarray data normalization.

Normalization methods	Dye effects	Slide effects	spatial effects	Scanner effects	between arrays
LOWESS correction for each print-tip [7, 15, 16]	■	■	■		■
Linear correction [16, 21]			■		
Total intensity normalization [19]	■	■			
Two dyes Cy3 and Cy5 [7, 15]	■				■
Double dye experimentation, dying a sample once with Cy5 and with Cy3 in the second experiment [19, 22]	■				■
Ratios values escalation across the slides [19, 22]					■
Housekeeping genes [15]					■

1.2.3 Microarray applications

The first microarray paper featured the small mustard plant *Arabidopsis thaliana* [23], but the technology quickly spread to yeast [24], mouse [25], and human [26, 27] studies.

Present main applications of microarrays [28] include the identification of genetic individuality of tissues or organisms (e.g. detection of single nucleotide polymorphisms, SNPs) [7, 29], the investigation of cellular states and processes (such

as the sporulation process) [30], the diagnosis of genetic and infectious diseases [31-33], the identification of the subtypes of a certain disease [34, 35], the detection of genetic warning signs [36] or the drug selection [37].

In the large number of investigation areas, oncology has become the main field of DNA microarray applications [38]. General aspects of cancer expression profiling have been extensively reviewed [39-41]. It has been shown that subclassification of tumours based on their molecular profiles may help to explain why these tumours respond differently to treatment. Golub *et al*. [34] were the first to use microarray gene expression data to distinguish between acute myeloid leukemia and acute lymphocytic leukemia. Posterior studies allowed distinguishing samples of adult versus paediatric leukemia [42], different subtypes of leukemia [43] and their molecular characterization [44]. Recently, Su *et al*. [45] and Ross *et al*. [46] used large-scale RNA profiling to construct a molecular classification of different carcinomas (prostate, lung, ovary, colorectum, kidney, liver, pancreas, bladder/urethra, and gastroesophagus). Additional research for diagnosis by genetic profiling has been done for different cancers [47, 48]. In breast cancer, microarrays permitted differentiating between tumour types, corresponding to BRCA1, BRCA2 and sporadic mutations [13, 49], the differentiation between the estrogen receptors [50] and the differentiation between the stages in the cancer progression [31]. In melanoma, most of the efforts have been applied to differentiate between metastasis and no metastasis tissues [51, 52] and in hepatocellular carcinoma the research has involved the pursuit of cancer progression [53]. In other types of tumours, the diagnosis has been the main target. This is the case of bladder cancer [54], cutaneous squamous cell cancer [55], and lung cancer [56]. In the fields of colon [57], prostate [23], liver [58], glioma [59] and epithelia [60] cancers, the research has focused on the differentiation between tumour and normal tissues and in the case of lymphoma [35], medulloblastoma [61] and adenocrinoma [62] on the differentiation between different subtypes of them.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

In non oncological clinical diagnosis, DNA microarrays are used to search for the expression pattern characteristics of complex genetic disorders [47] such as diabetes [33], obesity [63, 64], and schizophrenia [65]. Microarrays have also been used in transplantation research; for example in renal transplantation to generate gene expression profiles of renal biopsies for diagnoses of acute rejection [66], or in diagnosis of infectious diseases, to detect gene sequences in the genomes of Mycobacterium tuberculosis, HIV [67, 68], and other pathogens with the aim of providing a diagnostic tool that detects expression of antibiotic resistance genes or specified viral subtypes [38].

Another important application of DNA microarrays is the identification of the genes that are responsible of a certain disease [48]. One of the first papers that reported the use of microarrays for this purpose identified the genes differentially expressed between a rat strain with insulin resistance and a normal insulin sensitive control strain [69]. After this study, microarrays have been applied to identify genes involved in many different cancer expressions [70-72], tumour progressions [73] or in many other clinical fields such as neuronal diseases [74, 75]. In the last few years many methods have been developed to identify the most relevant genes for a certain diagnosis. Three major groups of methods exist: filters, wrappers and embedded techniques [76]. These methods have been based on Genetic Algorithms [77], Random Forests [78], weights of SVM [79], t-tests or the Wilcoxon test [80], to cite a few. Most of these criteria are univariate (i.e. each feature is evaluated independently), thus simple to interpret, but they omit interactions and correlations between genes during gene selection [81]. Anyhow, these interactions must be taken into account since it has been shown that there exist pairs of genes that are coexpressed. In a simple manner, if we find that the genetic expression levels for two genes are similar, we can hypothesize that the respective genes are co-regulated and possibly functionally related [82]. More

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

accurately, these coexpressions have been proved based on the correlation of expression profiles or on functional and chromosomal structural information [83, 84].

Linked with the identification of the genes that are responsible for a disease, microarrays have also been applied to find mutations that are responsible for the disease phenotype. Although there are numerous methods for identifying the mutations, microarrays may best satisfy a need for rapid, accurate and cost-effective method for genetic polymorphism identification [47]. This identification has been presented as the foundation of pharmacogenomics. In next future, pharmacogenomics aims to optimize the dose and drug formulation and to predict good and adverse clinical responses to individual drugs, using microarrays for personalized medicine [38, 47, 85].

The huge amount of data generated in each microarray experiment implies the use of multivariate techniques for their analysis. In one of the first studies with microarrays, Golub *et al*. [34] applied two cluster self-organizing maps to group 38 samples of leukemia into two classes. Eisen *et al*. [86] used hierarchical clustering to find out the genes with similar functions. Hierarchical clustering has also been used to discover two molecular distinct types of diffuse large B-cell lymphoma in which the patients in the two subgroups showed significant differences in overall survival [35], and to categorize breast cancer into its subtypes [87]. PCA has been applied to discriminate between different tumour tissues, including colon carcinoma, breast carcinoma, central nervous system tumour, lung cancer, leukemia, melanoma, ovarian carcinoma, and prostate cancer [88]. The same analysis has been performed with k-means clustering [88].

Multivariate supervised classification methods are probably the most important tools for microarray data analysis. Such methods can be used to identify different expressed genes, to find subgroups of samples, to differentiate between different states of a

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

tumour and to infer the class of a sample from its gene expression microarray data. In general terms, the aim of any classifier is to build a decision rule from a preclassified dataset and use it to assign a new unlabeled sample to one or none of the predefined classes. A large number of classification methods have been used in microarray data analysis. The main studies are summarized in Table 4.

Table 4. Classification references for microarray gene expression data**.**

Classification method	Objective of the study
SVM	Differentiate between ovarian cancer tissues, normal ovarian tissues and other normal tissues [89].
	Recognize five sets of genes in functional classes that were expected to be co-regulated: those mediating the tricarboxylic acid cycle, respiration, cytoplasmic ribosome biosynthesis, proteasome biosynthesis and histone biosynthesis [90].
TPCR	Discriminate between tumours from a variety of tissues and organs, e.g. between subtypes of leukemia and the mutations of breast cancer [91].
	Differentiate between round blue cell tumours of childhood (neuroblastoma, rhabdomysarcoma, non-Hodgkin lymphoma and Ewing family of tumours) [91].
NN	Classify cancer samples into the same four groups of childhood cancer [92].
	Investigate the gene expression patterns associated with estrogens receptor status in sporadic breast cancer [93].
MCR-ALS	Classify types of leukemia [94]. Differentiate between nine types of tumour samples (breast cancer, central nervous system tumour, colon carcinoma, lung cancer, leukemia, melanoma, ovarian carcinoma, prostate cancer and renal carcinoma) [94].
SOM + k-means clustering	Classify subtypes of oral cancer [95].

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

KNN	Select a subset of genes to classify subtypes of leukemia and a subset to discriminate between tumour and healty colon samples [96].
PLS (for dimension reduction) + LD or QLD	Differentiate between tumour and healthy samples of colon and ovarian [97]. Differentiate between the subtypes of a cancer such as lymphoma or leukemia [97].
PLS + PHR	Predict patient survival probabilities [98].
PLS + RPLR	Classify samples of two types of leukemia [99]. Differentiate between healthy and tumour colon samples [99].
DPLS	Differentiate between samples before and after chemotherapy [100]. Identify the estrogens receptor status [100].
	Differentiate the states of a breast cancer tumour [101].
	Predict the drug efficacy using expression data biomarkers [102].
	Identify the most relevant genes correlated with a certain tumour [103].
	Predict the quality of a DNA microarray spot [104].
	Classify tumour samples (different types of lymphoma and breast cancer) [105]. Differentiate between healthy samples and samples of carcinoma, colon and prostate tumour [105].
	Identify genes whose expression appears to be synchronized with cell cycling [106]. Identify genes with periodic fluctuations in expression levels coupled to the cell cycle in the budding yeast [106].
	Select a few gene expressions that are the most effective in discriminating tumoral types (melanoma, colon, leukemia and renal tumour cells) [103, 107].
	Identify new lung cancer molecular markers with diagnostic value [108].

Abbreviations. SVM: Support vector machines, TPCR: Total principal component regression, NN: Neural networks, MCR-ALS: Multivariate curve resolution alternating least squares, SOM: self-organizing maps , KNN: K-nearest neighbours, PLS: Partial least squares, LD: Logistic discrimination, QLD: Quadratic logistic discrimination, PHR: Proportional hazard regression, RPLR: Ridge penalized logistic regression, DPLS: Discriminant partial least squares.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Recently, the interest for using Discriminant Partial Least Squares (DPLS) has increased [109, 110]. This interest arises from the high computational efficiency, large flexibility and versatility of the method for the addressed microarray classification problems, and from the existence of a variety of algorithmic variants [110]. Hence, to improve the DPLS model in order to obtain better classification models and performances plays a key role in gene expression microarray data classification.

References

[1] Ting-Lee, M.L., *Analysis of microarray gene expression data*. 2004, USA: Kluwer Academic Publishers.

[2] Higgs, P.G. and T.K. Attwood, *Bioinformatics and Molecular Evolution*, ed. B.S. Ltd. 2006: Blackwell Publishing.

[3] *U.S. Department of health and human services, The New Genetics*, in *NIH Publication No.07-662*. 2006.

[4] Baldi, P. and G.W. Hatfield, *DNA microarrays and Gene expression. From Experiments to Data Analysis and Modeling.* 2002, Cambridge: Cambridge University Press.

][5] Primrose, S.B. and R.M. Twyman, *Principles of Gene Manipulation and Genomics (7th edition)*. 2006: Blackwell Publishing.

[6] Göhlmann, H. and W. Talloen, *Gene Expression Studies Using Affymetrix Microarrays*. Mathematical and Computational Biology Series, ed. C. Hall. 2009: Taylor & Francis Group, LLC.

[7] Pasanen, T., et al., *DNA Microarray Data Analysis*. 2003, Helsinki: Ed. CSC-The Finnish IT center for Science.

[8] Allison, D.B., et al., *Microarray data analysis: form disarray to consolidation and consensus.* Nature Reviews (Genetics), 2006. **7**: p. 55-65.

[9] Liew, A.W.-C., H. Yan, and M. Yang, *Pattern Recognition techniques for the emerging field of bioinformatics: A review.* . Pattern Recognition, 2005. **38**: p. 2055-2073.

[10] Southern, E., *Detection of specific sequences among DNA fragments separated by gel electrophoresis.* Journal of Molecular Biology, 1975. **98**: p. 503-507.

[11] Fodor, S.P., et al., *Multiplexed biochemical assays with biological chips.* Nature, 1993. **364**: p. 555-556.

[12] Shena, M., et al., *Quantitative monitoring of gene expression Patterns with complementary DNA microarray.* Science, 1995. **270**: p. 467-470.

[13] Hedenfalk, I., et al., *Gene Expression profiles in hereditary breast cancer.* The New England Journal of Medicine, 2001. **344**: p. 539-548.

[14] Mada, H., *Microarray Data Analysis (I), Part A: cDNA spotted Microarray. Material of Data Analysis Course*. http://www.sinica.edu.tw/~hmwu/CourseSMDA/index.htm, Academia Sinica: Institute of Statistical Science: Taiwan.

[15] Schuchhardt, J., et al., *Normalization strategies for cDNA microarrays.* Nucleic Acids Research, 2000. **28**: p. e47.

[16] Berrar, D., W. Dubitzky, and M. Granzow., *A practical approach to microarray data analysis*. 2004, USA: Kluwer Academic Publishers.

[17] http://www.moleculardevices.com/.

[18] Wang, D. and e. al., *Effects of replacing the unreliable cDNA microarray measurements on the disease classification based on gene expression profiles and functional modules.* Bioinformatics, 2006. **22**: p. 2883-2889.

[19] Quackenbush, J., *Extracting biology from high-dimensional biological data.* J Exp Biol, 2007. **210**: p. 1507-1517.

[20] Cleveland, W.S., *Robust Locally Weighted Regression and Smoothing Scatterplots.* Journal of the American Statistical Association, 1979. **74**: p. 829-836.

[21] Kepler, T.B., L. Crosby, and K.T. Morgan, *Normalization and analysis of DNA microarray data by self-consistency and local regression.* Genome Biology, 2002. **3**: p. 1-12.

[22] Yang, Y.H., et al., *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.* Nucleic Acids Research, 2002. **30**: p. e15.

[23] Singh, D., et al., *Gene expression correlates of clinical prostate cancer behavior.* Cancer Cell, 2002. **1**: p. 203-209.

[24] Shalon, D., S.J. Smith, and P.O. Brown, *1996.* Genome Research, A DNA microarray system for analyzing complex DNA samples using two- color fluorescent probe hybridization. **6**: p. 639-645.

[25] Lockhart, D.J., et al., *DNA Expression monitoring by hybridization to high density oligonucleotide arrays.* Nature Biotechnology, 1996. **14**: p. 1675-1680.

[26] Baldini, A. and D.C. Ward, *In situ hybridization banding of human chromosomes with Alu-PCR products: a simultaneous karyotype for gene mapping studies.* . Genomics, 1991. **9**: p. 770-774.

[27] Ried, T., et al., *Multicolor fluorescence in situ hybridization for the simultaneous detection of probe sets for chromosomes 13, 18, 21, X and Y in uncultured amniotic fluid cells.* Human Molecular Genetics, 1992. **1**: p. 307-313.

[28] Lesk, A.M., *Introduction to Bioinformatics (3rd Edition)*. 2008: Oxford University Press.

[29] Butcher, L.M., et al., *SNPs, microarrays and pooled DNA: identification of four loci associated with mild mental impairment in a sample of 6000 children.* Human Molecular Genetics, 2005. **14**: p. 1315-1325.

[30] Friedlander, G., et al., *Modulation of the transcription regulatory program in yeast cells committed to sporulation.* Genome Biology, 2006. **7**: article R20.

[31] Veer, L.J.v.t., et al., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002. **415**: p. 530-535.

[32] Thomas, R.S., et al., *Identification of toxicologically predictive gene sets using cDNA microarrays.* Molecular Pharmacology, 2001. **60**: p. 1189-1194.

[33] Mootha, V.K., et al., *PGC-1a-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.* Nature, 2003. **34**(3): p. 266-273.

[34] Golub, T.R., et al., *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring.* Science, 1999. **285**: p. 531-537.

[35] Alizadeh, A.A., et al., *Distinct types of diffuse large B-cell lymphoma identifyed by gene expression profyling.* Nature, 2000. **403** p. 503-511.

[36] Sebat, J., et al., *Strong Association of De Novo Copy Number Mutations with Autism.* Science, 2007. **316**: p. 445-449.

[37] Chavan, P., K. Joshi, and B. Patwardhan, *DNA Microarrays in Herbal Drug Research.* eCAM, 2006. **3**(7): p. 447-457.

[38] Aitman, T.J., *Science, medicine, and the future: DNA microarrays in medical practice.* The Brithish Medical Journal, 2001. **323**: p. 611-615.

[39] Cuperlovic-Culf, M., N. Belacel, and J. Ouellette, *Determination of tumour marker genes from gene expression data.* Drug Discovery Today Targets (Reviews), 2005. **10**: p. 429-437.

[40] MacGregor, P.F. and J.A. Squire, *Applications of microarrays to the analysis of gene expression in cancer.* Clinical Chemistry, 2002. **48**: p. 1170-1177.

[41] Wadlow, R. and S. Ramaswamy, *DNA microarrays in clinical cancer research.* Current Molecular Medicine, 2005. **5**: p. 111-120.

[42] Kohlmann, A., et al., *Pediatric acute lymphoblastic leukemia (ALL) gene expression signatures classify an independent cohort of adult ALL patients* Leukemia, 2004. **18**: p. 63-71.

[43] Haferlach, T., et al., *AML M3 and AML M3 variant each have a distinct gene expression signature but also share patterns different from other genetically defined AML subtypes.* Genes Chromosomes Cancer, 2005: p. 113-127.

[44] Kohlmann, A., et al., *Molecular characterisation of acute leukemias by use of microarray technology.* Genes Chromosomes Cancer, 2003. **37**: p. 396-405.

[45] Su, A.I., et al., *Molecular classification of human carcinomas by use of gene expression signatures.* Cancer Research, 2001. **61**: p. 7388-7393.

[46] Ross, D.T., et al., *Systematic variation in gene expression patterns in human cancer cell lines.* Nature Genetics, 2000. **24**: p. 227-235.

[47] Petrik, J., *Diagnostic applications of microarrays.* Transfusion Medicine. **16**: p. 233-247.

[48] Frolov, A.E., *Differential Gene Expression Analysis by DNA Microarray Technology and Its Application in Molecular Oncology.* Molecular Biology, 2003. **37**: p. 486-494.

[49] Sorlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* . Proceedings of the National Academy of Sciences, 2001. **98**: p. 10869-10874.

[50] West, M., et al., *Predicting the clinical status of human breast cancer by using gene expression profiles.* Proceedings of the National Academy of Sciences of the United States of America (PNAS), 2001. **98**: p. 11462-11467.

[51] Bittner, M., et al., *Molecular classification of cutaneous malignant melanoma by gene expression profiling.* Nature, 2000. **406**: p. 536–540.

[52] Clark, E.A., et al., *Genomic analysis of metastasis reveals an essential role for RhoC.* Nature, 2000. **406**: p. 532-535.

[53] Mao, H.J., et al., *Monitoring microarray-based gene expression profile changes in hepatocellular carcinoma.* World Journal of Gastroenterology, 2005. **11**: p. 2811-2816.

[54] Dyrskjot, L., *Classification of bladder cancer by microarray expression profiling: towards a general clinical use of microarrays in cancer diagnostics. Expert Reviews in Molecular Diagnostics.* 2003. **3**: p. 635–647.

[55] Dooley, T.P., et al., *Biomarkers of human cutaneous squamous cell carcinoma from tissues and cell lines identified by DNA microarrays and qRT-PCR.* . Biochemical and Biophysical Research Communications, 2003. **306**: p. 1026-1036.

[56] Gordon, G.J., R.V. Jensen, and L.L. Hsiao, *Translation of microarray data into clinically relevant cancer diagnostic tests using expression ratios in lung cancer and mesothelioma.* Cancer Research, 2002. **62**: p. 4963–4967.

[57] Alon, U., et al., *Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays.* Cell Biology, 1999. **96**: p. 6745-6750.

[58] Chen, X., et al., *Gene expression patterns in human liver cancers.* Molecular Biology of the Cell, 2002. **13**: p. 1929-1939.

[59] Boom, J.v.d., et al., *Characterization of Gene Expression Profiles Associated with glioma progression using Oligonucleotide-based microarray analysis and Real-Time Reverse Transcription-Polymerase Chain Reaction.* American Journal of Pathology, 2003. **163**: p. 1033-1043.

[60] Kitahara, O., et al., *Alterations of gene expression during colorectal carcinogenesis revealed by cDNA microarrays after laser-capture microdissection of tumour tissues and normal epithelia.* . Cancer Research, 2001. **61**: p. 3544-3549.

[61] Pomeroy, S.L. and e. al, *Prediction of central nervous system embryonal tumour outcome based on gene expression.* Nature, 2002. **415**: p. 436-442.

[62] Bhattacharjee, A., et al., *Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses.* Proceedings of the National Academy of Sciences, 2001. **98**: p. 13790–13795.

[63] Nadler, S.T., et al., *The expression of adipogenic genes is decreased in obesity and diabetes mellitus.* Proceedings of the National Academy of Sciences, 2000. **97**: p. 11371-11376.

[64] Permana, P.A., A.D. Parigi, and P.A. Tataranni, *Microarray gene expression profiling in obesity and insulin resistance.* Nutrition, 2004. **20**: p. 134-138.

[65] Hakak, Y., et al., *Genome-wide expression analysis reveals dysregulation of myelination-related genes in chronic schizophrenia.* Proceedings of the National Academy of Sciences, 2001. **98**: p. 4746-4751.

[66] Mayeux, R., *Mapping the new frontier: complex genetic disorders.* The Journal of Clinical Investigation, 2005. **115**: p. 1404-1407.

[67] Kozal, M.J., et al., *Extensive polymorphisms observed in the HIV1 cladeB protease gene using highdensity oligonucleotide arrays.* Nature Medicine, 1996. **2**: p. 753-759.

[68] Gingeras, T.R., et al., *Simultaneous genotyping and species identification using hybridization pattern recognition of generic mycobacterium DNA arrays.* Genome Research, 1998. **8**: p. 435-448.

[69] Aitman, T.J., et al., *Identification of Cd36 (Fat) as an insulinresistance gene causing defective fatty acid and glucose metabolism in hypertensive rats.* Nature Genetics,, 1999. **21**: p. 76-83.

[70] Otero, E., et al., *DNA microarrays in oral cancer.* Medicina Oral, 2004. **9**: p. 288-292.

[71] Graveel, C.R., et al., *Expression profiling and identification of novel genes in hepatocellular carcinomas.* Oncogene, 2001. **20**: p. 2704-2712.

[72] Brem, R., et al., *Global analysis of differential gene expression after transformation with the v-H-ras oncogene in a murine tumor model.* Oncogene, 2001. **20**: p. 2854-2858.

[73] Okabe, H., et al., *Genome-wide analysis of gene expression in human hepatocellular carcinomas using cDNA microarray: identification of genes involved in viral carcinogenesis and tumor progression.* Cancer Research, 2001. **61**: p. 2129–2137.

[74] Cavallaro, S., et al., *Gene expression profiles during long-term memory consolidation.* European Journal of Neuroscience, 2001. **13**: p. 1809-1815.

[75] Zirlinger, M., G. Kreiman, and D.J. Anderson, *Amygdala-enriched genes identified by microarray technology are restricted to specific amygdaloid subnuclei.* Proceedings of the National Academy of Sciences, 2001. **98**: p. 5270-5275.

[76] Saeys, Y., I. Inza, and P. Larrañaga, *A review of feature selection techniques in bioinformatics.* Bioinformatics, 2007. **23**: p. 2507-2517.

[77] Tang, E.K., P. Suganthan, and X. Yao, *Gene selection algorithms for microarray data based on least squares support vector machine.* BMC Bioinformatics, 2006. **7**: article 95.

[78] Díaz-Uriarte, R. and S.A.d. Andrés, *Gene selection and classification of microarray data using random forest.* BMC Bioinformatics, 2006. **7**: article 3.

[79] Guyon, I., et al., *Gene Selection for Cancer Classification using Support Vector Machines.* Machine Learning, 2002. **46**: p. 389-422.

[80] Troyanskaya, O.G., et al., *Nonparametric methods for identifying differentially expressed genes in microarrays.* Bioinformatics, 2002. **18**: p. 1454-1461.

[81] Li, G.-Z., et al., *Partial Least Squares based dimension reduction with gene selection for tumour classification.* 7th IEEE International Conference onBioinformatics and Bioengineering, 2007: p. 1439-1444.

[82] Brazma, A. and J. Vilo, *Gene expression data analysis.* Federation of European Biochemical Societies Letters, 2000. **480**: p. 17-24.

[83] Lee, H.K., et al., *Coexpression Analysis of Human Genes Across Many Microarray Data Sets.* Genome Research, 2004. **14:** p. 1085-1094.

[84] Kluger, Y., et al., *Relationship between gene co-expression and probe localization on microarray slides.* BMC Genomics, 2003. **4**: p. 49-54.

[85] Gunther, E.C. and e. al., *Prediction of drug efficacy by classification of drug-induced genomic expression profiles in vitro.* Proceedings of the National Academy of Sciences, 2003. **100**: p. 9608-9613.

[86] Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proceedings of the National Academy of Sciences, 1998. **95**: p. 14863-14868.

[87] Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**: p. 747-752.

[88] Crescenzi, M. and A. Giuliani, *The main biological determinants of tumor line taxonomy elucidated by a principal component analysis of microarray data.* FEBS Letters 2001. **507**: p. 114-118.

[89] Furey, T.S., et al., *Support Vector Machine classification and validation of cancer tissue samples using microarray expression data.* Bioinformatics, 2000. **16**: p. 906-914.

[90] Brown M.P.S, e.a., *Knowledge-based analysis of microarray gene expression data by using support vector machines.* Proceedings of the National Academy of Sciences, 2000. **97**: p. 262-267.

[91] Tan, Y., et al., *Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data.* Nucleic Acids Research 2005. **33**: p. 56-65.

[92] Khan, J., et al., *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.* Nature Medicine, 2001. **7**: p. 673-679.

[93] Gruvberger, S., et al., *Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns.* Cancer Research, 2001. **61**: p. 5979-5984.

[94] Jaumot, J., R. Tauler, and R. Gargallo, *Exploratory data analysis of DNA microarrays by multivariate curve resolution.* Analytical Biochemistry, 2006. **358**: p. 76-89.

[95] Warner, G.C., et al., *Molecular classification of oral cancer by cDNA Microarrays Identifies overexpressed genes correlated with nodal metastasis.* International Journal Cancer, 2004. **110**: p. 857-868.

[96] Li, L., et al., *Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method.* Bioinformatics, 2001. **17**: p. 1131-1142.

[97] Nguyen, D.V. and D.M. Rocke, *Tumor classification by partial least squares microarray gene expression data.* Bioinformatics, 2002. **18**: p. 39-50.

[98] Nguyen, D.V. and D.M. Rocke, *Partial least squares proportional hazard regression for application to DNA microarray survival data.* Bioinformatics, 2002. **18**: p. 1625-1632.

[99] Fort, G. and S. Lambert-Lacroix, *Classification using Partial Least Squares with Penalized Logistic Regression.* Bioinformatics, 2005. **21**: p. 1104-1111.

[100] Pérez-Enciso, M. and M. Tenenhaus, *Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach.* Human Genetics, 2003. **112**: p. 581-592.

[101] Modlich, O., et al., *Predictors of primary breast cancers responsiveness to preoperative Epirubicin/Cyclophosphamide-based chemotherapy: translation of microarray data into clinically useful predictive signature.* Journal of Translational Medicine, 2005. **3**: article 32.

[102] Man, M.Z., et al., *Evaluation methods for classifying Expression data.* Journal of Biopharmaceutical Statistics, 2004. **14**: p. 1065-1084.

[103] Musumarra, G., et al., *Potentialities of multivariate approaches in genome-based cancer research: identification of candidate genes for new diagnostics by PLS discriminant analysisy.* Journal of Chemometrics 2004. **18**: p. 125-132.

[104] Bylesjö, M., et al., *MASQOT: a method for cDNA microarray spot quality control.* BMC Bioinformatics, 2005. **6**: p. 250.

[105] Boulesteix, A.-L., *PLS dimension reduction for classification with microarray data.* Statistical Applications in Genetics and Molecular Biology, 2004. **3**: article 33.

[106] Johansson, D., P. Lindgren, and A. Berglund, *A multivariate approach applied to microarray data for identification of genes with cell cycle-coupled transcription. .* Bioinformatics. **19**: p. 467-473.

[107] Musumarra, G., et al., *A Bioinformatic Approach to the Identification of Candidate Genes for the Development of New Cancer Diagnostics.* Biol. Chem., 2003. **384**: p. 321-327.

[108] Musssumarra, G., et al., *Genome-based identification of diagnostic molecular markers for human lung carcinomas by PLS-DA.* Computational Biology and Chemistry, 2005. **29**: p. 183-195.

[109] Nguyen, D.V. and D.M. Rocke, *Multi-class cancer classification via partial least squares with gene expression profiles.* Bioinformatics, 2002. **18**: p. 1216-1226.

[110] Boulesteix, A.-L. and K. Strimmer, *Partial least squares: a versatile tool for the analysis of high-dimensional genomic data.* Briefings in Bioinformatics, 2007. **8**: p. 32-44.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

CHAPTER 2 | Thesis Objectives

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Microarrays allow the simultaneous analysis of thousands of gene expressions. Clinical diagnosis based on gene expression data has two main targets: 1) to achieve the correct diagnostic for a patient with the greatest confidence and 2) to identify the genes responsible for a particular disease. In data analysis words, these objectives imply developing the best classification model in order to classify a sample in its true class with a low risk of misclassification and to identify the relevant variables that allow discriminating among the classes under study.

Multivariate methods are required to analyse the huge amount of data generated in microarray experiments. Discriminant Partial Least Squares (DPLS) classification is commonly used in this field. The performance of this classification method depends on many settings such as the data pre-processing, the number of factors, the number of variables and the presence of outliers. Taking into account these considerations the aim of this thesis is to optimize the classification based on DPLS in order to classify clinical samples from their gene expression microarray data. More in detail the objectives of the present thesis are:

1. To discuss the limitation of *p*-DPLS classification following the Bayes rule, which forces the classifier to always assign a sample to one of the modeled classes, and propose different approaches to overcome this limitation.

2. To implement the reject option in the *probabilistic* Discriminant Partial Least Squares method (*p*-DPLS), used to classify the samples from their gene expression data. This gives to the classification rule the ability to reject to classify a sample when the risk of misclassification is too high, and avoids forcing the classification into one of the modelled classes.

3. To develop a new method for detecting ambiguous samples and outliers for *p*-DPLS, in order to improve the accuracy of the classification model. This will avoid

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

classifying samples that would be probably misclassified due: 1) they share characteristics of the two classes modelled, 2) they do not belong to any of the modelled classes 3) they have errors in instrumental data or 4) they have errors in their class codification.

4. To develop a new method for gene selection in order to reduce the data dimensionality – eliminating the redundant data and the noise – and to improve the classification model by decreasing the risk of misclassification.

5. To study the implications that the split of the datasets into training and test sets have on gene selection and on the performance of the classification models.

6. To extend the binary classification based on DPLS to multi-class classification. This should help to solve common clinical classification problems in which more than two subtypes of samples are involved.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

CHAPTER 3 Discussion of the implementation of the reject option in *p*-DPLS

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

3.1 Introduction

Microarray gene expression data are characterized by a set of *P* features or measurements obtained through observation, which are represented by the vector **x**. The objective of a classifier is to assign a class (category) label (*y*) to this sample based on its recorded **x**. In the *probabilistic* Discriminant Partial Least Squares classification method (*p*-DPLS) [1], the PLS model translates **x** into a predicted value $\hat{y}$. This $\hat{y}$ and the probability density function (PDF) that describes the distribution of the $\hat{y}$*'s* of the training samples of each class are used to calculate the *a posteriori* probability that the sample belongs to each modeled class. Classification is then decided using the Bayes rule for minimum error [2].

The Bayes rule is commonly used as a criterion for classification. Its drawback is that the unknown sample is always classified, even if the sample is either an outlier or is ambiguous (it has a similar *a posteriori* probability to belong to both classes). In such situations, it would be better to reject to classify the sample [3].

The objective of this chapter is to discuss the implementation of the reject option in *p*-DPLS. Section 3.2 introduces the formulation of the *p*-DPLS model. Then, the application of the Bayes rule for classifying in *p*-DPLS is shown in section 3.3. Section 3.4 discusses the limitations of using the Bayes rule in *p*-DPLS. Limitations that are overcome by implementing a reject option. Two approximations for implementing the reject option in *p*-DPLS are discussed in section 3.5. Finally, section 3.6 discusses the necessary changes in the interpretation of the measures of classification performance when the classifier includes the reject option.

3.2 *Probabilistic* discriminant partial least squares

3.2.1 The partial least squares model

One task in data analysis is to describe the relationship between the observations in the predictor space (**X**) and a dependent variable (**y**) [4]. Partial least squares (PLS) is a regression method that specifically searches a set of components (or factors) that perform a simultaneous decomposition of **X** and **y** with the constraint that these components explain as much as possible the covariance between **X** and **y**. Discriminant PLS (DPLS) applies PLS regression to binary classification problems, in which **y** codifies the class of the samples [5, 6]. With microarray gene expression data, **X** is an $N \times P$ matrix of N samples and P gene expressions and **y** is a $N \times 1$ vector of ones and zeros, where the integer 0 indicates that the sample belongs to class ω_0 (e.g. "cancer type I") and the integer 1 indicates that the sample belongs to class ω_1 (e.g. "cancer type II").

PLS decomposes **X** and **y** into:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

$$\mathbf{y} = \mathbf{u}q + \mathbf{f} \tag{2}$$

where **T** is the scores matrix, **P** is the loadings matrix, **u** is the vector of scores for **y** and q is the loading [7]. **E** is the (error) residual matrix of the **X**–matrix and **f** is the vector of (error) residual of the **y**–vector. An inner relationship is constructed that relates the scores of the **X** block to the scores of the **y** block.

$$\mathbf{u} = \mathbf{T}\mathbf{w} \tag{3}$$

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Once the model is calculated, the above equations can be combined to obtain a vector of regression coefficients for a given number of factors:

$$\hat{\mathbf{b}} = \mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1}q \tag{4}$$

where **W** is the matrix whose columns are the weights in Eq. (3).

The prediction for a sample is calculated as:

$$\hat{y} = \mathbf{x}^{\mathrm{T}}\hat{\mathbf{b}} \tag{5}$$

Note that if **b** has been calculated from mean-centered data, then **x** in Eq. (5) should be mean-centered, and the predicted $\hat{y}$ should be processed accordingly. Ideally, the prediction $\hat{y}$ for a sample of class ω_1 should be 1 and for a sample of class ω_0 should be 0. Since this is never the case, because of random variability and modelling error, a threshold is defined so that a sample whose prediction $\hat{y}$ is above this threshold is classified into class ω_1, and otherwise it is classified into class ω_0. The threshold can be defined with a different degree of rigour (e.g., the threshold is arbitrarily set at 0.5 or assuming that the $\hat{y}$'s of the training samples follow a Gaussian distribution and estimating the distribution using the mean and standard deviation of the $\hat{y}$'s of each class). In the following section, the threshold is defined from PDFs that describe each class. This has lead to a new version of DPLS called *probabilistic*-DPLS (*p*-DPLS).

3.2.2 Probability density function of a class

In *p*-DPLS, one PDF is calculated that represents the PLS predictions characterizing the samples of class ω_0 and one PDF is calculated that represents the range of predictions of samples of class ω_1. The PDFs are calculated as follows. For the PLS model with *A* factors, the training samples are predicted with Eq. (5). For each training sample *i*, a

ISBN:978-84-693-5427-8/DL:T-1418-2010

Gaussian function (also called kernel function) centred at the predicted value $\hat{y}_i$ is calculated as:

$$f(\hat{y}_i) = \frac{1}{SEP_i\sqrt{2\pi}} \cdot e^{\frac{1}{2}\left(\frac{y-\hat{y}_i}{SEP_i}\right)^2} \quad (6)$$

where

$$SEP_i = \text{RMSEC}\sqrt{1+h_i} \quad (7)$$

and

$$\text{RMSEC} = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N-A-\delta}} \quad (8)$$

SEP_i is the standard error of prediction for sample i, h_i is the leverage of the sample, RMSEC is the root mean square error of calibration, y_i is the known class of the training sample i (i.e. the value 0 for a sample of class ω_0 and the value 1 for a sample of class ω_1) and δ is 1 if the data has been centred and 0 otherwise. Figure 1 shows the Gaussian functions calculated for three training samples of class ω_0 and four samples of class ω_1.

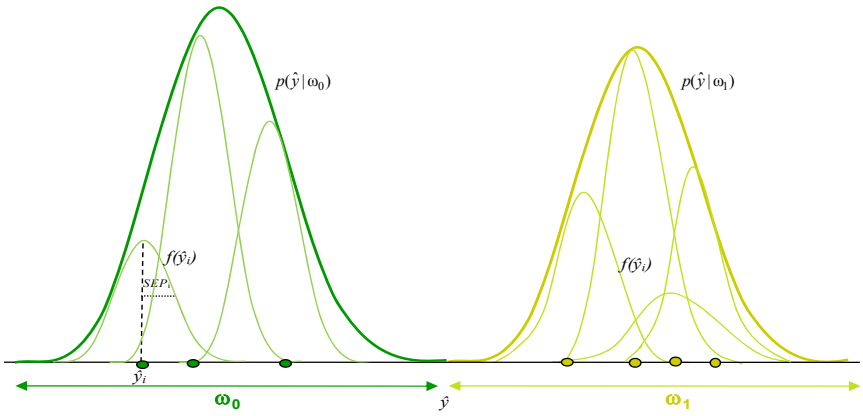


Figure 1. Gaussian functions ($f(\hat{y}_i)$) and PDFs ($p(\hat{y}|\omega_0)$, $p(\hat{y}|\omega_1)$) calculated for a hypothetical *p*-DPLS model. Note that the width of the Gaussian kernel for each sample is different, because it depends on the leverage of the sample, and, ultimately, on the relative position of the sample in the multivariate space.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

The PDFs for class ω_0 and ω_1 are calculated by averaging the individual kernel functions of the training samples of each class:

$$p(\hat{y}|\omega_0) = \frac{1}{n_0}\sum_{i=1}^{n_0} f(\hat{y}_i) \quad (9)$$

$$p(\hat{y}|\omega_1) = \frac{1}{n_1}\sum_{i=1}^{n_1} f(\hat{y}_i) \quad (10)$$

where n_0 and n_1 are the number of samples of class ω_0 and class ω_1 respectively.

For a test sample, the predicted value $\hat{y}_i$ is calculated with Eq. (5) for a DPLS model with *A* factors. Then, the sample is classified according to its probability to belong to each one of the classes, as it is shown in the next section.

3.3 Class prediction

3.3.1 Classification based on probabilities

Classification based on a priori probability

Let $\{\omega_1 \ldots \omega_c\}$ be a finite set of *C* classes. The *a priori* probability $P(\omega_c)$ is the probability of observing class *c* when a new sample arrives [8]. It reflects our prior knowledge of how likely we are to get a sample of one class (e.g. “cancer type I”) and not another kind of sample (e.g. “healthy” or “cancer type II”) [9]. *A priori* probabilities are often considered equal for all the classes [10, 11] or calculated from the number of samples in the training set assuming that this set is representative of the population [12-14], with the constraint that $\sum_{c=1}^{C} P(\omega_c) = 1$ [8].

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

In *p*-DPLS, the *a priori* probabilities are $P(\omega_0) = n_0/N$ and $P(\omega_1) = n_1/N$ for class ω_0 and class ω_1 respectively, where n_0 is the number of training samples of class ω_0, n_1 is the number of samples of class ω_1 and $N = n_0 + n_1$.

Based on the *a priori* probability only, the classification rule in *p*-DPLS that minimizes the probability of error is to assign a sample to class ω_c if

$$P(\omega_c) > P(\omega_{c\prime}) \quad c' = 1 \dots C; c \neq c' \tag{11}$$

The drawback of this rule is that it will always assign any new sample to the same class (the one with the highest *a priori* probability), although we know that samples from different classes may arrive. The information about the sample contained in **x** is ignored.

Classification based on probability density functions

A better classification decision can be made by using the measurement vector **x** that characterizes the incoming sample; in our case, the data **x** from a microarray experiment. In *p*-DPLS, **x** is first converted into the prediction $\hat{y}$ with Eq. (5) for the PLS model with *A* factors. Then, the rule is to assign the sample *i* with prediction $\hat{y}_i$ to the class ω_c if

$$p(\hat{y}_i|\omega_c) > p(\hat{y}_i|\omega_{c\prime}) \quad c' = 1 \dots C; c \neq c' \tag{12}$$

where $p(\hat{y}_i|\omega_c)$ is the class-conditional PDF for class *c* obtained from the $\hat{y}$'s of the training samples evaluated at position $\hat{y}_i$ (section 3.2.2). Note that, if for a certain sample, $p(\hat{y}_i|\omega_0) = p(\hat{y}_i|\omega_1)$, the value of the PDF will not decide. Figure 2 shows different PDFs for two classes, ω_0 and ω_1 for different hypothetical *p*-DPLS models (e.g. calculated with different number *A* of factors). For a given sample with the predicted value $\hat{y}_i$ (▪), the classification is done by comparing the values of each PDF at such $\hat{y}_i$ (arrows in Figure 2a). The sample is classified into the class with the largest $p(\hat{y}_i|\omega_c)$. Note that in the zone where the PDFs overlap, the values $p(\hat{y}_i|\omega_c)$ are similar for both classes (see the

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

first two PDFs images in Figure 2a). Hence, a small variation in $\hat{y}_i$ due to random error in **x** may change the class that has the largest $p(\hat{y}_i|\omega_c)$, and hence changes the classification decision. The classification of samples in that zone (called ambiguous samples) will be discussed later.

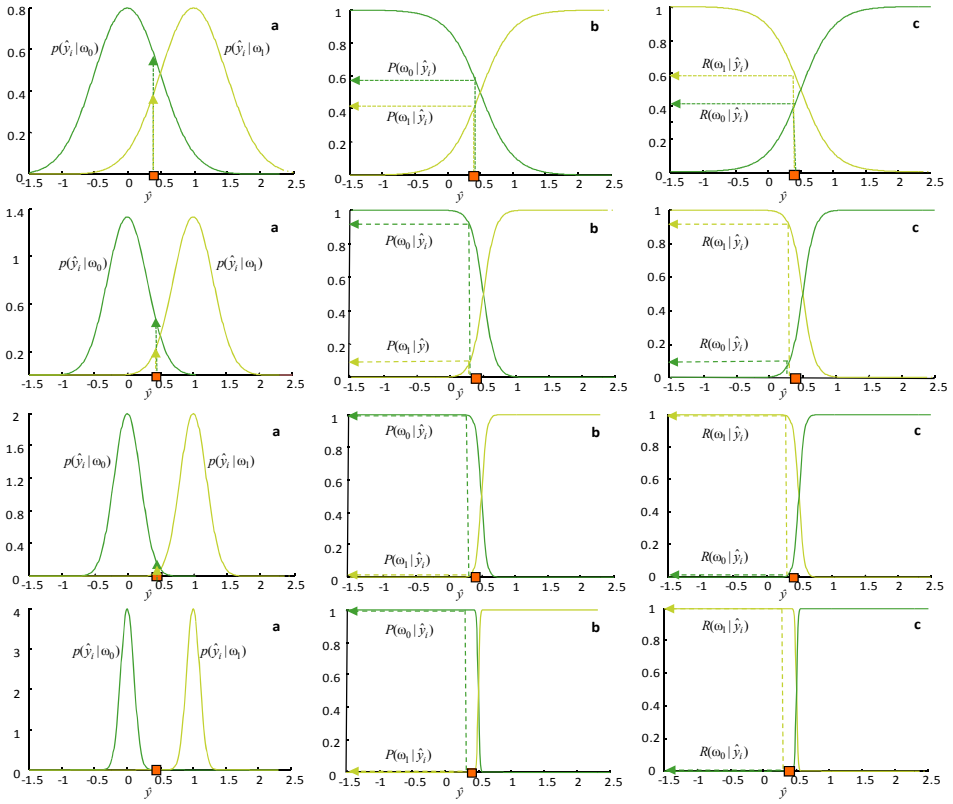


Figure 2. Example for hypothetical *p*-DPLS models **a.** PDFs **b.** *a posteriori* probabilities **c.** risk functions assuming λ_{cc}=0 and $\lambda_{cc'}$=1.

Classification based on a posteriori probability

A more elaborated classification decision combines the *a priori* probability and the prediction $\hat{y}_i$ of the incoming sample. The probability that this new sample belongs to class *c* in a *C*-class problem is given by the Bayes' *a posteriori* probability expression:

$$P(\omega_c|\hat{y}_i) = \frac{p(\hat{y}_i|\omega_c)P(\omega_c)}{p(\hat{y}_i)} \quad (13)$$

When applied to microarray data classification, $P(\omega_c|\hat{y}_i)$ is the probability that a cell or tissue characterized by its gene expression data **x** (from which $\hat{y}_i$ is obtained) is either from the "healthy" class or, otherwise, from the "tumour" class.

The denominator (known as evidence or unconditional probability density function) is a scale factor that measures how frequently we will measure a sample with such $\hat{y}_i$:

$$p(\hat{y}_i) = \sum_{c=1}^{C} p(\hat{y}_i|\omega_c)P(\omega_c) \quad (14)$$

The rule assigns the sample to the class with the largest *a posteriori* probability. So, a sample will be assigned to class ω_c if:

$$P(\omega_c|\hat{y}_i) > P(\omega_{c'}|\hat{y}_i) \quad c' = 1 \dots C; c \neq c' \quad (15)$$

Or, since the evidence is the same for all the classes, if

$$p(\hat{y}_i|\omega_c)P(\omega_c) > p(\hat{y}_i|\omega_{c'})P(\omega_{c'}) \quad c' = 1 \dots C; c \neq c' \quad (16)$$

For a two-class classification problem, as in *p*-DPLS, the *a posteriori* probabilities $P(\omega_c|\hat{y}_i)$ are:

$$P(\omega_0|\hat{y}_i) = \frac{p(\hat{y}_i|\omega_0)P(\omega_0)}{p(\hat{y}_i)} \quad (17a)$$

$$P(\omega_1|\hat{y}_i) = \frac{p(\hat{y}_i|\omega_1)P(\omega_1)}{p(\hat{y}_i)} \quad (17b)$$

where:

$$p(\hat{y}_i) = p(\hat{y}_i|\omega_0)P(\omega_0) + p(\hat{y}_i|\omega_1)P(\omega_1) \quad (18)$$

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Figure 2b shows the *a posteriori* probabilities calculated from the PDFS of Figures 2a for two classes along the $\hat{y}$ domain. The arrows indicate the *a posteriori* probability of the sample (▪) in each class.

Note that since the *a posteriori* probability is calculated as a ratio (Eq. 17a-b), it increases for one class as the $\hat{y}$ is far away from the PDF of the other class. Hence, for a sample with $\hat{y}_i$ predicted value (▪) the classification is more risked when PDFs overlap (first two rows of images of Figure 2). Instead when the distributions are more separated (images on third and fourth rows in Figure 2), the classification action is taken with higher probability of being correct.

3.3.2 Classification based on risk

Classification costs

Each classification decision has an associated cost. Let $\{\alpha_1 \ldots \alpha_c\}$ be the possible decisions, where α_c indicates that the sample is classified in class ω_c. Let $\lambda(\alpha_c|\omega_{c'})$ be the cost incurred for making the decision α_c (classify in ω_c) when the true class is $\omega_{c'}$. For short $\lambda(\alpha_c|\omega_{c'})$ is represented as $\lambda_{cc'}$.

In practice, to decide the right costs for the classification problem is difficult and requires an expert opinion. Costs result from combining several factors measured in different units – money, time or quality of life [8] –, but a general approach is to consider that a correct classification has cost 0 (i.e., when a sample of class *c* has been classified in class *c*, $\lambda_{cc} = 0$) and an incorrect classification has cost 1 (i.e., when a sample of class *c* has been classified in class *c'*, $\lambda_{cc'} = 1$) [15-17]. Other approaches have been used. Santos-Pereira [18] proposed seven different combinations of costs to optimize the classification, based on the work published by Tortorella [19]. They introduced negative costs for correct classifications and positive costs for misclassifications. Deceux

ISBN:978-84-693-5427-8/DL:T-1418-2010

[15] presented costs of classifying the samples in three different classes, with values from 0.5 to 3 to penalize each classification. Another strategy is to assign different costs to each type of error and classification, i.e. to classify a sample as "healthy" when it is "tumor" is penalized different, with a higher cost, than to classify a sample as "tumor" when it is "healthy" [9, 20].

The risk of classification

The risk of classification, called the *conditional* risk, $R(\alpha_c|\hat{y}_i)$ is defined as the expected loss (cost). Conditional means that the risk depends on the value that characterizes the sample (here $\hat{y}_i$, that derives from the observed **x** through the PLS model) in which the classification is based on. Depending on $\hat{y}_i$, we may run a higher or a lower risk. For a particular $\hat{y}_i$ and the action α_c taken, the loss incurred is $\lambda(\alpha_c|\omega_{c'})$, where $\omega_{c'}$ is the possible true class (i.e. classes in wihch the samples may be classified). Since $P(\omega_{c'}|\hat{y}_i)$ is the probability that the true class for such $\hat{y}_i$ is $\omega_{c'}$, the expected loss associated with taking action α_c is [9]:

$$R(\alpha_c|\hat{y}_i) = \sum_{c'=1}^{C} \lambda(\alpha_c|\omega_{c'}) \cdot P(\omega_{c'}|\hat{y}_i) \tag{19}$$

For two classes, the risk of classification becomes:

$$R(\alpha_0|\hat{y}_i) = \lambda_{00}P(\omega_0|\hat{y}_i) + \lambda_{01}P(\omega_1|\hat{y}_i) \tag{20a}$$

$$R(\alpha_1|\hat{y}_i) = \lambda_{11}P(\omega_1|\hat{y}_i) + \lambda_{10}P(\omega_0|\hat{y}_i) \tag{20b}$$

Here action α_0 is "classify the sample into class ω_0" and action α_1 is "classify the sample into class ω_1". λ_{01} is the loss incurred for deciding ω_0 when the true class is ω_1, λ_{10} is the loss incurred for deciding ω_1 when the true class is ω_0 and λ_{00} and λ_{11} are the costs of correctly classifying the samples into class ω_0 and class ω_1, respectively.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Whenever we have a prediction $\hat{y}_i$ we can minimize the expected loss by selecting the action that minimizes the conditional risk. The decision rule based on risk is known as Bayes' theorem of the minimum error [2]. The rule for the Bayes minimum risk classifies the sample in class ω_c if

$$R(\alpha_c|\hat{y}_i) < R(\alpha_{c'}|\hat{y}_i) \;\; c' = 1 \dots C; c \neq c' \tag{21}$$

For binary classifiers like *p*-DPLS, Eq. (21) becomes to classify the sample into class:

$$\begin{aligned} &\omega_0 \; if \; R(\alpha_0|\hat{y}_i) < R(\alpha_1|\hat{y}_i) \\ &\omega_1 \; if \; R(\alpha_1|\hat{y}_i) < R(\alpha_0|\hat{y}_i) \end{aligned} \tag{22}$$

with $R(\alpha_0|\hat{y}_i)$ and $R(\alpha_1|\hat{y}_i)$ evaluated with equations 20a-20b.

If we consider cost zero for a correct classification and cost one for any error (i.e., λ_{00} = λ_{11} = 0 and λ_{01} = λ_{10} =1), the risk of classification becomes:

$$R(\alpha_0|\hat{y}_i) = \lambda_{01} P(\omega_1|\hat{y}_i) = P(\omega_1|\hat{y}_i) \tag{23a}$$

$$R(\alpha_1|\hat{y}_i) = \lambda_{10} P(\omega_0|\hat{y}_i) = P(\omega_0|\hat{y}_i) \tag{23b}$$

and the classification decision may be expressed in terms of *a posteriori* probabilities as

$$\text{decide } \omega_0 \;\; if \;\; \lambda_{10} P(\omega_0|\hat{y}_i) > \lambda_{01} P(\omega_1|\hat{y}_i) \;\; \text{otherwise decide } \omega_1 \tag{24}$$

Figure 2c shows the risk over the $\hat{y}$ domain for a binary classifier with λ_{cc}=0 and $\lambda_{cc'}$=1. Note that the risk curves are opposite to the *a posteriori* probability curves, i.e., a high *a posteriori* probability involves a low risk, and vice-versa. Also note that the risk of classification in one of the classes decreases the furthest away the prediction is from the PDF of the other class. For a test sample (▪), in the top two models, the risk taken to classify the sample into class ω_0, $R(\alpha_0|\hat{y}_i)$, is similar to the risk to classify the sample into class ω_1, $R(\alpha_1|\hat{y}_i)$. In such a situation the chance of misclassification is high. By contrast, when the PDFs are not overlapped (Figure 2c, bottom) the risk taken when classifying

this sample in class ω_0 is much higher than the risk taken when classifying it in class ω_1 (i.e $R(\alpha_0|\hat{y}_i) >> R(\alpha_1|\hat{y}_i)$). Hence the sample will be classified in class ω_1 with a low risk of classification.

The classification based on risk is a general rule from which the previous rules derive. To be meaningful, the classification based on risks requires the costs to be set objectively (e.g. in monetary units). If they are not known and the cost of misclassification is set equal to one and the cost of correct classification is set equal to zero, the classification based on risk is equivalent to the classification based only on *a posteriori* probabilities.

3.4 Discussion of class prediction

The Bayes rule is optimal in the sense that no other rule can yield a lower error probability. However, when the $\hat{y}_i$ lies in the *ambiguity region* and when the sample lies in the *limits of the classes' domains* this rule may lead to questionable results. These situations are commented below.

It is common that in binary classification the PDFs of class ω_0 and class ω_1 overlap (Figure 3a). The overlap arises because either the classification algorithm has a limited discriminative power, or because some samples of both classes have similar measured **x**. A sample whose prediction is in that region has similar values of the PDFs $p(\hat{y}_i|\omega_0) \approx p(\hat{y}_i|\omega_1)$ and, assuming that the *a priori* probabilities are equal, has also similar values of the *a posteriori* probabilities $P(\omega_0|\hat{y}_i) \approx P(\omega_1|\hat{y}_i)$. Since there is not a clear difference, the sample could well belong to any of the two classes and the probability of misclassification is high. The overlap zone (dashed region in Figure 3) is called *ambiguity region*.

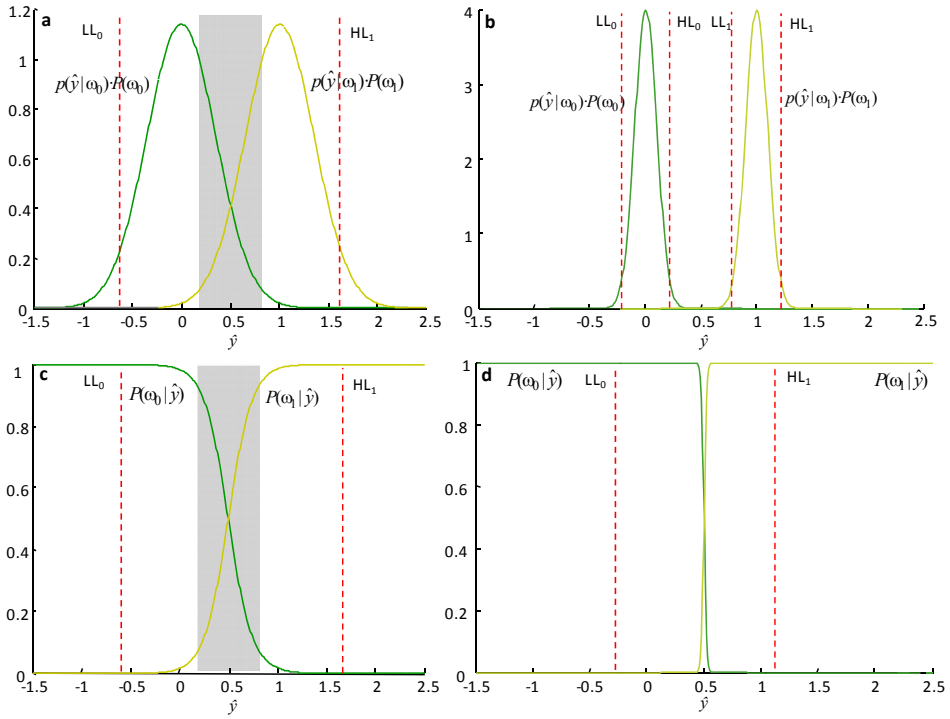


Figure 3. Hipothetic *p*-DPLS model. **a-b.** PDF's **c-b.** *a posteriori* probability functions. Class ω_0 is represented by the green line and class ω_1 by the yellow line. The dashed region is the *ambiguity* region.

Another common situation arises when the sample's prediction is outside the range of the predictions of the training samples. This situation may happen at the extremes of the PDFs (Figure 3a and 3b) and also in the region between the PDFs if the PDFs do not overlap (Figure 3b). In these regions, the class-conditional probabilities $p(\hat{y}_i|\omega_c)$ are very low for both classes and also the products $p(\hat{y}_i|\omega_0)\cdot P(\omega_0)$ and $p(\hat{y}_i|\omega_1)\cdot P(\omega_1)$ are low. However, note in the limits of the PDFs, the *a posteriori* probability for one of the classes is high (Figures 3c and 3d) because it is calculated as a ratio. For example, for $p(\hat{y}_i|\omega_0)\cdot P(\omega_0) = 10^{-7}$ and $p(\hat{y}_i|\omega_1)\cdot P(\omega_1) = 10^{-10}$, the *a posteriori* probability is $P(\omega_0|\ \hat{y}_i) = 10^{-7} / (10^{-7}+10^{-10}) \approx 1$. By letting the *a posteriori* probability decide, the sample would be classified into class ω_0 with a high *a posteriori* probability. This result is satisfactory if the

sample must necessarily belong to one of the two possible classes and the classification model has been designed to do so. However, the fact that the prediction of the sample is in the tail of the PDF, where only a very low percentage of training samples are, suggests that the sample may be an outlier and even not belong to the class. Hence, allowing the classifier to reject to classify, instead of forcing it to make a classification decision, might be advantageous. This possibility is not considered neither in the two-class Bayes rule of *a posteriori* probability (Eq. 15) nor in the *minimum risk* of classification rules (Eq. 21), which will always classify the sample.

3.5 *Probabilistic* discriminant partial least squares with reject option

In many cases, such as in clinical diagnosis, the cost of a wrong classification may be so high that it may be better to suspend the decision (to reject to classify the sample), and call for a further test [21], than to risk to obtain a wrong classification. The reject option is introduced in a classification rule to preserve against excessive misclassifications [3] and to obtain the accuracy required by the user of the classification system [22]. The reject option avoids classifying the samples with a high probability to be wrongly classified [22], and only the classifications with a low risk are performed. Hence, the reject option converts potential misclassifications into rejections [23], which reduces the error rate. The reject option, however, has two limitations:

1. Some samples that would be correctly classified by the classification model may be converted into rejections.

2. The classification model becomes useless if too many samples are rejected.

Undoubtedly a tradeoff between errors and rejects must be achieved [18]. Several strategies have been developed to define the optimal reject option [11, 18, 21, 23, 24].

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

These strategies basically reduce to two approximations, either by defining the reject option as a new class (reject class) to which the objects are assigned to or by defining the reject option as a threshold so the object is only classified if its *a posteriori* probability is higher than the threshold. These two approaches are commented below.

3.5.1 Reject option as a class

The reject option may be introduced in the classification process as an additional class, the reject class (ω_r). In such a case, the possible classification actions of the *p*-DPLS classifier are: classify the sample into class ω_0 (α_0), classify the sample into class ω_1 (α_1) and classify the sample into the reject class ω_r (α_r).

Classification based on a posteriori probability

The *a posteriori* probabilities when the reject option is implemented as a class are defined as:

$$P(\omega_0|\hat{y}_i) = \frac{p(\hat{y}_i|\omega_0)P(\omega_0)}{p(\hat{y}_i)} \tag{25a}$$

$$P(\omega_1|\hat{y}_i) = \frac{p(\hat{y}_i|\omega_1)P(\omega_1)}{p(\hat{y}_i)} \tag{25b}$$

$$P(\omega_r|\hat{y}_i) = \frac{p(\hat{y}_i|\omega_r)P(\omega_r)}{p(\hat{y}_i)} \tag{25c}$$

where the scale factor defined in Eq. (14) becomes:

$$p(\hat{y}_i) = p(\hat{y}_i|\omega_0)P(\omega_0) + p(\hat{y}_i|\omega_1)P(\omega_1) + p(\hat{y}_i|\omega_r)P(\omega_r) \tag{26}$$

The rule is to classify into:

$$\begin{aligned} &class\ \omega_0\ if\ P(\omega_0|\hat{y}_i) > \max\ (P(\omega_1|\hat{y}_i), P(\omega_r|\hat{y}_i)) \\ &class\ \omega_1\ if\ P(\omega_1|\hat{y}_i) > \max\ (P(\omega_0|\hat{y}_i), P(\omega_r|\hat{y}_i)) \\ &class\ \omega_r\ if\ P(\omega_r|\hat{y}_i) > \max\ (P(\omega_0|\hat{y}_i), P(\omega_1|\hat{y}_i)) \end{aligned} \tag{27}$$

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

If the reject class is defined in this way, the *a priori* probabilities for class ω_0 and class ω_1 are calculated from the proportion of samples of each class in the training set. For the reject class, $P(\omega_r)$ is the *a priori* probability that a new sample that should be rejected arrives and $p(\hat{y}_i|\omega_r)$ defines the distribution of the $\hat{y}_i$ of any sample that should be rejected. Both $p(\hat{y}_i|\omega_r)$ and $P(\omega_r)$ are clearly difficult to calculate. Usually it is assumed that the reject class has a uniform distribution over the $\hat{y}$ domain [25] and since the *a priori* probability has only a multiplicative effect, only the product $p(\hat{y}_i|\omega_r)\cdot P(\omega_r)$ must be calculated. One criterion is to define $p(\hat{y}_i|\omega_r)\cdot P(\omega_r)$ as a threshold so that the 5% of the area in the tails of the PDFs is below this threshold [11] (dashed regions in Figure 4). In this way a sample whose $\hat{y}_i$ is at the tails of the PDFs is rejected. Figure 4 shows the PDFs for class ω_0 and class ω_1 for overlapped and non overlapped classes. The red horizontal line is the uniform distribution calculated for the reject class. Note that this reject class defines two kinds of regions, the acceptance and the reject ones.

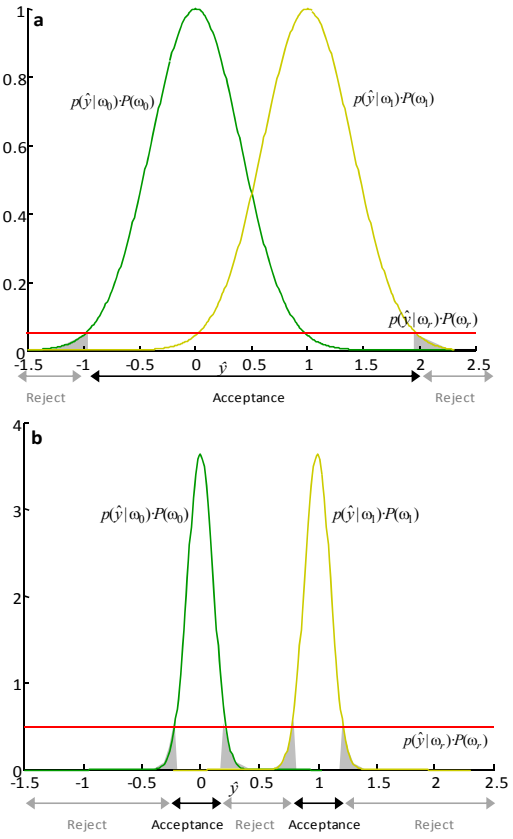


Figure 4. PDFs for overlapped and non-overlapped classes (ω_0 and ω_1) and the reject class (ω_r). The reject class is defined as a uniform distribution. This is set as the 5% area in the tails of the PDFs.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

The *a posteriori* probabilities (Eq. 25a-25c) derived from the PDFs in Figure 4 are shown in Figure 5. When the PDFs are overlapped (Figure 5a), the reject class is useful at the ends of the PDFs and also in the ambiguous zone if the distribution defining the reject class is higher than the PDFs of the classes. However, samples in the ambiguous zone will not be rejected if the reject class is below the PDFs of the classes (as usually happens) because the *a posteriori* probability of the reject class will always be smaller than the probability of classification (i.e. $P(\omega_r|\hat{y}_i) < \max(P(\omega_0|\hat{y}_i), P(\omega_1|\hat{y}_i))$. For non overlapped distributions (Figure 5b) between the PDFs the probability of the sample to belong to the reject class is the largest of the three *a posteriori* probabilities, so a sample in that zone would be rejected. This is the behaviour to be expected because there are no training samples with such $\hat{y}_i$ values. The same happens at the extreme of these distributions (i.e. equally to the extremes of overlapped distributions). In which the samples whith such $\hat{y}_i$ will be rejected to classify.

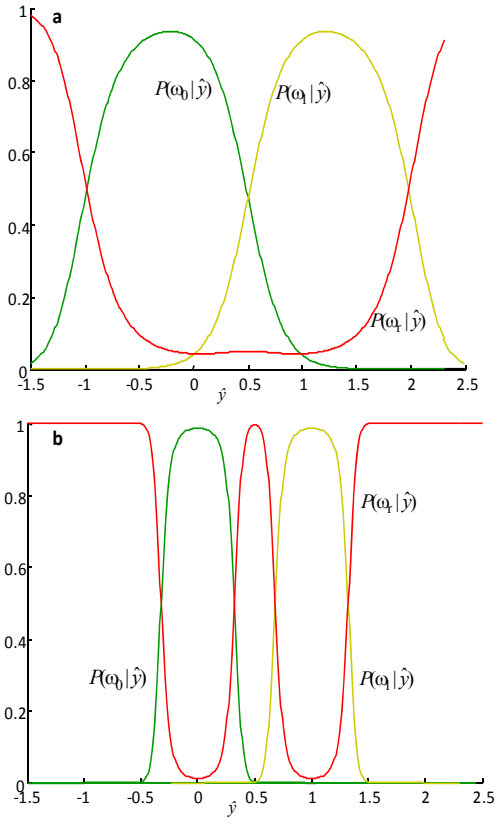


Figure 5. *A posteriori* probabilities for class ω_0 (green), class ω_1 (yellow), and the reject class ω_r (red) for **a.** overlapped classes and **b.** non overlapped classes presented on Figure 4.

Different adaptations of the reject class have been described. Pereira *et al.* [18] introduced an *indecision class* in order to reject the samples, but this reject class is not

introduced in the evaluation of *a posteriori* probabilities nor in the conditional risk. Their approach may be assimilated to introduce a reject threshold. Landgrebe *et al.* [10] considered the problem as one in which there is a well-defined *target* class and a poorly defined *outlier* class, and introduced the reject class only in the prediction step. In other words, in the training step there are two classes (target and outlier) and in the prediction or classification step an additional class is used, the reject class. This class is assumed to be uniformly distributed across the training classes' domains, and it is included in the evaluation of the probabilities. The criticism arises because in this approach the *a priori* probabilities used in the training step are different than the *a priori* probabilities used in the prediction step. Muzzolini *et al.* [11] introduced an ambiguous *class* to reduce the probability of an erroneous classification. This class identifies those samples that are classified as belonging to two or more classes with (near) equal probability. In addition, they introduced the reject distance to identify those samples that have little or no similarity with the predefined classes. The reject thresholds to identify such samples are determined by fixing the probability in which the samples are classified as belonging to the distance reject class. This is equivalent to reject the samples predicted outside a confidence interval fixed around each PDF (reject distance) [11].

3.5.2 Reject option as a threshold

A second alternative to introduce the reject option is to introduce a *reject threshold*.

Classification based on a posteriori probability

The *a posteriori* probabilities for each class over the $\hat{y}$ domain are calculated (Eqs. 17a - 17b) using the PDFs (Figure 6a-6b). For such *a posteriori* probabilities a threshold of rejection is set at (1–*t*) (Figure 6c-6d), so that a sample is rejected if the maximum *a posteriori* probability is lower than this threshold value [22].

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

The classification rule based on *a posteriori* probabilities with reject option becomes classify into:

$$class\ \omega_0\ \ if\ P(\omega_0|\hat{y}_i) > max\ \ (P(\omega_1|\hat{y}_i), (1-t))$$
$$class\ \omega_1\ \ if\ P(\omega_1|\hat{y}_i) > max\ \ (P(\omega_0|\hat{y}_i), (1-t)) \quad (28)$$

and reject the sample if:

$$(1-t) > max\ \ (P(\omega_0|\hat{y}_i), P(\omega_1|\hat{y}_i)) \quad (29)$$

If the PDFs of the two classes are overlapped, the reject threshold divides the $\hat{y}$ domain into two regions: acceptance region and reject region (Figure 6c and 6d).

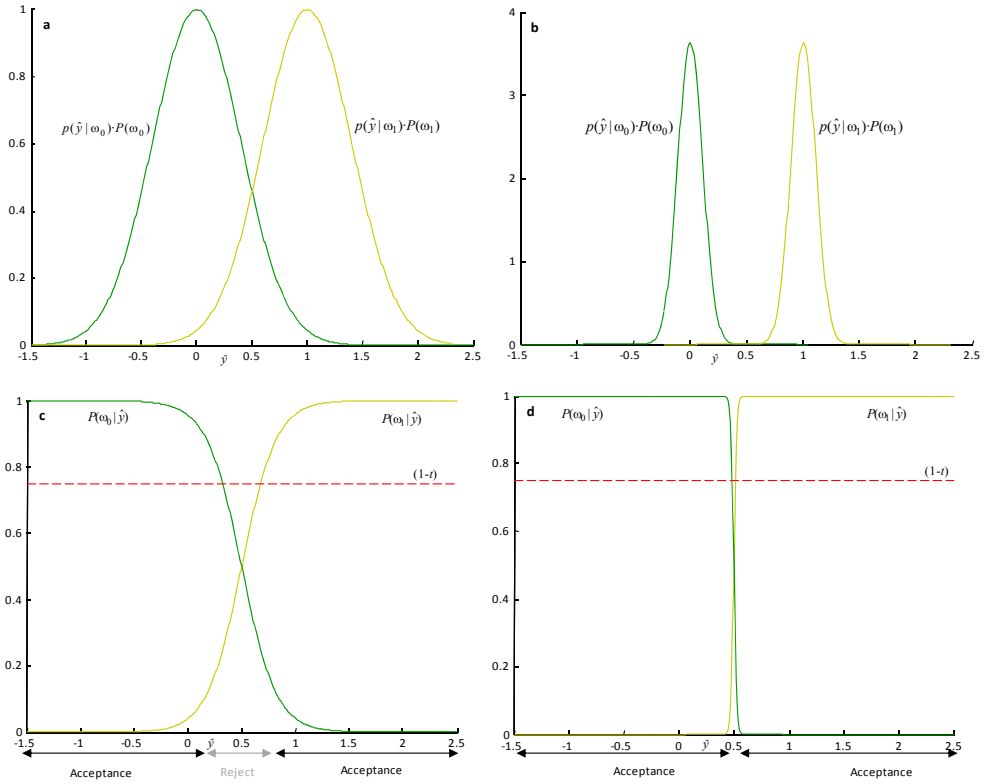


Figure 6. (a-b) PDFs for overlapped and no overlapped classes. **(c-d)** *a posteriori* probabilities with reject threshold derived from a-b PDFs.

ISBN:978-84-693-5427-8/DL:T-1418-2010

According to Chow in [23], the optimal reject threshold (t) is given by

$$t = (\lambda_r - \lambda_c)/(\lambda_m - \lambda_c) \quad (30)$$

where λ_r is the cost of rejecting a sample and λ_c and λ_m are the costs of a correct classification or a misclassification, respectively. Generally $\lambda_m > \lambda_r > \lambda_c$, and in most cases $\lambda_c=0$ (i.e. there is no cost if the classification is correct) [23, 26].

A limitation of classification based on Eq. (28) and Eq. (29) is that the threshold has no effect if the PDFs are not overlapped (Figures 6b and 6d), since there is not a significant ambiguous region. Note also that for the reject option work properly, (1–t) must be higher than 0.5. If (1–t) is lower than 0.5 the probability to classify the sample in one of the classes will always be higher than the reject threshold so that the classification rule based on *a posteriori* probabilities with reject option is simply the classical Bayes rule (see Figure 6a). In addition, the use of the *a posteriori* probability of the class and the reject threshold for rejection ignores the possibility of having samples from unknown classes. This situation may be partially overcomed by setting limits on the PDFs (High Limit and Low Limit in Figure 3 as will be discussed on chapter 4). These limits avoid classifying samples that lie on the extremes of the classes.

Other approaches have been proposed to implement rejection based on thresholds. Fumera *et al.* [22] proposed to set an individual threshold for each class, thus avoiding rejecting too many samples of one of the classes if the number of samples of both classes is not balanced. Tortorella *et al*. [19, 21] considered also two thresholds, which were optimized by maximizing the classification *utility function.* This is an alternative to the Chow's approach. Chow takes into account costs and minimizes the risk [18]. In order to optimize the reject threshold, Li *et al.* [27] proposed to control the error instead of finding a trade-off between rejection rate and error rate. They reformulated the problem as: given an error rate for each class, design a classifier with the smallest

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

rejection rule. A similar alternative was proposed by Hanczar *et al.* [28], although they controlled the conditional error rate of the classifier, not the error rate. Kressel *et al.* [29] optimized the reject threshold to get a minimal false positive rate and Herbei *et al.* [30] presented the rejection cost *t* as the upper bound on the conditional probability of misclassification, optimized by minimizing the error rate for also a minimal reject rate. These approaches often ignore the detection of outliers and the rejection of samples when the classes are not overlapped.

Further improvements on the application of the reject option are discussed in chapter 4.

3.6 Implications of reject option in classification performance evaluation

When a classifier involves the reject option, the performance measures the classifier must be properly interpreted in order to take into account that samples can be rejected.

p-DPLS is a binary classifier. This means that the classification decision is to choose between two classes, ω_1 and ω_0, that can be generically called Positive (P) class and Negative (N) class respectively. Hence, the result from *p*-DPLS can be that the sample is correctly classified in its class, either in class ω_1 (True Positive, TP, i.e., a positive sample that is classified as positive) or in class ω_0 (True Negative, TN, i.e., a negative sample that is classified as negative) or incorrectly classified, either in class ω_1 (False Positive, FP, i.e., a negative sample that is incorrectly classified as positive) or in class ω_0 (False Negative , FN, i.e., a positive sample that is classified as negative) (Table 1). When the reject option is implemented, the possible outputs of the classifier include that the sample may be rejected. A positive object that is rejected is called Reject Positive (RP) and, equivalently, a negative object that is rejected is called Reject Negative (RN). A sample may have been

rejected because its classification was not reliable enough (the risk was too high) or because it was pointed out as outlier (see chapters 4 and 5).

Table 1. Confusion matrix, outcomes of a binary classifier, as described by Kohavi and Provost in [31].

		True class	
		Positive (ω_0)	Negative(ω_1)
Predicted	Positive (ω_0)	TP	FP
	Negative (ω_1)	FN	TN
	Rejected (ω_r)	RP	RN

The objective of *p*-DPLS or any other classifier is to classify correctly as many future samples as possible, i.e., minimize the number of false positives, false negatives and rejections. For simplicity, this is generally evaluated by the accuracy or the error rate of the classification model.

Accuracy is defined as the percentage of samples correctly classified:

$$Accuracy = \frac{\mathrm{TN+TP}}{\mathrm{TN+FN+TP+FP}} \tag{31}$$

If rejection is not an option, all samples are classified and the denominator of Eq. (31) is equal to the number of samples *I* submitted to the classifier (i.e. *I* = TN+FN+TP+FP). Hence, classically, accuracy is calculated by dividing the number of samples correctly classified by the total number of samples, *I*. When rejection is an option, Eq. (31) is still valid but note that the denominator is no longer equal to the total number of samples *I*, since some of them may have been rejected (i.e. *I* = TN+FN+TP+FP+RP+RN). Hence, the accuracy must be interpreted as the percentage of correctly classified samples with respect to the number of samples for which the classifier issued a class label [22]. Note that this is the most meaningful interpretation, although it is rarelly considered in the works with reject option, in which the accuracy is calculated by dividing the number of

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

samples classified correctly by the number of samples submitted to the classifier, either rejected or not [21].

This significance resides in that the experimenter wants that the class label issued by the classifier be correct. Hence, the performance measure should reflect the percentage of the samples for which the classifier assigned a class and if it has been done correctly or wrongly. In this way, the accuracy of the classifier with reject option can be higher than the accuracy of the classifier without reject option (note that if the accuracy were defined over the total number of samples, classifiers with reject option would always performe worse than models without reject option, because the number of samples well classified using the reject option would be equal or lower).

Similarly, the error rate is defined as the percentage of samples that are assigned to the wrong class [32]:

$$\text{Error rate} = \frac{\text{FN+FP}}{\text{TN+ FN+TP+FP}} \tag{32}$$

The error rate must be also reinterpreted like the accuracy parameter when rejection is an option. Hence, the denominator of Eq. (32) is the total number of samples classified (without taking into account the rejected ones).

The sensitivity and the specificity are defined in similar terms [33].

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \tag{33}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} \tag{34}$$

The sensitivity is evaluated as the number of positive samples (class ω_1) correctly classified respect to the number of positive samples classified. Note that, while the denominator expression must be maintained, without reject option the number of

ISBN:978-84-693-5427-8/DL:T-1418-2010

positive samples classified is the total number of positive samples, with reject option, the number of positive samples classified (TP+FN) may differ from the total number of positive samples (TP+FN+RP). An analogous situation happens for the negative samples with the Specificity. In this measure, the number of negative samples classified may differ from the number of total negative samples since some of them may be rejected when the reject option is implemented.

Furthermore, when the reject option is introduced, new performance parameters appear [33]:

$$\text{Positive Predictive Value} = \frac{\text{TP}}{\text{RP}} \tag{35}$$

$$\text{Negative Predictive Value} = \frac{\text{TN}}{\text{RN}} \tag{36}$$

However, the redefinition of the performance parameters is not enough to accurately evaluate the classifiers with reject option. Note that a model that rejects to classify most of the samples but classifies correctly the remaining few will have a high accuracy; however, it is not useful. In addition, the drawback of using parameters like accuracy is that individually they are not enough to evaluate all the aspects that summarize the performance of the classifier (i.e. correct classifications, misclassifications and rejections). For that purpose, the cost is a more useful parameter. It is defined as:

$$Cost = \lambda_m N_m + \lambda_r N_r + \lambda_c N_c \tag{37}$$

where λ_m is the cost of a wrong classification, λ_r is the cost of rejecting a sample, λ_c is the cost of a correct classification and N_m, N_r, N_c are the number of samples misclassified, rejected or correctly classified, respectively. The *Cost* allows taking into account the rejections and, in addition, the cost that each classification implies [3, 17]. These costs (λ) must be optimized to keep the efficiency of the classifier.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

3.7 Conclusions

Probabilistic Discriminant Partial Least Squares (*p*-DPLS) is a binary classifier that has some advantages over other versions of DPLS: 1) it assumes neither an arbitrary classification threshold for the $\hat{y}$'s nor a Gaussian distribution for the $\hat{y}$'s of each class, and 2) it assigns the class label based on the Bayes classification rule of the *a posteriori* probability, or, more generally, of minimum risk.

However, the strict application of the Bayes rule forces the classifier to always assign the sample to one of the predefined classes. This is a limitation for those samples that may be outliers or ambiguous, and hence with a large chance to be misclassified. The danger of misclassification can be reduced by implementing the reject option. In this chapter, two approximations to implement the reject option in *p*-DPLS have been discussed. One of them introduces reject option as a reject class. The second one introduces the reject option as a threshold. The best approach to introduce the reject option is to set a reject threshold. With this approach, the *a priori* probabilities or shapes of an extra class do not need to be assumed. However, the reject option set by the reject threshold alone is not able to reject outliers; so, additional constraints must be considered.

It is also essential for any classifier to evaluate correctly the classification performance. A general approach is to use the accuracy or the error rate. These parameters, however, have the weaknesses that they consider all incorrect decisions (or correct decisions) equally risky and they treat all outcomes as equally likely [26]. Since the rejections are not evaluated, these parameters are not useful to evaluate classifiers with reject option. For such classifiers, the *Cost* parameter is a better approach.

References

[1] Pérez, N.F., J. Ferré, and R. Boqué, *Calculation of the reliability of classification in Discriminant Partial Least-Squares Classification.* Journal of Chemometrics and Intelligent Laboratory Systems, 2009. **95**: p. 122-128.

[2] Bayes, T., *An Essay towards solving a Problem in the Doctrine of Chances.* Philosophical Transactions of the Royal Society of London, 1763. **53**: p. 370-418.

[3] Chow, C.K., *An optimum character recognition system using decision functions.* IRE Trans. Electronic Computers, 1957. **16**: p. 247-254.

[4] Eriksson, L., et al., *Multi- and Megavariate Data Analysis. Principles and Applications.* 2001: Umetrics AB.

[5] Boulesteix, A.-L. and K. Strimmer, *Partial least squares: a versatile tool for the analysis of high-dimensional genomic data.* Briefings in Bioinformatics, 2007. **8**: p. 32-44.

[6] Wold, H., *Partial least squares*, in *Encyclopedia of Statistical Sciences* K.a.N.L. Johnson, Editor. 1985, Wiley: New York. p. 581-591.

[7] Gemperline, P.J., L.D. Webber, and F.O. Cox, *Raw Materials Testing Using Soft Independent Modelling of Class Analogy Analysis of Near-Infrared Reflectance.* Anal. Chem, 1989. **61**: p. 138-144.

[8] Webb, A., *Statistical Pattern Recognition, 2n edition*, ed. Wiley. 2002, Malvern, UK.

[9] Duda, R.O., P.E. Hart, and D.G. Store, *Pattern Classification (2nd edition)*, ed. W. Intersicence. 2001, New York.

[10] Landgrebe, T., et al., *The interaction between classification and reject performance for distance-based reject-option classifiers.* Pattern Recognition Letters, 2006. **27**: p. 908-917.

[11] Muzzolini, R., Y.-H. Yang, and R. Pierson, *Classifier desing with incomplete knowledge.* Pattern Recognition, 1998. **31**: p. 345-369.

[12] Botella, C., J. Ferré, and R. Boqué, *Classification from microarray data using probabilistic discriminant partial least squares with reject option* Talanta, 2009. **80**: p. 321-328.

[13] Hills, M., *Allocation Rules and their Error Rates.* Journal of the Royal Statistical Society. Series B (Methodological), 1966. **28**: p. 1-31.

[14] Bishop, C.M., *Pattern Recognition and Machine learning*, ed. Springer. 2006, New York.

[15] Denceux, T., *Analysis of evidence-Theoretic Decision rules for pattern classification.* Pattern Recognition, 1997. **30**: p. 1095-1107.

[16] Lachenbruch, P.A. and M. Goldstein, *Discriminant Analysis.* Biometrics, 1979. **35**: p. 69-85.

[17] Anderson, T.W., *Introduction to Multivariate Statistical Analysis*. 1958, New York: John Wiley and Sons.

[18] Santos-Pereira, C.M. and A.M. Pires, *On optimal reject rules and ROC curves.* Pattern Recognition Letters, 2005. **26**: p. 943-952.

[19] Tortorella, F., *An optimal reject rule for binary classifiers.* In: Ferri, F.J et al. (Eds.), Advances in Pattern Recognition: Joint IAPR International Workshops, SSPR 2000 and SPR 2000, Lecture Notes in Computer Science, vol 1876. Springer-Verlag, Heidelberg. 2000: p. 611-620.

[20] Bishop, C.M., *Pattern Recognition and Machine Learning*. Springer Science+Bussiness Media. 2006, Singapore.

[21] Tortorella, F., *A ROC-based reject rule for dichotomizers.* Pattern Recognition Letters, 2005. **26**: p. 167-180.

[22] Fumera, G., F. Roli, and G. Giacinto, *Multiple Reject Thresholds for Improving Classification Reliability*, in *Advances in Pattern Recognition*, SSPR&SPR, Editor. 2000, Springer: Berlin - Heidelberg. p. 863-871.

[23] Chow, C.K., *On optimum recognition error and reject tradeoff.* IEEE -Transactions on information theory, 1970. **16**: p. 41-46.

[24] Fumera, G., I. Pillai, and F. Roli, *Classification with Reject Option.* Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP'03), 2003.

[25] Landgrebe, T., et al. *A combining strategey for ill-defined problems*. in *Fifteenth Ann. Sympos. of the Pattern Recognition Association of South Africa*. 2004.

[26] Brown, C.D. and H.T. Davis, *Receiver operating characteristics curves and related decision measures: A tutorial.* Chemometrics and Intelligent Laboratory Systems, 2006. **80**: p. 24-38.

[27] Li, M. and I.K. Sethi, *Confidence-based classifier design.* Pattern Recognition, 2006. **39**: p. 1230-1240.

[28] Hanczar, B. and E.R. Dougherty, *Classification with reject option in gene expression data.* Bioinformatics, 2008. **24**: p. 1889-1895.

[29] Kressel, U., F. Lindner, and C. Wöler, *Classification System with reject class*. 2004, DaimlerChrysler AG (DE): United States.

[30] Herbei, R. and M.H. Wegkamp, *Classification with reject option.* The Canadian Journal of Statistics, 2006. **34**: p. 709-721.

[31] Kohavi, R. and F. Provost, *Glossary of Terms* Machine Learning - Kluwer Academic Publishers, 1998. **30**: p. 271-274.

[32] Smith, C.A.B., *Some examples fo discrimination.* Ann. Eugen., 1974. **13**: p. 272-282.

[33] Bradley, A.P., *The use of the area under the ROC curve in the evaluation of machine learning algorithms.* Pattern Recognition, 1997. **30**: p. 1145-1159.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

CHAPTER 4 Classification from microarray data using *p*-DPLS with reject option

Talanta, 2009, Vol.80 (1): 321-32

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Microarrays allow evaluating simultaneously the expression of thousands of genes in a cell. One of the most relevant applications of these gene expressions is to classify the samples (e.g. cell or tissues) into one of the classes of interest. Discriminant Partial Least-Squares (DPLS) is often used for such a purpose. However, most published results report the straight application of this method, with disregard to the quality of each individual prediction and the possibility of detecting prediction outliers. The aim of this chapter is to improve DPLS for classifying microarray data. Firstly, we implement a new version of DPLS called *probabilistic Discriminant Partial Least Squares* (*p*-DPLS). This method bases the classification of a sample on kernel probability density functions (PDFs) and the Bayes rule of *a posteriori* probability. Secondly, a reject option is introduced so that the classifier can reject samples in the ambiguity region, based on Chow's rule, and can reject samples outside the defined limits of the classes. The ambiguity region is the zone where the PDFs that characterize each one of the classes overlap. In that zone, the model cannot discriminate well enough whether a sample belongs to one class or to the other, either because of limitation of the PLS model, or because the samples actually share characteristics of the modeled classes. Hence, there is high risk that any attempt of classifying that sample could result in a misclassification. The second possibility of rejection is implemented at the ends of the classes' domains and also between PDFs for non overlapped classes. Samples in those regions have extreme predictions, outside the limits set for the classes, so they may be considered as outliers. For such samples, we prefer to reject to classify them instead of taking the risk of misclassifying them. These two approaches will be detailed and discussed in the methods section.

The existence of a reject option increases the experimenter's confidence in the classification rule and improves the accuracy of the final classification models. Note that with reject option only those samples whose classification is reliable are actually classified, while the samples either outside the limits or in the ambiguity region that could lead to misclassifications are rejected to classify.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

The *p*-DPLS with reject option was tested with two public datasets. With the Human Cancers dataset, the accuracy measured by leave-one-out cross-validation was improved from 97% to 99% when compared to *p*-DPLS without reject option. For the Breast Cancer dataset, the method could reject 100% of the test samples submitted to the classifier that did not belong to any of the modelled classes. These samples would have been misclassified if the reject option had not been considered.

This work is presented in paper form published in *Talanta* 2009, Vol. 8 (1) 321-328.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Classification from microarray data using *probabilistic* discriminant partial least squares with reject option

Cristina Botella, Joan Ferré*, Ricard Boqué
Department of Analytical Chemistry and Organic Chemistry, Rovira i Virgili University.
Marcel·lí Domingo s/n, 43007. Tarragona, Spain
*Corresponding author: joan.ferre@urv.cat

***Talanta* 2009, Vol. 8 (1) 321-328** (Edited for format)

Abstract

Microarrays are used to simultaneously determine the expressions of thousands of genes. An important application of microarrays is in the classification of samples into classes of interest (e.g. either healthy cells or tumour cells). Discriminant Partial Least-Squares (DPLS) has often been used for this purpose. In this paper, we describe an improvement to DPLS that uses kernel-based probability density functions and the Bayes rule to classify samples whilst keeping the option of not classifying the sample if this cannot be done with sufficient confidence. With this approach, those samples outside the boundaries of the known classes or from the ambiguity region between classes are rejected and only samples with a high probability of being correctly classified are indeed classified. The optimal model is found by simultaneously minimizing the misclassification and rejection costs. The method (*p*-DPLS with reject option) was tested with two datasets. For the Human Cancers dataset the accuracy (obtained by leave-one-out cross-validation) was improved from 97% to 99% when compared to *p*-DPLS without reject option. For the Breast Cancer dataset, *p*-DPLS with reject option was able to reject 100% of the test samples that did not belong to any of the modelled classes. These samples would have been misclassified if the reject option had not been considered.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

4.1 Introduction

Supervised classification is increasingly being applied to microarray gene expression data in order to predict tumour types [1-3], to differentiate between healthy and tumour samples [4-6] and to differentiate between pharmacological mechanisms [7], among other applications. Microarray data are characterized by thousands of variables (genes) and few samples, resulting in high redundancy and a high number of non-informative measurements. There has been a lot of interest in using factor-based multivariate classification methods such as Discriminant Partial Least Squares (DPLS) to analyze these data [8, 9]. The DPLS uses a few latent variables rather than a lot of measured variables and this brings with it a series of advantages. DPLS takes variable correlations into account, filters noise and leads to classification rules with good predictive performance, especially when DPLS is implemented together with variable selection methods. DPLS has been used to differentiate between samples before and after chemotherapy [10], to determine the different states of a breast cancer tumour [11], to predict the efficacy of a drug by using expression data biomarkers [12], and to predict the quality of DNA-microarray spots [13].

Like other classification rules, DPLS must have two main qualities: it must provide reliable classifications of forthcoming samples and it must minimize the number of misclassifications (i.e. the expected error rate). Both of these are improved if the classifier is allowed to reject doubtful samples instead of always being forced to classify them in one of the modelled classes. By classifying only the most well-defined cases, both the accuracy of the classifier and the reliability of each classification are improved.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

In this paper we implement the reject option in the recently developed *p*-DPLS classifier and show how it can be used for microarray data classification. *p*-DPLS is a variant of DPLS, which uses kernel functions to calculate a probability density function (PDF) for each class. This allows a flexible implementation of the Bayes rule for classification, and also provides a measure of the reliability of the classification. Reliability is a primary concern in statistical classification, especially when this classification is used in critical health applications such as cancer diagnosis [14], an issue which has also led to several other studies [15].

In classification, rejection is advantageous when: (a) the new sample does not belong to any of the trained classes, (b) the new sample belongs to one of the classes but is very different from the samples used for training the classifier, or (c) the sample is in the boundary region between classes. Situation (a) occurs when the sample is an outlier. Forcing the classifier to decide among one of the modelled classes will produce a classification error (e.g. a cell does not belong to any of the modelled cell types but it is classified as one of them). Situation (b) typically arises when the sampling of the training samples is incomplete or not representative. Finally, situation (c) may arise, for example, because of the limited discriminative power of the measured variables or because the classification algorithm has limited discriminative power. Although samples in situations (b) and (c) might finally be classified correctly, they might also be classified incorrectly because either they are unique samples or they are ambiguous samples and can belong to either of the classes, respectively.

The reject option aims to overcome situations (a) to (c) by rejecting the sample and not classifying it when the probability of error is too high. This is a safeguard against errors and improves the accuracy of the classifier, which is evaluated as the percentage of samples correctly classified among the number of samples classified [16]. This in turn leads to greater confidence in the samples that are finally classified. The reject option

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

can be fine-tuned in order to avoid rejecting too many samples that would otherwise be classified correctly. Since too high a rejection rate would decrease the usefulness of the classifier, a compromise must be reached between improving the accuracy and reducing the usefulness of the classifier. There has been extensive research into the theoretical aspects of the reject option [14, 17-28], most of which relates to Chow's reject option [29], which implemented the reject option for the Bayes rule. Chow's reject option has recently been used to microarray expression data [30].

There are still two limitations to jointly applying the Bayes and Chow rules. First, they are not adequate for the extreme (outlying) samples (situations (a)-(b)) which are typically found at the extremes of the probability density functions (PDFs). These samples must be rejected according to a different criterion. Second, both rules require knowledge of the *a priori* probabilities and the PDFs of the classes [31], which makes applying these rules more difficult. In this paper, the first limitation is overcome by including distance based thresholds, which is equivalent to selecting a confidence interval around each class and rejecting samples outside this interval [17]. The second limitation is overcome by the calculating PDF-like functions in *p*-DPLS [32], which makes an approximate Bayesian classification easier.

4.2 Methods

4.2.1 *Probabilistic* DPLS

The DPLS method applies Partial Least-Squares (PLS) regression to binary classification problems, in which the dependent variable *y* codifies the class of each sample [8, 33]. A DPLS model is calculated by regressing **y** on **X** using the adequate number of factors.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

For microarray gene expression data, **X** is an *N*×*P* matrix of *N* samples and *P* gene expressions and **y** is a *N*×1 vector of ones and zeros, where the integer 0 codifies the sample as belonging to class ω_0 (e.g. "cancer of type I") and the integer 1 codifies the sample as belonging to class ω_1 (e.g. "cancer of type II"). For a sample *i*, the value predicted by the PLS model is $\hat{y}_i = \mathbf{x}_i^T\mathbf{b}$, where the *b*'s are the regression coefficients for the model of *A* factors and the adequate pre-processing is implicit (e.g. if the *b*'s had been calculated from mean-centered data, then $\mathbf{x}_i$ should be mean-centered, and the predicted $\hat{y}_i$ should be unprocessed accordingly). With the coding of *y*, the prediction for a sample should be close to 0 if the sample belongs to class ω_0, and it should be close to 1 if the sample belongs to class ω_1. In order to better define the cut-off value between classes, Pérez *et al*. [32] developed *p*-DPLS, a probabilistic version of the DPLS in which the uncertainty of the predicted value $\hat{y}$ is accounted for in the calculation of the model. This method is described here for completeness. The method starts by calculating a DPLS model of *A* factors with **X** and **y**. Then, this model is used to predict the training samples and, for each training sample *i*, a Gaussian function centred at the predicted value $\hat{y}_i$ is calculated as:

$$F(\hat{y}_i) = \frac{1}{SEP_i\sqrt{2\pi}} \cdot e^{\frac{1}{2}\left(\frac{y-\hat{y}_i}{SEP_i}\right)^2} \quad (1)$$

$$SEP_i = \mathrm{RMSEC}\sqrt{1+h_i} \quad (2)$$

$$\mathrm{RMSEC} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n-A-\delta}} \quad (3)$$

where SEP_i is the standard error of prediction for sample *i*, h_i is the leverage of the sample, RMSEC is the root mean square error of calibration, y_i is the known class of the training sample *i* (i.e. value 0 for a sample of class ω_0 and value 1 for a sample of class ω_1) and δ is 1 if the data has been centred and 0 if it has not. Figure 1 shows the

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Gaussian functions calculated from the predictions of three training samples of class ω_0 and four samples of class ω_1. Note that the width of the Gaussian kernel for sample *i* depends on SEP_i, which is particular to that sample, and depends on the relative position of the sample in the multivariate space. Then, for classes ω_0 and ω_1, a PDF is calculated as the average of the individual kernel functions of the training samples of each class:

$$p(\hat{y}|\omega_0) = \frac{1}{n_0}\sum_{i=1}^{n_0} f_i(\hat{y}) \quad (4)$$

$$p(\hat{y}|\omega_1) = \frac{1}{n_1}\sum_{i=1}^{n_1} f_i(\hat{y}) \quad (5)$$

where n_0 and n_1 are the number of samples of class ω_0 and class ω_1 respectively.

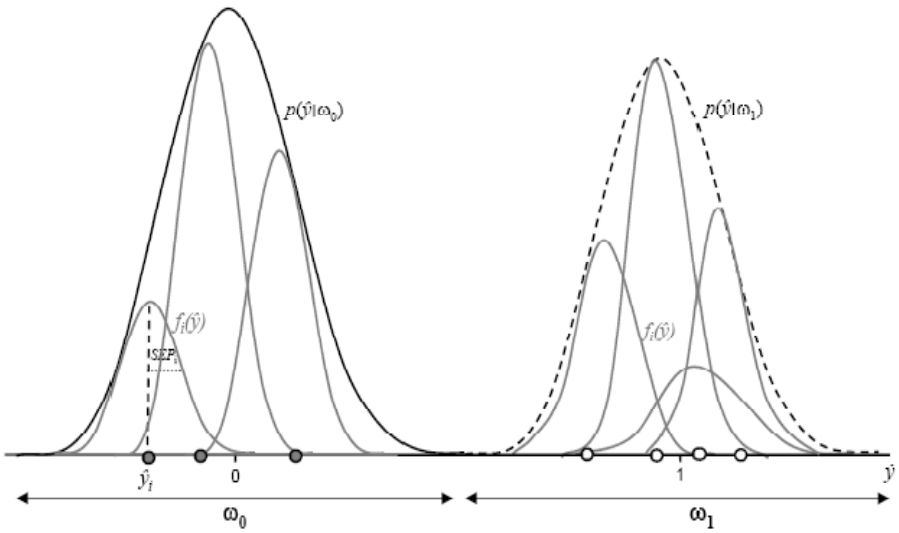


Figure 1. Simulated PDFs of class ω_0 and class ω_1 obtained from Equations (4) and (5). The kernel functions (Eq. (1)) are centred on prediction $\hat{y}_i$ of each training sample. According to the code assigned to the classes, the sample predictions of class ω_0 and class ω_1 should be located around the values 0 and 1 respectively.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

4.2.2 Bayes rule for classification

In *p*-DPLS, the predicted class for sample *i* is obtained using the Bayes rule. The prediction $\hat{y}_i$ for that sample is used to obtain the *a posteriori* probabilities $P(\omega_0|\hat{y}_i)$ and $P(\omega_1|\hat{y}_i)$. These are the probabilities that the sample belongs either to class ω_0 or to class ω_1, once it is known that the sample's prediction is $\hat{y}_i$. For the two-class classification problem:

$$P(\omega_0|\hat{y}_i) = \frac{p(\hat{y}_i|\omega_0)P(\omega_0)}{p(\hat{y}_i)} \quad (6a)$$

$$P(\omega_1|\hat{y}_i) = \frac{p(\hat{y}_i|\omega_1)P(\omega_1)}{p(\hat{y}_i)} \quad (6b)$$

where $p(\hat{y}_i|\omega_0)$ and $p(\hat{y}_i|\omega_1)$ are the *conditional* probabilities evaluated from the PDFs of classes ω_0 and ω_1 and $P(\omega_0)$ and $P(\omega_1)$ are the *a priori* probabilities. Both *a priori* probabilities may be estimated as the proportion of samples of each class in the training set, provided that the set is representative of the overall population. That is, $P(\omega_0)=n_0/N$ and $P(\omega_1) = n_1/N$ where $N=n_0+n_1$. The denominator of Equation (6a) and (6b) is:

$$p(\hat{y}_i) = p(\hat{y}_i|\omega_0)P(\omega_0) + p(\hat{y}_i|\omega_1)P(\omega_1) \quad (7)$$

The Bayes rule assigns the sample to the class in which it has the highest *a posteriori* probability [31]. The rule is:

$$\begin{aligned} &\text{Assign the sample to} \\ &class\ \omega_0\ \ if\ P(\omega_0|\hat{y}_i) > P(\omega_1|\hat{y}_i) \\ &class\ \omega_1\ \ if\ P(\omega_1|\hat{y}_i) > P(\omega_0|\hat{y}_i) \end{aligned} \quad (8)$$

Although this rule is optimal in the sense that no other rule can yield a lower error probability, it is not always satisfactory. For example, when the $\hat{y}_i$ is at one of the extremes of the PDF (Figure 2), both $p(\hat{y}_i|\omega_0)$ and $p(\hat{y}_i|\omega_1)$ are low, and the products $p(\hat{y}_i$

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

$|\omega_0)\cdot P(\omega_0)$ and $p(\hat{y}_i|\omega_1)\cdot P(\omega_1)$ are also low but the *a posteriori* probability for one of the classes (the ratio in Equations 6a and 6b) is high. This means that the further the prediction $\hat{y}_i$ is from one class, the more likely it will be allocated to the other class. This is a reasonable result since the classifier only expects to receive samples from the two modelled classes. In most multivariate applications, however, samples from non-modelled classes (outliers) may also be inadvertently submitted to the classifier. The predictions for those samples will most probably be found at the tails of a PDF, and, hence give a misleading high *a posteriori* probability for one of the classes. Consequently, forcing the two-class Bayes rule to classify any input sample may involve a high risk because outliers may be erroneously classified in one of the modelled classes.

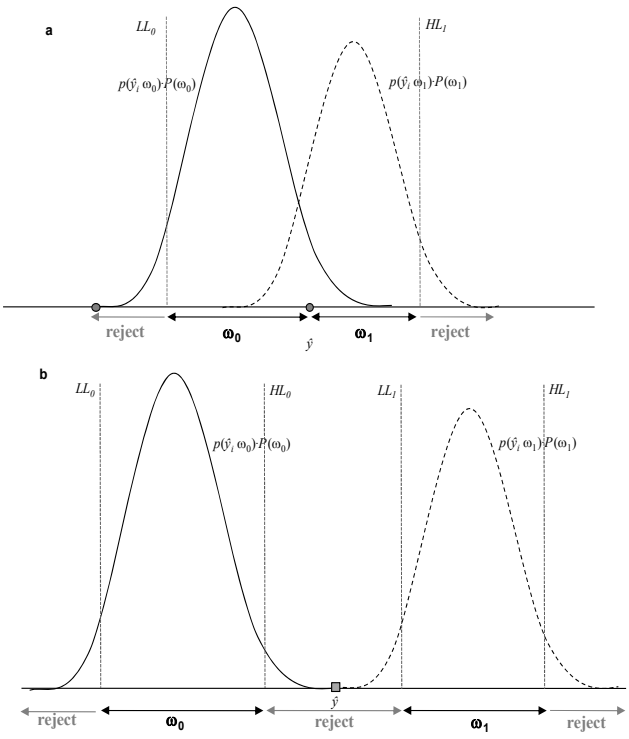


Figure 2. Possible distributions of class ω_0 and class ω_1 with the distance reject limits. **a.** Overlapped classes. **b**. Well separated classes. ● and □ indicate possible predictions of unknown samples for which the Bayes rule gives questionable results. LL_0, HL_0, LL_1 and HL_1 are the limits for rejection based on the distance reject option.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Another situation in which the usefulness of the Bayes rule is limited occurs when the predicted value $\hat{y}$ is in the boundary between classes (the ambiguity region). The dot in the centre of Figure 2a represents a sample whose characteristics are similar for both classes, with the result that the model cannot clearly distinguish whether it belongs to one class or to the other. Again, the sample will be assigned to the class to which, according to the Bayes rule, it has the highest probability of belonging. However, since the probability that the sample belongs to class ω_0 is similar to the probability that it belongs to class ω_1, there is a high risk of misclassification and the reliability of the classification is low.

These situations show that the reject option might be an advantageous addition to the decision rule. In this paper, we implement the reject option in *p*-DPLS. Both the classification reliability and accuracy of the *p*-DPLS model are improved by identifying unreliable classifications and rejecting the sample instead of running the risk of misclassifying it.

4.2.3 Implementation of the reject option in *p*-DPLS

The reject option in the case of classification ambiguity (i.e. for overlapped PDFs) can be derived by adapting Chow's rule to the PDFs obtained in *p*-DPLS. Chow's rule sets a threshold *t* so that the sample is rejected if the highest *a posteriori* probability is lower than (1– *t*). In other words, the sample is classified only if:

$$max\left(P(\omega_0|\hat{y}_i), P(\omega_1|\hat{y}_i)\right) > (1-t) \quad (9)$$

Thus, only those samples whose classification is reliable enough are indeed classified. The other samples are rejected because they could be misclassified. The threshold that

optimizes the trade-off between the error rate and reject rate can be derived from the costs associated with each classification result [29]:

$$t = (\lambda_r - \lambda_c)/(\lambda_m - \lambda_c) \quad (10)$$

where λ_m, λ_r, and λ_c, are the costs of incorrect classification, of rejection and of correct classification, respectively. The values that are assigned to these costs make the reject option tuneable. The cost of being wrong is higher than the cost of both rejecting and classifying correctly ($\lambda_m > \lambda_r > \lambda_c$). In fact, it is preferable to reject a sample and gather additional information than to classify the sample incorrectly. It is also generally assumed that classifying correctly has no cost ($\lambda_c = 0$). Note that Equation 9 is a generalization of the standard Bayes rule. In particular, for the extreme case in which the cost of rejection λ_r equals the cost of misclassification λ_m, the reject threshold is t = 1 and Chow's rule is reduced to the standard Bayes rule, in which samples are never rejected. A sample is also not rejected if $t > 1/C$, where C is the number of possible classes (C=2 for a binary classification) [34].

The second reason for using the reject option is to avoid classifying extreme samples that have a large *a posteriori* probability but low values at both PDFs. In order to solve this problem, Dubuisson and Masson [18] added a distance reject criterion to Chow's ambiguity reject option. This idea is implemented here for the *p*-DPLS model by imposing limits on the $\hat{y}$ values, which define the extreme regions in which the samples will be rejected. The limits are chosen so that the sum of the area in the tails of each PDF is five percent of the total area of the distribution (i.e. the distance reject probability equals 0.05 for each class, see Figure 2) [19]. Since the limits depend on the shape of the distributions of each class, they are particular for each *p*-DPLS model with a given number of factors. In practice, when the PDFs are overlapped we have two

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

operative limits, a High Limit (HL) and a Low Limit (LL) and when the PDFs are separated we have four limits (HL and LL for each class, see Figure 2).

Assuming the constraints for the distance reject and the ambiguity reject, the Bayes rule with reject option is:

$$\begin{aligned}
&\text{Reject} \quad if\ \hat{y}_i < \mathrm{LL}_0 \\
&\qquad or\ if\ (\mathrm{HL}_0 < \hat{y}_i < \mathrm{LL}_1) \\
&\qquad or\ if\ \hat{y}_i > \mathrm{HL}_1 \\
&\qquad or\ if\ \ max\left(P(\omega_0|\hat{y}_i), P(\omega_1|\hat{y}_i)\right) < (1-t) \\
&\text{Otherwise} \\
&classify\ into\ \omega_0\ \ if\ P(\omega_0|\hat{y}_i) > P(\omega_1|\hat{y}_i) \\
&\qquad or\ into\ \ \omega_1\ \ if\ P(\omega_1|\hat{y}_i) > P(\omega_0|\hat{y}_i)
\end{aligned} \tag{11}$$

4.2.4 Evaluation of the classification method performance

The *p*-DPLS models can be calculated for a different number of factors that are needed to explain the relevant information. Thus, every *p*-DPLS model will produce different $\hat{y}$ predictions for the calibration samples and, therefore, for the different PDFs, which, in turn, will influence the performance of the classifier. The performance of a classifier is commonly characterized by its error rate (or the classification rate, which is the percentage of correctly classified samples) when classifying a test set of unseen samples that were not used during the training phase. The actual class of every sample in the test set is compared to the class to which it is assigned by the classifier. In general terms, however, it is not the misclassification (and rejection) rate that we want to minimize, but the misclassification (and rejection) cost [35], since the cost more accurately reflects the objective of the classification rule [36]. The *Cost* is here defined as:

$$Cost = \lambda_m \mathrm{N}_m + \lambda_r \mathrm{N}_r \quad (12)$$

where N_r is the number of rejected samples and N_m is the number of misclassified samples. The cost of correctly classifying a sample has been set to zero. Here, the minimization of the cost will be used to decide on the optimal number of factors in the *p*-DPLS model.

4.3 Results and Discussion

4.3.1 Datasets

The proposed classification rule (Eq. 11) was applied to two datasets, the Human Cancers dataset [37] and the Breast Cancer dataset [38]. These datasets have been studied extensively in the literature [39, 40] and also used to evaluate the performance of classification models [41-44]. The Human Cancers dataset consists of 282 microRNA (miRNA, non coding RNA species) normalized expression profiles for 218 samples, including 46 healthy samples (class ω_0) and 172 tumour samples (class ω_1) from several healthy and tumour tissues (ovary, colon and lung to mention a few). The dataset was divided into a training set and a test set by applying the Kennard-Stone algorithm [45] to the scores of the first 20 Principal Components (PCs), which were obtained from the Principal Component Analysis (PCA) of the raw gene expression matrix. For this dataset, the training set contained 153 samples (116 samples of class ω_1 and 37 samples of class ω_0), and the test set had 65 samples (56 of class ω_1 and 9 of class ω_0). The Breast Cancer dataset consists of 5361 normalized gene expression ratios. These were used in [38] to prove that a hereditable mutation influences the gene expression profile of breast cancer. Seven samples of the BRCA1 mutation were used as class ω_0, eight samples of BRCA2 mutation were used as class ω_1, and six samples of Sporadic mutation were used as test samples.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

4.3.2 Human Cancers dataset

Although *p*-DPLS is a full-variable method, it can often be improved by carefully selecting the variables and removing irrelevant miRNA expressions that interfere with the discriminative power of the relevant miRNA [46]. For this dataset, the 100 variables with the highest VIP values (variable importance for the projection) were considered. VIP values were calculated as described in [8, 47]. These values quantify how each variable influences the response summed over all components and classes.

For the selected variables, six *p*-DPLS models were calculated with 1 to 6 factors using mean-centered miRNA expression patterns (we will denote each model as $p\text{-DPLS}_A$, where A is the number of factors). The *a priori* probabilities for these six models were $P(\omega_0) = 37/153 = 0.24$ and $P(\omega_1) = 116/153 = 0.76$.

The PDF of classes ω_0 and ω_1 were calculated for each *p*-DPLS model (Eqs. 1 to 5). The test sample was classified by obtaining its $\hat{y}$ prediction and then calculating the *a posteriori* probabilities (Eqs 6a, 6b). Finally, the sample was either rejected or classified in the class with the highest *a posteriori* probability (Eq. 11). In this dataset, the high and low limits (HL and LL) for $\hat{y}_i$ were defined so as to retain the five percent of the total area of the PDF in the tails of the distributions. The costs were arbitrarily set to $\lambda_c = 0$, $\lambda_r = 0.25$, and $\lambda_m = 1$ because no information was available about the costs of each classification decision. Note that these costs are relative, and indicate that it is preferable to reject four samples than to classify one wrongly. These values are illustrative and should be adjusted for each particular classification problem. With these values, the threshold value for rejection in the ambiguity zone is $t = 0.25$ (Eq. 10). The models with A=1 to A=6 factors were validated by leave-one-out cross-validation (CV). In this process, sample *i* was left out of the training set, the $p\text{-DPLS}_A$ model was calculated, and the prediction $\hat{y}_i$ for the left-out sample was obtained (note that the *a priori* probabilities were recalculated to take into account that one sample had been

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

left out). This procedure was repeated for all the samples of the training set and for all the *p*-DPLS models.

Figure 3 and Table 1 show the cross-validation results obtained for the different *p*-DPLS models when the reject option (Eq. 11) is considered. Note that predictions for samples in class ω_0 are around 0 and predictions for samples in class ω_1 are around 1, but that the predictions partially overlap in models with less than four factors (underfitted models). As a result of the overlap, many samples are either rejected or wrongly classified and the cost of these models (Table 1) is high. For example, for *p*-DPLS$_2$, 51% of the samples in class ω_0 were rejected by CV and 27% were misclassified. On the other hand, the predictions from the models with four to six factors are grouped tighter together. Consequently, these models have fewer misclassifications, fewer rejections, and lower classification costs.

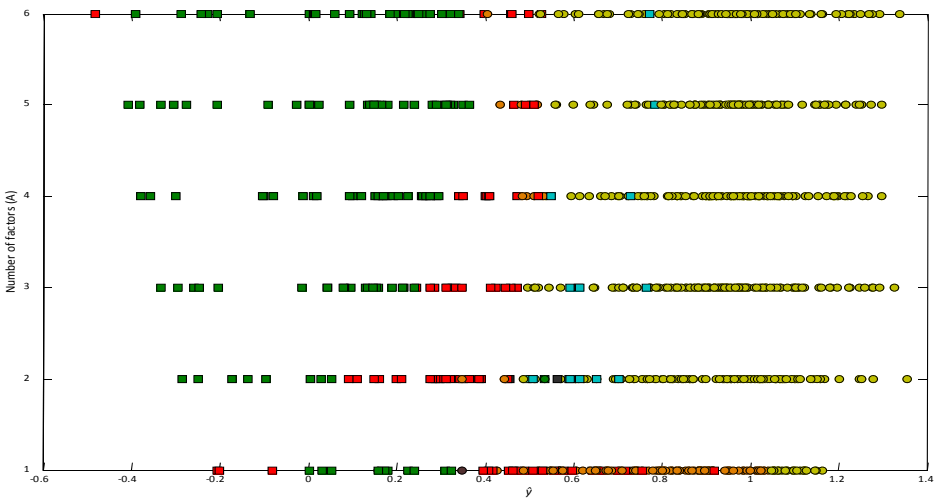


Figure 3. Prediction of the training samples by CV for the different *p*-DPLS models with reject option. Squares: healthy samples (class ω_0), – Green: correctly classified, Blue: misclassified, Red: rejected to classify–. Circles: tumour samples (class ω_1), – Yellow: correctly classified, Brown: misclassified, Orange: rejected to classify–.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Table 1. Classification of validation samples by leave-one-out cross-validation and test samples for the different *p*-DPLS models for t =0.25. In brackets, classifications performed without considering the reject option.

Factors	Cost	Cross-Validation Samples					
		Wrongly classified	Rejected	Correctly classified	Wrongly classified	Rejected	Correctly classified
		Samples of class ω_0			Samples of class ω_1		
1	30.3	0 (6)	24	10 (31)	2 (19)	89	25 (97)
2	15.5	10 (22)	19	8 (15)	0 (0)	3	113 (116)
3	7.3	4 (9)	13	20 (28)	0 (0)	0	116 (116)
4	5	2 (4)	8	27 (33)	0 (0)	4	112 (116)
5	3	1 (4)	4	32 (33)	0 (1)	4	112 (115)
6	3	1 (5)	7	29 (32)	0 (1)	1	115 (115)

Factors	Test Samples					
	Wrongly classified	Rejected	Correctly classified	Wrongly classified	Rejected	Correctly classified
	Samples of class ω_0			Samples of class ω_1		
1	0 (2)	6	3 (7)	0 (0)	17	39 (55)
2	2 (4)	5	2 (5)	0 (0)	0	56 (56)
3	0 (2)	4	5 (7)	0 (0)	0	56 (56)
4	0 (0)	2	7 (9)	0 (0)	0	56 (56)
5	0 (0)	0	9 (9)	0 (0)	0	56 (56)
6	0 (0)	0	9 (9)	0 (0)	0	56 (56)

In terms of classification cost, the optimal model is p-DPLS$_5$ since no further improvement is obtained for the model of six factors. The PDFs for this model are presented in Figure 4a and the a posteriori probabilities across the $\hat{y}$ domain are presented in Figure 4b. The limits were found to be $LL_0 = -0.43$ and $HL_1 = 1.42$. Thus, samples with a predicted value $\hat{y}_i < -0.43$ or $\hat{y}_i > 1.42$ would be flagged as outliers and rejected. These limits were different for each p-DPLS$_A$ model because the training sample predictions changed. According to the rejection criterion, eight samples (four from class ω_0 and four from class ω_1) were rejected, all of them in the ambiguity region (Table 1). As an example, the dot in Figure 4 corresponds to the sample T_BRST_2 (tumour sample, class ω_1) during the leave-one-out process. The prediction is $\hat{y}_i = 0.44$ and the calculated *a posteriori* probabilities are $P(\omega_0 | \hat{y}_i) = 0.59$ and $P(\omega_1 | \hat{y}_i) = 0.41$ (Eqs. 6a, 6b). Since both probabilities are similar, the confidence (reliability) that the classification is correct is low because a slight shift in $\hat{y}_i$ due to measurement errors could have changed the assigned class. The application of the classic Bayes rule (Eq. 8) would assign the sample to the class with the highest *a posteriori* probability, meaning that the sample would be wrongly classified into class ω_0. By allowing the reject option, defined here by Chow's rule (with t=0.25), the sample was rejected and not classified because the highest *a posteriori* probability was below $1-t$ (i.e. $\max(P(\omega_1 | \hat{y}_i), P(\omega_0 | \hat{y}_i) < 0.75$). In this case, the reject option prevented us from classifying a tumour sample as a healthy sample, and the expert would be prompted to make more tests before the final diagnosis. It is interesting to note, as we indicated before, that the reject option's performance depends on the relative costs assigned to the classification results. Thus, by setting different costs, the threshold (and hence the number of samples rejected) will be tuned to meet the experimenter's needs.

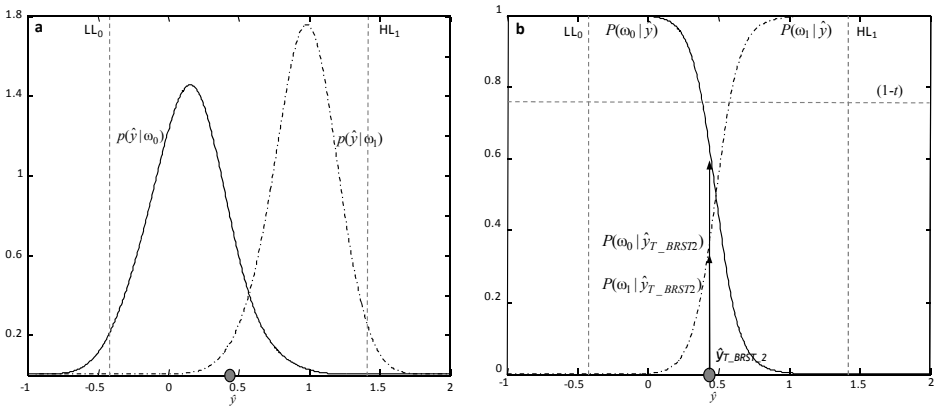


Figure 4. a. PDFs for the five factor *p*-DPLS model obtained from the training samples during the LOOCV process when T_BRST_2 is used as the validation sample. **b.** *A posteriori* probabilities across the $\hat{y}$ domain (Eq. 6a and 6b) derived from the PDFs in **a-b.** The prediction and the *a posteriori* probability for sample T_BRST_2 during the LOOCV process are also shown.

For comparison, Table 1 shows in brackets the classification results when the classical Bayes rule is applied. For the model that best minimizes the cost, that is, *p*-DPLS$_5$, five samples were misclassified if the reject option was not applied, whereas only one sample was misclassified (a healthy sample) when the reject constraints were applied. Thus, this model's classification accuracy (i.e. the ratio of samples well classified and the number of samples classified) was improved from 97% (148/153) to 99% (144/148). Notice, however, that the reject option also rejected some samples that would otherwise be correctly classified: the number of samples well classified decreased from 148 to 144. This reduction in the number of well classified samples is the price to pay for safeguarding against errors, and follows the trend of the suggested costs of classifications, in which rejecting four samples was preferable to misclassifying one.

Different reject thresholds were tested by varying the classification costs (Table 2). When λ_r=0.10, λ_m=1 and λ_c=0, the threshold was *t*=0.10. As expected, the number of rejected samples increased because the cost of doing so decreased (i.e. we preferred

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

to reject ten samples rather than classify one wrongly). However, the number of misclassified samples did not change, which means that the rule rejected samples that would have been correctly classified with t=0.25. Thus, decreasing t to below 0.25 did not improve the model's classification performance for this dataset. On the other hand, when t was set to 0.35 (i.e. λ_r=0.35, λ_e=1, λ_c=0), the results (not shown) were the same as those obtained for t = 0.25. Hence, t = 0.25 was considered optimal for this p-DPLS$_5$ model.

The samples of the test set were also classified according to Eq. 11. For the p-DPLS$_5$ model using t=0.25, 100% of the samples were well classified and there were no rejects (Figure 5 and Table 1). By setting the threshold to t=0.10, two correctly classified healthy samples turned into rejects (Table 2). This was seen in the classification of the training samples above and highlights the need to set an adequate reject threshold in order to obtain an adequate trade-off between the rejects and the misclassifications. This will depend on the needs of the experimenter and the cost constraints in each particular application.

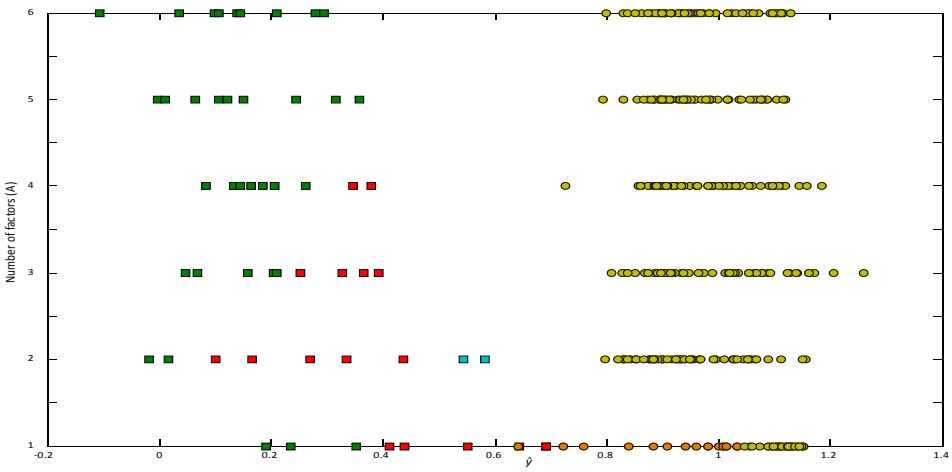


Figure 5. Classification of test samples for the different p-DPLS models with reject option. Squares: healthy samples (class ω_0), – Green: correctly classified, Blue: misclassified, Red: rejected to classify–. Circles: tumour samples (class ω_1), – Yellow: correctly classified, Brown: misclassified, Orange: rejected to classify–.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Table 2. Classification of validation samples via leave-one-out cross-validation and test samples for the different *p*-DPLS models for *t* =0.10.

Factors	Cost	Cross-Validation Samples					
		Wrongly classified	Rejected	Correctly classified	Wrongly classified	Rejected	Correctly classified
		Samples of class ω_0			Samples of class ω_1		
1	16.1	0	36	1	1	115	0
2	6.5	2	31	4	0	14	102
3	5.9	3	23	11	0	6	110
4	3.5	1	17	19	0	8	108
5	3.3	1	14	22	0	9	107
6	2.9	1	13	23	0	4	110

Factors	Test Samples					
	Wrongly classified	Rejected	Correctly classified	Wrongly classified	Rejected	Correctly classified
	Samples of class ω_0			Samples of class ω_1		
1	0	9	0	0	56	0
2	0	9	0	0	0	56
3	0	7	2	0	0	56
4	0	3	6	0	0	56
5	0	2	7	0	0	56
6	0	2	7	0	0	56

4.3.3 Breast Cancer dataset

This dataset demonstrates the rejection of test samples that are outside the class limits. The same methodology as for the Human Cancers dataset was used except that the Kennard Stone algorithm was not applied. Instead, the samples of mutations BRCA1 and BRCA2 were used as a training set and the Sporadic mutation samples were used as a test set. The aim was to show that the classification rule could reject prediction samples from non-modelled classes. This would prevent the classification error that would otherwise occur if the classifier had to assign the samples to one of the two modelled classes. Detecting this type of outlier is fundamental to the application of any classification rule.

Probabilistic DPLS models were calculated for one to three factors by using $\log_2$ mean-centred gene expression data from BRCA1 (class ω_0) and BRCA2 (class ω_1) mutation samples. This data consisted of the 51 most relevant gene expressions according to [38]. These genes were found to be the most discriminative between the three mutations. The costs of classifying correctly, rejecting and misclassifying were set at $\lambda_c = 0$, $\lambda_r = 0.25$, and $\lambda_m = 1$ respectively. The one factor *p*-DPLS model (*p*-DPLS$_1$) was the optimal model with the lowest cost (i.e. cost of 0.5). Models with two and three factors were overfitted, with costs of 2.5 and 3.25 respectively. These higher costs are due to the fact that most of the samples are rejected and, although there are no misclassifications, the classifiers become useless. For example, for the *p*-DPLS$_2$ model, 10 of the 15 training samples were rejected during LOOCV. Similarly, the *p*-DPLS$_3$ model rejected 13 of the training samples.

The *p*-DPLS$_1$ calculated with the 51 gene expressions selected in the bibliography was able to distinguish the samples of class ω_0 from those of class ω_1, thus providing well separated PDFs (Figure 6). Only the sample *s1252_P2*, of class ω_0, and the sample *s1816_P13*, of class ω_1, were rejected during LOOCV. The predictions of both samples

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

were outside the limits of the classes (i.e. $\hat{y}_{s1252\_P2} > HL_0$ and $\hat{y}_{s1816\_P13} > HL_1$). Notice that because the PDFs were not overlapped, there was no ambiguity region and the limits of the classes were defined by four operative limits. The classification performance of the *p*-DPLS$_1$ did not change when the reject threshold was varied to t=0.35 and t=0.10.

The *p*-DPLS$_1$ model was used to classify the six test samples of sporadic mutation of breast cancer. This mutation was not modelled in the training step; hence, all these samples should be pointed as outliers and not classified. Classifying these samples in any of the two modelled classes would result in a classification error. Figure 6 shows the PDFs (Eqs. 4 and 5) of class ω_0 and class ω_1 for *p*-DPLS$_1$ together with the predictions for the test samples. According to Eq. 11, all test samples were correctly detected as outliers and rejected since their predictions $\hat{y}_i$ were between the limits HL_0 ($\hat{y}$ = 0.24) and LL_1($\hat{y}$ = 0.54). If the reject constraints had not been applied, the classifier would have assigned the test samples to the class with the highest *a posteriori* probability. In this case, the samples *s1572_P16* and *s1324_P17* would have been incorrectly classified into class ω_1 (i.e. as BRCA2 mutation samples) and the remaining samples (*s1649_P15*, *s1320_P18*, *s1542_P19* and *s1281_P21*) would have been incorrectly classified into class ω_0 (as BRCA1 mutation samples). For these samples, the *a posteriori* probability for one class was near 1. For example, sample *s1572_P16* had $p(\hat{y}_i|\omega_0) = 6 \cdot 10^{-6}$ and $p(\hat{y}_i|\omega_1) = 2 \cdot 10^{-3}$ which results in $P(\omega_0|\hat{y}_i) \approx 0$ and $P(\omega_1|\hat{y}_i) \approx 1$. Therefore, if it is believed that the *a posteriori* probability demonstrates the classification's reliability, then the high values of probability obtained for the test samples would suggest that we can trust the classifications, despite the fact that all of them were incorrect. This shows that the classic Bayes rule is unreliable when both conditional probabilities $p(\hat{y}_i|\omega_0)$ and $p(\hat{y}_i|\omega_1)$ are low. Moreover, it should be noted that the predicted values are not as extreme as those expected for outliers. Hence, these are not directly suspicious samples because of their $\hat{y}_i$ values. It was the reject option, which set limits on the classes, which allowed these samples to be detected.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

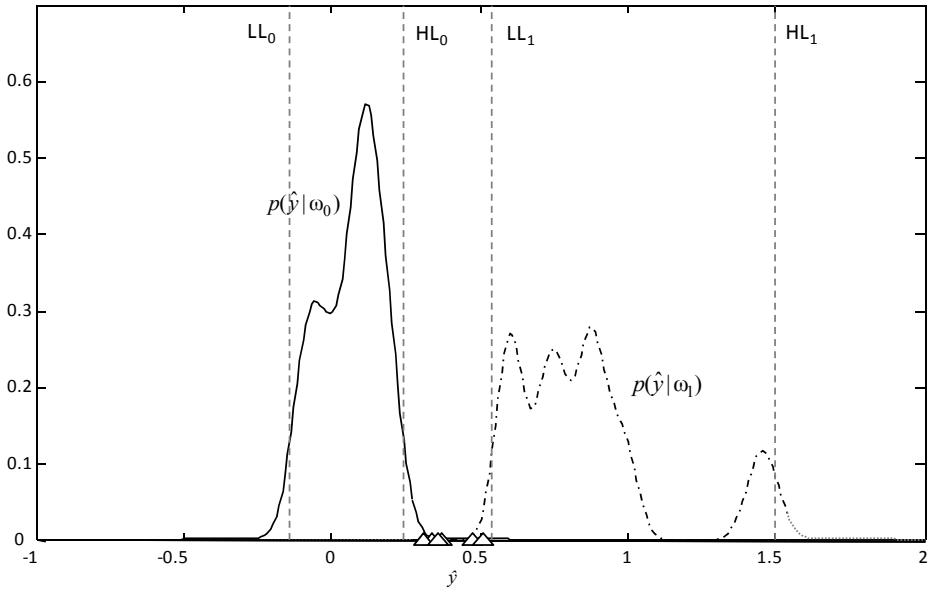


Figure 6. PDFs for the one factor *p*-DPLS model. Limits on LL_0=-0.15, HL_0=0.24, LL_1=0.54 and HL_1=1.50. Triangles represent the test samples classified with *p*-$DPLS_1$.

4.4 Conclusions

Recently, the DPLS method has received much attention in the field of gene expression data analysis. We have applied a new version of DPLS, namely *probabilistic* DPLS (*p*-DPLS), to classify biological samples using their microRNA (miRNA) expression patterns and cDNA microarray data. *p*-DPLS takes into account the uncertainty of the PLS predictions in the definition of the classification model. In this version, the possibility of rejection has been introduced. *p*-DPLS with reject option performs better than the original *p*-DPLS, because only those samples that have the highest probability of being correctly classified are indeed classified, whereas doubtful cases are rejected. The methodology involves evaluating the probability of each classification together with

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

the overall cost of the classifications performed for each model. In addition, the reject option allows us to deal with situations in which the results of the Bayes rule may be questioned. Moreover, the classification rule with reject option can help the experimenter to check that a sample does not belong to any of the classes modelled in the training step and therefore to ensure that it is rejected rather than misclassified. Thus, the reject option enables the classifier to detect outliers, and this in turn provides a new approach for improving outlier detection methods in the near future.

Acknowledgements

The authors thank the Department of Universities, Research and the Information Society of the Catalan Government for providing Cristina Botella's doctoral fellowship, and of the Spanish Ministry of Education and Science (project CTQ2007-66918/BQU). The authors would like also to acknowledge the useful comments of the referees.

References

[1] Alizadeh, A.A., et al., *Distinct types of diffuse large B-cell lymphoma identifyed by gene expression profyling.* Nature, 2000. **403** p. 503-511.

[2] Golub, T.R., et al., *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring.* Science, 1999. **285**: p. 531-537.

[3] Li, L., et al., *Gene Assessment and Sample Classification for Gene Expression Data Using a Genetic Algorithm/k-nearest Neighbor Method.* Combinatorial Chemistry & High Throughput Screening, 2001. **4**: p. 727-734.

[4] Brown M.P.S, et al., *Knowledge-based analysis of microarray gene expression data by using support vector machines.* Proceedings of the National Academy of Sciences, 2000. **97**: p. 262-267.

[5] Furey, T.S., et al., *Support Vector Machine classification and validation of cancer tissue samples using microarray expression data.* Bioinformatics, 2000. **16**: p. 906-914.

[6] Nguyen, D.V. and D.M. Rocke, *Multi-class cancer classification via partial least squares with gene expression profiles.* Bioinformatics, 2002. **18**: p. 1216-1226.

[7] Gunther E.C., et al., *Prediction of drug efficacy by classification of drug-induced genomic expression profiles in vitro.* Proceedings of the National Academy of Sciences, 2003. **100**: p. 9608-9613.

[8] Boulesteix, A.-L. and K. Strimmer, *Partial least squares: a versatile tool for the analysis of high-dimensional genomic data.* Briefings in Bioinformatics, 2007. **8**: p. 32-44.

[9] Nguyen, D.V. and D.M. Rocke, *Tumor classification by partial least squares microarray gene expression data.* Bioinformatics, 2002. **18**: p. 39-50.

[10] Pérez-Enciso, M. and M. Tenenhaus, *Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach.* Human Genetics, 2003. **112**: p. 581-592.

[11] Modlich, O., et al., *Predictors of primary breast cancers responsiveness to preoperative Epirubicin/Cyclophosphamide-based chemotherapy: translation of microarray data into clinically useful predictive signature.* Journal of Translational Medicine, 2005. **3**: article 32.

[12] Man, M.Z., et al., *Evaluation methods for classifying Expression data.* Journal of Biopharmaceutical Statistics, 2004. **14**: p. 1065-1084.

[13] Bylesjö, M., et al., *MASQOT: a method for cDNA microarray spot quality control.* BMC Bioinformatics, 2005. **6**: p. 250.

[14] Tax, D.M.J. and R.P.W. Duin, *Growing a multi-class classifier with a reject option.* Pattern Recognition Letters, 2008. **29**: p. 1565-1570.

[15] Knauthe, B., et al., *Visualization of quality parameters for classification of spectra in shooting crimes.* Journal of Chemometrics, 2008. **22**: p. 252-258.

[16] Fumera, G., F. Roli, and G. Giacinto, *Multiple Reject Thresholds for Improving Classification Reliability*, in *Advances in Pattern Recognition*, SSPR&SPR, Editor. 2000, Springer: Berlin - Heidelberg. p. 863-871.

[17] Devarakota, P.R.R., B. Mirbach, and B. Ottersten, *Reliability estimation of a statistical classifier.* Pattern Recognition Letters, 2008. **29**: p. 243-253.

[18] Dubuisson, B. and M. Masson, *A statistical decision rule with incomplete knowledge about classes.* Pattern Recognition, 1993. **26**: p. 155-165.

[19] Muzzolini, R., Y.-H. Yang, and R. Pierson, *Classifier desing with incomplete knowledge.* Pattern Recognition, 1998. **31**: p. 345-369.

[20] Ripley, B.D., *Statistical ideas for selecting network architectures*, in *Neural Networks: Artificial Intelligence and Industrial Applications*, B.K.a.S. Gielen, Editor. 1995, Springer. p. 183-190.

[21] Ripley, B.D., *Pattern Recognition and Neural Networks*. 2000, Cambridge, Unitet Kingdom: Cambridge University Press.

[22] Tortorella, F., *An optimal reject rule for binary classifiers.* In: Ferri, F.J et al. (Eds.), Advances in Pattern Recognition: Joint IAPR International Workshops, SSPR 2000 and SPR 2000, Lecture Notes in Computer Science, vol 1876. Springer-Verlag, Heidelberg., 2000: p. 611-620.

[23] Fumera, G., I. Pillai, and F. Roli, *Classification with Reject Option.* Proceedings of the 12th International Conference on Image Analysis and Processing (ICIAP'03), 2003.

[24] Fumera, G. and F. Roli. *Error Rejection in Linearly Combined Multiple Classifiers*. in *Proceedings of 2nd Int. Workshop on Multiple Classifier Systems (MCS 2001)*. 2001. Robinson College, Cambridge, UK.

[25] Fumera, G., F. Roli, and G. Giacinto, *Reject option with multiple thresholds.* Pattern Recognition 2000. **33**: p. 165-167.

[26] Cordella, L.P., et al., *A method for improving classification reliability of multilayer perceptrons.* IEEE Transactions on neural networks, 1995. **6**: p. 1140-1147.

[27] Landgrebe, T., et al., *The interaction between classification and reject performance for distance-based reject-option classifiers.* Pattern Recognition Letters, 2006. **27**: p. 908-917.

[28] Landgrebe, T., et al. *A combining strategey for ill-defined problems*. in *Fifteenth Ann. Sympos. of the Pattern Recognition Association of South Africa*. 2004.

[29] Chow, C.K., *On optimum recognition error and reject tradeoff.* IEEE -Transactions on information theory, 1970. **16**: p. 41-46.

[30] Hanczar, B. and E.R. Dougherty, *Classification with reject option in gene expression data.* Bioinformatics, 2008. **24**: p. 1889-1895.

[31] Duda, R.O., P.E. Hart, and D.G. Store, *Pattern Classification (2nd edition)*, ed. W. Intersicence. 2001, New York.

[32] Pérez, N.F., J. Ferré, and R. Boqué, *Calculation of the reliability of classification in Discriminant Partial Least-Squares Classification.* Journal of Chemometrics and Intelligent Laboratory Systems, 2009. **95**: p. 122-128.

[33] Wold, H., *Partial least squares*, in *Encyclopedia of Statistical Sciences* K.a.N.L. Johnson, Editor. 1985, Wiley: New York. p. 581-591.

[34] Webb, A., *Statistical Pattern Recognition, 2n edition*, ed. Wiley. 2002, Malvern, UK.

[35] Bradley, A.P., *The use of the area under the ROC curve in the evaluation of machine learning algorithms.* Pattern Recognition, 1997. **30**: p. 1145-1159.

[36] Li, M. and I.K. Sethi, *Confidence-based classifier design.* Pattern Recognition, 2006. **39**: p. 1230-1240.

[37] Lu, J., et al., *MicroRNA expression profiles classify human cancers.* Nature Letters, 2005. **435**: p.834-838.

[38] Hedenfalk, I., et al., *Gene Expression profiles in hereditary breast cancer.* The New England Journal of Medicine, 2001. **344**: p. 539-548.

[39] Zheng, Y. and C.K. Kwoh, *Informative microRNA expression patterns for cancer classification.* Data mining for biomedical applications, Proceedings, 2006. **3916**: p. 143-154.

[40] Lin, J. and M. Li, *Molecular profiling in the age of cancer genomics.* Expert Review of molecular diagnostics, 2008. **8**: p. 263-276.

[41] Boulesteix, A.-L., *PLS dimension reduction for classification with microarray data.* Statistical Applications in Genetics and Molecular Biology, 2004. **3**: article 33.

[42] Raza, M., et al., *Comparative Study of Multivariate Classification Methods using Microarray Gene Expression Data for BRCA1/BRCA2 Cancer Tumors.* Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05), IEEE., 2005. **2**: p. 475-480.

[43] Branden, K.V. and S. Verboven, *Robust data imputation.* Computational Biology and Chemestry, 2009. **33:** p. 7-13.

[44] Pochet, N., et al., *Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction.* Bioinformatics, 2004. **20**: p. 3185-3195.

[45] Kennard, R.W. and L.A. Stone, *Computer Aided Design of Experiments.* Technometrics, 1969. **11**: p. 137-148.

[46] Lu, Y. and J. Han, *Cancer classification using gene expression data.* Information Systems, 2003. **28**: p. 243-268.

[47] Musumarra, G., et al., *Potentialities of multivariate approaches in genome-based cancer research: identification of candidate genes for new diagnostics by PLS discriminant analysisy.* Journal of Chemometrics 2004. **18**: p. 125-132.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

CHAPTER 5 Outlier detection and ambiguity detection for microarray data in *p*-DPLS regression

Journal of Chemometrics 2010, Accepted

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Microarray data are obtained after a complex series of experimental steps that go from hybridization to image analysis. Microarray manufacturing errors like dye instability, different incorporation of the dyes, slide, spatial and print-tip effects together with scanning errors may introduce unspected data variability, which can make the collected data for one sample very different than the data from other samples of the same class. Additionally, the experimenter may be confronted with new samples that are not like any of the other samples that have been modelled (e.g., samples that do not belong to any of the modelled classes). All these samples are considered as outliers, and can have a degrading impact in the calculated classification model (if they are training samples), can produce wrong evaluations of the classification performance of the model (if the samples are validation samples) and can lead to wrong classifications (if the samples are new samples to be classified).

Outlier detection is often unnoticed in microarray data classification. However it is essential that any classificatition method that is intended to have a real practical use be implemented together with appropriate outlier detection tools.

Basically, all the outliers can be detected either because they have errors in the recorded data (**x**), because they have been identified erroneously (with erroneous *y*), because they have abnormal **x**-y relation or because they belong to a different population than the samples we are trying to classify. In this work we develop outlier detection for *probabilistic* discriminant partial least squares (*p*-DPLS) method by combining diagnostics based on leverage and *x*-residuals (common in PLS) and the reject option approach developed in chapter 4.

The method was tested on two datasets: the prostate cancer dataset and the small round blue cell tumours of childhood dataset. Results showed that without outliers the *p*-DPLS classification models have better classification abilities and samples from

classes not modelled during the training step are rejected to classify, thus avoiding their misclassification.

The removal of outliers in the prostate cancer dataset reduced the *Cost* of classification per sample from 0.11 to 0.06, and the model increased the proportion of correct classifications of test samples from 95% to 100%. In the small round blue cell tumours of childhood dataset the *p*-DPLS with outlier detection method implemented is able to flag correctly as outliers the 95% of the samples in the prediction step. These samples did not belong to any of the classes modelled. When the outlier detection method was not implemented in the training step, only the 5% of the test samples were pointed as outliers, misclassifiying the remaining 95%.

This work is presented in paper form published in *Journal of Chemometrics* 2010 (Accepted).

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Outlier detection and ambiguity detection for microarray data in *probabilistic* Discriminant Partial Least Squares Regression

C. Botella*, J. Ferré, R. Boqué

Department of Analytical Chemistry and Organic Chemistry, Rovira i Virgili University.
Marcel·lí Domingo s/n, 43007. Tarragona, Spain
*Corresponding author: cristina.botella@urv.cat

***Journal of Chemometrics*, 2010 Accepted**. (Edited for format)

Abstract

The reject option plays an important role in the classification of microarray data. In this work, a reject option is implemented in the discriminant partial least squares (*p*-DPLS) method in order to reject to classify both outliers and ambiguous samples. Microarray data are highly susceptible to present outliers because of the many steps involved in the experimental process. During the development of the classifier, outliers in the training data may strongly influence the model and degrade its performance. Some future samples to be classified may also be outliers that will most probably be misclassified. Ambiguous samples are samples that cannot be clearly assigned to any of the classes with a high confidence. In this work outlier detection and ambiguity detection are implemented taking into account the *x*-residuals, the leverage and the predicted $\hat{y}$. The method was applied to oligonucleotide microarray data and cDNA microarray data. For the first dataset (prostate cancer data set), the outlier detection criteria allowed us to remove nine samples from the training set. The model without those samples had better classification ability, with a decrease in the classification *Cost* per sample from 0.10 to 0.07. The method was also used in a second dataset (small round blue cell tumours of childhood dataset) to detect prediction outliers so that most of the outliers were rejected to classify and misclassifications were reduced from 100% to 5%.

5.1 Introduction

Outlier detection plays a fundamental role in the development and application of multivariate classification methods for microarray data. Outliers are either samples, variables, or certain variables in certain samples that have a different behaviour than the rest of the data. This paper focuses on sample outliers. Sample outliers may be training samples, validation samples, or future samples to be classified. The experimenter is interested in flagging them for different reasons. Outliers in the training set may have an excessive influence on the classification rule, unless robust methods of classification are used. Hence, it is interesting to know whether the classification rule is dominated by a few special samples, and discover if this influence can be adverse. Samples with large measurement errors or samples that belong to a different population than the samples we are trying to classify will degrade the rule. These "bad" outliers should be detected, removed and the rule recalculated. Training outliers may also contain "good" samples with unique information. These must be kept, since they will improve the model by expanding its application domain. Their detection will warn the experimenter that more samples of the similar type should be obtained in order to model that variability better. Study of the good outliers may also lead to discover special variables (gene expressions) that may have a high discriminative power [1]. Outlier detection must also be applied when future samples are to be classified, which is the ultimate objective of the classification rule. Unknown samples that do not belong to any of the classes for which the classification rule was trained or samples with large data errors will be misclassified. The experimenter wants to be warned about these samples so that they can be rejected to classify until more information is available. In this sense, outlier detection increases the confidence the experimenter has in the classification protocol, since the samples that might be misclassified will hopefully be flagged. Finally, outlier detection must also be used to detect outliers in the validation set. Samples not representative of the future samples

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

to be classified will likely produce an erroneous classification result that will worsen the classification ability of the model. Hence, these samples should be detected, as it is done for the unknown samples, and not considered to evaluate the performance of the model.

The particularities of microarray gene expression data and the many levels of variation introduced at the complex experimental stages, from hybridization to image analysis, make necessary the use of outlier diagnostics [2, 3]. First of all, the recorded microarray data depend on the biological variations of the population under study (intrinsic to all organisms and influenced by genetic or environmental factors). Technical variations introduced during the extraction, labelling or hybridization of samples, scanner settings and measurement errors associated with the reading of the fluorescent signals (which may be affected, for example, by dust on the array [4]) will also increase the data variability. Moreover, the large number of variables (gene expressions) compared to the relatively low number of objects, make the data analysis and the classification a nontrivial task. Fortunately, the combined use of data pre-processing and multivariate algorithms can extract the main systematic variation in the data and lead to satisfactory classification results. For example, normalization methods, such as the lowess correction [5] or the total intensity normalization [6] can remove inconsistencies of the microarray data. However, not all the errors in the data may be mathematically removed and outlier diagnostics are still needed in order to prevent misclassifications due to new unexpected data variations. Outlier detection is also needed to flag those samples from new unexpected classes (biological outliers) and those that present extreme biological variability.

Several methods have been used for detecting outliers in microarray data for particular classification rules. Paoli [7] improved the performance of Support Vector Machines (SVM) by selecting the optimal number of genes and treating the most relevant as

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

outliers. Moffitt [8] constructed the SVM model by removing outliers via re-validation. Olsen [9] analysed the intensity scores of tissue microarrays of sarcoma phenotypes with Euclidean hierarchical cluster analysis, presenting as outliers those samples that did not cluster into any of the defined groups. The VizRank tool, which combines the *k*-nearest neighbours (*k*-NN) method with a range of visualizations was also used to detect outliers [10]. Model *et al*. [11] pointed out that outliers in microarray data cannot always be detected visually and proposed a robust version of Principal Component Analysis (rPCA). Their objective was to exclude single outlier chips from the analysis and to detect systematic changes in experimental conditions as early as possible in order to facilitate a fast recalibration of the production process. Shieh [12] addressed outlier detection with highly different expression patterns in microarray data using also PCA and a robust estimation of Mahalanobis distance. Tomlins *et al.* [13] proposed the cancer outlier profile analysis (COPA) method for detecting translocations from microarray data. For gene selection, genetic algorithms were proposed for outlier detection using a grid count tree [14]. Liu *et al*. studied different statistical methods to detect genes with differential expressions across the different class samples (1). And Loo *et al.* used with the same objective, filter-based methods [15]. In contrast, Tibshirani [16] and Wu [17] proposed alternative cancer outlier differential expression detection methods for detecting genes that, inside a disease group, exhibit unusually high gene expression in some but not all samples.

In this work, we develop outlier detection for discriminant partial least squares (DPLS). DPLS is one of the preferred methods for classification of microarray data [18]. In DPLS, the assigned class is decided from the predicted value $\hat{y}$ when the measured microarray data are submitted to a PLS model. Hence, the outlier detection approaches that exist for PLS (already applied in multivariate calibration in chemical and industrial fields) can be applied. Pell used the studentized residuals versus leverage plot to detect outliers in PLS, which was successful when either masking or

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

swamping occurred [19]. Pell [20] also, based on the work by Martens and Næs [21], detected outliers from an *F*-ratio which compared the validation samples *x*-residuals to the *x*-residuals of the calibration samples. Chiang and Pell [22] presented the closest distance to center (CDC), a multiple outlier detection algorithm applied together with ellipsoidal multivariate trimming (MVT), taking into account only the *x*-data. A methodology to detect prediction outliers in PLS was applied by projecting the new objects on the Sammon's mapping space containing the convex hull which defines a boundary around each cluster and another around the whole calibration data [23]. Q and Hotelling's T^2 statistics [24] were also used to detect outliers in PLS, although the authors indicated that in some cases these indexes would not be enough.

Most of these mentioned approaches take into account only the *x*-response data to point a sample as a potential outlier since it is the only information available for unknown samples. Note also that the predicted value $\hat{y}$ is rarely used to detect prediction outliers in PLS, since it is often difficult to set limits on the lowest and highest values of $\hat{y}$ that can be accepted. Only those predictions that are really extreme can warn the sample being an outlier. DPLS, however, has the particularity that the $\hat{y}$ values (from which the class is decided) are located around the value that codifies the class (around 0 or 1 in the DPLS scheme used in this paper) and that probability density functions of the predictions can be established. This fact has been previously used to define a reject option for DPLS and microarray data [25]. The reject option allowed to reject to classify those samples that had extreme $\hat{y}$ values or those with "normal" $\hat{y}$ values but whose classification was ambiguous (i.e., samples that have a very similar probability to belong to any of the modelled classes). In this paper, we provide a unified approach for outlier detection in DPLS for microarray data. This approach combines the new criterion based on the predicted value $\hat{y}$ particularly developed for DPLS, with the well-known diagnostics based on the leverage and the *x*-residuals commonly used in PLS.

5.2 Theory

5.2.1 *Probabilistic* Discriminant Partial Least Squares

DPLS is the application of PLS regression to classification problems. A DPLS model is calculated by regressing **y**, which codifies the class of the samples, on **X** using *A* latent variables (factors) [18, 26]. For microarray gene expression data, **X** is an $N \times P$ matrix of *N* samples and *P* gene expressions and **y** is a $N \times 1$ vector of ones and zeros, where the 0 codifies that the sample belongs to class ω_0 and the 1 codifies that the sample belongs to class ω_1. For an unknown sample with measured *x*-data, $\mathbf{x}_t$, the value predicted by the DPLS model taking into account *A* factors is given by $\hat{y}_t = \mathbf{x}_t^T \mathbf{b}$, where **b** is the vector of regression coefficients and the pre-processing is implicit in the formula. With the mentioned coding, $\hat{y}_t$ should ideally be close to zero if the sample belongs to class ω_0 and close to one if the sample belongs to class ω_1. The criterion for deciding the class from $\hat{y}_t$ will influence the performance of the classification rule. The criterion used in this work is based on the probabilistic version of DPLS, *p*-DPLS [27]. The *p*-DPLS procedure starts by calculating a PLS model of *A* factors relating **X** and **y**. Then, the training samples are predicted with this model. For each training sample *i*, a potential function $f(\hat{y}_i, \mathrm{SEP}_i)$ is calculated with the shape of a Gaussian centred at the predicted value $\hat{y}_i$ and with standard deviation the standard error of prediction (SEP_i) of that sample. Next, the individual potential functions of all the samples of class ω_0 are averaged to obtain the probability density function (PDF) that describes the predictions of class ω_0 (Eq. 1):

$$p(\hat{y}_i|\omega_0) = \frac{\sum_{i=1}^{n_0} f(\hat{y}_i, \mathrm{SEP}_i)}{n_0} \quad (1)$$

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

where n_0 is the number of samples of class ω_0. The PDF for class ω_1 is calculated likewise, using n_1, the number of samples of class ω_1. A sample is classified by calculating its prediction $\hat{y}$ and applying the Bayes theorem to the two PDFs so that the sample is allocated in the class with the highest *a posteriori* probability. A consequence of the straight application of this rule is that a sample is always classified in one of the classes. So, samples from new unexpected classes will be misclassified, and those samples with either extremely low or extremely high values of $\hat{y}$ (which may be outliers) will be assigned to one of the classes with a very large probability.

5.2.2 Reject option in *p*-DPLS

The purpose of the reject option in *p*-DPLS is to allow the classifier to reject a sample if this will likely be misclassified. In other words, a class label is assigned only to those samples with the highest probability of being correctly classified. By not forcing the classifier to always make a decision in one of the two modelled classes, the misclassification rate of the model (measured as the number of correctly classified samples with respect to the number of samples for which the classifier assigns a class) decreases, and gives confidence to the experimenter on the outputs of the classification rule. The reject option in *p*-DPLS is implemented here for two main types of samples: outliers and ambiguous samples.

5.2.2.1 Rejection of outliers

Outliers are samples whose *x*-data have different features than the bulk of the training samples. Several reasons for this behaviour are (a) the sample belongs to a class that was not modelled, (b) the sample belongs to one of the modelled classes but the *x*-data have gross errors or contain unmodelled interferences, and (c) the sample belongs to one of the modelled classes but has correct extreme values of some variables. Samples in situation (a) should be detected and rejected otherwise they will

be wrongly classified in one of the two modelled classes. Samples in situations (b) and (c) will not necessarily be misclassified, but this uncommon behaviour will likely affect the classification result and hence we might prefer to reject the samples and ask for extended analysis instead of running the risk of misclassifying them. Outliers (hence, candidates to be rejected to classify) in *p*-DPLS are flagged based on the following four criteria: limits on the $\hat{y}$**,** leverage, ratio of residual variances and classification error.

a. Limits on the $\hat{y}$

In *p*-DPLS, the predictions $\hat{y}$ of the training samples are used to calculate a distribution of predictions for each class (see Figure 1). These distributions are ideally centered on 0 and 1, the reference values used at the training stage. Uncommon x-data will produce $\hat{y}$ values at the extremes of the PDF's of a class. Hence, limits for $\hat{y}$ are set around the majority of the $\hat{y}$ of the training data. These limits define regions in the $\hat{y}$ axis in which the sample is either classified in one class, in the other class, or rejected to classify [25]. The limits are defined such that the area in the tails of each distribution is five percent of the total area of the distribution (i.e. 2.5% in each tail of the PDF of each class). These limits depend on the PDFs. Hence, they are different for *p*-DPLS models with a different number of factors. In practice, when the PDFs are overlapped (Figure 1) there are two limits, a High Limit (HL) and a Low Limit (LL) and when the PDFs are separated there are four limits (a HL and a LL for each class) (Figure 4b). A sample with a $\hat{y}$ predicted outside the limits will be flagged as outlier. If the PDFs are not overlapped, a sample with $\hat{y}$ between HL_0 and LL_1 will be flagged as inlier. This criterion improves the direct application of the Bayes rule in the sense that, at the extremes of the PDFs, the *a posteriori* probability for one class is high, and hence the Bayes rule would assign the sample to that class with a high probability. By imposing the limits, the sample will now be rejected to classify. Note also that the limits on the $\hat{y}$ values will not account for all the outlier situations in *p*-DPLS, since they will not detect those outliers whose unusual *x*-data makes $\hat{y}$ be inside a classification region, e.g.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

when the $\hat{y}$ of a sample of class ω_0 falls within the classification region of class ω_1. These samples might be detected by the criteria described next.

b. Leverage

The leverage of sample *t* for a DPLS model calculated with mean-centered *x*-data is given by [28]:

$$h_t = \frac{1}{n} + \boldsymbol{t}_t{}^{\mathrm{T}}(\mathbf{T}^{\mathrm{T}}\mathbf{T})^{-1}\boldsymbol{t}_t \qquad (2)$$

where $\boldsymbol{t}_t$ denotes the score vector and **T** is the scores matrix of the mean centered training data. The leverage measures the distance from the sample to the center (mean) of the training set taking into account the correlation in the data. A low value of h_t indicates that the sample is similar to the average of the training samples. A high leverage indicates that the sample has an unusual ***x***-vector (or score vector) relative to the training samples, so it is an *x*-outlier. In that case, the experimenter should suspect about the reliability of the classification and wait for additional studies. Although no strict rules exist, it is common to declare as a high-leverage sample the one with $h_t > 3\bar{h}$ where $\bar{h}$ is the average leverage value for the training samples ($\bar{h} = 1/\mathrm{N} + A/N$) [29, 30].

c. Ratio of residual variances

In DPLS, there is a vector of *x*-residuals for each sample and number of factors *A* used in the model. The residuals are the difference between the measured *x*-data and the data predicted by the model with *A* factors. While the leverage refers to the position of the sample in the subspace of the factors used for regression, the residual refers to the orthogonal subspace, i.e., the factors not used for regression. Residuals that are much larger than most of the residuals of the training samples indicate that the sample is

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

poorly described by the model for that number of factors and, hence, it is an *x*-outlier. It must be pointed out, however, that large *x*-residuals do not necessarily imply a wrong $\hat{y}$, and, hence, a wrong classification. Actually, one of the advantages of the factor-based methods such as PLS is that the factors retained in the model should account for the relevant variability in the *x*-data, while the remaining factors not used in the model should account for the irrelevant variability (the *x*-residuals). Hence, a large *x*-residual simply indicates that some part of the measured *x*-data is not modelled. However, there is a large chance that the source of these unmodelled data had also a contribution in the model space and influenced the $\hat{y}$. These outliers are detected by comparing the unmodelled parts of the test sample to the unmodelled parts of the training samples using the *A*-factor *p*-DPLS model [31] with the ratio of residual variances:

$$\mathrm{V} = \frac{S_t^2}{S_T^2} \tag{3}$$

where s_t^2 is the residual variance for the test sample :

$$S_t^2 = \frac{\sum_{j=1}^{P}(x_{tj} - \hat{x}_{tj})^2}{(P - A)} \tag{4}$$

and s_T^2 is the total variance for the training samples [21]:

$$S_T^2 = \frac{\sum_{i=1}^{N}\sum_{j=1}^{P}(x_{tj} - \hat{x}_{tj})^2}{(N \cdot P - P - A \cdot (\max(N,P))} \tag{5}$$

An object with *V*>3 is considered to be an outlier. A similar criterion was used in [31] to detect outliers in PLS. Note that the usual comparison of V with a tabulated F-value is not useful. The very large number of degrees of freedom involved [32] makes the tabulated F-value be low and most of the samples be flagged as outliers, which is meaningless.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

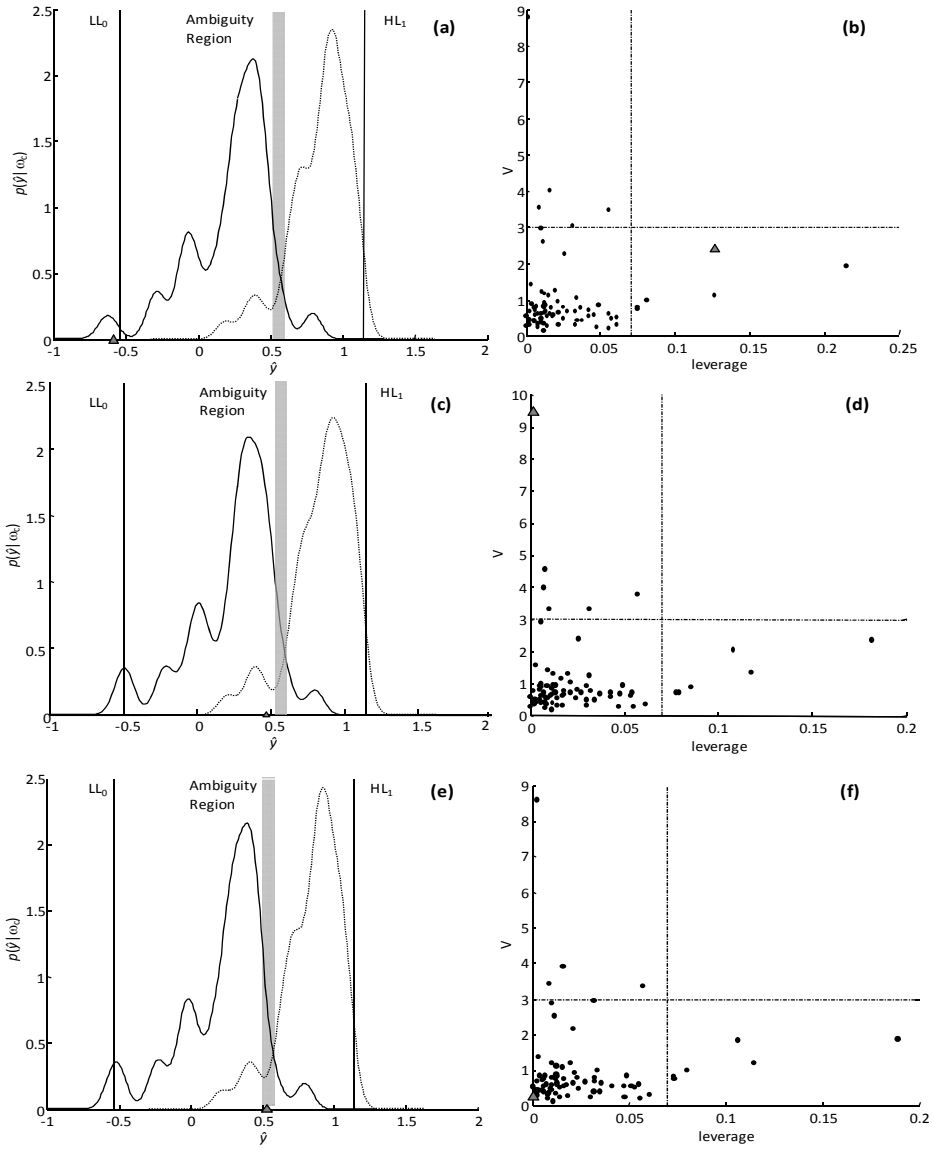


Figure 1. Probability density functions (PDFs) for the *p*-DPLS model with two factors obtained during leave-one-out cross-validation and influence plots for the training samples when a sample is used as a test. **a-b.** PDFs and influence plot when sample N43_normal is left out **c-d.** PDFs and influence plot when sample T11_tumour is left out, **e-f.** PDFs and influence plot when sample N41_normal is left-out. In **a**, **c** and **e**, the triangle (▲) identifies the prediction of the left-out-sample. In **b**, **d** and **f**, the triangle (▲) identifies the left-out-sample as compared to the rest of the training data. The vertical and the horizontal dotted lines indicate the limits for outlier detection.

d. Classification error

Classification error is an easy-to-use outlier diagnostics during the training stage of a classification rule. When the **x**-*y* relation of a sample does not agree with the **x**-*y* relation described by the model, the sample is misclassified, and this is used to flag the sample as an outlier. This is the equivalent to a large prediction error in regression models. Different from the criteria (a) to (c), the classification error can only be used to detect outliers in the training and validation sets because it requires the true class to be known. Despite it cannot be applied to new samples, the criterion is still very helpful to refine the classification model.

5.2.2.2 Rejection of ambiguous samples

Ambiguous samples are samples that share characteristics of both class ω_0 and class ω_1 because the measured *x*-variables are not discriminative enough for the algorithm used. When these samples are predicted by the DPLS model, their $\hat{y}$ values are in the boundary between classes (ambiguity region, Figure 1) so the Bayesian probability of belonging to any of the classes $P(\omega_c|\hat{y}_t)$ is similar. Even small variations in the measured *x*-data can make the classifier assign the sample to either one class or the other. This increases the uncertainty of the classification result, so it may be preferable to reject that sample. This rejection is defined by the rule:

$$\text{reject } if\ max\left(P(\omega_c|\hat{y}_i)\right) < (1-t) \quad c = 0{,}1 \tag{6}$$

so that the sample is rejected if the *a posteriori* probability of belonging to any of the classes is lower than a reject threshold (1–*t*). Note that the threshold can be set to reject any slightly doubtful sample. This improves the error rate of the classifier, since less samples will be misclassified, but, in turn, more samples will be rejected that otherwise could be correctly classified, which reduces the usefulness of the classifier. Chow [33] derived an optimum rejection scheme that gives a trade off between reject rate and error rate. This rule was recently described for *p*-DPLS [25].

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

5.3 Results

5.3.1 Data

The prostate cancer data set [34] consists of 50 non-tumour samples (class ω_0) and 52 tumour samples (class ω_1) with 12600 gene expressions (variables). From these gene expressions (variables), the 150 with the highest variance weight [35] were selected to avoid irrelevant genes from interfering with the discrimination power of the relevant genes [36]. The dataset was divided into a training set (82 samples, 42 of class ω_0 and 40 of class ω_1) and a test set (20 samples, 8 of class ω_0 and 12 of class ω_1) using the Kennard-Stone algorithm [37]. This dataset is used to show the ability of the methodology to detect outliers in the training set and to show that the final classification model and the prediction of the test set improve when these outliers are deleted.

The small round blue cell tumours of childhood dataset [38] includes 2308 gene expressions of 12 samples of neuroblastoma (NB), 8 samples of non-Hodgkin lymphoma (BL), 23 samples of Ewing family of tumours (EWS) and 20 samples of rhabdomyosarcoma (RMS). EWS samples (class ω_0) and RMS samples (class ω_1) were used for training and the remaining, NB and BL samples, as test samples. This dataset was used to show how the proposed method can reject new samples that do not belong to any of the modelled classes. Since, the test samples do not belong to any of the modelled classes, they would be misclassified unless the reject option is implemented.

5.3.2 Prostate data set

Briefly, the procedure was as follows. First, the *p*-DPLS model was calculated for a given number of factors using mean-centered gene expressions of the training samples. Then, the training samples were predicted and the predictions, $\hat{y}$, were used to calculate kernel Gaussians, which, in turn, defined a PDF for each class (Eq. 1). From the PDFs, the reject option limits for $\hat{y}$ were set. The leverage and *x*-residuals of the training samples were also calculated. An unknown sample with measured $\mathbf{x}_t$, was first predicted ($\hat{y}_t=\mathbf{x}_t^{\mathrm{T}}\mathbf{b}$) and the *x*-residuals, the leverage and the probability of classification for each modelled class (evaluated as the Bayes *a posteriori* probability detailed in [25]) were calculated. The sample was then either classified or rejected to classify if it was flagged as outlier (section *2.2.1*) or ambiguous (section *2.2.2*, Eq. 6). Before classifying unknown samples, the optimal model was selected by leave-one-out cross-validation (LOOCV). In LOOCV, a sample is left out and the model is calculated using the remaining samples. Each left-out sample was treated as an unknown sample and was either classified or rejected as described above. Of the rejected samples, outliers were removed from the training set and the model was recalculated; ambiguous samples, however, were maintained in the model since they introduced relevant variability.

The performance of the model was evaluated with the classification cost per sample:

$$Cost = (\lambda_r \mathrm{N}_r + \lambda_m \mathrm{N}_m)/\mathrm{N} \quad (7)$$

where N_r is the number of samples rejected, N_m is the number of samples misclassified, λ_r and λ_m, are the costs of rejecting a sample or misclassifying it respectively and N is the total number of samples used to validate the model. The cost criterion, calculated during LOOCV, was used to compare the *p*-DPLS models with a different number of factors and to select the optimal model. Note that λ_r and λ_m may be fine-tuned to meet the requirements of the classification problem. Since for this dataset there is no reference in the literature about the associated costs of rejecting or misclassifying a

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

sample, we used λ_r =0.25 and λ_m =1, which indicates that we prefer to reject four samples instead of classifying one wrong. In this case, λ_r values lower than 0.25 did not improve the classification performance of the model.

For this dataset, preliminary *p*-DPLS models using 1 to 4 factors were calculated. Taking into account the cost per sample calculated by LOOCV, the optimal model had two factors. Samples N43_normal and N25_normal (both of class ω_0) were pointed out as outliers because their predictions were outside the accepted region for $\hat{y}$ for its corresponding cross-validation segment. The prediction of sample N43_normal (Figure 1a-b) was $\hat{y}$=–0.59, lower than LL_0=–0.56, while sample N25_normal had $\hat{y}$ =–0.66, lower than the limit LL_0=–0.50 established for its *p*-DPLS model (note that the limits *HL* and *LL* vary for each cross-validation segment since the *p*-DPLS is calculated with different samples). These extreme predictions suggested the possibility of an unusual **x** vector. This was later confirmed because the leverage of these samples exceeded three times the average leverage of the training set: sample N43_normal had h = 0.13 while $\bar{h}$ =0.024 and sample N25_normal had h = 0.23 while $\bar{h}$ =0.037. The reason for the high leverage is that five genes, those with Accession Numbers 36785_at, 221_s_at, 774_g_at, 31449_at, 38411_at, had higher intensities than the rest of the samples of class ω_0. The five variables (genes) differentially expressed, in this case, were not considered relevant since the different intensities were only present in a few samples so they did not seem to respond to a differential characteristic of one class.

In addition to the samples N43_normal and N25_normal, the leverage criterion also flagged sample N33_normal as outlier (h = 0.12, while $\bar{h}$ =0.025), despite this sample did not have an unusual $\hat{y}$.

Six additional samples (N04_normal, T02_tumour, T05_tumour, T11_tumour, T15_tumour and T25_tumour) were rejected for having high *x*-residuals (V > 3). These

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

six samples had most of the gene expressions with higher intensities than the mean of the intensities of the training samples, so the samples were not well-modelled by the factors of the DPLS model. The T11_tumour sample (class ω_1) (Figure 1c-d), for example, was rejected because V = 9.45. Its prediction $\hat{y}$=0.47 was closer to the predictions for class ω_0 than to the predictions of class ω_1 so the sample would have been classified wrongly (i.e., non-tumour) if it had not been rejected. Notice that the prediction for this sample is not an extreme value, so the sample had not been labelled as suspicious based only on the prediction.

In addition to the previous samples flagged as outliers, five samples were wrongly classified (Table 1). In these samples, the relation of **x**-y did not agree with the trend modelled by the *p*-DPLS model. The reason for the wrong classification is that the intensities of the samples of class ω_0 (non-tumour) are lower than those of class ω_1 (tumour) for the majority of the samples of this dataset. The sample N38_normal (class ω_0), however, had intensities in some of the variables higher than expected, more similar to the intensities of tumour samples (class ω_1) than to the intensities of the samples of its true class (Figure 2a). For this reason, the sample was misclassified. The opposite happened with the misclassified samples of class ω_1 (T39_tumour, T21_tumour, T49_tumour and T34_tumour). Some intensities were lower than most of the intensities of class ω_0 (Figure 2b). This situation may result from either an incorrect codification of the samples (mislabelling), experimental problems (e.g. bad intensity acquisition) or because these samples were truly different from the rest of samples of their class (which would indicate that more representative samples of this type should be collected before they are included in the model).

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

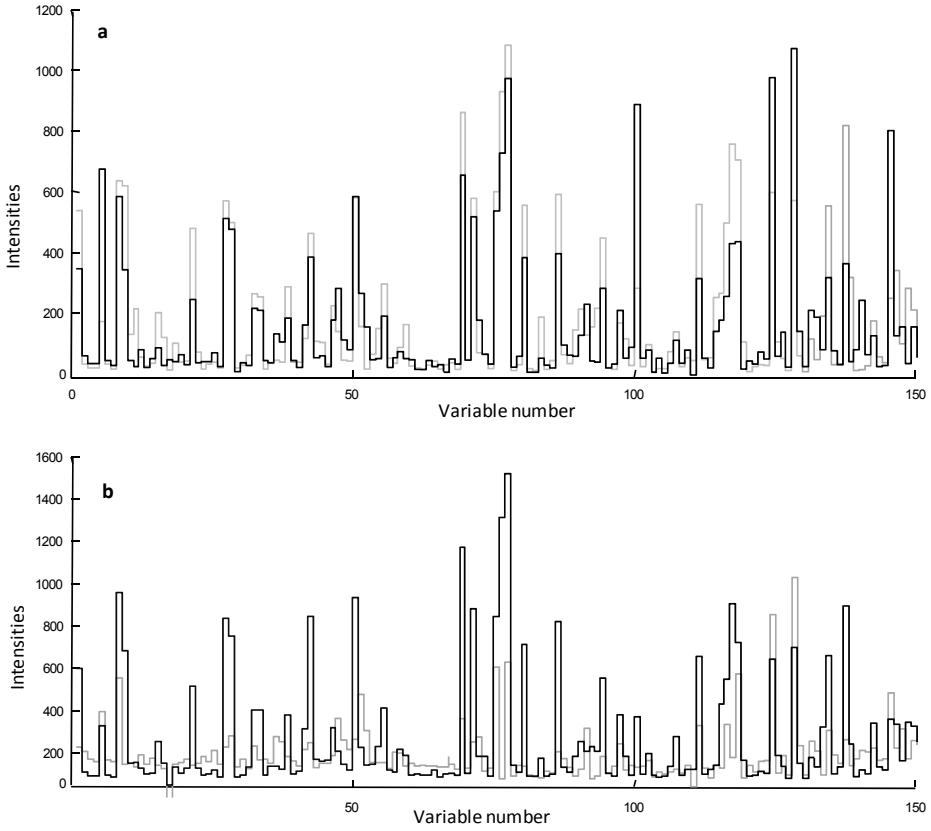


Figure 2. a. Intensities of sample N38_normal of class ω_0 (grey) and mean of intensities of class ω_0. **b.** Intensities of sample T21_tumour of class ω_1 (grey) and mean of intensities of class ω_1 (black).

During the cross-validation process, four additional samples were rejected to classify because they were ambiguous. These samples did not have extreme values, so they were not likely to influence the model excessively and they were kept in the training set. However, since in the LOOCV process these samples acted as test samples, they were considered as rejects for the calculation of the performance of the classifier. An example is shown in Figure 1e and 1f. The figure shows the acceptance and reject regions for the cross-validation model when sample N41_normal is left out. Because its $\hat{y}$ was in the

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

ambiguity zone, this sample would have been rejected to classify if it had been an unknown sample.

It is to note also that the outlier detection process could suffer from the masking effect, so that the presence of several outliers could hide the presence of some other outlier. Despite this, extreme samples could still be detected and the model was recalculated without those samples. The optimal model was again the *p*-DPLS model with 2 factors, with a decrease of the *Cost* of classification per sample from 0.11 to 0.06 (Figure 3).

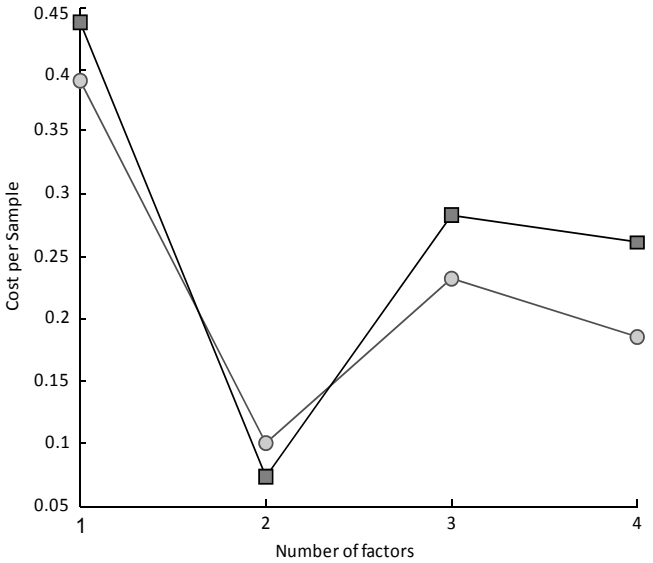


Figure 3. *Cost* per sample for the training samples with λ_r=0.25 and λ_m=1. (●) *p*-DPLS model with all the samples. (■) *p*-DPLS models after removing outliers.

Table 1 shows the classification results for the models calculated with the original dataset and with the dataset after removing the rejected training samples. The LOOCV and test set classifications are first presented for the initial dataset using the *p*-DPLS model for two factors (columns 2 and 3) without reject option (i.e., there are no rejected samples). Columns 4 and 5 show the classifications when the reject option is

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

enabled. Note that one false negative and one false positive of the classical model become rejects. In turn, five true negatives and six true positives become also rejects. This is because the high certainty required in the classification results makes the samples with uncertain classification be rejected. Columns 6 to 9 show the results after the outliers in the training set had been removed. Comparing the classical *p*-DPLS models with and without outliers (columns 2-3 and 6-7), it is seen that the model without outliers misclassifies one sample less. This improvement is more notable when rejection is allowed (columns 4–5 versus 8–9). In this case, the LOOCV error rate (calculated as the ratio of samples misclassified divided by the samples classified), for the model with outliers is 5/69=0.07, higher than the error rate for the depurated model (2/56=0.04). The reduction of misclassified samples is also observed in the test set. Columns 8 and 9 show the results of the depurated model with reject option. This depurated model predicts better than the models calculated with all the training samples without reject option. This optimal model classifies wrongly only two samples, and also has fewer rejections, so the classification *Cost* per sample is lower (Figure 3).

The prediction of the test set is also better. The two misclassifications of *p*-DPLS calculated with the initial dataset are now rejections (based on the ambiguity rejection rule, Eq. 6). Compared with the *p*-DPLS model with reject option calculated with the initial dataset, the number of misclassifications and of rejections decreased, so the classification cost per sample decreases from 0.10 to 0.07. Hence, the removal of the outliers of the training set improved the *p*-DPLS model in the sense of classifying better both the training samples via LOOCV and the test samples.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Table 1. Prostate cancer dataset. Classification of validation and test samples for the *p*-DPLS model with two factors calculated with initial training samples and after removing outliers.

	Initial dataset				**Dataset after removing outliers from the training set**			
	***p*-DPLS**		***p*-DPLS with reject option**		***p*-DPLS**		***p*-DPLS with reject option**	
	LOOCV	test	LOOCV	test	LOOCV	test	LOOCV	test
FN	2	2	1	1	6	2	2	0
FP	5	0	4	0	0	0	0	0
TN	42	5	37	5	38	5	33	5
TP	33	13	27	13	24	13	21	13
RN	0	0	6	0	0	0	5	0
RP	0	0	7	1	0	0	7	2

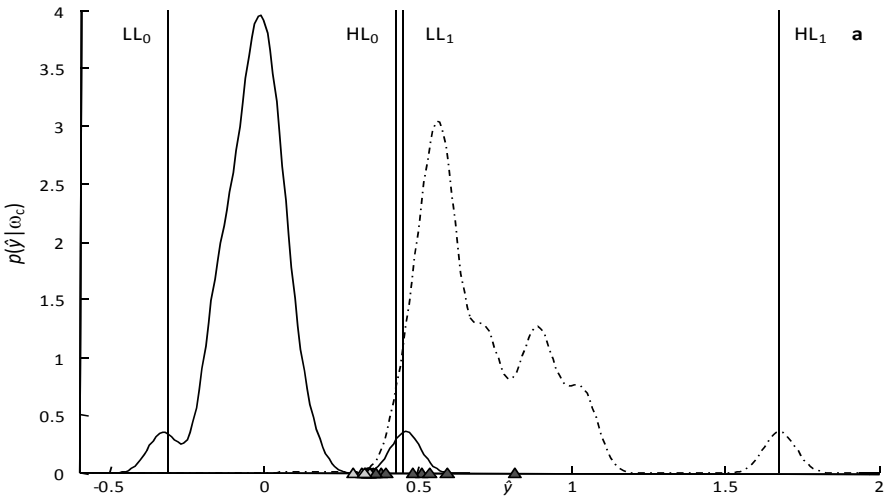
** False Negative (FN): samples of class ω_0 classified in class ω_1, False Positive (FP): samples of class ω_1 classified as class ω_0, Reject Negative (RN): samples of class ω_1 rejected, Reject Positive (RP): samples of class ω_0 rejected, True Negative (TN): samples of class ω_1 correctly classified, True Positive (TP): samples of class ω_0 correctly classified.

5.3.3 Small round blue cells tumour dataset

The same strategy as for the prostate cancer dataset was followed. In this case the *p*-DPLS models were calculated using the 96 most significant gene expressions according to reference [38]. Preliminary *p*-DPLS models were calculated with 1 to 3 factors using mean-centered gene expression data and then validated by LOOCV. The optimal model, with the lowest cost of classification per sample was the one factor model. For this model, four training samples were detected as outliers. Three of them had large *x*-residuals with values s_t^2/s_T^2 of 4.98 (sample EWS_T13), 6.11 (sample RMS_T7), and 8.43 (sample RMS_T11) larger than the cut-off value of 3. Moreover, the prediction of sample RMS_T11 was $\hat{y}$ = 1.91, higher than the class limit HL_1=1.35. The fourth outlier, sample EWS_T12, had a prediction $\hat{y}$=–0.18, lower than the limit LL_0=–0.072. After deleting

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

these four samples, the *p*-DPLS model was recalculated and used to predict the test samples. Without reject option, all the test samples would have been incorrectly classified by the model. With the reject option, 19 out of the 20 test samples were pointed as outliers by the $\hat{y}$ limits because they were inliers. The other sample had the prediction $\hat{y}$ in the acceptance region and hence it was classified, but erroneously. The classification performance would have been worse if the *p*-DPLS model had not been depurated from outliers. Without excluding the training outliers, the PDFs of the model varied, and hence the $\hat{y}$ limits for rejection (Figure 4). In that case (i.e., the *p*-DPLS model calculated with all the training samples) only 13 of the 20 test samples were rejected and the remaining 7 were considered valid by the model and classified either in class ω_0 or in class ω_1 (hence, wrongly classified). Note that the test samples have intermediate values of the *x*-variables between the two modelled classes EWS and RMS. Since the samples are close to the centre of the multivariate space, their predictions were around 0.5, in the middle of the PDFs of the two modelled classes. In this case, none of test samples could have been rejected neither by the leverage criterion (the maximum leverage was *h*=0.01, while the $3\bar{h}$ was 0.15 for this model) nor by the ratio of variances (all had *V*<3). This shows the complementary information that the $\hat{y}$ limits, the leverage and the ratio of variances offer.



UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

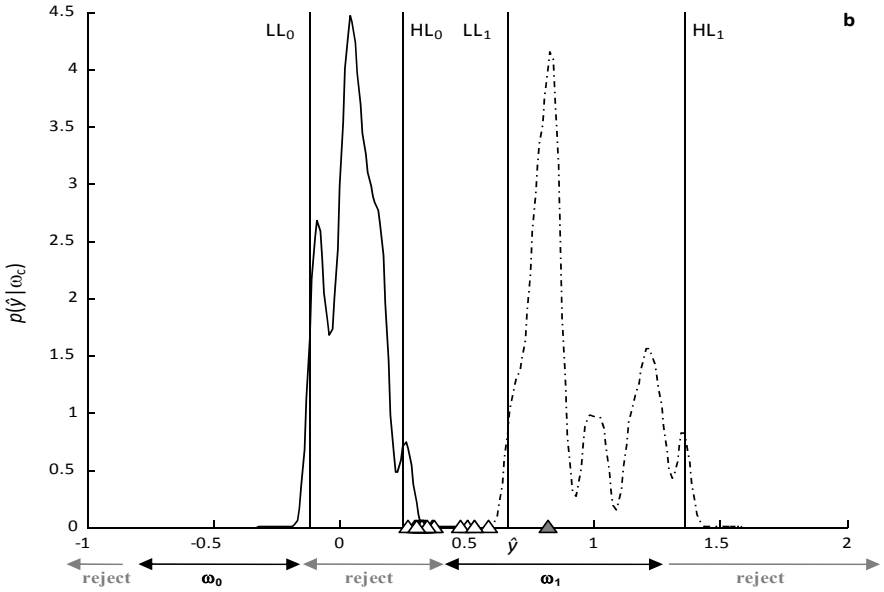


Figure 4. Small round blue cells tumour dataset. PDFs of *p*-DPLS model with one factor **a.** with all the training samples, **b.** without the training outliers. Note how PDFs (and hence, the $\hat{y}$ limits and the rejection and acceptance zones) change when outliers in training set are removed.

5.4 Conclusions

Classification rules for microarray data require appropriate rejection diagnostics. The several steps involved in the generation and measurement of microarray data, that may introduce important errors in the data, as well as the possibility of submitting to the classifier samples from a non-modelled class, make it necessary the use of diagnostics to prevent misclassifications. Rejection diagnostics act both in the training stage of the rule, by identifying those outliers than can degrade the performance of the rule, and in the prediction of new incoming samples, by identifying those samples that will likely be misclassified. Within this approach, the classification model is not forced to classify any

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

future sample that arrives. This work extends the previous work on reject option for *p*-DPLS that was based only on the predicted $\hat{y}$, which has been shown to be not always sufficient to detect outliers. Both training and prediction outliers were now detected by taking into account the *x*-residuals, the leverage and the predicted $\hat{y}$. The possibility of using *x*-residuals is an advantage of classification methods based on latent variables such as *p*-DPLS. The deletion of the training outliers from the training set improved the classification model. At the prediction stage, samples were rejected to classify either because they were outliers, or because they were ambiguous.

Acknowledgements

The authors thank the support of the Departament d'Universitats, Recerca i Societat de la Informació de Catalunya for providing Cristina Botella's doctoral fellowship, an of the Spanish Ministerio de Educación y Ciencia (project CTQ2007-66918/BQU).

References

[1] Liu, F. and B. Wu, *Multi-group cancer outlier differential gene expression detection.* Computational Biology and Chemistry, 2007. **31**: p. 65-71.

[2] Li, C. and W.H. Wong, *Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection.* Proceedings of the National Academy of Sciences, 2001. **98**: p. 31-36.

[3] Gottardo, R., et al., *Quality Control and Robust Estimation for cDNA Microarrays With Replicates.* Journal of the American Statistical Association, 2006. **101**: p. 30-40.

[4] Churchill, G.A., *Fundamentals of experimental design for cDNA microarrays.* Nature Genetics, 2002. **32**: p. 490-495.

[5] Cleveland, W.S., *Robust Locally Weighted Regression and Smoothing Scatterplots.* Journal of the American Statistical Association, 1979. **74**: p. 829-836.

[6] Yang, Y.H., et al., *Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.* Nucleic Acids Research, 2002. **30**: p. e15.

[7] Paoli, S., et al., *Integrating gene expression profiling and clinical data.* International Journal of Approximate Reasoning, 2008. **47**: p. 58-69.

[8] Moffitt, R., et al., *Effect of Outlier Removal on Gene Marker Selection Using Support Vector Machines.* Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, 2005. **1**: p. 917-920.

[9] Olsen, S.H., D.G. Thomas, and D.R. Lucas, *Cluster analysis of immunohistochemical profiles in synovial sarcoma, malignant peripheral nerve sheath tumor, and Ewing sarcoma.* Modern Pathology, 2006. **19**: p. 659-668.

[10] Mramor, M., et al., *Visualization-based cancer microarray data classification analysis.* Bioinformatics, 2007. **23**: p. 2147-2154.

[11] Model, F., et al., *Statistical process control for large scale microarray experiments.* Bioinformatics, 2002. **18**: p. S155-S163.

[12] Shieh, A.D. and Y.S. Hung, *Detecting Outlier Samples in Microarray Data.* Statistical Applications in Genetics and Molecular Biology, 2009. **8**: article 13.

[13] Tomlins, S.A., et al., *Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer.* Science, 2005. **310**: p. 644-648.

[14] Bandyopadhyay, S. and S. Santra, *Agenetic approach for efficient outlier detection in projected space.* Pattern Recognition, 2008. **41**: p. 1338-1349.

[15] Loo, L.-H., et al., *New Criteria for Selecting Differentially Expressed Genes.* IEEE Engineering in Medicine and Biology Magazine, 2007. **26**: p. 17-26.

[16] Tibshirani, R. and T. hastie, *Outlier sums for differential gene expression analysis.* Biostatistics, 2007. **8**: p. 2-8.

[17] Wu, B., *Cancer outlier differential gene expression detection.* Biostatistics, 2007. **8**: p.566-575.

[18] Boulesteix, A.-L. and K. Strimmer, *Partial least squares: a versatile tool for the analysis of high-dimensional genomic data.* Briefings in Bioinformatics, 2007. **8**: p. 32-44.

[19] Pell, R.J., *Multiple outlier detection for multivariate calibration using robust statistical techniques.* Chemometrics and Intelligent Laboratory Systems, 2000. **52**: p. 87-104.

[20] Pell, R.J., L.S. Ramos, and R. Manne, *The model space in partial least squares regression.* Journal of Chemometrics, 2007. **21**: p. 165-172.

[21] Martens, H. and T. Naes, *Multivariate Calibration*. 1989, New York: John Willey & Sons.

[22] Chiang, L.H., R.J. Pell, and M.B. Seasholtz, *Exploring process data with the use of robust outlier detection algorithms.* Journal of Process Control, 2003. **13**: p. 437-449.

[23] Pierna, J.A.F., et al., *A methodology to detect outliers/inliers in prediction with PLS.* Chemometrics and Intelligent Laboratory Systems, 2003. **68**: p. 17-28.

[24] Lleti, R., et al., *Outliers in partial least squares regression Application to calibration of wine grade with mean infrared data.* Analytica Chimica Acta 2005. **544**: p. 60-70.

[25] Botella, C., J. Ferré, and R. Boqué, *Classification from microarray data using probabilistic discriminant partial least squares with reject option* Talanta, 2009. **80**: p. 321-328.

[26] Wold, H., *Partial least squares*, in *Encyclopedia of Statistical Sciences* K.a.N.L. Johnson, Editor. 1985, Wiley: New York. p. 581-591.

[27] Pérez, N.F., J. Ferré, and R. Boqué, *Calculation of the reliability of classification in Discriminant Partial Least-Squares Classification.* Journal of Chemometrics and Intelligent Laboratory Systems, 2009. **95**: p. 122-128.

[28] Faber, N.K.M. and R. Bro, *Standard error of prediction for multiway PLS: 1. Background and a simulation study.* Chemometrics and Intelligent Laboratory Systems, 2002. **61**: p. 133-149.

[29] Faber, N.K.M., *Estimating the uncertainty in estimates of root mean square error of prediction: application to determining the size of an adequate test set in multivariate calibration.* Chemometrics and Intelligent Laboratory Systems, 1999. **49**: p.79-89.

[30] Faber, N.K.M., *A closer look at the bias-variance trade-off in multivariate calibration.* Journal of Chemometrics, 1999. **13**: p. 185-192.

[31] Fernández-Pierna, J.A., et al., *Methods for outlier detection in prediction.* Chemometrics and Intelligent Laboratory Systems, 2002. **63**: p. 27- 39.

[32] Maesschalck, R.D., et al., *Decision criteria for soft independent modelling of class analogy applied to near infrared data.* Chemometrics and Intelligent Laboratory Systems, 1999. **47**: p. 65-77.

[33] Chow, C.K., *On optimum recognition error and reject tradeoff.* IEEE -Transactions on information theory, 1970. **16**: p. 41-46.

[34] Singh, D., et al., *Gene expression correlates of clinical prostate cancer behavior.* Cancer Cell, 2002. **1**: p. 203-209.

[35] Sharaf, M.A., D.L. Illman, and B.R. Kowalski, *Chemometrics*. 1986: Wiley-IEEE.

[36] Lu, Y. and J. Han, *Cancer classification using gene expression data.* Information Systems, 2003. **28**: p. 243-268.

[37] Kennard, R.W. and L.A. Stone, *Computer Aided Design of Experiments.* Technometrics, 1969. **11**: p. 137-148

[38] Khan, J., et al., *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.* Nature Medicine, 2001. **7**: p. 673-679.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

CHAPTER 6 Gene selection based on selectivity ratio for probabilistic discriminant partial least squares

Submitted April 2010

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Microarray data are often used to determine if a cell or a tissue is healthy or tumour, or if it belongs to a subtype of a certain tumour. The quality of these classifications depends on the discriminating ability of the multivariate classification model. This ability decreases if irrelevant genes are included in the training data. Hence, gene selection plays a key role in the analysis of microarray data. In fact, gene selection acomplishes several purposes: 1) the identification of genes that are biologically relevant for the development of a certain disease 2) the discovery of coexpressed genes in order to build metabolic pathways and 3) the reduction of the dimensionality of the data in order to make data analysis easier.

Many gene selection methods have been developed. Some are based on biological inferences and some have been developed from other type of data. In some cases, gene selection is based on criteria that can be valid for different types of classification models, such as using genetic algorithms to select the genes that minimize the prediction error of a certain classifier [1]. Different classification strategies can be plugged into this selection scheme, as long as the model takes in certain selected genes and gives out a prediction error that characterizes the selected subset of genes. Others, such as selecting the genes that are most correlated with the class label [2] or based on statistical tests [3-4] ignore how classification algorithms processes the data, so it may not favour the same systematic variations in the data that the algorithm will do.

Since the basis of this thesis has been the application of DPLS, we sought for gene selection that could enhance the characteristics that the DPLS algorithm uses from the data. Hence, in this work, we implement the selectivity ratio (*SR*) index in order to choose the most relevant subset of genes for classification with *p*-DPLS models. The selectivity ratio evaluates specifically the most relevant variables in PLS models. For each variable, this index is the ratio of the explained variance with respect to the

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

residual variance. The best genes are those with a high explained variance and a low residual variance. Hence, the genes with the highest *SR* are selected as significant and the remaining genes are discarded from the analysis.

This paper also discusses another important aspect related to gene selection, namely the influence that the split of the dataset into training and test sets has on the subset of selected genes and on the evaluation of the classifier performance. It is a common practice that the goodness of a gene selection algorithm is checked by classifying a test set. For that purpose, the initial data set is split into a training set and a test set either randomly or using an algorithm such as the Kennard-Stone algorithm. Then, based on the training set, a subset or several subsets of genes are selected, and the classification model is calculated using only these genes. Next, the test set is classified. The subset of genes with the highest classification ability indicates the best goodness of the selection. These selected genes may be relevant only to discriminate the samples of this particular training set and the test accuracy may be overoptimistic since the genes were selected based on the accuracy of classification of this particular test set.

In this chapter it is shown that the split of the data intro training and test subsets influences the accuracy of the classification. Certain splits can lead to classify correctly 100% of the test samples while other splits can only classify correctly 80% of the test set, thus giving a false indication of the true ability of the gene selection algorithm for selecting the best genes. In this work, many random splits of training and test sets have been used for defining the final accuracy of the classification models.

These aspects are discussed and implemented for two datasets, prostate cancer dataset and non-small cell lung cancer dataset. For the prostate cancer dataset, the mean of the accuracies (by cross-validation) of classification increased from 85% (all 5966 genes used) to 94% when only 17 selected genes were used. Equivalently, the

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

mean of accuracies for the test samples increased from 84% to 94%. For the non-small cell lung cancer dataset, the model calculated with only 17 of the 54675 original genes, provided a cross-validation classification accuracy of 93%.

This work has been submitted in April 2010.

References

[1] Tang, E.K., P. Suganthan, and X. Yao, Gene selection algorithms for microarray data based on least *squares support vector machine.* BMC Bioinformatics, 2006. **7**: article 95.

[2] Mao K. Z. and W. Tang, Correlation-Based Relevancy and Redundancy. Measures for Efficient Gene Selection. Pattern Recognition in Bioinformatics, 2007, **4774**: p.230-241.

[3] Dai, J.J., L. Lieu, and D. Rocke, *Dimension reduction for classification with gene expression microarray data.* Statistical Applications in Genetics and Molecular Biology, 2006. **5**: article 6.

[4] Huang, X., et al., *Borrowing information from relevant microarray studies for sample classificatio using weighted partial least squares.* Computational Biology and Chemistry, 2005. **29**: p. 204–211.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Gene selection in microarray data based on selectivity ratio index

C. Botella, J. Ferré*, R. Boqué
Department of Analytical Chemistry and Organic Chemistry, Rovira i Virgili University.
Marcel·lí Domingo s/n, 43007. Tarragona, Spain
*Corresponding author: joan.ferre@urv.cat

***Submitted* April 2010**. (Edited for format)

ABSTRACT

Most of the gene expressions measured in a microarray experiment are irrelevant for the final application of the data. Irrelevant genes may confound the classification models and decrease their performance. In this work, a gene selection method based on the selectivity ratio index is used. This index is specific for the DPLS method and has been used to select the best genes that discriminate between healthy and tumour prostate cancer tissues and that discriminate between different subtypes of non small cell lung cancers. It is also shown that the split of the dataset into training and test sets influences both the genes selected and the estimated accuracy of the classification model. A wrong assessment of the accuracy of the model may lead to either reject a good subset of genes or accept a suboptimal subset. To overcome this influence a repetitive strategy including data split, gene selection, validation and prediction is performed. For the prostate dataset, models calculated with only 17 selected genes were able to classify the samples with accuracies around the 94%, better than models calculated with all the gene expressions (5966) whose accuracies varied between 50 and 100% depending on the data split. For the non-small cell lung cancer dataset the models calculated with the genes selected following the selectivity ratio index had

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

better classification abilities, independently to the split of the data (accuracies from 94 to 98% for leave one out cross-validation) than the models calculated with all the genes.

6.1 Introduction

DNA microarrays simultaneously provide gene expressions for thousands of genes. Usually, only a few of the measurements describe informative genes either overexpressed or underexpressed, while the rest describe unspecific variations or noise. Discovering the co-expressed genes is interesting in order to build metabolic pathways, to know the biological relevance of genes for clinical diagnosis and also to enhance the performance of classification algorithms [1]. Classification of cells and tissues according to their gene expression profiles is one of the main uses of microarray data. Multivariate classification is adversely affected by irrelevant genes, which interfere with the discriminative power of the relevant genes. Hence, gene selection is needed to enhance the accuracy of the classifiers, and it is especially relevant when the biochemical importance of the selected genes will be sought.

In the last years, many methods have been developed to identify the most relevant genes for certain types of diagnoses. Three major groups of methods have been described: filters, wrappers and embedded techniques [2]. Some methods have been based on genetic algorithms [3], random forests [4], weights of support vector machines [5] and statistical tests such as the t-test or the Wilcoxon test [6] to cite a few.

DPLS is one of the most used classification methods for gene expression data [7]. DPLS's most important feature is that it uses linear combinations of the original

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

variables, which enables dimensionality reduction, noise filtering and outlier detection. Although DPLS does not necessarily require variable selection, it is preferable to input only the relevant variables and to discard those that can distort the calculated factor space. Several approaches for gene selection in PLS have been described. Tan *et al.* [8] selected genes using the sum of squared correlation coefficients between the gene expressions and the response variables. Czekaj and Walczak [1] used the stability of regression coefficients, and Li Shen [9], following the work of Guyon *et al.* [5], selected the genes with a high absolute value of the regression coefficient using a recursive feature elimination system. Petterson [10], based on Trygg *et al.* approach [11], used the first weight vector of a PLS model with one factor to estimate the importance of a gene for describing the dependent variable. Other criteria often used to select genes is the Variable Importance on Projection (VIP) [12], which is based on the weighs of the DPLS model and t– or F–statistics [13, 14].

Since each classification method enhances particular features of the data, gene selection based on general criteria (e.g., selecting the genes that are most correlated with the class label) does not always provide optimal solutions. Recently, Rajalahti *et al.* [15] used the selectivity ratio (*SR*) index to discover the relevant variables in a mass spectral profile, detecting peptides in the low molecular mass range without problems of false biomarker candidates. The advantage of this index is that it can be calculated specifically for DPLS so that the variables pointed as relevant have also the largest discriminative power for this type of classification model.

In this work we show the use of the selectivity ratio index to choose the most relevant genes when the classification is carried out using DPLS and microarray gene expression data. It is shown that the initial split of the dataset into a training and a test set may influence significantly the estimated classification performance of the classifiers, and hence the conclusion about the goodness of the selection criterion and of the selected

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

subset of genes. An approach based on repetitive data split, gene selection, training of the classifier and validation is used in order to better estimate the ability of the selected genes for providing a good classifier.

6.2 Methods

6.2.1 *Probabilistic* discriminant partial least squares (*p*-DPLS)

Probabilistic Discriminant Partial Least Squares (*p*-DPLS) is a new version of Discriminant Partial Least Squares (DPLS) regression [16]. Briefly, *p*-DPLS starts by calculating a PLS model of A factors relating a $N \times P$ gene expression matrix (**X**) and a $N \times 1$ vector of ones and zeros that codifies the samples' class (**y**). Next, the training samples are predicted with this model. For each training sample, a potential function is calculated as a gaussian centred at the predicted value $\hat{y}$ and with standard deviation equal to the standard error of prediction (SEP) of that sample. Next, the potential functions of the samples of the same class are averaged to obtain the probability density function (PDF) of class ω_0 and of class ω_1. The classification of a test sample is done by calculating the *a posteriori* probability on each class, based on the prediction $\hat{y}$ of the sample. The performance of DPLS depends on the relevance of the input genes. Below, the selectivity ratio index is introduced as a method for gene selection.

6.2.2 Selectivity ratio index

The selectivity ratio (*SR*) index [15] is based on Kvalheim and Karstang target rotation approach [17]. It is defined as the ratio of the explained variance ($v_{ex,p}$) to the residual variance ($v_{res,p,}$) of a variable:

$$SR_p = \frac{v_{ex,p}}{v_{res,p}} \quad (1)$$

A target projection model is calculated as

$$\mathbf{X} = \mathbf{t}_{TP}\mathbf{p}_{TP}^{T} + \mathbf{E}_{TP} = \mathbf{X}_{TP} + \mathbf{E}_{TP} \quad (2)$$

where $\mathbf{t}_{TP}$ (N×1) are the target-projected scores and $\mathbf{p}_{TP}$ (P×1) and the target-projected loadings. These are obtained as

$$\mathbf{t}_{TP} = \mathbf{X}\,\mathbf{b}_{PLS}/\|\mathbf{b}_{PLS}\| \quad (3)$$

$$\mathbf{p}_{TP}^{T} = \mathbf{t}_{TP}^{T}\mathbf{X}/(\mathbf{t}_{TP}^{T}\mathbf{t}_{TP}) \quad (4)$$

where $\mathbf{b}_{PLS}$ (P×1) are the regression coefficients of the DPLS model calculated for A factors. From Eq. (2), the explained variance for variable p, $v_{ex,p}$, is calculated from the pth column of X_{TP} and the residual variance for variable p, $v_{res,p}$, is calculated from the pth column of E_{TP}.

The genes with the highest SR are the ones that best define the relevant variations in the data.

6.2.3 Effect of data split on performance evaluation

Commonly, gene selection starts by splitting the dataset into a training set and a test set [18-20], either randomly or using a sample selection algorithm such as the Kennard and Stone algorithm [21]. Then, genes are selected so as to optimize a criterion calculated from the training set, and the goodness of the selected genes, and hence of the selection criterion, is checked either by cross validation [22, 23] or by predicting a test set [18-20, 24]. Other debatable approaches, such as selecting the genes that best classify a test set have also been used [25]. The limitation of the single-split approach is that a selection algorithm or a set of selected genes may be discarded because an

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

unfortunate split of the dataset leads to low classification accuracies for the test set. Or the other way round, a suboptimal set of genes can be accepted if the classification ability of that particular test set is high.

In order to overcome this situation, gene selection is done in this work from one thousand different training subsets selected randomly. For each training set, a DPLS model is evaluated and the selectivity ratio index *SR* for each gene is evaluated. After the one thousand iterations, the mean of the *SR*'s of each gene is calculated and the genes with the largest mean *SR* are selected. The usefulness of the genes selected is then checked by calculating the classification accuracy of new five hundred DPLS models calculated using the selected genes after randomly selecting the training and test sets again.

6.3 Results

6.3.1 Datasets

The prostate dataset [26] consists of 50 non-tumour samples (class ω_0) and 52 tumour samples (class ω_1) with 12.600 gene expressions analysed for each sample. This dataset has been previously studied in gene selection studies and used to evaluate the performance of a classification method [4, 27, 28] to cite a few.

The non-small cell lung cancer (NSCLC) dataset [29] consists of 58 samples of the two major histological subtypes of lung cancer, 40 from adenocarcinoma (class ω_0) and 18 from the squamous cell carcinoma (class ω_1) with 54675 gene expressions analysed for each sample.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

6.3.2 Discussion

6.3.2.1 Prostate cancer dataset

The dataset was pre-processed like in [26]. The floor value was set at 10, the ceil value at 16000 and the genes with ($I\max_p$-$I\min_p$) < 50 and ($I\max_p$/$I\min_p$) <5 were removed, where $I\max_p$ and $I\min_p$ are the maximum and minimum intensities of the gene respectively. The intensities of the final 5966 genes left were then $\log_2$ transformed.

This dataset was randomly split into a training set and a test set with the only constraint that the training set should contain 50% of the samples of each class from the initial dataset. Then two-factor DPLS models were calculated with mean-centered data and the *SR* index was calculated for each gene. The number of factors was initially determined as the one with the lowest root mean square error of cross validation using all the genes. It was latter checked that a different reasonable number of factors of the DPLS model did not affect the genes that were selected as relevant after the one thousand repetitions. The procedure was repeated one thousand times and the average *SR* index of each gene was calculated. The 10, 17 and 35 genes with the highest average of *SR* for these models were selected as potentially relevant (Table 1). Figure 1 shows the mean *SR* for the fifty genes with the highest index. After the first 17 selected genes, the remaining genes have similar *SR.* Hence, the discriminative power for the rest of the genes is not relevant enough to justify their inclusion in the model. Anyway, the best 10, 17 and 35 genes were selected in order to compare them with previous selection results using the random probabilistic model building genetic algorithm (RPMBGA) criterion [25].

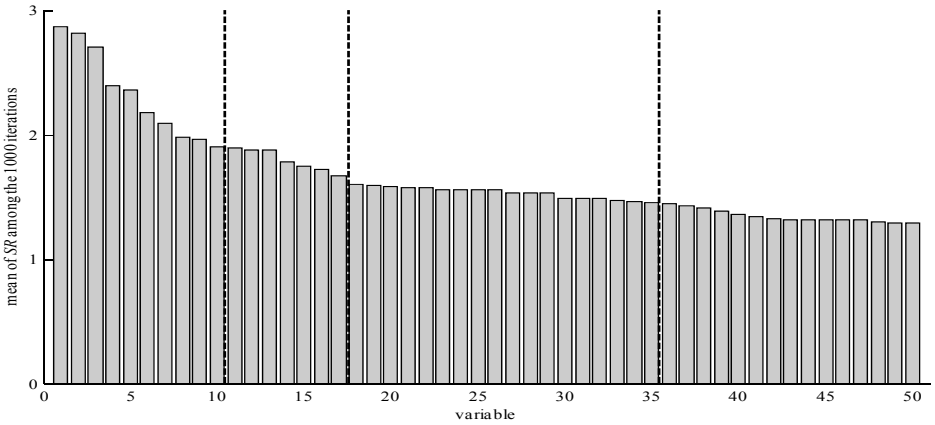


Figure 1. Mean of *SR* among the 1000 iterations for the fifty genes with highest *SR*.

Table 1. The 35 most relevant genes accordingly with the *SR* index calculated from 1000 *p*-DPLS models.

Id of genes selected			
10 genes	**17 genes**	**35 genes**	
37639_at	1767_s_at	39756_at	33137_at
32598_at	36601_at	769_s_at	32076_at
40282_s_at	37720_at	36491_at	1521_at
41468_at	575_s_at	38410_at	35742_at
38406_f_at	39315_at	38087_s_at	32206_at
41288_at	34840_at	40024_at	1740_g_at
38634_at	31444_s_at	38051_at	34407_at
32243_g_at		33904_at	33198_at
37366_at		33362_at	1513_at
40856_at			

*Note that to avoid redundancy, the 17 genes are the 10 in the first column plus the 7 in the second column, and analogously the 35 genes are the 10 in the first column plus the 7 in the second and the 18 in the third and the fourth columns.

The ability of the selected genes to discriminate between tumour and non-tumour samples was evaluated for models calculated using the 10, 17 and 35 relevant genes only [24]. In order to make the results less dependent on data split, the classification

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

performance was calculated for five-hundred *p*-DPLS models. These models were calculated from randomly generated training and test sets with 50% of samples of each class in each set. For a better comparison, the samples in the training set and test set in each repetition are the same for the models calculated with 10, 17 and 35 genes.

The histograms in Figure 2 summarize the validation accuracies of the five hundred models calculated with 10, 17 or 35 genes. For each model (a selected subset of training samples and genes) the leave-one-out cross-validation (LOOCV) accuracy and the test set accuracy were evaluated. If the subset of genes is adequate, one would expect both accuracies be high and similar, independently on the samples used to calculate the model.

Figure 2a shows that the models calculated with 10 genes had LOOCV accuracies from 85% to 100% depending on the data split. Test set accuracies also ranged from 85% to 100%. Most of the models had a LOOCV accuracy of 96% and test set accuracy of 92%. These high values of both accuracies indicate that the subset of genes accounted for the main differences between non-tumour and tumour prostate cancer samples. The fact that the histogram is sharp indicates that the high accuracy was maintained for most of the models and it was quite independent on the split of the samples into training and test sets. Note also that a single unfortunate split can lead to low values of both LOOCV accuracy (88%) and test accuracy (90%), which could lead to reject the selected subset of genes in front of previously reported subsets as they did not improve the performance. Also note that some data splits can lead to models with a large difference between the LOOCV classification accuracies and the test sets classification accuracies (e.g. LOOCV accuracy of 88% and test set accuracy of 100%). These results highlight the relevance that the data split may have when determining the usefulness of a selected subset of genes or the usefulness of a given classification rule.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Similar remarks can be drawn from Figures 2b and 2c for models calculated with the optimal 17 and 35 genes following the *SR* criterion. For 17 genes, the most frequent LOOCV accuracy is 94%, with a test accuracy of 92% (Figure 2b). For the subset of 35 genes, most of the models have high LOOCV and test accuracies of 94% (Figure 2c).

The models calculated with genes selected with the selectivity ratio index were compared with models calculated from genes selected in the bibliography. Figure 2d shows the accuracies when the five-hundred models calculated using optimal genes reported in reference [25]. Note that although the subsets of genes were chosen with a different criterion (RMPMGA) and for a different classifier (support vector machines), they can also give DPLS models with high accuracies. However, the histograms are not as sharp as in Figure 2(a-c), so the quality of the models depends much more on the data into training and test sets than when the genes are selected with the SR index. Reference [25] reported test set accuracies of 98% calculated for one single dataset split. Note that for DPLS those genes can give accuracies as high as 100% for certain dataset splits, but most of them have around 92% accuracy. This suggests an inferior performance for p-DPLS than when the subset selected with the SR index is used.

For the subsets of 17 and the 35 genes, the accuracies varied from 85% to 100% (Figure 2e-2f). Note that in that case the accuracies obtained depended even more on the training and test sets in which the dataset was split and the histograms were more flat.

When using the raw dataset without gene selection (5966 genes), the validation accuracies range from 50% to 100% for different data splits (Figure 3). The mean of LOOCV accuracy was 85% and the mean of test accuracy was 84%. The lower accuracies as compared to using subsets of selected genes can be attributed to the fact that the models are taking into account false correlations. Given the large number of

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

genes, some uninteresting genes may become correlated with the class label for a certain data split, so that the model will assign a high modelling importance to those genes. The test set, which does not show the same correlation pattern, is then classified with a high error. The almost flat histogram suggests that the accuracies change often depending on the split into training and tests set and hence that using all the genes are not able to provide models that systematically perform well.

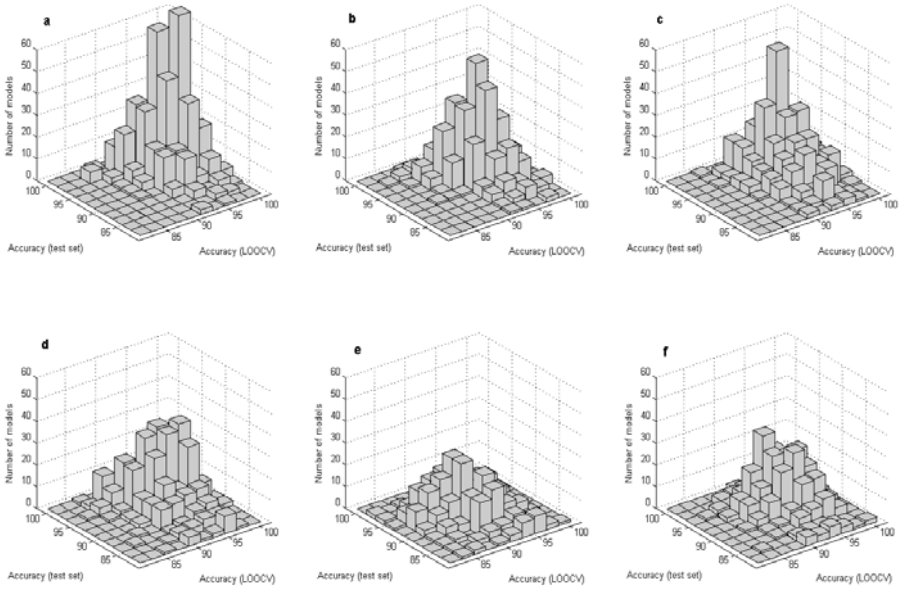


Figure 2. Prostate dataset training (LOOCV) and test accuracy frequences (per unit) for the five hundred *p*-DPLS models calculated with 10 (a, d), 17 (b, e) and 35 (c, f) genes chosen with the SR criterion (a-c) and by RPMBGA (d-f).

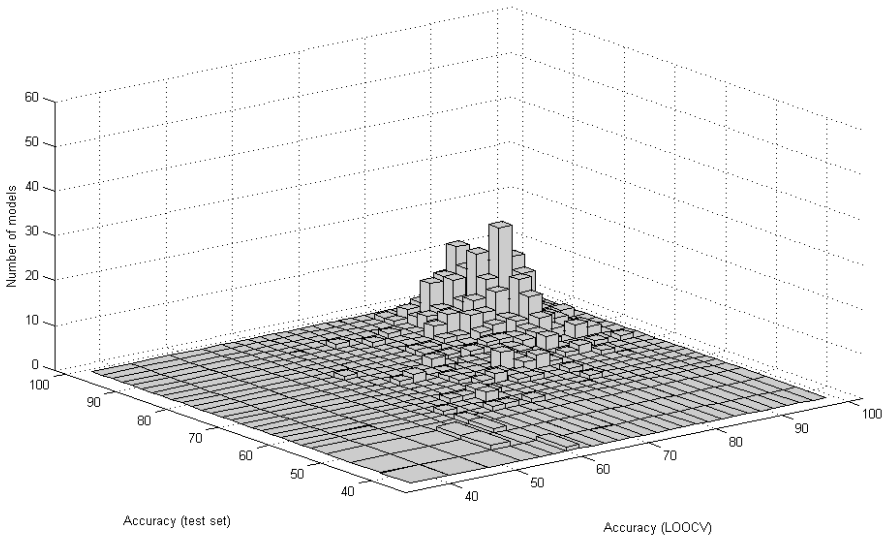


Figure 3. Prostate cancer dataset. Training and test accuracy frequences (per unit) for the five hundred *p*-DPLS models calculated with all final genes the genes after preprocessing (5966 genes).

6.3.2.2 Non-small cell lung cancer dataset

The non small cell lung cancer dataset consists of 54675 gene expressions from 58 samples of asenocarcinoma (AC) and squamous cell carcinoma (SCC). Following the procedure described for the prostate dataset, one thousand randomly training and test subsets were generated and the *SR* index for each gene was calculated for each of the models to discriminate between AC and SCC samples. The 17 and 30 genes with the highest average *SR* index over the one thousand models were selected as relevant (Table 2). This number of genes was decided in order to compare the results with previously reported results [29].

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Table 2. The 30 most relevant genes accordingly with the *SR* index 1000 *p*-DPLS models.

Id of genes selected			
17 genes		**30 genes**	
206032_at	204455_at	1559606_at	1555501_s_at
206033_at	217528_at	205595_at	219507_at
211194_s_at	206164_at	217272_s_at	228806_at
216918_s_at	225822_at	221796_at	206156_at
244107_at	226832_at	235075_at	
207382_at	221795_at	214680_at	
57703_at	222892_s_at	204136_at	
206266_s_at	230464_at	203097_s_at	
206165_s_at		201818_at	

* The 30 genes are the 17 in the first two columns plus the 13 in the third and fourth columns.

The selected genes were used to calculate five hundred *p*-DPLS models using random training and test sets. These models were also compared with the models calculated with the genes selected in a previous work [29].

Figure 4 summarizes the validation accuracies of the five hundred models obtained by LOOCV and by predicting the test set for subsets of 17 and 30 genes. The accuracies for LOOCV and for test data ranged from 85% to 100%. Note that most of the models with the 17 genes selected having maximal *SR* have LOOCV and test accuracies from 94 to 98% (Figure 4a). This fact is even more notable when the 30 genes are used (Figure 4b), for which the number of models with test accuracies out of this range is insignificant. In contrast, for the 17 and 30 genes selected in [29] the *p*-DPLS models have varying accuracies, from 88% to 98%, without dominant training and test accuracy values (Figures 4c-4d). Again, this points out the importance that the data split has on the evaluated accuracies. Note also that the mean of the test accuracies obtained by the models calculated with the genes selected following the *SR* criterion

are slightly better than those obtained with the genes selected in [29] (from the 92% to 93% for the 17 genes subset or from 92% to 93% for the 30 genes subset).

selected in [29] (from the 92% to 93% for the 17 genes subset or from 92% to 93% for the 30 genes subset).

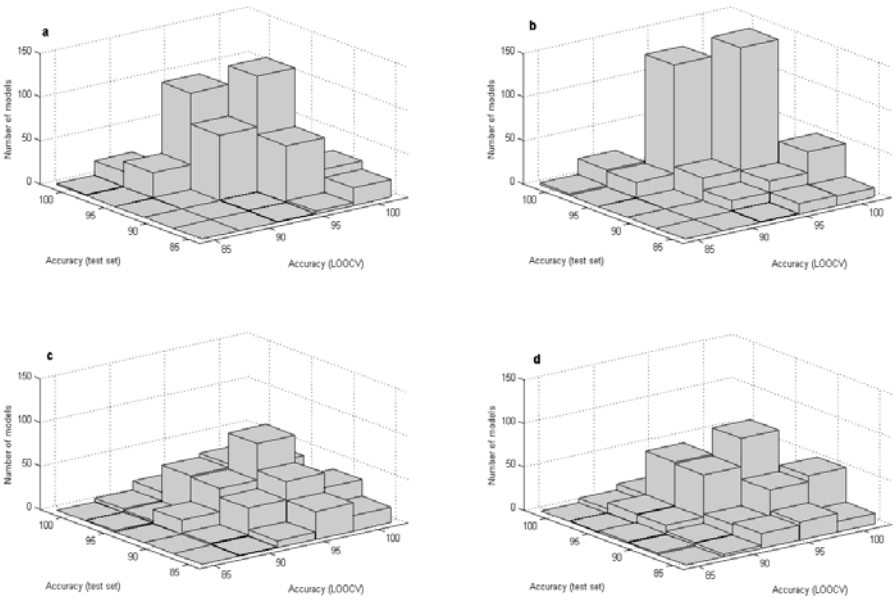


Figure 4. Training and test accuracy frequences (per unit) for the five hundred *p*-DPLS models calculated with 17 (a, c) and 30 (b, d) genes selected by the *SR* criterion (a-b) or in the reference work (c-d).

6.4 Conclusions

The selectivity ratio index has been used to select the best subset of discriminant genes for microarray data classification with *p*-DPLS. The methodology reduces the influence of the samples selected as training samples on the final classification accuracies, and the genes selected give models with very similar classification abilities

independent of the data split. We have also shown that the accuracies of the models may depend to a large extent on the particular samples in the training set and that using a single test set to validate the gene subset may result in either too optimistic or pessimistic conclusions.

Acknowledgements

The authors thank the support of the Departament d'Universitats, Recerca i Societat de la Informació de Catalunya for providing Cristina Botella's doctoral fellowship, an of the Spanish Ministerio de Educación y Ciencia (project CTQ2007-66918/BQU).

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

References

[1] Czekaj, T., W. Wu, and B. Walczak, *Classification of genomic data: Some aspects of feature selection.* Talanta, 2008. **76**: p. 564-574.

[2] Saeys, Y., I. Inza, and P. Larrañaga, *A review of feature selection techniques in bioinformatics.* Bioinformatics, 2007. **23**: p. 2507-2517.

[3] Tang, E.K., P. Suganthan, and X. Yao, *Gene selection algorithms for microarray data based on least squares support vector machine.* BMC Bioinformatics, 2006. **7**: article 95.

[4] Díaz-Uriarte, R. and S.A.d. Andrés, *Gene selection and classification of microarray data using random forest.* BMC Bioinformatics, 2006. **7**: article 3.

[5] Guyon, I., et al., *Gene Selection for Cancer Classification using Support Vector Machines.* Machine Learning, 2002. **46**: p. 389-422.

[6] Troyanskaya, O.G., et al., *Nonparametric methods for identifying differentially expressed genes in microarrays.* Bioinformatics, 2002. **18**: p. 1454-1461.

[7] Boulesteix, A.-L. and K. Strimmer, *Partial least squares: a versatile tool for the analysis of high-dimensional genomic data.* Briefings in Bioinformatics, 2007. **8**: p. 32-44.

[8] Tan, Y., et al., *Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data.* Nucleic Acids Research 2005. **33**: p. 56-65.

[9] Shen, L., *PLS and SVD based penalized logistic regression for cancer classification using microarray data.* Proceedings of the 3rd Asia-Pacific Bioinformatics conference, 2005: p. 219-228.

[10] Pettersson, F. and A. Berglund, *Interpretation and validation of PLS models for microarray data.* Chemometrics and Chemoinformatics ACS Symposium series, 2005. **894**: p. 31-40.

[11] Trygg, J., *O2-PLS for qualitative and quantitative analysis in multivariate calibration.* Journal of Chemometrics, 2002. **16**: p. 283-293.

[12] Musumarra, G., et al., *Potentialities of multivariate approaches in genome-based cancer research: identification of candidate genes for new diagnostics by PLS discriminant analysisy.* Journal of Chemometrics 2004. **18**: p. 125-132.

[13] Dai, J.J., L. Lieu, and D. Rocke, *Dimension reduction for classification with gene expression microarray data.* Statistical Applications in Genetics and Molecular Biology, 2006. **5**: article 6.

[14] Huang, X., et al., *Borrowing information from relevant microarray studies for sample classification using weighted partial least squares.* Computational Biology and Chemistry, 2005. **29**: p. 204–211.

[15] Rajalahti, T., et al., *Biomarker discovery in mass spectral profiles by means of selectivity ratio plot.* Chemometrics and Intelligent Laboratory Systems, 2009. **95**: p. 35-48.

[16] Botella, C., J. Ferré, and R. Boqué, *Classification from microarray data using probabilistic discriminant partial least squares with reject option* Talanta, 2009. **80**: p. 321-328.

[17] Kvalheim, O.M. and T.V. Karstang, *Interpretation of latent-variable regression models* Chemometrics and Intelligent Laboratory Systems, 1989. **7**: p. 39-51.

[18] Horng, J.-T., et al., *An expert system to classify microarray gene expression data using gene selection by decision tree* Expert Systems with Applications, 2009. **36**: p. 9072-9081

[19] Yoon, Y., et al., *Direct integration of microarrays for selecting informative genes and phenotype classification.* Information Science, 2008. **178**: p. 88-105.

[20] Li, L., et al., *Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method.* Bioinformatics, 2001. **17**: p. 1131-1142.

[21] Kennard, R.W. and L.A. Stone, *Computer Aided Design of Experiments.* Technometrics, 1969. **11**: p. 137-148

[22] Hossain, A., et al., *A flexible approximate likelihood ratio test for detecting differential expression in microarray data* Computational Statistics & Data Analysis, 2009. **53**: p. 3685-3695

[23] Li, G.-Z., et al., *Selecting subsets of newly extracted features from PCA and PLS in microarray data analysis.* BMC Genomics, 2008. **9**: p. S24-S38.

[24] Paul, T.K. and H. Iba, *Prediction of Cancer Class with Majority Voting Genetic Programming Classifier Using Gene Expression Data.* EEE/ACM Transactions on Computational Biology and Bioinformatics, 2009. **6**: p. 353-367.

[25] Paul, T.K. and H. Iba, *Gene selection for classification of cancers using probabilistic model building genetic algorithm.* BioSystems, 2005. **82**: p. 208-225.

[26] Singh, D., et al., *Gene expression correlates of clinical prostate cancer behavior.* Cancer Cell, 2002. **1**: p. 203-209.

[27] Dettling, M., *BagBoosting for tumour classification with gene expression data.* Bioinformatics, 2004. **20**: p. 3583-3593.

[28] Jeffery, I.B., D.G. Higgins, and A.C. Culhane, *Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data.* BMC Bioinformatics, 2006. **7**: p. 359-375.

[29] Kuner, R., et al., *Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes.* Lung Cancer, 2009. **63**: p. 32-38.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

CHAPTER 7

Multi-class classification of microarray gene expression data

Submitted May 2010

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Microarray gene expression data were initially used for binary differentiation e.g., to classify a sample or a cell as healthy or tumour. Commonly, however, the diseases with genetic origin have more than two subtypes, so the problem of classifying a sample from gene expression data is more often than not a multi-class classification problem.

Although some classification algorithms can easily handle many classes (e.g., k-nearest neighbours classification), others (e.g. some versions of DPLS) are designed to deal with two classes only. In order to be able to use for multiclass-classification the powerful binary classifiers available, new strategies have to be devised. One of these strategies is to perform binary classifications between pairs of classes, and then combine the results to obtain the final class label. This one-versus-one strategy is often better than to model one class against all the others (the one-versus-all strategy). The reason is that in the one-versus-all strategy, different subtypes of samples are grouped into the same class, which must be differentiated from the target class. In contrast, the one-versus-one strategy allows the model to focus on the genes that actually differentiate one particular class from another particular class.

A difficulty in the one-versus-one strategy is that a new sample will be submitted to all the binary models that make the classification system. For the binary models that modelled the class, the prediction should be that the sample belongs to the modelled class. For all the other models, the sample is an outlier and should be detected as such. Hence, the combination of the results of the binary classifiers in order to obtain the final assigned class is a fundamental step.

In the present work multi-class classification is performed in two steps by combining partial least squares (PLS) regression and the linear discriminant analysis (LDA). In the initial step, one-versus-one PLS models allow obtaining the predictions for each sample

(a single value) and for each model. Each one-versus-one PLS model can only discriminate between two different classes. However, the predictions of samples from the classes not modelled by each PLS model may span all the domain, and hence misclassified. So, the multi-classification is done in a second step with the LDA classifier applied over the predictions of the samples for all the one-versus-one PLS models.

The methodology was used to classify samples of leukemia and small round blue cell tumours datasets. The accuracies of classification were 97%, using only 15 genes, and 100% with 17 genes, respectively.

This paper was submitted in May 2010.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Multi-class classification of microarray gene expression data

C. Botella*, J. Ferré, R. Boqué

Department of Analytical Chemistry and Organic Chemistry, Rovira i Virgili University.
Marcel·lí Domingo s/n, 43007. Tarragona, Spain
*Corresponding author: cristina.botella@urv.cat

Submitted **May 2010**. (Edited for format)

ABSTRACT

When classification from microarray gene expression data is a multi-class problem, the outputs of binary classifiers such as discriminant partial least squares (DPLS) must be combined to obtain the final classification result. In this work a new methodology for multi-class classification that combines partial least squares (PLS) and linear discriminant analysis (LDA) has been developed. The method also includes a gene selection step based on the selectivity ratio index so that the best performing genes for each binary PLS model are selected. When the methodology was applied to the leukemia dataset, that has three classes, 97% of the samples were correctly classified using only 15 genes in the PLS models. For the round blue cell tumour dataset, that has four classes, 100% of the samples were correctly classified using only 17 genes in the PLS models.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

7.1. Introduction

An important challenge in the use of large-scale gene expression data for biological classification occurs when the dataset involves multiple classes [1]. So far, most of the research on classification of microarray data has focused on two major classes only (e.g. normal versus cancer tissue, response to treatment versus no response). However, practical cancer diagnosis requires differentiating among more than two types or subtypes and, hence, multi-class classification techniques are needed [2].

Multi-class classification can be approached in two ways. One way is the use of algorithms that treat multi-class problems directly, such as *k*-Nearest Neighbours (*k*NN), Linear Discriminant Analysis (LDA) or Neural Networks (NN). A second way is to decompose the multi-class problem into multiple binary classification problems and use binary classification algorithms, such as Discriminant Partial Least Squares (DPLS) or Total Principal Component Regression (TPCR). These binary classification models can be calculated by modelling either one class versus the others (one versus all, OVA), one class versus each other class (one versus one, OVO) or using hierarchical partitioning [3, 4]. Then, the results of the binary classifiers are combined to obtain the assigned class label.

Several novel methods have been developed for multi-class classification with microarray data. Tan *et al*. in [5] used TPCR, which takes into account the information of the dependent variables and also the errors in the dependent and independent variables. Ooi *et al*. in [1] used genetic algorithms (GA) for gene selection and classification was based on the maximum likelihood. They obtained better classification accuracies than previously published methods and reduced the number of genes needed for classification. Leng *et al*. in [6] proposed Sparse Optimal Score (SOS), based on Fisher LDA, as a multicategory classifier and classified three public

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

datasets satisfactorily. Tibshirani *et al.* [7] proposed the nearest shrunken centroid method, for cancer class prediction. With the same multi-classification objective, some studies proposed derivations of SVM for multi-classification. Lee *et al*. designed an optimal multicategory SVM [8], Peng *et al*. in [2] and Liu *et al*. in [9] combined GA and one versus one SVM. In contrast, de Souza *et al*. in [10] applied GA and one versus all SVM.

DPLS has proven useful for binary classification of microarray data but it has not been much studied for multiclass classification. Nguyen *et al*. [11] used PLS as a dimension reduction technique for a posterior classification with Logistic discrimination or Quadratic Discriminant Analysis. DPLS2 was used by Tan *et al.* [12] to classify multiclass public datasets using the OVA strategy. However, this strategy may lack biological sense for microarray data analysis when, for instance, healthy samples must be grouped together with tumour samples and discriminated from other tumour types.

In this work we describe the application of PLS combined with LDA for multi-class classification. Several OVO PLS models are calculated and LDA is applied to the predictions of the samples on each of these models. The advantage of using OVO models is that each model maximizes the differences between the two modelled classes. Additionally, gene selection is performed for each PLS model to increase the discriminant ability. The selection is based on the highest selectivity ration index [13] that is specially suited for PLS. The method has been applied to two datasets, the leukemia dataset [14] and small round blue cell tumour dataset [15].

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

7.2 Methods

7.2.1 Multi-class classification method: Partial Least Squares - Linear discriminant analysis

The multi-class classification in *C* classes is done by combining PLS regression and LDA (Figure 1) and may be validated by leave one out cross validation (LOOCV) or by a test set.

PLS is a regression method based on maximizing the covariance between **X** and **y** [16]. The gene expression microarray data, **X** is an *N*×*p* matrix of *N* samples and *P* gene expressions and **y** is a vector of zeros and ones that codifies the classes of the samples. In this paper, one-versus-one PLS models are calculated, so **X** only contains samples from two modelled classes, for instance class ω_1 (e.g. "tumour type I") and class ω_2 (e.g. "tumour type II"). The zeros in **y** codify the samples of class ω_1 and the ones in **y** codify the samples of class ω_2. With these settings, PLS models for every combination of two classes ω_i vs. ω_j *i* = 1,…*C*, *j*>*i* are calculated (Figure 1(c)).

For a sample to be classified in one of the *C* classes, its prediction in each DPLS model is calculated as:

$$\hat{y} = \mathbf{x}^{\mathrm{T}}\hat{\mathbf{b}} \qquad (1)$$

where **b** is the vector of regression coefficients for the model of *A* factors and **x** is the gene expression vector for such sample. Note that if **b** has been calculated from mean-centered data then **x** should be mean-centered and $\hat{y}$ should be processed accordingly. The sample to be classified is predicted in all the OVO PLS models (Figure 1(c)), thus obtaining a vector, of predictions $\hat{\mathbf{y}}$ (Figure 1(d)). For instance, if there are three subtypes of samples, three PLS models are calculated: class ω_1 versus class ω_2, class ω_1 versus class ω_3 and class ω_2 versus class ω_3. The prediction of a sample in these three

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

models generates $\hat{\mathbf{y}} = [\hat{y}_{12}\ \hat{y}_{13}\ \hat{y}_{23}]$ (the subscripts indicate the classes accounted for in each model) that describes the behaviour of the sample in the multiclass-classifier. Ideally, if the sample belongs to class ω_1, $\hat{y}_{12}$ and $\hat{y}_{13}$ should be close to zero and $\hat{y}_{23}$ should be far above 1 or far below 0 so that the sample could be detected as an outlier in the model of class ω_2 vs. class ω_3. Actually, this is not always the case and outliers may have predictions along the entire $\hat{y}$ domain and mixed with the predictions of the modelled classes. Similarly, a sample of class ω_2 should have a $\hat{y}_{12}$ close to one, a $\hat{y}_{23}$ close to zero and an undetermined value of $\hat{y}_{13}$. Finally, a sample of class ω_3 should have $\hat{y}_{13}$ and $\hat{y}_{23}$ close to one and an undetermined value of $\hat{y}_{12}$. LDA is then applied to $\hat{\mathbf{y}}$.

LDA finds discriminant functions (directions) such that the distance between the classes' mean vectors is maximized when the data are projected onto such functions. Let $\hat{\mathbf{y}}$ be the vector of predictions obtained for the sample that must be classified. A discriminant score (*m*) is calculated for that sample in each discriminant function as:

$$m(\hat{\mathbf{y}}) = (\hat{\mathbf{y}} - \boldsymbol{\mu}_c)^{\mathrm{T}} \mathbf{S}_{pooled}^{-1} (\hat{\mathbf{y}} - \boldsymbol{\mu}_c) - 2\ln\pi_c \quad (2)$$

where $\boldsymbol{\mu}_c$ is the mean vector of the predictions of the training samples of class *c*, π_c is the *a priori* probability of class *c* calculated as the number of samples of the class over the total number of samples.

$$\pi_c = \frac{n_c}{N} \quad (3)$$

and $\mathbf{S}_{pooled}$ is the covariance matrix evaluated as:

$$\mathbf{S}_{pooled} = \frac{1}{\mathrm{N}} \sum_{\mathrm{c}=1}^{\mathrm{C}} n_{\mathrm{c}}\, \mathbf{S}_{\mathrm{c}} \quad (4)$$

where $\mathbf{S}_c$ is

$$\mathbf{S}_{\mathrm{c}} = \frac{1}{n_{\mathrm{c}}} \sum_{n=1}^{n_{\mathrm{c}}} (\hat{\mathbf{y}} - \boldsymbol{\mu}_c)(\hat{\mathbf{y}} - \boldsymbol{\mu}_c)^{\mathrm{T}} \quad (5)$$

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Then the sample is classified in the class for which it has the lowest classification score (Figure 1(e)).

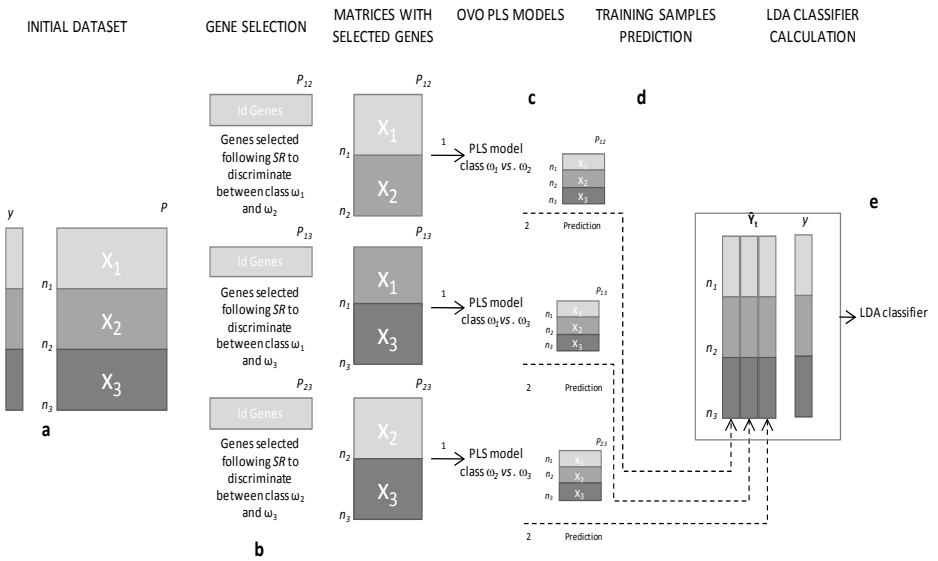


Figure 1. Scheme of a three class PLS-LDA training classification process: **a**. Initial dataset. **b** OVO PLS model with an *A* factors (initial guess) are calculated and genes are selected with the SR index for each model. **c**. The optimal OVO PLS models are calculated with the selected genes. **d**. All the training samples are predicted in each OVO PLS model obtaining a $\hat{\mathbf{Y}}$ matrix. **D**. LDA classifier is calculated, using $\hat{\mathbf{Y}}$ as independent variables and **y** as the class code. Note the P_{12}, P_{23} and P_{13} represent the same number of genes but not necessarily the same genes. The optimal number of factors in the OVO PLS models is those that minimize the RMSECV criterion.

7.2.2 Selectivity ratio index

The Selectivity Ratio (*SR*) index flags the most relevant variables for PLS. It is based on a target rotation approach [17] and is detailed in reference [13]. The *SR* index is defined as the ratio of the explained variance ($v_{ex,p}$) to the residual variance ($v_{res,p}$) of a variable (p):

$$SR_p = v_{ex,p} / v_{res,p} \quad (6)$$

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Taking into account that PLS decomposes **X** as:

$$\mathbf{X} = \mathbf{t}_{TP}\mathbf{p}_{TP}^{T} + \mathbf{E}_{TP} = \mathbf{X}_{TP} + \mathbf{E}_{TP} \quad (7)$$

where $\mathbf{t}_{TP}$ ($P \times 1$) are the target-projected scores and $\mathbf{p}_{TP}$ ($P \times 1$) and the target-projected loadings.The explained variance for each variable p is calculated from the p column of the reconstructed $\mathbf{X}_{TP}$**,** and the residual variance is calculated from the p column of the residual matrix **E**. Note that $\mathbf{t}_{TP}$ and $\mathbf{p}_{TP}$ on equation 7 are calculated following the procedure in [13]. The genes with a highest SR_p index are selected as the more relevant to discriminate between the two classes modelled by the PLS model. Note that each OVO PLS model has its optimal subset of genes that best discriminate between the two modelled classes. The number of genes in the subset may differ from one PLS model to another. To avoid an additional optimization step, the methodology implemented here used the same number of genes for all the PLS models, although the genes were not necessarily the same.

7.3 Datasets

The leukemia dataset [14] consists of 72 samples of acute lymphoblastic leukemias carrying a chromosomal translocation that derives on three subtypes of samples, acute lymphoblastic leukemia (ALL, 24 samples class ω_1), mixed lineage leukemia (MLL, 20 samples, class ω_2) and acute myeloid leukemia (AML, 28 samples, class ω_3). For each sample 12582 gene expressions were obtained. This dataset was pre-processed as described in [14].

The small round blue cell tumour (SRBCT) dataset [15] consists of 63 training samples from four different cell subtypes. 23 samples are from Ewing family of tumours (EWS, ω_1), 20 are rhabdomyosarcomas (RMS, class ω_2), 12 are neuroblastomas (NB, ω_3) and

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

the remaining 8 are Burkitt lymphomas (BL, ω_4). The independent test set has 20 samples, 6 of class ω_1, 5 of class ω_2, 6 of class ω_3 and 3 of class ω_4. For each training and test sample 2308 genes were analysed. The algorithms were run in Matlab® software.

7.4 Results

7.4.1 Leukemia dataset

Three OVO PLS models were calculated for *A* factors: a model of ALL vs. MLL, a model of ALL vs. AML, and a model of MLL vs. AML. For each PLS model, genes having the highest SR index were selected. Three groups of 15, 50 and 100 genes were tested so that the results could be compared with previous results [14, 18]. The OVO PLS models were recalculated using the selected genes and the optimal number of factors was the one that minimized the root mean square error of leave-one-out cross-validation (RMSECV). Note that this number of factors may differ from the ones used in the preliminary model used for selecting the genes.The three optimal PLS models were used to predict all the training samples. A matrix $\hat{\mathbf{Y}}$ (72×3) of predictions was then obtained and used for training the LDA classifier. A sample to be classified was first predicted with the three PLS models, thus obtaining a vector, $\hat{\mathbf{y}}$ (3×1) of predictions. This vector was supplied to the LDA classifier to obtain the final classification.

In this dataset, a test set was not available, so leave-one-out cross-validation (LOOCV) was carried out. Hence, all the samples were used once as a test sample, obtaining for each one a (3×1) vector of predictions, and yielding matrix $\hat{\mathbf{Y}}_t$ (72×3) of predictions in total. This matrix was used to predict the class with the LDA classifier calculated with the training samples in the previous step.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Modelling with the 15 most relevant genes

Figure 2a shows the LOOCV predictions from the three binary PLS models calculated with only the 15 most discriminant genes, selected according to the *SR* index. Note that the model of ALL vs. MLL can discriminate correctly samples from the class ALL (whose predictions are around 0) from the samples of class MLL (whose predictions are around 1). However, it cannot differentiate the samples from class AML. Ideally, these samples should behave differently and have extreme predictions, so that they could be detected as outliers. Instead, their predictions are between the values 0 and 1, so the predictions of the PLS model only are not enough for correctly classifying all the samples. A similar situation happened with the MLL samples in the model ALL vs. AML and with the AML samples in the model ALL vs. MLL (Figure 2a).

Next, LDA was applied to the predictions $\hat{\mathbf{y}}$ of each LOOCV sample. Figure 2b shows the validation samples already classified by LDA in the space of the PLS. The LOOCV classification accuracy was 97.2%, higher than the 95% accuracy by LOOCV previously reported for this dataset [14] using *k*NN and selecting the genes following a signal to noise criterion. A 97.2% of accuracy means that the method only misclassified 2 of the 72 samples. These two misclassified samples, MLL_2 and MLL_15, are samples of class MLL that were assigned to class AML. Figure 3 shows the discriminant scores of the LDA classifier for the first two discriminant functions. Note that for these samples the discriminant score in the second discriminant function is not high enough to be assigned to their true class MLL. Both samples have raw intensities lower than the intensities of the samples of their true class MLL and more similar to the intensities of the samples of class AML. As a consequence, the discriminant scores and the predicted $\hat{y}$'s for these two samples were more similar to the $\hat{y}$'s for class AML. More concretely (Table 1) MLL_2 has predictions $\hat{y}_{12} = 0.62$ and $\hat{y}_{13} = 0.94$, which are almost equal to the mean of the predictions of the samples of class AML ($\bar{\hat{y}}_{12}= 0.68$ and $\bar{\hat{y}}_{13} = 0.97$) and differ considerably from the mean of the predictions for the samples of its true class

($\bar{\hat{y}}_{12}$= 0.94 and $\bar{\hat{y}}_{13}$ = 0.61). The predictions for the model of MLL vs. AML did not contribute significantly to the classification of the MLL_2 sample, having a value between the predictions of both classes.

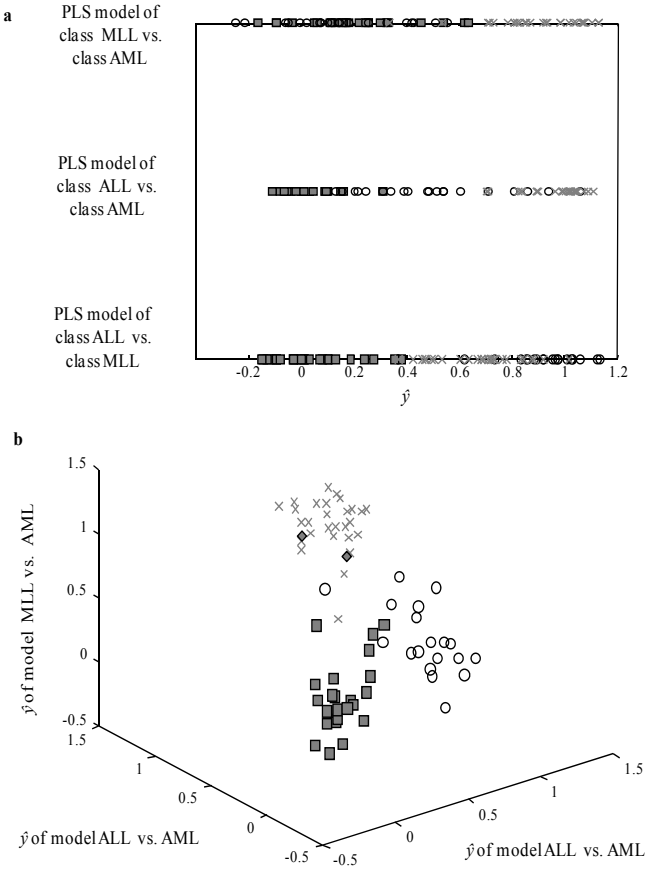


Figure 2a.Predictions of LOOCV samples for OVO PLS models **2b.** Samples classified according to LDA based on the LOOCV predictions of the OVO PLS models calculated with the 15 genes selected with the highest *SR* index. (▪) samples of class ALL correctly classified, (○) samples of class MLL correctly classified, (×) samples of class AML correctly classified, and (♦) misclassified samples.

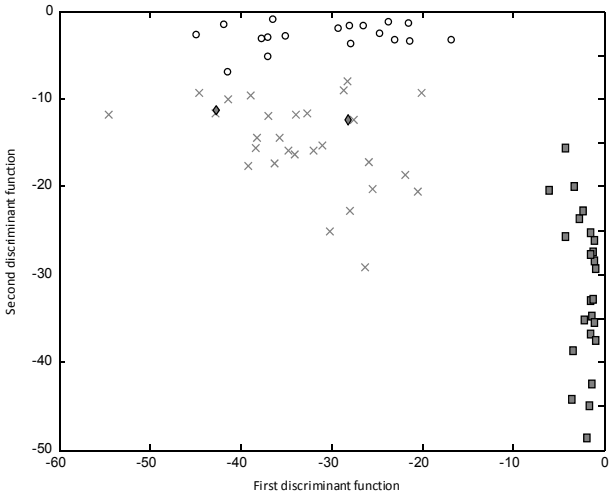


Figure 3. Discriminant scores of the LDA classifier calculated for the first two discriminant functions. (▪) samples of class ALL (○) samples of class MLL, (×) samples of class AML, and (♦) misclassified samples.

Modelling with the 50 most relevant genes

When the number of genes selected to calculate the PLS models was 50, the classification performance was similar as for 15 genes, except for one additional sample that was misclassified. The predictions of each class are more clustered around their target values, which should improve the discrimination between the classes. However, the two outliers detected when the classification was performed with 15 genes, MLL_2 and MLL_15, were again outliers. In addition, the sample AML_11 was also pointed out as outlier. This resulted in a 95.8% of LOOCV classification accuracy. In this case, then, increasing the number of genes worsened the classification. This contrasts with previous results where the best accuracies were obtained with 50 genes [18].

Figures 4a and 4b show the predictions and the LOOCV results for the models calculated with 50 genes. The two samples of class MLL misclassified (MLL_2 and MLL_15) behave like in the models calculated with 15 genes. AML_11 is an AML

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

sample whose intensities for these 50 selected genes are higher than the expected for a sample of class AML. This did not happen when only 15 genes were used. These high intensities influenced the predicted **y**, which was similar to the predictions of the MLL samples and very different from the predictions of the samples of its true class.

When the number of genes increased to 100 the classification performance was like the performance of the models with 50 genes, and the three samples pointed above as outliers were again misclassified.

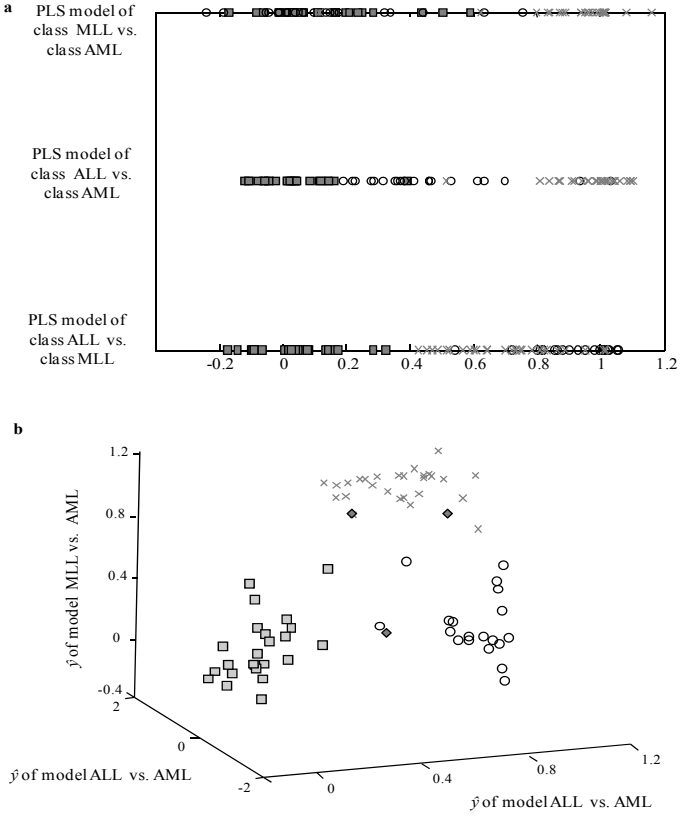


Figure 4a. Predictions of LOOCV samples for OVO PLS models calculated with the 50 genes with highest *SR* index. **4b**. Classification of LDA from the OVO PLS predictions. (▪) samples of class ALL correctly classified, (○) samples of class MLL correctly classified, (×) samples of class AML correctly classified and (♦) misclassified samples.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

7.4.2 Small round blue cell tumour dataset

Following the procedure described for the leukemia dataset, OVO PLS models were calculated. By combining the four different classes, six PLS models were calculated. For each one, the best 17 discriminant genes, obtained using the *SR* index, were selected. The optimal number of factors for each one of the six PLS models was determined based on the minimum RMSECV. The optimal PLS models were used to predict all the training samples, which where then submitted to the LDA classifier. Figure 5 shows the predictions for the test samples for three of the six PLS models, along with the classification performed by LDA from those predictions. From the OVO PLS predictions LDA was able to classify correctly all test samples.. Note that in reference [15] a 100% of test accuracy was achieved using 96 genes. With PLS-LDA, the same performance is achieved using only 17 genes, selected independently for each one of the OVO PLS models.

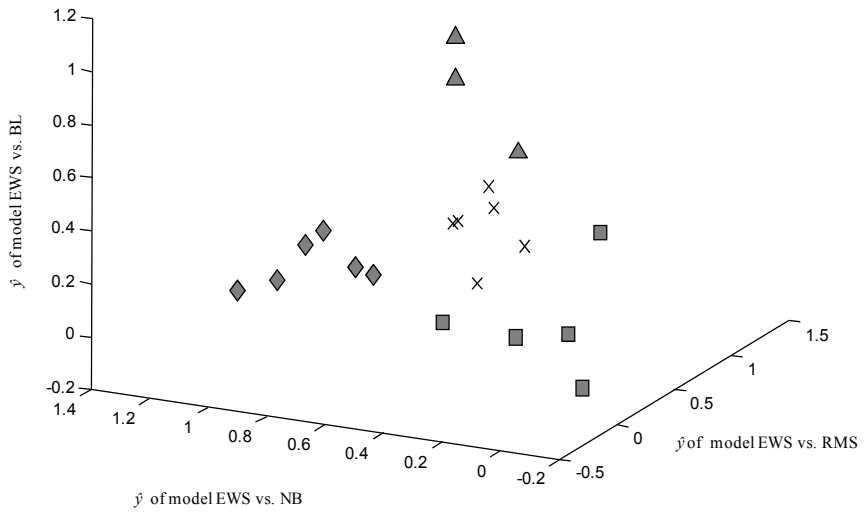


Figure 5. Predictions from three of the six PLS models and the classification performed by LDA. (×) samples of class EWS (▪) samples of class RMS (♦) represents samples of class NB (▲) represents samples of class BL). All of the samples are correctly classified.

7.5 Conclusions

LDA applied on the predictions of one-versus-one PLS models allows multi-class classification of microarray gene expression data with good performance. By selecting the most discriminant genes independently for each PLS model, the accuracies are similar to those previously published but using less genes. In addition, the use of only a few genes allows a better posterior interpretation of the biological sense of the genes and their relation with a particular illness.

Acknowledgements

The authors thank the support of the Departament d'Universitats, Recerca i Societat de la Informació de Catalunya for providing Cristina Botella's doctoral fellowship, an of the Spanish Ministerio de Educación y Ciencia (project CTQ2007-66918/BQU).

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

References

[1] Ooi, C.H. and P. Tan, *Genetic algorithms applied to multi-class prediction for the analysis of gene expression data.* Bioinformatics, 2003. **19**: p. 37-44.

[2] Peng, S., et al., *Molecular classi¢cation of cancer types from microarray data using the combination of genetic algorithms and support vector machines.* FEBS Letters, 2003. **555**: p. 358-362.

[3] Statnikov, A., et al., *A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis.* Bioinformatics, 2005. **21**: p. 631-643.

[4] Yeang, C.H., et al., *Molecular classification of multiple tumour types.* Bioinformatics, 2001. **17**: p. S316-S322.

[5] Tan, Y., et al., *Multi-class cancer classification by total principal component regression (TPCR) using microarray gene expression data.* Nucleic Acids Research 2005. **33**: p. 56-65.

[6] Leng, C., *Sparse optimal scoring for multiclass cancer diagnosis and biomarker detection using microarray data.* Computational Biology and Chemistry, 2008. **32**: p. 417-425.

[7] Tibshirani, R., et al., *Diagnosis of multiple cancer types by shrunken centroids of gene expression.* PNAS, 2002. **99**: p. 6567-6572.

[8] Lee, Y., Y. Lin, and G. Wahba, *Multicategory Support Vector Machines: Theory and Application to the Classification of Microarray Data and Satellite Radiance data.* Journal of the American Statistical Association, 2004. **99**: p. 67-81.

[9] Liu, J.J., et al., *Multiclass cancer classification and biomarker discovery using GA-based algorithms.* Bioinformatics, 2005. **21**: p. 2691-2697.

[10] Souza, B.F.d. and A.P.d.L.F.d. Carvalho, *Gene selection based on multi-class support vector machines and genetic algorithms.* Genetics and molecular research, 2005. **4**: p. 599 -607.

[11] Nguyen, D.V. and D.M. Rocke, *Multi-class cancer classification via partial least squares with gene expression profiles.* Bioinformatics, 2002. **18**: p. 1216-1226.

[12] Tan, Y., et al., *Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models.* Computational Biology and Chemistry 2004. **28**: p. 235–244.

[13] Botella, C., J. Ferré, and R. Boqué, *Gene selection in microarray data based on the selectivity ratio index.* Submitted, 2010.

[14] Armstrong, S.A., et al., *MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia.* Nature Genetics, 2002. **30**: p. 41-47.

[15] Khan, J., et al., *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.* Nature Medicine, 2001. **7**: p. 673-679.

[16] Wold, H., *Partial least squares*, in *Encyclopedia of Statistical Sciences* K.a.N.L. Johnson, Editor. 1985, Wiley: New York. p. 581-591.

[17] Kvalheim, O.M. and T.V. Karstang, *Interpretation of latent-variable regression models* Chemometrics and Intelligent Laboratory Systems, 1989. **7**: p. 39-51.

[18] Yang, T.Y., *Efficient multi-class cancer diagnosis algorithm, using a global similarity pattern.* Computational Statistics and Data Analysis, 2009. **53**: p. 756-765.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

CHAPTER 8 Conclusions

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

1. *Probabilistic* Discriminant Partial Least Squares (*p*-DPLS) has been applied to the binary classification of microarray gene expression data.

The probabilistic Discriminant Partial Least Squares (p-DPLS) method has been successfully applied to classification of microarray gene expression data. In the training step, a PLS model is calculated from the microarray data matrix X and the vector y of 0's and 1's that codifies two classes. Next, the training data are predicted with the PLS model for a selected number of factors and their predictions ŷ are used to estimate two probability density functions (PDFs), one for each modelled class. These PDFs define the range of predictions that characterizes each class. In the prediction step, the prediction ŷ of the sample to be classified and the PDFs are used to calculate the a posteriori probability that the sample belongs to each one of the modelled classes. The sample is then assigned to the class with the highest probability.

There are several reasons that make p-DPLS suitable for classifying microarray data. Microarray data involve thousands of variables and a much smaller number of samples. Many of these variables are redundant, falsely correlated or irrelevant to distinguish between classes. The PLS model compresses the large data matrix X into a few latent variables by focussing on the variables in X that are most correlated with the vector of class codes y. Hence, the classifier uses the systematic relevant data variability, so that the prediction ŷ of a sample and the final classification result are minimally affected by irrelevant genes. In addition, since only a few latent variables are used, a noise filtering effect is achieved.

Another advantage of p-DPLS in front of other algorithms that perform discriminant PLS lies in the calculation of the PDFs of each class and in how the class label is assigned. The classical discriminant PLS approach decides the class label based only on whether ŷ is higher or lower than an arbitrary threshold (e.g. 0.5). More elaborated procedures assume that the ŷ's of each class are normally distributed, and the mean and standard

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

deviation of the ŷ's are used to estimate a Gaussian distribution for each class. The threshold is then the ŷ where the PDFs of both classes coincide or (if a priori probabilities are taken into account) where the a posteriori probabilities are the same. None of these approaches has been useful for microarray data. First, there is not reason for setting an arbitrary threshold. Second, the number of samples available for analysis is usually limited and often one class may have many more samples than the other. This makes the prediction of the PLS model be usually not clustered around the target values 0 and 1 that codify the classes, but slightly biased and not normally distributed (see, for example, the predictions in Figure 6 of chapter 4). In p-DPLS the type of distribution of the ŷ's does not need to be assumed and the PDFs are calculated by combining kernel functions. Hence, the PDFs better describe the distribution of the predictions of each class. In addition, the kernel functions use as smoothing parameter the uncertainty of the predictions, so that the relative position of the samples in the multivariate space also contributes to the calculated PDFs through the leverage and the fit of the model. Another advantage of the p-DPLS method used in this thesis is that limits for the range of possible ŷ's of each class can be set, which allows outlier detection (see section 2 below) and the implementation of a reject option that allows rejecting to classify a sample when the a posteriori probabilities for both classes are too similar (see section 2 below). The latent variable structure of the PLS model also offers enhanced outlier detection capabilities based on the leverage and residual variance (see section 4 below). A final advantage of p-DPLS is that diverse variable selection methodologies, already used in PLS regression, can be used to select the most relevant genes for classification. One of these methodologies has been implemented in the p-DPLS, as it is explained in the section 5 below.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

2. A reject option was implemented in *probabilistic* Discriminant Partial Least Squares (*p*-DPLS).

The classification in p-DPLS is based on the Bayes Theorem so that the sample is assigned to the class with the highest a posteriori probability. The straight application of this rule makes produces that a sample will always be assigned to one of the modelled classes even when the sample may be suspected. One of these situations occurs when the prediction of the new sample is at the extremes of the PDF of one class. Such a sample is so different from the training samples (it is an outlier) that it might be misclassified. The second situation occurs when the PDFs of the two classes are partially overlapped, and the sample has a prediction $\hat{y}$ in the overlap zone (called ambiguous region). That sample has characteristics of both classes, so the a posteriori probability to belong to any of the classes is similar and its classification is not reliable enough. While the samples in the two mentioned situations should preferably be not classified, the strict application of the Bayes Theorem forces its assignement into one of the modelled classes. In this thesis, the possibility of not classifying a sample has been implemented in p-DPLS. This is called the reject option. The reject option in p-DPLS is generally overlooked. However, it allows avoiding classifications with a low reliability, by rejecting to classify both outliers and ambiguous samples. This increases the confidence of the experimenter that the classification model yields correct results when a class label is issued for a new sample.

In this work, the reject option for ambiguous samples has been implemented in p-DPLS as a reject threshold (following Chow's rule), and the reject option for outliers has been implemented by setting limits to the allowed $\hat{y}$ values for each class.

An inconvenient of the reject option is that some samples rejected would be classified correctly if reject option is not implemented. Hence, when the reject threshold and

limits are set, a trade-off between the number of samples incorrectly classified, correclty classified and rejected must be achieved.

In this thesis, p-DPLS with reject option has been successfully applied to classify oligonucleotide and miRNA microarray data by rejecting samples that would have been classified incorrectly. With the reject option, for the Small Round Blue Cell Cancer dataset the misclassification rate of the model was reduced from 100% to 10% for test samples from classes not modelled during the training step, and for the Human Cancers dataset from 3% to less than 1% for the training samples classified by cross-validation.

3. The performance evaluation of classifiers must be reconsidered when a reject option is allowed.

A p-DPLS classifier must be evaluated to assure its quality. Common measures of a classifiers' performance are the accuracy or the error rate. These parameters are usually calculated as the number of correct (or erroneous) classifications over the total number of samples classified.

When rejection is not an option, the total number of samples classified is equal to the number of samples that have been submitted to the classifier. In contrast, when rejection is an option the calculation of performance values such as the accuracy or the error rate are still useful but must be reinterpreted to be meaningful. They are equally calculated as the number of correct (or erroneous) classifications over the total number of samples classified. However, the number of samples for which the classifier has given a class label (classified) may be different than the total number of samples submitted to the classifier (the difference is the number of samples that have been rejected).

The reasoning of this reinterpretation is that the analyst wants, first of all, that the class label issued by the classifier is correct. Hence, the performance measure should reflect

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

the percentage of the samples for which the classifier assigned a class, which are the ones for which a decision is taken (e.g., 'tumour type 1', 'tumour type 2'). After that, the analyst may accept the classifier to reject some “difficult” samples (of course, the classifier should classify as many samples as possible and reject as few as possible). In addition, if the accuracy were defined over the total number of samples, classifiers with reject option would always perform worse than models without reject option, because the number of samples correctly classified using the reject otpion would be equal or lower).

The performance measures are also used to decide among several classifiers. For example, in p-DPLS, different classifiers are obtained by selecting a different number of factors in the PLS model. When the reject option is allowed, the error rate alone may not be a sufficient criterion to compare classifiers, since the rejected samples are not included in the count. In that sense, a classifier that rejects most of the samples and classifies correctly the remaining will have a high accuracy, although it is clearly not useful for classification.

A better criterion for evaluating the performance of a classifier is to use the Cost parameter, which takes into account the number of rejected samples. The Cost evaluates the number of correct classifications, the misclassifications and also the rejections of the model, taking into account the individual cost of each of these actions and providing a single value representative of the performance of the classifier or the classification model. The Cost has been used in this thesis to compare the performance of the p-DPLS with reject option and to determine the optimal number of factors for the p-DPLS model. The Cost has also been used to evaluate if removing outliers improves the p-DPLS models (see section 4).

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

4. Outlier detection in *p*-DPLS has been implemented as a reject option.

Microarray data may contain outliers caused by the many steps involved in obtaining the data. Moreover, samples that belong to classes that have not been modelled may also be submitted to the p-DPLS classifier. Hence, outlier detection is a necessary tool for the practical implementation of p-DPLS. Outliers in p-DPLS were detected in this work by combining leverage, variances and predicted values (ŷ) of the p-DPLS model. This method for outlier detection allows to reject not only samples with errors in the instrumental data (x), in the codification (y) or samples with an erroneous x-y relation but also to identify that an incoming sample does not belong to any of the classes in the training set.

In the Small Blue Round Cell tumours dataset, 90 % of the samples of a class not modelled in the training step were detected as outliers using this method. These samples would have been all misclassified if the reject option had not been used. In the prostate dataset, outlier elimination improves the classification model, decreasing the Cost per classification from 0.11 to 0.06. The outlier elimination has also a beneficial effect on the accuracy of the classification of unknown (test) samples, which increases from 95% to 100%, rejecting to classify a sample that had been wrongly classified.

5. Gene selection was implemented in *p*-DPLS with reject option

Most of the thousands of gene expressions in microarray datasets are irrelevant to classify samples. Irrelevant data may degrade the classifier's performance and difficult the understanding of the genes that are discriminating the classes. For these reasons, variable selection is required.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

In this work the selectivity ratio index has been applied as a gene selection method to select the relevant variables in PLS. This allowed pointing out the most relevant genes to discriminate subtypes of prostate cancer and non-small cell lung types of cancer with high accuracy independently on the training and test sets used.

For the prostate dataset, models with only 17 selected genes had a mean LOOCV accuracy of 94%, compared to the 85% accuracy obtained for the p-DPLS model without gene selection (5966 genes). Equivalently, the mean of the accuracies for the test set improved to 92% from the 84% obtained without gene selection. When the number of selected genes increases from 17 to 35, the accuracy did not improve. Similarly for the non-small cell lung cancer dataset, the genes used in the classification were reduced from 54675 to 17, achieving a mean of LOOCV accuracy of 93%. In this case the increase in the number of genes selected from 17 to 30 neither improved the classification accuracy.

The most adequate method for proving the validity of a selected subset of genes (and, in turn, the validity of the gene selection algorithm, and of the gene selection criterion) has also been studied. Most variable selection methods start by initially splitting the dataset into a training and a test set. Such an split influences the calculated accuracy of the classification model and also influences the conclusion about the validity of the selected subset of genes. If the selected genes and the conclusions are based on a single split, underoptimistic or overoptimistic results can be found. A single unfortunate split can lead to low accuracies (around 88%) and, by contrast, a fortunate split can lead to overoptimistic accuracies (around 100%). For this reason, a repetitive strategy of training set and test set splits, gene selection, p-DPLS model calculation and validation was carried out to measure the performance of the selected genes. The genes selected following this strategy provided models much less influenced by the split of the data.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

6. Linear Discriminant Analysis has been combined with PLS to solve multi-class classification problems.

Multi-class classifiers are required for microarray data classification since most of the cells or tissues to be classified may belong to more than two classes.

p-DPLS is suitable to analyse microarray data due to advantages like the use of latent variables or the noise reduction (detailed in section 1), which are important in order to improve the multiclass classification. However, p-DPLS is a binary classifier, hence, it can only discriminate between two classes at a time. One usual option is to reduce the multiclass classification problems to binary classification ones, following a one-versus-one or a one-versus-all strategy; but these strategies are not always enough to achieve an adequate muticlass classification. The inconvenient resides that the DPLS allows discriminating between two modelled classes, but the $\hat{y}$ predicted values of the incoming samples (that may not belong to any of these two classes) present values that span all the $\hat{y}$ domain (i.e. Figure 2a chapter 7). Hence, these samples are confused among the samples of the modelled classes, assigned to any of them and, so, misclassified.

In this thesis a method that combines PLS and linear discriminant analysis (LDA) has been developed for multi-class classification. The method involves also a selection of the most discriminant genes for each of the PLS models. This strategy allows reducing the data dimension and performing the multi-class classification with high accuracy with a few genes. This method has been applied to the leukemia and the small round blue cell tumour dataset. Leukemia data consist on three different types of samples (AML, ALL and MLL) that generally have poor prognosis and the small round blue cell tumour includes four subtypes (NB, RMS, NHL and EWS) the accurate diagnosis of which is essential because the treatment options, responses to therapy and prognoses vary widely depending on it. For both datasets, the accuracies achieved were very high, a

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

97% and a 100% of classification accuracy, respectively, using 15 genes to classify the leukemia dataset and 17 genes for the small round blue cell tumour dataset.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Appendix

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Datasets

Human cancers dataset

The Human Cancers dataset was published by Lu *et al*. in [1]. The normalized dataset is available at [2] together with supplementary information [1]. The dataset consists of 282 microRNA (miRNA, non coding RNA species) of 218 samples (46 healthy and 172 tumour) from twenty tissues (ovary, colon, lung, prostate, bladder, breast, follicular lymphoma, kidney, liver, brain, melanoma, mesothelioma, stomach, uterus, acute myelogenous leukaemia, diffuse large-B cell lymphoma, B-cell ALL, mycosis fungoides, mixed lineage leukaemia and T-cell ALL).

The published dataset had been normalized as detailed in the Supplementary_Notes document:

1. Well-to-well scaling – the reading from each well was scaled such that the total of the two post-labeling controls, in that well, became 4500 (a median value based on a pilot study).
2. Sample scaling – the normalized readings were scaled such that total of the 6 pre-labeling controls in each sample reached 27,000 (a median value based on a pilot study).
3. Floor threshold was set at 32.
4. Data were log2 transformed.

The normalized downloadable data file is a tab-delimited text file (miGCM_218.gct), of 218 samples and 217 gene expression (left after filtering). The first row of the matrix indicates the tissue ID, and the first and the second column detail the gene name and genes description respectively.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

In this original work, the dataset was used to demonstrate the feasibility and utility of monitoring the expression of miRNAs in human cancer tissue. This dataset has been used in other studies. Lodes *et al.* [3] used the miRNA as markers for cancer detection and it has been pointed that miRNAs may be the future of pharmacogenomics [4].

In this thesis it has been used to evaluate the performance of the *probabilistic* DPLS with reject option classifier.

Breast cancer dataset

The Breast Cancer dataset was published by Hedenfalk *et al.* in [5]. The dataset after filtering (3226 genes) is available in [6].

The downloadable data are the normalized gene expression ratios of 21 samples and from three different mutations (BRCA1, BRCA2 and sporadic mutation). The *format description document*, in the same web page, describes the downloadable data. The downloadable data file is a tab-delimited text file, in which the first row indicates the Patient ID for each experiment (1to 21). The second row provides the mutation classification for each experiment, (BRCA1,BRCA2, Sporadic) and the third row provides the experiment ID, (s1996, s1822, etc). Columns 1 to 3 are related to the genes ID and their localization in the plate. Columns 4 to 24 contain gene expression ratios for each gene in each experiment.

The gene expression ratios are derived from the fluorescent intensity (proportional to the gene expression level) of a tumor sample (BRCA1, BRCA2, or Sporadic) divided by the fluorescent intensity of a common reference sample (MCF-10A cell line). The common reference sample is used in all 21 microarray experiments.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

The genes are filtered based on: (a) average fluorescent intensity (level of expression) greater than 2,500 (gray level) across all 21 samples, (b) average spot area greater than 40 pixels across all 21 samples, and (c) no more than one sample in which the spot area is zero pixels.

The ratios, included in the downloadable data file, for each experiment were normalized such that the majority of the gene expression ratios from a pre-selected internal control gene set were around 1.0. No log transformation was done in the downloadable data.

This dataset was previously used to evaluate the performance of classification models [7, 8], for gene selection methods testing [9, 10], for multiclass classification models evaluation [6] and to check imputation methods [11], to cite a few.

We have used this dataset to demonstrate the usefulness of *p*-DPLS with reject option to reject to classify samples from classes not modeled in the training step.

Prostate dataset

The prostate cancer dataset was published by Singh *et al* in [12] and it is available on [13].After filtering, it has 50 non-tumour samples and 52 tumour samples with 12600 gene expressions.

The pre-processing was detailed in the supplementary information document (SuppInfo_CCv3.pdf). Briefly, the data was scaled to reference intensity (mean average difference of all genes present in the microarrays). The genes with average differences below 10 were filtered. Equivalently, the maximum threshold was set at 16000. After thresholding, the relative variation of expression for each gene was determined by dividing the maximum expression (Max) of the gene among all samples by the minimum

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

expression (Min). The absolute variation in expression was determined by subtracting the (Min) from the maximum (Max). The genes with (Max/Min) <5 or (Max-Min) <50 were also filtered.

The downloadable matrix is a tab-delimited text file that contains expression values in Affymetrix's scaled average difference units. Rows 1 to 3 contain the identification of the samples, the scale factor of each microarray (sample) and the number of genes respectively. Associated to each average difference expression number there is a P, M, or A label that indicates whether RNA for the gene is present, marginal, or absent, respectively (as determined by the GeneChip software), based upon the matched and mismatched probes for the genes.

This dataset was previously studied in gene selection studies and used to evaluate the performance of classification methods. To cite a few, Dettling *et al.* [14] used this dataset (and others) to demonstrate that when bagging was used as a module in boosting, the resulting classifier consistently improved the predictive performance; Diaz-Uriarte *et al.* in [15] used this dataset to check gene selection and the performance of a classification using random forest; and Jeffery *et al.* in [16] used this dataset to compare different gene selection methods (and the lists of genes generated by each one) and different classifiers.

In this thesis, this has been used to check the outlier detection and gene skeleton methods implemented to *p*-DPLS classifier.

Small round blue cells tumour dataset

The small round blue cell tumours of childhood dataset was published by Khan *et al*. in [17] and it is available at [18]. The pre-processing of the data is detailed in the *Supplemental Methods* document.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Initially, the expression levels from 6567 genes were measured for each one of the 88 analyzed samples (of which 63 were labelled as calibration samples and 25 were blind tests). In the analysis the red intensity (*ri*) and the relative red intensity (*rri*) were used. Genes were omitted if for any of the samples *ri* was less than 20. This main removed spots for which the image analysis failed. With this cut only 2308 genes were left.

The final downloadable dataset is a tab-delimited text file that contains the natural algorithm of the relative red intensity (*rri*) for all of all the 88 samples and 2308 genes.

This dataset was previously used to check gene selection methods [19, 20], to compare between different linear discriminant methods [21] or to evaluate multi-class classification methods [22].

We have used this to check the ability of the proposed outlier detection method of detecting samples from classes not modeled in the training step of the *p*-DPLS models. Furthermore it has been used to demonstrate the ability of the PLS combined with linear discriminant analysis (LDA) to multi-class classification.

Non-small cell lung cancer

The non-small cell lung cancer (NSCLC) dataset was published by Kuner *et al.* in [23]. The dataset consists of 58 samples of the two major histological subtypes of lung cancer, 40 from adenocarcinoma and 18 from the squamous cell carcinoma. For each one, 54675 gene expressions were analysed. The data were normalized by the gcRMA method published by Wu *et al*. in [24]. From the initial 60 hybridizations two microarray hybridizations (PatID 42 and 421) failed the quality criteria due to local hybridization artefacts and were excluded from further analysis.

The data are available at NCBI GEO database [25] with the dataset identification GSE10245. Raw data are provided as supplementary files, one for each sample. All samples are grouped in a matrix in the *Series Matrix File.* This is a tab delimited file with the hybridizations of the 58 samples.

This dataset was recently published (year 2009) and, as far as we known, it has not been used yet to check classifiers or gene selection method. It has been only used as a reference in biological studies of lung cancer.

We have used non-small cell lung cancer dataset to verify the usefulness of the gene selection method proposed and to show the influence over the accuracies of the classification models that have the initial divisions of the datasets (i.e. the splits of the dataset into a training and a test set).

Leukemia dataset

The leukemia dataset was published by Armstrong et al in [26] and it is available on [27]. This dataset consists of 72 samples of acute lymphoblastic leukemias carrying a chromosomal translocation that derives on three subtypes of samples, 24 samples of acute lymphoblastic leukemia (ALL), 20 samples of mixed lineage leukemia (MLL) and 28 samples of acute myeloid leukemia (AML). For each sample 12582 gene expressions were analysed.

The downloadable data is a tab delimited file text. The file contains Affymetrix "average difference" expression values for all samples. The data are already scaled as detailed in the *File info* document. Linear scaling is used to reduce technical noise due to global intensity differences between scans. Linear regression of all "Present" genes (Affymetrix "P" calls) was used to determine the scaling factor for each scan (the first ALL scan used as a reference). The scaling factor was applied to expression values (regardless of A/P

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

call). Scaling factors ranged from 0.93 to 2.1; all scaling factors are shown in the *scan id* file.

Then once the dataset obtained, user must pre-process it according to the authors in [26] as follows: a floor threshold and a ceiling threshold were set at 100 units and at 16000 units respectively. After this pre-processing, gene expression values were subjected to the variation filter. The variation filter tests for a fold-change and absolute variation over samples, by comparing max/min and max-min intensities. The max/min filter was set at 5 and the max-min at 500 for all experiments.

This dataset had been previously used to compare different gene selection methods [20] and to check different multi-class classification methods and strategies [20, 28, 29].
We have used this dataset to show the ability of the multi-class classifier proposed in this thesis by combining PLS and LDA.

References

[1] Lu, J., et al., *MicroRNA expression profiles classify human cancers.* Nature Letters, 2005. **435**: p.834-838.

[2] http://www.broadinstitute.org/cgibin/cancer/publications/pub_paper.cgi?mode=view&paper_id=114.

[3] Lodes, M.J., et al., *Detection of Cancer with Serum miRNAs on an Oligonucleotide Microarray.* PLOS One, 2009. **4**: p. e6229.

[4] Mishra, P.J. and J.R. Bertino, *MicroRNA polymorphisms: the future of pharmacogenomics, molecular epidemiology and individualized medicine.* Pharmacogenomics, 2009. **10**: p. 399-416.

[5] Hedenfalk, I., et al., *Gene Expression profiles in hereditary breast cancer.* The New England Journal of Medicine, 2001. **344**: p. 539-548.

[6] http://research.nhgri.nih.gov/microarray/NEJM_Supplement/

[7] Boulesteix, A.-L., *PLS dimension reduction for classification with microarray data.* Statistical Applications in Genetics and Molecular Biology, 2004. **3**: p. article 33.

[8] Raza, M., et al., *Comparative Study of Multivariate Classification Methods using Microarray Gene Expression Data for BRCA1/BRCA2 Cancer Tumors.* Proceedings of the Third International Conference on Information Technology and Applications (ICITA'05), IEEE., 2005. **2**: p. 475-480.

[9] Pettersson, F. and A. Berglund, *Interpretation and validation of PLS models for microarray data.* Chemometrics and Chemoinformatics ACS Symposium series, 2005. **894**: p. 31-40.

[10] McLachlan, G.J., R.W. Bean, and L.B.-T. Jones, *A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays.* Bioinformatics, 2006. **22**: p. 1608-1615.

[11] Branden, K.V. and S. Verboven, *Robust data imputation.* Computational Biology and Chemistry, 2009. **33**: p. 7-13

[12] Singh, D., et al., *Gene expression correlates of clinical prostate cancer behavior.* Cancer Cell, 2002. **1**: p. 203-209.

[13] http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi.

[14] Dettling, M., *BagBoosting for tumour classification with gene expression data.* Bioinformatics, 2004. **20**: p. 3583-3593.

[15] Díaz-Uriarte, R. and S.A.d. Andrés, *Gene selection and classification of microarray data using random forest.* BMC Bioinformatics, 2006. **7:** article 3.

[16] Jeffery, I.B., D.G. Higgins, and A.C. Culhane, *Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data.* BMC Bioinformatics, 2006. **7**: p. 359-375.

[17] Khan, J., et al., *Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks.* Nature Medicine, 2001. **7**: p. 673-679.

[18] http://research.nhgri.nih.gov/microarray/Supplement/.

[19] Zhu, S., et al., *Feature Selection for Gene Expression Using Model-Based Entropy.* IEEE/ACM Transactions on computational biology and bioinformatics, 2010. **7**: p. 25-36.

[20] Mohamad, M.S., et al., *Three-Stage Method for Selecting Informative Genes for Cancer Classification.* IEEJ Transactions on Electrical and Electronic Engineering, 2009. **4**: p. 725-730.

[21] Huang, D., et al., *Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data.* Journal of Experimental & Clinical Cancer Research, 2009. **28**: p. 149:156.

[22] Chetty, G. and M. Chetty, *Multiclass Microarray Gene Expression Analysis Based on Mutual Dependency Models.* Pattern Recognition in Bioinformatics, Proceedings. Lecture notes in bioinformatics, 2009. **5780**: p. 46-55.

[23] Kuner, R., et al., *Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes.* Lung Cancer, 2009. **63**: p. 32-38.

[24] Wu, Z., et al., *A model-based background adjustment for oligonucleotide expression arrays.* Journal of the American Statatistical Association, 2004. **99**: p. 909-17.

[25] http://www.ncbi.nlm.nih.gov/geo/.

[26] Armstrong, S.A., et al., *MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia.* Nature Genetics, 2002. **30**: p. 41-47.

[27] http://research.dfci.harvard.edu/korsmeyer/MLL.htm.

[28] Anand, A. and P.N. Suganthan, *Multiclass cancer classification by support vector machines with class-wise optimized genes and probability estimates.* Journal of Theoretical Biology, 2009. **259**: p. 533-540.

[29] Wang, X. and O. Gotoh, *Accurate molecular classification of cancer using simple rules.* BMC Medical Genomics, 2009. **2**: p. 64-87.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Abreviations

AC	Adenocarcinoma
ALL	Acute lymphoblastic leukemia
AML	Acute myeloid leukemia
BL	Burkitt lymphomas
BRCA1	Breast cancer gene 1
BRCA2	Breast cancer gene 2
CDC	Closest distance to center
cDNA	Complementary deoxyribonucleic acid
CV	Cross validation
Cy3	Cyanine 3
Cy5	Cyanine 5
DA	Discriminant analysis
DNA	Deoxyribonucleic acid
DPLS	Discriminant partial least squares
EWS	Ewing family of tumours
FN	False negative
FP	False positive
GA	Genetic algorithms
HL	High limit
KNN	K nearest neighbours
LDA	Linear discriminant analysis
LL	Low limit
LOOCV	Leave one out cross-validation
LOWESS	Locally weighted scatterplot smoothing
MA plot	Ratio - intensity plot
miRNA	MicroRNA, non coding RNA species
MLL	Mixed lineage leukemia
mRNA	Messenger ribonucleic acid
MVT	Ellipsoidal multivariate trimming
NB	Neuroblastoma
NN	Neural networks
NSCLC	Non-small cell lung cancer
OVA	One versus all
OVO	One versus one
PCA	Principal component analysis
Pcs	Principal components
PDF	Probability density function

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

p-DPLS	*Probabilistic* discriminant partial least squares
RMS	Rhabdomyosarcoma
RMSEC	Root mean square of calibration
RMSECV	Root mean square of cross validation
RMSEP	Root mean square of prediction
RN	Reject negative
RNA	Ribonucleic acid
RP	Reject positive
RPMBGA	Random probabilistic model building genetic algorithm
rRNA	Ribosomal ribonucleic acid
SCC	Squamous cell carcinoma
SEP	Standard error of prediction
SOS	Sparse optimal score
SR	Selectivity ratio
SRBCT	Small round blue cell tumour
SVM	Support vector machines
TN	True negative
TNR	True negative rate
TP	True positive
TPCR	Total principal component regression
TPR	True positive rate
tRNA	Transfer ribonucleic acid
VIP	Variable importance on projection

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Publications

Cristina Botella, Joan Ferré, Ricard Boqué. *Classification from microarray data using probabilistic discriminant partial least squares with reject option.*Talanta, 2009, 80(1): 321-329.

Cristina Botella, Joan Ferré, Ricard Boqué. *Outlier detection and ambiguity detection for microarray data in probabilistic Discriminant Partial Least Squares Regression.* Journal of Chemometrics, 2010, Accepted.

Cristina Botella, Joan Ferré, Ricard Boqué. *Gene selection in microarray data based on selectivity ratio*. 2010, Submitted.

Cristina Botella, Joan Ferré, Ricard Boqué. *Multi-class classification of microarray gene expression data.* 2010, Submitted.

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Communications

Cristina Botella, Joan Ferré and Ricard Boqué

A new criterion for selecting the optimal number of factors in Discriminant-Partial Least Squares (DPLS). Application to microarray gene expression data.

VI Colloquium Chemiometricum Mediterraneum, Saint-Maximin. France. 2007

Poster communication

Cristina Botella, Joan Ferré and Ricard Boqué

A new performance criterion for classification methods for microarray gene expression data.

CAMDA (Critical Assessment of Microarray Data Analysis) ,Valencia, Spain. 2007

Poster communication

Cristina Botella, Joan Ferré and Ricard Boqué

Classification of tumour cells from gene expression data using Probabilistic DPLS with reject option.

III Workshop de Quimiometria, Burgos, Spain. 2008

Oral communication

Cristina Botella, Joan Ferré and Ricard Boqué

Reject option implementing outlier detection and ambiguity detection in the classification of microarray gene expression data.

11th Scandinavian Symposium on Chemometrics, Loen, Norway. 2009

Poster communication

UNIVERSITAT ROVIRA I VIRGILI
MULTIVARIATE CLASSIFICATION OF GENE EXPRESSION MICROARRAY DATA
Cristina Botella Pérez
ISBN:978-84-693-5427-8/DL:T-1418-2010

Cristina Botella, Joan Ferré and Ricard Boqué

Gene selection in microarray data based on selectivity ratio.

VII Colloquium Chemiometricum Mediterraneum, Granada. Spain. 2010

Poster communication