KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT INGENIEURSWETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

MICROARRAY DATA ANALYSIS USING SUPPORT VECTOR MACHINES AND KERNEL METHODS

Promotors:
Prof. dr. ir. B. De Moor
Prof. dr. ir. J. Suykens

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Nathalie POCHET

Mei 2006

KATHOLIEKE UNIVERSITEIT LEUVEN
FACULTEIT INGENIEURSWETENSCHAPPEN
DEPARTEMENT ELEKTROTECHNIEK
Kasteelpark Arenberg 10, 3001 Leuven (Heverlee)

MICROARRAY DATA ANALYSIS USING SUPPORT VECTOR MACHINES AND KERNEL METHODS

Jury:
Prof. Dr. Ir. A. Haegemans, voorzitter
Prof. Dr. Ir. B. De Moor, promotor
Prof. Dr. Ir. J. Suykens, promotor
Prof. Dr. Ir. Y. Moreau
Prof. Dr. M. Hubert
Prof. Dr. D. Timmerman
Dr. Ir., Dr.(med) F. De Smet
Prof. Dr. P. Neven
Prof. Dr. I. Vergote

Proefschrift voorgedragen tot
het behalen van het doctoraat
in de ingenieurswetenschappen

door

Nathalie POCHET

UDC 681.3*J3 Mei 2006

D/2006/7515/39
ISBN 90-5682-705-7

Voorwoord

In de zomer van 2002 stuurde ik een mail naar Prof. Bart De Moor met de vraag of ik bij hem een doctoraat zou kunnen beginnen waarbij ik mijn studies bioinformatica en artificiële intelligentie zou kunnen combineren. Zo bracht Prof. Kathleen Marchal mij in contact met Dr. Frank De Smet, de expert in de analyse van microroostergegevens en de klinische interpretatie ervan, en met Prof. Johan Suykens, de expert in LS-SVMs en kernel methoden. Op die manier was ik vertrokken: een onderwerp waar ik me op kon uitleven en goede begeleiding en ruime expertise om me heen. Daarom zou ik in de eerste plaats mijn promotor Prof. Bart De Moor willen bedanken voor de kansen en steun die hij me gegeven heeft om binnen zijn nieuwe en interdisciplinaire bioinformatica groep mijn onderzoek uit te voeren. Ook mijn promotor Prof. Johan Suykens zou ik van harte willen bedanken voor al die keren dat we hebben samengezeten om het een en ander te bespreken en ook voor alles wat ik van je geleerd heb. Speciale dank zou ik willen richten naar mijn begeleider Dr. Frank De Smet die me gedurende deze vier jaren ongelooflijk veel gesteund en geholpen heeft en tegelijkertijd ook een heel goede en enthousiaste collega was. Frank, ook heel erg bedankt om me te betrekken bij de vele interessante samenwerkingsprojecten met U.Z.Leuven. Ook Prof. Kathleen Marchal, heel erg bedankt om mijn onderwerp te helpen uitpuzzelen. Bovendien zou ik ook Prof. Joos Vandewalle, als hoofd van onze afdeling, willen bedanken voor alle steun.

Graag zou ik ook de leden van mijn begeleidingscommissie, Prof. Yves Moreau, Prof. Mia Hubert en Prof. Dirk Timmerman, oprecht willen bedanken voor de steun die ze me tijdens dit onderzoek hebben gegeven en voor het doornemen van dit proefschrift. Verder zou ik ook Prof. Ann Haegemans, als voorzitter, Dr. Frank De Smet, Prof. Patrick Neven en Prof. Ignace Vergote willen bedanken dat zij deel willen uitmaken van de jury van dit doctoraatsproefschrift.

Zonder financiële steun was dit onderzoek nooit mogelijk geweest. Daarom wil ik het IWT bedanken voor de financiële steun gedurende deze vier jaren.

Verder wil ik verschillende leden van de afdeling gynaecologie-verloskunde van het U.Z.Leuven bedanken voor de vele boeiende samenwerkingen. Prof. Ignace Vergote, Prof. Dirk Timmerman en Toon Van Gorp zou ik heel erg willen bedanken voor de interessante discussies tijdens de vele vergaderingen die we hadden binnen ons samenwerkingsproject rond ovariumtumoren. Hierbij zou ik ook Prof. Paul Van Hummelen (Microarray Facility van het V.I.B.) willen vermelden voor de samenwerking tijdens dit project. Prof. Dirk Timmerman wil ik ook graag bedanken voor de vele vergaderingen die we daarnaast hadden en ook voor de interessante besprekingen met uw collega's uit Londen. Verder wil ik ook Prof. Patrick Neven en zijn collega's bedanken voor de enthousiaste samenwerkingen in verband met borsttumoren. Ook Prof. Thomas D'Hooghe en Dr. Attila Mihalyi wil ik bedanken in verband met het onderzoek naar endometriose. Als laatste wil ik Prof. Sabine Van Huffel, Prof. Bart De Moor, Prof. Dirk Timmerman, Prof. Patrick Neven, Prof. Ignace Vergote, Dr. Frank De Smet, Olivier Gevaert en Ben Van Calster bedanken voor de samenwerkingen binnen Biopattern.

Uiteraard wil ik ook mijn collega's binnen de bioinformaticagroep en SCD heel erg bedanken voor de aangename samenwerkingen en ook de vele leuke momenten tussendoor, zowel vroeger als nu. Vooreerst wil ik de mensen bedanken waar ik nauw mee samengewerkt heb, namelijk Frank, Kristof, Olivier, Fabian, Frizo, Jos, Kristiaan, Marcelo, Carlos en Tijl. Verder ook veel dank aan Bert P., Raf, Tom, Ruth, Wouter, Tim, Bert C., Gert, Steven, Pieter, Steffen, Joke, Cynthia, Karen, Thomas, Niels en alle anderen. Ook mag ik Bart, Ida, Ilse en Pela niet vergeten voor alle hulp bij de administratie. Een speciaal bedankje zou ik nog willen richten aan Bert P. en Marcelo voor de vele steun bij onze parallelle sprint naar de eindmeet.

Graag wil ik ook iedereen bedanken die heeft bijgedragen tot het mogelijk maken van mijn geweldige toekomstperspectieven. In de eerste plaats wil ik Prof. Kevin Verstrepen bedanken voor de kans die hij mij gegeven heeft om het komende jaar postdoctoraal onderzoek te verrichten in zijn lab binnen het Bauer Center for Genomics Research op Harvard University. Bovendien wil ik ook Dr. Frank De Smet, Prof. Bart De Moor, Prof. Johan Suykens, Prof. Sabine Van Huffel, Prof. Ignace Vergote, Prof. Dirk Timmerman en Prof. Patrick Neven heel erg bedanken voor de steun die ze mij gegeven hebben bij het halen van de titel Henri Benedictus - BAEF Fellow (of the King Baudouin Foundation and the Belgian American Educational Foundation), welke onontbeerlijk was hiervoor.

Zelfs tijdens een doctoraat is ontspanning een vereiste. Daarom zou ik Valery Oistrakh en Samuel Barsegian heel erg willen bedanken voor de geweldige vioollessen die ik met heel veel plezier bij hen heb gevolgd en voor alles wat ik van hen heb geleerd. Hierbij wil ik ook mijn muzikale broer Christophe bedanken om mij met hen in contact te brengen.

Ook zou ik mijn vrienden en vriendinnen willen bedanken voor de vele leuke momenten de voorbije jaren: Marian, An V.D., An R., Birger, Tinne, Dieter, Nathalie, Bert P., Gert, Ruth, Raf, Frizo, Ben,...

Tot slot zou ik mijn ouders en mijn broer Christophe speciaal willen bedanken omdat ze me steeds weer alle steun en alle kansen geven en ze er steeds voor me zijn.

En Xander, ik ben vooral heel blij dat ik jou heb leren kennen en wil je bedanken voor de leuke noot waarmee je er de laatste maanden steeds in sloeg om me de drukte even te doen vergeten.

Abstract

In this thesis, we investigated how microarray data can be optimally used in clinical management decisions in oncology. For this purpose, we used machine learning techniques like Least Squares Support Vector Machines (LS-SVM) and kernel methods, capable of both handling the high dimensionality and discovering nonlinear relationships in the data.

These methods were studied and fine-tuned to make them more suitable for microarray data in clinical decision-making problems. We performed a systematic benchmarking study to investigate the influence of regularization, nonlinearity and dimensionality reduction on the performance of clinical predictions. We concluded that regularization or dimensionality reduction is required for the classification of microarray experiments. Furthermore, a nonlinear LS-SVM model with a Radial Basis Function (RBF) kernel is a first choice for the classification of microarray experiments.

These methods were incorporated into an interface called M@CBETH (a MicroArray Classification BEnchmarking Tool on a Host server) that is freely available (http://www.esat.kuleuven.be/MACBETH/) and can easily be used by clinicians for making optimal two-class predictions. This web service aims at finding the best prediction among different classification methods by using randomizations of a benchmarking dataset.

We applied these techniques to solve several diagnostic problems on a set of gene expression patterns originating from ovarian tumors. We applied a broad range of classical and linear techniques on these experiments, followed by the set of more advanced nonlinear techniques incorporated in M@CBETH. Both studies indicated that it is possible to distinguish between stage I without recurrence, platin-sensitive advanced-stage and platin-resistant advanced-stage ovarian tumors.

Finally, we compared classical and kernel clustering algorithms on several microarray datasets. This revealed that very good results can be obtained with spectral clustering in terms of internal validation criteria. To realize this, we showed how these internal validation measures can be extended in feature space.

Korte inhoud

In dit proefschrift onderzochten we hoe microroostergegevens optimaal gebruikt kunnen worden bij klinische beleidsbeslissingen in de oncologie. Hiertoe maakten we gebruik van leeralgoritmen zoals Least Squares Support Vector Machines (LS-SVM) en kernelmethoden, welke in staat zijn hoogdimensionaliteit te hanteren alsook niet-lineaire relaties in de data te ontdekken.

Deze methoden werden bestudeerd en verfijnd om ze meer geschikt te maken voor microroostergegevens bij problemen uit de klinische oncologie. We voerden een systematische benchmarkingstudie uit om de invloed van regularisatie, niet-lineariteit en dimensionaliteitsreductie op de performantie van klinische voorspellingen te onderzoeken. We besloten dat regularisatie of dimensionaliteitsreductie vereist is voor de classificatie van microroosterexperimenten. Bovendien geeft een niet-lineair LS-SVM model met een Radiale Basis Functie (RBF) kernel over het algemeen de beste resultaten voor de classificatie van microroosterexperimenten.

Deze methoden werden opgenomen in een interface genaamd M@CBETH (a MicroArray Classification BEnchmarking Tool on a Host server) die vrij beschikbaar is (http://www.esat.kuleuven.be/MACBETH/) en die eenvoudig gebruikt kan worden door clinici voor het maken van optimale voorspellingen. Deze webservice vindt het beste voorspellingsmodel uit verschillende classificatiemethoden door gebruik te maken van randomisaties van een dataset die voor benchmarking doeleinden gebruikt wordt.

We pasten deze technieken toe op een verzameling van genexpressiepatronen afkomstig van ovariale tumoren om in deze context verschillende diagnostische problemen op te lossen. We pasten een brede waaier aan klassieke en lineaire technieken toe op deze experimenten, gevolgd door de meer geavanceerde niet-lineaire technieken beschikbaar in M@CBETH. Dit toonde aan dat het mogelijk is om onderscheid te maken

tussen stadium I ovariumtumoren zonder herval, platinumsensitieve en platinumresistente ovariumtumoren in een vergevorderd stadium (stadium III/IV).

Tenslotte vergeleken we klassieke en kernelclusteringalgoritmen op verschillende datasets gegenereerd met microroosters. Hieruit besloten we dat zeer goede resultaten bekomen kunnen worden met spectrale clustering in termen van interne validatiecriteria. Bij de realisatie hiervan toonden we aan hoe deze interne validatiematen uitgebreid kunnen worden naar de kenmerkenruimte.

Nederlandse samenvatting

Support vector machines en kernel methoden voor analyse van microrooster gegevens

Hoofdstuk 1: Inleiding

Motivatie

Technologische ontwikkelingen in de moleculaire biologie hebben geleid tot het ontstaan van microroosters. Deze zijn in staat expressieniveaus van duizenden genen tegelijkertijd te bepalen. Een van de toepassingsdomeinen van deze technologie is de oncologie (Friend, 1999; Nasir, 2001; Patten-Hitt, 2001; Ahr *et al.*, 2001; Ahr *et al.*, 2002; Nielsen *et al.*, 2002; Perou *et al.*, 2000; Sørlie *et al.*, 2001). Omdat de ontregelde expressie van genen een centrale rol speelt bij het gedrag van tumoren, kan de meting hiervan zeer waardevol zijn om het klinisch gedrag van kwaadaardige processen te voorspellen of te modelleren. Hierdoor worden de fundamentele processen die aan de basis liggen van de carcinogenese mee betrokken in de klinische besluitvorming.

De microroostertechnologie heeft geleid tot het ontstaan van enorme hoeveelheden gegevens. Om hieruit klinisch en biologisch relevante informatie te extraheren, moeten specifieke procedures uit de statistiek gevolgd worden. Deze informatie kan typisch bekomen worden op drie verschillende manieren: door klinische voorspellingen te maken (Furey *et al.*, 2000), door diagnostische klassen te herschikken en te verfijnen (Ben-Dor *et al.*, 2001), en door genen of groepen van genen te selecteren die belangrijk zijn om het onderscheid te maken tussen verschillende groepen tumoren (Guyon *et al.*, 2002). Dit is een interessante uitgangspositie voor de relatief nieuwe wetenschap die bioinformatica wordt genoemd.

Voor klinische toepassingen kan een dataset met microroostergegevens typisch voorgesteld worden door een expressiematrix waarbij de rijen de genexpressieprofielen en de kolommen de expressiepatronen van de patiënten voorstellen. Datasets gegenereerd met de microroostertechnologie bestaan uit een groot aantal genexpressieniveaus per patiënt en een relatief klein aantal patiënten (behorend tot verschillende klassen tumoren). Het grote aantal expressieniveaus per patiënt is een probleem voor de meeste methoden uit de klassieke statistiek. Daarom wordt vaak eerst dimensionaliteitsreductie toegepast op de gegevens voordat ze gebruikt worden (Alter *et al.*, 2000; Guyon *et al.*, 2002; Müller *et al.*, 2001). Support Vector Machines (SVMs) daarentegen, blijken ondanks de hoge dimensies wel in staat deze microroostergegevens goed te leren en te generaliseren, dankzij het regularisatieprincipe waarop ze gebaseerd zijn (Burges, 1998; Cristianini en Shawe-Taylor, 2000; Müller *et al.*, 2001; Marron en Todd, 2002; Schölkopf *et al.*, 1999; Schölkopf *et al.*, 2001; Schölkopf en Smola, 2002; Suykens *et al.*, 2002; Van Gestel *et al.*, 2004; Vapnik, 1998). Bovendien zal naar de toekomst toe de hoeveelheid data afkomstig van microroosters verder toenemen, wat gevolgen heeft voor het verwerken van de gegevens aangezien dan complexe relaties zichtbaar zouden kunnen worden in de datasets. De meeste traditionele methoden uit de statistiek steunen echter op lineaire functies om de relaties in de gegevens weer te geven. Deze functies zijn niet in staat om complexe niet-lineaire relaties te ontdekken. Dit probleem kan opgelost worden door gebruik te maken van complexere kernelfuncties, welke ervoor zorgen dat de gegevens beter gemodelleerd kunnen worden.

De algemene doelstelling in deze thesis is na te gaan hoe microroostergegevens optimaal gebruikt kunnen worden in het klinisch beleid bij neoplastische aandoeningen (tumoren) met de nadruk op een goede wiskundige fundering. Meer specifiek is het doel leeralgoritmen (‘machine learning’) zoals Least Squares Support Vector Machines (LS-SVM) en kernelmethoden, welke in staat zijn zowel de hoogdimensionaliteit te hanteren alsook niet-lineaire relaties in de gegevens te ontdekken, te gebruiken om de analyse van microroostergegevens in een klinische context te optimaliseren. In deze thesis worden methoden bestudeerd en verfijnd zodat ze geschikt zijn om microroostergegevens te kunnen gebruiken bij klinische beslissingsproblemen. Verder worden deze methoden geïntegreerd in een interface die gemakkelijk bruikbaar en vrij toegankelijk is voor clinici. Tenslotte worden deze technieken ook toegepast op verschillende klinische toepassingen.

Ondersteuning van het klinisch beleid bij kanker met hoge-doorvoer technologieën

De klassieke aanpak voor het beleid bij kanker is gewoonlijk gebaseerd op klinische informatie afkomstig van de anamnese, het klinisch, technisch of histopathologisch onderzoek, en van de ervaring van de geneesheer. De fundamentele mechanismen die de diagnostische categorieën, prognose, en therapeutische keuze bepalen worden echter vaak genegeerd. Het in beschouwing nemen van deze mechanismen is van groot belang voor het maken van de juiste beleidsbeslissingen. Hiertoe kan men steunen op gegevens afkomstig van recentelijk ontwikkelde hoge-doorvoer ('high-throughput') technologieën zoals de microrooster- en proteoomtechnologieën. Het transcriptoom is de verzameling van de genen (of het overeenkomstige mRNA) die tot expressie kunnen komen in weefsels. Dit kan gemeten worden met microroostertechnologie. Het proteoom is de verzameling van alle proteïnen aanwezig in weefsels of klinische stalen en kan gemeten worden onder meer met massaspectrometrie. Aangezien deze databronnen meer informatie over het klinisch gedrag zouden kunnen bevatten dan de meer traditionele klinische gegevens, omvat het hoofddoel van deze thesis te onderzoeken hoe microroostergegevens optimaal kunnen gebruikt worden in het klinisch beleid bij tumoren. Op basis van microroostergegevens kunnen meer geoptimaliseerde voorspellingen gemaakt worden voor een individuele patiënt, bijvoorbeeld over het antwoord op therapie, de prognose en over de aanwezigheid van metastasen.

Expressiepatronen zijn parallelle metingen van expressieniveaus voor duizenden genen tegelijk. Dit resulteert in gegevensvectoren die duizenden waarden bevatten. Een microrooster bestaat uit een reproduceerbaar patroon van verschillende DNA-sondes vastgehecht op een drager. Gemerkt cDNA, bereid uit mRNA, wordt gehybridiseerd met de complementaire DNA-sondes aanwezig op het microrooster. De hybridisaties worden gemeten door een laserscanner en kwantitatief omgezet. Twee belangrijke types van microroosters zijn op dit ogenblik beschikbaar: cDNA-microroosters en oligonucleotideroosters. cDNA-microroosters bestaan uit een tienduizendtal gekende cDNA's (bekomen na PCR-amplificatie) geordend in een matrix op een glasplaatje. Oligonucleotideroosters (of DNA-chips) worden gemaakt door de synthese van oligonucleotiden op siliciumchips. Figuur 1.2 toont een schematisch overzicht van een experiment met de cDNA technologie. Beide technieken hebben hun eigen karakteristieken die hier niet zullen besproken worden.

Leeralgoritmen en methoden uit de statistiek: klinische context

Klinische microroostergegevens kunnen geanalyseerd worden vanuit verschillende standpunten. De drie belangrijkste perspectieven zijn: (1) het

maken van klinische voorspellingen (classificatie), (2) het ontdekken van diagnostische klassen (clusteren van experimenten), en (3) het selecteren van relevante genen (of groepen van genen) of dimensionaliteitsreductie (extractie van kenmerken). Tabel 1.1 geeft een overzicht van deze doelstellingen en de respectievelijke traditionele statistische methoden en leeralgoritmen. Elk van deze doelstellingen wordt hieronder besproken. Ondanks het feit dat de selectie van kenmerken gewoonlijk de eerste stap is bij de analyse van microroostergegevens, en hierbij dus classificatie en clustering voorafgaat, bespreken we eerst classificatie en clustering aangezien hierop de nadruk ligt in deze thesis.

Classificatie van experimenten: In een klinische omgeving is het belangrijk dat, aan de hand van metingen met microroosters (eventueel aangevuld met andere klinische gegevens), voor individuele patiënten voorspellingen kunnen worden gedaan i.v.m. prognose, antwoord op therapie, stadiumbepaling, histopathologische diagnose,… Dit gebeurt door middel van modellen die aan de hand van geselecteerde kenmerken de tumor trachten te classificeren. De parameters van het model moeten worden bepaald aan de hand van een verzameling patiënten van wie reeds geweten is tot welke klasse ze behoren (m.a.w. patiënten voor wie bijvoorbeeld de stadiumbepaling, histopathologische diagnose, prognose, effect van therapie,… reeds gekend zijn). Deze verzameling van patiënten wordt de training set genoemd, waarvan gesteld wordt dat ze gebruikt wordt om het model te trainen (of ook bepaling van de parameters van het model). Het getrainde model kan achteraf aangewend worden om voorspellingen te doen voor patiënten van wie de classificatie nog niet gekend is (deze zijn de patiënten van de test set).

Clustering van experimenten: Door gebruik te maken van het uitgebreid arsenaal aan klinische en morfologische parameters kan men een kwaadaardig proces indelen in verschillende categorieën of clusters. De manier van groeperen zal in de meeste gevallen ook het beleid bepalen. Zoals reeds vermeld kunnen patiënten met een gelijkaardige diagnose en therapie (dus patiënten die volgens de huidige kennis tot dezelfde categorie behoren) een variabel verloop kennen. Door het herschikken of opdelen van diagnostische categorieën door middel van clustering kan gepoogd worden deze variabiliteit binnen eenzelfde klasse te verminderen en kan het in sommige gevallen mogelijk zijn om de therapie te verfijnen en het verloop van de ziekte beter te voorspellen. Hier is het dus niet de bedoeling om voorspellingen te gaan maken voor individuele patiënten, maar om te bepalen welke de verschillende tumorklassen en hun eigenschappen zijn.

Kenmerkenselectie: Hier gaat het over het verminderen van het aantal gegevens (of waarden) per patiënt of per meting op het microrooster. Dit wordt ook het probleem van de afname van de dimensionaliteit genoemd. Deze afname is meestal noodzakelijk vooraleer gestart kan

worden met het maken van voorspellingen of het ontdekken van klassen. Belangrijk is echter dat deze afname gepaard gaat met een minimaal verlies aan essentiële informatie. Enkel de meest essentiële kenmerken, nodig voor het bestuderen van een bepaald probleem, moeten worden geselecteerd.

Voornaamste bijdragen van deze thesis

In deze thesis werden verschillende onderzoeksthema's behandeld. Hier geven we een algemeen overzicht van de onderzoeksonderwerpen die besproken worden in deze uiteenzetting. In deze thesis realiseerden we de volgende uitdagingen:

1. We bestudeerden en onderzochten het gedrag van statistische methoden en leeralgoritmen rekening houdend met de specifieke problemen inherent aan microroostergegevens;
2. We optimaliseerden en verfijnden deze methoden voor biomedische toepassingen;
3. We stelden deze methoden ter beschikking in zodanige vorm dat deze eenvoudig bruikbaar zijn voor clinici;
4. We pasten deze methoden toe op verschillende klinische applicaties.

In deze thesis worden methoden bestudeerd en verfijnd zodat ze geschikt zijn om microroostergegevens te kunnen gebruiken bij klinische beslissingsproblemen. Verder worden deze methoden geïntegreerd in een interface die gemakkelijk bruikbaar en vrij toegankelijk is voor clinici. Tenslotte worden deze technieken ook aangewend in verschillende klinische toepassingen.

Hoofdstuk 2: Least Squares Support Vector Machines en kernelmethoden

De traditionele methoden uit de statistiek en de leeralgoritmen beschouwd in dit werk worden wiskundig uitgewerkt in dit hoofdstuk (zie ook Tabel 1.5).

Classificatiemethoden

Een van de meest gebruikte classificatiemethoden is Fisher Discriminant Analyse (FDA). Deze techniek is echter niet geschikt voor het hanteren van hoogdimensionale gegevens. Daarom zijn methoden voor dimensionaliteitsreductie vereist ter voorbereiding van de data. In dit werk beschouwen we FDA als de standaard traditionele statistische methode waarmee de leeralgoritmen vergeleken worden. De kernelversie van FDA kan gezien worden als een speciaal geval van Least Squares Support Vector Machines (LS-SVM). Doordat zij gebaseerd zijn op het regularisatieprincipe hebben zij reeds getoond dat zij in staat zijn direct om te gaan met hoogdimensionale data in een aantal andere toepassingsgebieden zoals tekstontginning (Joachims *et al.*, 2002) en beeldverwerking (Gupta *et al.*, 2002). We zullen verder in dit werk aantonen dat dit leeralgoritme ook een beter gedrag vertoont voor microroostergegevens. Aangezien LS-SVM kan beschouwd worden als een kernelmethode, kan deze techniek bovendien gebruikt worden om zowel lineaire als niet-lineaire modellen te bouwen.

Clustering

Door de conceptuele simpliciteit en de beschikbaarheid in standaard software pakketten, zijn de traditionele clusteringtechnieken zoals K-means (Tavazoie *et al.*, 1999; Rosen *et al.*, 2005) en hiërarchische clusteringalgoritmen (Eisen *et al.*, 1998) de voornaamste clusteringmethoden in een brede waaier van toepassingen. Daarom focust dit werk zich op een klas van lineaire en niet-lineaire kernelclusteringtechnieken gebaseerd op de traditionele K-means clustering. Kernelclusteringmethoden zijn reeds nuttig gebleken onder meer in toepassingen van tekstontginning (De Bie *et al.*, 2004; Dhillon *et al.*, 2004) en beeldverwerking (Zhang en Rudnicky, 2002). Deze kernelclusteringmethoden zijn recentelijk ontstaan voor het clusteren van data waarbij de clusters niet lineair scheidbaar zijn en om niet-lineaire relaties in de data te vinden. Bovendien laten deze technieken toe efficiënter te werken met hoogdimensionale data in termen van de rekencomplexiteit, wat dus interessant is voor toepassingen met microroostergegevens. De kernel K-means en spectrale clusteringalgoritmen worden voor dit doel in dit werk beschouwd.

Dimensionaliteitsreductie

Dimensionaliteitsreductie wordt vaak toegepast op een niet-gesuperviseerde multivariate manier met behulp van Principale Component Analyse (PCA). Deze techniek alsook de kernelversie ervan worden toegepast verder in dit werk ter voorbereiding van de data op classificatie en clustering.

Hoofdstuk 3: Voorspellingsmodellen voor classificatie van klinische microroostergegevens

De voornaamste doelstelling van deze thesis is klinische beleidsbeslissingen te ondersteunen aan de hand van voorspellingsmodellen gebaseerd op gegevens van microroosters. Daarom is het van groot belang optimale modellen te ontwikkelen voor elk classificatieprobleem in de klinische oncologie. In dit hoofdstuk onderzoeken we daarom hoe een optimale performantie bekomen kan worden met voorspellingsmodellen op basis van microroostergegevens.

De dimensies van datasets gegenereerd met microroosters zijn een cruciale factor bij het bepalen welke methoden al dan niet kunnen toegepast worden voor het maken van voorspellingen. Op dit ogenblik is het genereren van microroostergegevens kostelijk gezien deze technologie nog vrij recent en experimenteel is. Daarom is het aantal experimenten dat haalbaar is in economische zin beperkt. Datasets gegenereerd met microroosters zijn typisch gekenmerkt door een hoge dimensionaliteit vanwege een klein aantal patiënten en een groot aantal genexpressieniveaus per patiënt. De meeste classificatiemethoden ondervinden problemen met deze hoogdimensionale natuur van de microroostergegevens en vereisen daarom vooraf dimensionaliteitsreductie (Alter *et al.*, 2000; Guyon *et al.*, 2002; Müller *et al.*, 2001). SVMs daarentegen zijn wel in staat deze data goed te leren en te generaliseren dankzij het regularisatieprincipe waarop zij gebaseerd zijn (Mukherjee *et al*., 1999; Furey *et al*., 2000). Naar de toekomst toe kan verwacht worden dat het aantal patiënten zal toenemen wanneer deze technologie minder duur wordt. Bovendien zijn de meeste classificatiemethoden, zoals bijvoorbeeld FDA, gebaseerd op lineaire functies en daarom niet in staat eventuele niet-lineaire relaties in de microroostergegevens te ontdekken. Door gebruik te maken van kernelfuncties streeft men naar een beter begrip van deze data (Brown *et al*., 2000), vooral wanneer gegevens van meer patiënten beschikbaar worden naar de toekomst toe.

Teneinde een optimale strategie te vinden voor het maken van klinische voorspellingen, voeren we een systematische benchmarkingstudie uit om zo lineaire versies van de standaard technieken te vergelijken met hun kernelversie tegenhangers (gebruik makende van lineaire en RBF kernels). Merk op dat – zelfs met een lineaire kernel – LS-SVM technieken meer geschikt zouden kunnen zijn aangezien zij regularisatie bevatten en dus geen dimensionaliteitsreductie vereisen omwille van toepassing in de duale ruimte. Het toepassen van complexere kernelfuncties zou echter nuttig kunnen zijn voor het bouwen van voorspellingsmodellen op grotere datasets

gegenereerd met microroosters aangezien verwacht kan worden dat datasets in de toekomst meer microroosterexperimenten zullen bevatten. Daarom onderzoeken we in dit hoofdstuk systematisch wat de invloed is van regularisatie, dimensionaliteitsreductie en niet-lineariteit op een grote variëteit van datasets gegenereerd met microroosters. De resultaten op een specifieke verdeling van training, validatie en test set (zoals vaak gerapporteerd in de literatuur) zou echter gemakkelijk kunnen leiden tot misleidende resultaten, zeker in het geval van een klein aantal patiëntgegevens. In plaats van deze studie op een *ad hoc* manier uit te voeren, worden randomisaties op alle datasets gegenereerd om zo een betrouwbaarder idee te krijgen van de te verwachten performantie en de variatie erop.

Vergelijkende studie

1. Datasets:

Deze studie beschouwt 9 classificatieproblemen uit de klinische oncologie, elk 2 klassen omvattend. Hiertoe werden 7 publiek beschikbare datasets gebruikt: tumoren van het colon (Alon *et al*., 1999), acute leukemie (Golub *et al*., 1999), borsttumoren (Hedenfalk *et al*., 2001), hepatocellulaire carcinomen (Iizuka *et al*., 2003), hersentumoren (Nutt *et al*., 2003), prostaattumoren (Singh *et al*., 2002) en borsttumoren (van ‘t Veer *et al*., 2002). Tabel 3.1 toont een overzicht van de karakteristieken van deze datasets.

Systematische benchmarkingstudies zijn belangrijk om betrouwbare resultaten te bekomen teneinde verschillende numerieke experimenten te kunnen vergelijken en te herhalen. Daarom maken we in deze studie niet enkel gebruik van de originele verdeling van elke dataset in training en test set, maar herverdelen (randomiseren) we alle datasets. Alle numerieke experimenten worden vervolgens ook uitgevoerd op 20 randomisaties van de 9 originele datasets.

2. Methoden:

De methoden die gebruikt werden om de numerieke experimenten samen te stellen zijn de klassieke en kernel PCA voor dimensionaliteitsreductie, en FDA en LS-SVM voor classificatie.

3. Numerieke experimenten:

Negen numerieke experimenten worden toegepast op de hierboven beschreven datasets. Deze experimenten kunnen onderverdeeld worden in 2 groepen, afhankelijk van de vereiste procedure voor het optimaliseren van de parameters. De eerste 3 experimenten zijn zonder dimensionaliteitsreductie, namelijk LS-SVM met lineaire kernel, LS-SVM met Radiale Basis Functie (RBF) kernel en LS-SVM met lineaire kernel zonder regularisatie ($\gamma \rightarrow \infty$).

Vervolgens worden 6 experimenten met dimensionaliteitsreductie toegepast. De eerste 2 zijn gebaseerd op klassieke PCA gevolgd door FDA voor het bouwen van het classificatiemodel. Selectie van de principale componenten wordt zowel op een niet-gesuperviseerde als een gesuperviseerde manier gedaan. Dezelfde strategie wordt toegepast bij de laatste 4 experimenten, mits gebruik van kernel PCA met lineaire kernel alsook met RBF kernel in plaats van de klassieke lineaire PCA. Niet-gesuperviseerde selectie van principale componenten maakt eenvoudigweg gebruik van de eigenwaarden van de principale componenten komende van PCA. De gesuperviseerde manier berekent de absolute waarden van de score geïntroduceerd door Golub *et al*. (1999), alsook gebruikt door Furey *et al*. (2000), welke oorspronkelijk toegepast werd op de individuele genexpressieprofielen, voor de principale componenten komende van PCA.

Aangezien het bouwen van een voorspellingsmodel goede generalisatie vereist voor het maken van voorspellingen voor ongeziene test patiënten, is het tunen van de parameters heel belangrijk. Het klein aantal patiënten dat kenmerkend is voor datasets gegenereerd met microroosters beperkt de keuze van de methode om de generalisatieperformantie te schatten. Het optimalisatiecriterium dat hier gebruikt wordt is de leave-one-out cross-validatie (LOO-CV) performantie. In elke LOO-CV iteratie (aantal iteraties is gelijk aan het aantal patiënten) wordt een patiënt uit de dataset verwijderd, een classificatiemodel wordt getraind op de rest van de data en dit model wordt dan geëvalueerd op de verwijderde patiënt. Als maat voor de evaluatie wordt de LOO-CV performantie [(Aantal correct geclassificeerde patiënten)/(Aantal patiënten in de data) · 100]% gebruikt. Een overzicht van de optimalisatieprocedure die gevolgd wordt in het meest complexe geval van kernel PCA met RBF kernel gevolgd door FDA kan gezien worden in Tabel 0.1. Andere optimalisatieprocedures zijn vereenvoudigingen van deze procedure.

Om de resultaten weer te geven, worden 3 maten gebruikt: de LOO-CV performantie (enkel gebaseerd op de training datasets voor het tunen van de parameters), de classificatieperformantie voor training en test sets, en de oppervlakte onder de Receiver Operating Characteristic (ROC) curve (AUC) (Hanley en McNeil, 1982) voor training en test sets. Indien deze laatste 2 gemeten worden op onafhankelijke test sets, geeft dit een idee van de generalisatieperformantie. Wanneer deze ook gemeten worden op de training sets, dan krijgt men een idee van de graad van overfitting door de performanties op training en test sets met elkaar te vergelijken. Overfitting kan gezien worden in het geval van een hoge training set performantie en een lage test set performantie. Hypothesetesten worden uitgevoerd om zo tot een correcte interpretatie van de resultaten te komen, rekening houdend met alle randomisaties. Hiervoor wordt gebruik gemaakt van een niet-

parametrische gepaarde test: de Wilcoxon signed rank test (Dawson-Saunders en Trapp, 1994).

Optimalisatieprocedure:
kernel PCA met RBF kernel gevolgd door FDA

(1) Genereren van het rooster met parameters

voor elke waarde van de kernel parameter binnen het geselecteerde interval
voor elk mogelijk # principale componenten
voor elke LOO-CV iteratie
- neem een experiment apart
- standaardisatie
- dimensionaliteitsreductie (kernel PCA)
- selectie van de principale componenten (niet-gesuperviseerd of gesuperviseerd)
- classificatie (FDA)
- classificeer het apart genomen experiment

einde

bereken de LOO-CV performantie
einde
einde

(2) Optimalisatie van de parameters

voor elke waarde van de kernel parameter binnen het geselecteerde interval
optimaal # principale componenten:
1. beste LOO-CV performantie
2. kleinste # principale componenten *

einde

optimale waarde voor de kernel parameter:
1. beste LOO-CV performantie
2. kleinste # principale componenten *
3. kleinste waarde van de kernel parameter *

* indien meerdere

Tabel 0.1 : *Optimalisatieprocedure voor het tunen van de parameters in het geval van kernel PCA met RBF kernel gevolgd door FDA.*

Conclusie

In het verleden is het gebruik van classificatiemethoden in combinatie met microroosters reeds veelbelovend gebleken voor het ondersteunen van klinische beleidsbeslissingen in de oncologie. In deze studie werden verschillende belangrijke onderzoeksvraagstellingen geformuleerd om de performantie van klinische voorspellingen gebaseerd op microrooster-gegevens te optimaliseren. Deze zijn gebaseerd op niet-lineaire technieken, dimensionaliteitsreductie en regularisatietechnieken, hierbij rekening houdend met de mogelijke toename in grootte en complexiteit van de datasets gegenereerd met microroosters naar de toekomst toe.

Een eerste belangrijke conclusie van deze studie die 9 probleemstellingen gebaseerd op datasets gegenereerd met microroosters omvat, is dat wanneer classificatie met LS-SVM (zonder dimensionaliteitsreductie) wordt uitgevoerd, goed afgestelde RBF kernels kunnen toegepast worden zonder het risico op overfitting op alle bestudeerde datasets. Een tweede conclusie is dat het gebruik van LS-SVM zonder regularisatie (zonder dimensionaliteitsreductie) leidt tot slechte resultaten, wat het belang van regularisatie benadrukt, zelfs in het lineaire geval. Een laatste belangrijke conclusie is dat wanneer kernel PCA wordt uitgevoerd voor classificatie, het gebruik van een RBF kernel bij kernel PCA de neiging heeft te leiden tot overfitting, vooral in het geval van gesuperviseerde selectie van kernmerken. Ook is waar te nemen dat een optimale selectie van een groot aantal kenmerken vaak een indicatie is voor overfitting. Kernel PCA met een lineaire kernel geeft betere resultaten.

Ook al was het mogelijk deze belangrijke algemene conclusies af te leiden uit deze studie, toch kan de beste classificatiemethode om het meest optimale voorspellingsmodel te bouwen verschillen voor elk classificatieprobleem bij kanker. Aangezien het logisch is dat het ontwikkelen van een optimaal voorspellingsmodel van groot belang is met het oog op het gebruik van deze modellen in de klinische praktijk, is het vinden van de beste classificatiemethode in elk specifiek geval van onmiskenbaar belang. Dit idee wordt verder uitgewerkt in het volgende hoofdstuk.

Hoofdstuk 4: Webservice M@CBETH als tool voor classificatie van microroostergegevens

In het vorige hoofdstuk eindigden we met de waarneming dat de beste classificatiemethode om het meest optimale voorspellingsmodel te bouwen, kan verschillen voor het classificeren van elke dataset gegenereerd met microroosters bij kanker. Daarom is het essentieel het beste classificatiemodel voor elke dataset op een individuele basis te ontwikkelen. Dit omvat niet alleen het vinden van de beste classificatiemethode voor elke dataset, maar ook het tunen van alle parameters (bijvoorbeeld de regularisatieparameter, de kernel parameter, en het aantal principale componenten), wat belangrijk is in het proces waarbij het model ontwikkeld wordt. Het exploreren van alle combinaties om het meest optimale classificatiemodel te vinden is complex. Het vinden van dit optimale model voor elke dataset kan een vervelende en niet voor de hand liggende taak zijn voor gebruikers die niet vertrouwd zijn met deze classificatietechnieken. Daarom ontwerpen we de webservice M@CBETH ('a MicroArray Classificatie BEnchmarking Tool on a Host server') om de microroostergemeenschap een eenvoudige tool aan te bieden voor het maken van optimale voorspellingen gebaseerd op twee klassen. In dit hoofdstuk stellen we deze webservice voor die voor elke dataset gegenereerd met microroosters, verschillende classificatiemodellen met elkaar vergelijkt en de beste selecteert in termen van gerandomiseerde onafhankelijke test set performanties.

Website M@CBETH

De website van M@CBETH kan gevonden worden op http://www.esat.kuleuven.be/MACBETH/ en deze biedt twee services aan: 'benchmarking' en 'prediction'. Na registratie en inloggen op de webservice kunnen gebruikers analyses aanvragen voor benchmarking of het maken van voorspellingen. De gebruikers worden per email op de hoogte gehouden van de status van hun analyses die op de server op ESAT draaien. Ze kunnen dit ook volgen op de pagina met de resultaten van de analyses, welke een overzicht toont van al de analyses en welke links bevat naar de bijhorende pagina's met de resultaten.

De belangrijkste service aangeboden op de website van M@CBETH is de benchmarking service. Benchmarking omvat het selecteren en trainen van een optimaal model gebaseerd op de ingegeven benchmarking en de bijhorende klasselabels. Dit model wordt dan bewaard voor direct of later gebruik op prospectieve data. Deze benchmarking service resulteert in een

tabel die de samenvattende statistieken weergeeft (de LOO-CV performantie, de training set classificatieperformantie en AUC performantie, en de test set classificatieperformantie en AUC performantie) voor alle geselecteerde classificatiemethoden, waarbij de beste methode in het rood getoond wordt. Prospectieve data kan ook ingegeven en onmiddellijk geëvalueerd worden tijdens dezelfde benchmarking analyse.

Met de service voor het maken van voorspellingen biedt de website van M@CBETH de mogelijkheid tot latere evaluatie van prospectieve data door het hergebruiken van een bestaand optimaal voorspellingsmodel (dat gebouwd werd tijdens een voorgaande benchmarking analyse door dezelfde gebruiker). Voor beide services is het zo dat als de bijhorende prospectieve labels ook ingegeven worden, de prospectieve performantie berekend wordt. Zoniet, worden de labels voorspeld voor alle prospectieve experimenten. Dit laatste is handig om nieuwe ongeziene patiënten te classificeren in de klinische praktijk.

Gebruikers kunnen de classificatiemethoden selecteren die vergeleken moeten worden (de beste en meest efficiënte methoden van de benchmarkingstudie worden automatisch geselecteerd), het aantal randomisaties aanpassen (automatisch op 20, waarbij niet uit het oog verloren mag worden dat de resultaten betrouwbaarder zijn indien het aantal randomisaties groot is) en normalisatie afzetten.

Algoritme

Een overzicht van het algoritme achter deze webservice is weergegeven in Figuur 4.1. Het algoritme verloopt als volgt. De benchmarking dataset wordt herschikt tot het aantal aangevraagde randomisaties bereikt is. Iteratief worden alle geselecteerde classificatiemethoden toegepast op alle randomisaties. In elke iteratie worden eerst de parameters geselecteerd met behulp van LOO-CV, vervolgens wordt het model getraind op basis van de training set, en tenslotte wordt dit model toegepast op de test set, wat resulteert in een test set performantie. De gemiddelde gerandomiseerde test set performantie wordt dan berekend voor elke classificatiemethode. De best generaliserende methode – met de beste test set performantie – wordt dan gebruikt voor het bouwen van het optimale classificatiemodel op basis van de volledige benchmarking dataset, welke dan bewaard wordt voor toepassing op prospectieve datasets. Negen verschillende classificatiemethoden – gebaseerd op LS-SVM (met lineaire en RBF kernels), FDA, PCA en kernel PCA (met lineaire en RBF kernels) – werden beschouwd.

Conclusie

Aangezien het vergelijken van classificatiemodellen en het selecteren van het beste model voor elke dataset gegenereerd met microroosters een vervelende en niet voor de hand liggende taak is, werd een webservice ontwikkeld in dit hoofdstuk. De webservice M@CBETH biedt de microroostergemeenschap een eenvoudige tool aan om optimale voorspellingen te maken voor twee klassen. Deze webservice genereert het beste voorspellingsmodel op basis van verschillende classificatiemethoden en gebruik makende van randomisaties van de benchmarking dataset. Op deze manier laat de webservice M@CBETH een optimaal gebruik van de classificatie van klinische microroostergegevens toe. De website van M@CBETH is vrij beschikbaar en vertoonde reeds internationale impact na de recente introductie ervan.

Hoofdstuk 5: Classificatie van ovariumtumoren op basis van microroostergegevens

In de vorige hoofdstukken bestudeerden we verschillende manieren om voorspellingsmodellen te bouwen op basis van microroostergegevens en we stelden ook een webservice voor die toelaat dat gebruikers gemakkelijk modellen kunnen genereren op een statistisch verantwoorde manier. In dit hoofdstuk worden de eerder beschreven algemene principes toegepast op microroostergegevens van ovariale tumoren gegenereerd in een samenwerkingsproject.

Dit hoofdstuk is voornamelijk gewijd aan ons ovariumtumorenproject met Prof. I. Vergote en Prof. D. Timmerman van het Departement Gynaecologie-Verloskunde en Gynaecologische Oncologie van de Universitaire Ziekenhuizen, Leuven. In deze context bestuderen we eerst de experimenten gebruik makende van een brede waaier aan klassieke lineaire technieken. Verder passen we ook de webservice M@CBETH toe op deze experimenten gebruik makende van ook niet-lineaire leeralgoritmen voor het ontwikkelen van voorspellingsmodellen. De microroosterexperimenten werden gegenereerd in samenwerking met de MicroArray Facility (VIB).

Expressiepatronen van ovariale tumoren: algemene analyse

Binnen dit project beogen we te onderzoeken en trachten we geschikte voorspellingsmodellen te bouwen op basis van microroostergegevens om te voorspellen of:

1. Een patiënt met een stadium III of IV (FIGO) ovariale tumor zal hervallen binnen de 6 maanden na de laatste therapeutische interventie. Aangezien de standaard chemotherapie bij ovariale tumoren in een vergevorderd stadium gewoonlijk gebaseerd is op platinumderivaten (bijvoorbeeld carboplatinum + paclitaxel), zal dit model in staat zijn platinumresistentie (of chemosensitiviteit van de tumor) te voorspellen. Dit is vooral van belang op het gebied van prognose, maar zou ook kunnen toelaten nieuwe therapeutische strategieën te ontwikkelen naar de toekomst toe voor tumoren die voorspeld worden niet gepast te reageren op de standaard chemotherapeutische behandeling.
2. Een patiënt in een vroeg (stadium I) of een vergevorderd (stadium III of IV) stadium van de ziekte is aan de hand van de primaire tumor. In de klinische praktijk brengt dit model uiteraard niet veel bij, maar we willen nagaan of de expressiepatronen verschillen vertonen.

3. In een latere fase van het project zullen we nagaan of een patiënt met een stadium I ovariale tumor al dan niet zal hervallen na initiële chirurgie. Vrouwen in een vroeg stadium van de ziekte voor wie voorspeld wordt dat ze een hoge kans op herval hebben zijn geschikte kandidaten die maximaal baat zouden hebben bij een adjuvante behandeling (chemotherapie of lymfadenectomie), terwijl de vrouwen voor wie voorspeld wordt dat ze een lage kans op herval hebben gespaard zouden kunnen blijven van de neveneffecten van adjuvante therapie.

Binnen een pilootstudie onderzochten we of de verschillen tussen de groepen van patiënten besproken in eerste 2 doelstellingen effectief gereflecteerd worden in de genexpressiepatronen. Hiertoe genereerden we cDNA microroosterexperimenten voor 20 ovariumtumoren, welke 7 stadium I ovariumtumoren zonder herval (klasse I), 7 stadium III/IV platinumsensitieve (klasse A_s) en 6 stadium III/IV platinumresistente (klasse A_r) ovariumcarcinoma's omvat. Bovendien bouwen we aan de hand van deze 20 ovariumtumoren van de pilootstudie ook voorspellingsmodellen, welke we in de nabije toekomst beogen te valideren op cDNA microrooster-experimenten van 50 ovariumtumoren in een prospectieve studie.

Binnen deze pilootstudie kwantificeerden we eerst de graad van differentiële expressie van elk gen tussen klasse I en A_r, klasse I en A_s, en klasse A_s en A_r met behulp van de Wilcoxon rank sum test. Op de website (http://www.esat.kuleuven.be/~fdesmet/ovarian/) kunnen de 3 respectievelijke lijsten met de 500 best scorende genen (kleinste p-waarde) gevonden worden. Deze lijsten bevatten echter valse positieven terwijl valse negatieven ontbreken. Ondanks het feit dat de p-waarden niet gebruikt kunnen worden om de individuele vals positieve en vals negatieve genen te identificeren, zijn er procedures voorhanden om de proporties ervan te schatten (De Smet *et al.*, 2004) en om het aantal genen te berekenen dat verwacht wordt *echt* differentieel tot expressie te komen (namelijk de som van de echte positieven en valse negatieven). Gebruik makende van de methode voorgesteld door Storey en Tibshirani (2003), werd het aantal genen dat echt differentieel tot expressie komt geschat op 7059 tussen klasse I en A_r, op 4943 tussen klasse I en A_s, en op 2028 tussen klasse A_s en A_r. Deze aantallen suggereren dat de expressiepatronen effectief de verschillen tussen de klassen onder studie reflecteren en dat de hoeveelheid van differentiële expressie groter is tussen klasse I en A_r dan tussen klasse I en A_s. Merk echter op dat – aangezien niet alle weefsels van klasse I sereuse carcinomen waren (in tegenstelling tot de weefsels van klassen A_s en A_r) – we niet kunnen uitsluiten dat de differentiële expressie tussen klasse I en A_s en tussen klasse I en A_r gedeeltelijk veroorzaakt wordt door het verschil in histopathologieën.

De 20 expressiepatronen werden ook geanalyseerd met PCA. De richtingen van de 3 principale componenten die in staat zijn de grootste variatie in de data te verklaren (namelijk deze die geassocieerd zijn met de grootste eigenwaarden) werden geselecteerd en elk van de 20 expressiepatronen werd geprojecteerd op deze 3 vectoren (Figuur 5.1). Deze analyse toont een duidelijke scheiding tussen de patiënten van klasse I en klasse A_r met daartussen liggend de patiënten van klasse A_s, wat de volgorde van overgang tussen de verschillende types van tumoren aangeeft. Om de scheiding tussen de verschillende types van tumoren te verbeteren, herhalen we PCA na selectie van de 3000 genen met de grootste hoeveelheid van differentiële expressie (bepaald met de Kruskal-Wallis test) tussen de 3 klassen (Figuur 5.2). Dit resulteert in 3 duidelijk gescheiden clusters die bijna perfect overeenkomen met de gekende klassen. Deze observatie suggereert dat de 3 verschillende types van ovariale tumoren nauwkeurig kunnen geïdentificeerd worden op basis van de expressiepatronen.

De quasi perfecte scheiding tussen de 3 klassen bekomen met gesuperviseerde genselectie die PCA voorafgaat, zou echter veroorzaakt kunnen zijn door random effecten (de 3000 geselecteerde genen kunnen mogelijk een hoog aantal valse positieven bevatten) die niet bevestigd zouden worden op nieuwe experimenten. Om te bepalen of het mogelijk is om onafhankelijke stalen afkomstig van ovariale tumoren toe te kennen aan de juiste klasse, pasten we LS-SVM toe op de expressiegegevens. Dit resulteerde in een geschatte LOO classificatieperformantie van 100% voor het onderscheid tussen stadium I en een vergevorderd stadium van de ziekte en 76.92% voor het onderscheid tussen klasse A_s en A_r (2 patiënten van klasse A_r en 1 patiënt van klasse A_s werden fout geclassificeerd). Als bovendien een enkel model getraind werd (gebruik makende van alle 13 tumoren met vergevorderd stadium) voor het onderscheid te maken tussen klasse A_s en A_r en vervolgens toegepast op de 7 stadium I patiënten, dan werden deze allen toegekend aan klasse A_s. Dit toont weeral aan dat stadium I ovariumtumoren meer gelijken op platinumsensitieve dan op platinumresistente ovariumtumoren in een vergevorderd stadium.

Deze resultaten suggereren dat genexpressiepatronen effectief gebruikt kunnen worden om met een behoorlijke performantie het onderscheid te maken tussen stadium I ovariumtumoren, platinumsensitieve en platinumresistente ovariumtumoren in een vergevorderd stadium. Microroostertechnologie zou dus bij voorbeeld nuttig kunnen zijn om clinici te helpen de stadium I patiënten met een heel laag risico op herval te selecteren en deze te besparen van adjuvante chemotherapie, of om patiënten met een vergevorderde ovariale tumor te selecteren om zo platinumresistentie te voorspellen. Merk op dat het eerste geval overeenkomt met het derde klinische geval geformuleerd in het begin van dit hoofdstuk en dat we dit in de toekomst zullen onderzoeken.

Expressiepatronen van ovariale tumoren: M@CBETH

In het vorige hoofdstuk werd de webservice M@CBETH ontwikkeld die voor elke dataset verschillende classificatiemodellen met elkaar vergelijkt en het beste model selecteert in termen van gerandomiseerde onafhankelijke test set performanties. Hier passen we deze tool toe op de microroosterexperimenten afkomstig van ovariale tumoren. Tot nu toe gebruikten we enkel klassieke en lineaire technieken om deze data te bestuderen. Hier onderzoeken we de invloed van het gebruik van de niet-lineaire leeralgoritmen beschikbaar in M@CBETH. Op deze manier zouden de in de pilootstudie reeds ontwikkelde voorspellingsmodellen eventueel nog verder geoptimaliseerd kunnen worden.

Toepassing van de benchmarking service in M@CBETH op beide klinische probleemstellingen van de pilootstudie resulteerde automatisch in het opslaan van een optimaal model voor elk van beide gevallen. Op deze manier is het mogelijk in een later stadium beide modellen te evalueren op de patiënten van de prospectieve studie. Merk op dat er geen genselectie werd gedaan bij gebruik van M@CBETH, wat betekent dat alle genen in deze voorspellingsmodellen geïncludeerd zijn.

Met deze nieuwe studie bevestigden we onze vroegere bevindingen omtrent het feit dat de verschillen tussen stadium I ovariumtumoren en ovariumtumoren in een vergevorderd stadium, en tussen platinumsensitieve en platinumresistente ovariumtumoren in een vergevorderd stadium effectief gereflecteerd worden in de expressiepatronen. Vanuit methodologisch standpunt kan besloten worden dat het belangrijk is een optimale classificatiemethode te kiezen voor elke dataset.

Deze studie toont aan dat verdere optimalisatie van de voorspellingsmodellen ontwikkeld in de context van de pilootstudie mogelijk zou kunnen zijn door ook niet-lineaire technieken te beschouwen. We zouden echter willen benadrukken dat een directe vergelijking tussen de performanties van de voorspellingsmodellen van beide studies niet zonder meer is toegelaten. Hieronder zullen we de verschillen tussen beide studies verklaren om deze bewering te ondersteunen. Deze discussie wordt afgesloten met het voorstel tot een oplossing waardoor het mogelijk wordt deze vergelijkingen te maken. De eigenlijke implementatie en de bepaling van het meest optimale classificatiemodel voor beide klinische gevallen zal echter pas gedaan worden in de context van de prospectieve studie.

In de pilootstudie selecteerden we slechts 3000 genen (met de Kruskal-Wallis test) voor inclusie in het model. Door gebruik te maken van de webservice M@CBETH voerden we een dergelijke genselectie niet uit. Dit impliceert ook dat de finale modellen gegenereerd in beide studies verschillen wat betreft het aantal genen waarop zij gebaseerd zijn.

Bovendien was de test set performantie van de modellen gegenereerd in de pilootstudie geschat aan de hand van een LOO procedure waarbij het apart genomen experiment verwijderd werd *voor* genselectie. De test set performance berekend in de webservice M@CBETH daarentegen is een gemiddelde test performantie van de 20 test set gedeelten (1/3 van de experimenten in elke herschikking of randomisatie van de dataset). Verder is het ook niet toegelaten om vergelijkingen te maken met de LOO-CV performantie berekend door M@CBETH aangezien dit een gemiddelde LOO-CV performantie is van de 20 training set gedeelten (2/3 van de experimenten in elke herschikking of randomisatie van de dataset), welke in feite gebruikt werd voor het optimaliseren van de parameters en is dus zeker geen test performantie.

Een eenvoudige en voor de hand liggende oplossing om deze ongemakken te omzeilen bij het vergelijken van verschillende classificatiemodellen is de in de pilootstudie gevolgde classificatiestrategie te implementeren in M@CBETH. Merk echter op dat het vast aantal genen (3000) dat geselecteerd wordt voor het bouwen van de modellen mogelijk niet optimaal is voor algemene toepassing op andere datasets. Idealerwijze zou ook het aantal genen dat geselecteerd wordt geoptimaliseerd moeten worden, maar dit zou zeer rekenintensief worden.

Conclusie

In dit hoofdstuk concentreerden we ons op de analyse van genexpressiepatronen afkomstig van ovariale tumoren. In een pilootstudie bestudeerden we deze experimenten gebruik makende van een brede waaier aan klassieke en lineaire technieken. Vervolgens pasten we de meer geavanceerde niet-lineaire technieken beschikbaar in M@CBETH toe op deze experimenten.

Beide studies onderzochten of het mogelijk is onderscheid te maken tussen stadium I ovariumtumoren zonder herval, platinumsensitieve en platinumresistente ovariumtumoren in een vergevorderd stadium (stadium III/IV). De resultaten in de pilootstudie werden bekomen door het aantal genen te bestuderen dat niet bij toeval differentieel tot expressie komt tussen de verschillende tumorklassen, door niet-gesuperviseerde PCA uit te voeren, en door gebruik te maken van een LOO strategie in combinatie met LS-SVM voor het ontwikkelen van classificatiemodellen. Deze resultaten toonden aan dat genexpressiepatronen nuttig zouden kunnen zijn in het klinisch beleid bij ovariale tumoren. Deze bevindingen werden bevestigd door het toepassen van de benchmarking service van M@CBETH op deze experimenten. Verder toont deze studie ook aan dat verdere optimalisatie van de voorspellingsmodellen ontwikkeld in de pilootstudie eventueel mogelijk zou zijn door ook niet-lineaire technieken te beschouwen.

Hoofdstuk 6: Kernelclustering van microrooster-experimenten

In tegenstelling tot alle vorige hoofdstukken die handelden over het ontwikkelen van voorspellingsmodellen voor klinische toepassingen, wordt dit hoofdstuk volledig gewijd aan het ontdekken van diagnostische klassen. Clusteringtechnieken worden algemeen toegepast op microrooster-experimenten voor de identificatie van klinische klassen, wat zou kunnen leiden tot het verfijnen van het klinisch beleid. Clusteranalyse van volledige microroosterexperimenten (expressiepatronen van patiënten of weefsels) biedt de mogelijkheid om nog ongekende diagnostische categorieën te ontdekken zonder de eigenschappen van deze klassen op voorhand te kennen. Deze clusters zouden dan de basis kunnen vormen van nieuwe diagnostische schema's waarbij de verschillende categorieën patiënten bevatten met een kleinere klinische variabiliteit.

Clustering van microroosterexperimenten is reeds nuttig gebleken in tal van studies in de oncologie. Hiertoe worden gewoonlijk methoden gebruikt zoals de klassieke K-means clustering en de hiërarchische clustering (Handl *et al.*, 2005; Bolshakova *et al.*, 2005). Deze methoden zijn gebaseerd op eenvoudige afstands- of similariteitsmaten (bijvoorbeeld de Euclidische afstand). Daarom kunnen enkel lineaire afstandsmaten toegepast worden op de data gebruik makende van deze technieken. Recentelijk werden methoden ontwikkeld voor het clusteren van data waarvan de clusters niet lineair scheidbaar zijn. Twee belangrijke methoden zijn kernel K-means clustering (Dhillon *et al.*, 2004a; Dhillon *et al.*, 2004b; Zhang en Rudnicky, 2002) en de gerelateerde spectrale clustering (Cristianini *et al.*, 2002; Ng *et al.*, 2001). Het introduceren van deze technieken voor de analyse van microroostergegevens zou toelaten niet-lineaire relaties in de data te ontdekken alsook de rekencomplexiteit veroorzaakt door de hoogdimensionale data te verbeteren.

Validatietechnieken worden gebruikt om de performantie van verschillende clusteringmethoden te beoordelen en te vergelijken. Deze methoden kunnen ook gebruikt worden om de clusterinstellingen te tunen (bijvoorbeeld om het aantal clusters te optimaliseren en de kernelparameters te tunen). Twee belangrijke types van validatietechnieken zijn interne en externe validatie (Handl *et al.* 2005; Bolshakova en Azuaje, 2003; Halkidi *et al.*, 2001). Interne validatie beoordeelt de kwaliteit van een clusterresultaat gebaseerd op statistische karakteristieken (bijvoorbeeld het beoordelen van de compactheid van een cluster, of het maximaliseren van de intercluster afstanden en het minimaliseren van de intracluster afstanden). Externe validatie reflecteert de graad van overeenkomst van een clusterresultaat met

een externe partitie (bijvoorbeeld bestaande diagnostische klassen die gebruikt wordt door dokters in de klinische praktijk). De Globale Silhouette index, de Distortie score en de Calinski-Harabasz index (F-statistiek) worden vaak gebruikt voor interne validatie, de Rand index en de Adjusted Rand index voor externe validatie.

Dit hoofdstuk bestudeert de voor- en nadelen van de klassieke K-means, de kernel K-means en de spectrale clusteringalgoritmen in de context van clusteranalyse van microroosterexperimenten.

Experimenten

Aangezien de rekencomplexiteit van de klassieke K-means clustering nadelen ondervindt van de hoogdimensionale microroosterexperimenten, voerden we PCA uit als een voorafgaande dimensionaliteitsreductie stap. Indien geen selectie van principale componenten uitgevoerd wordt, geeft dit gelijkaardige resultaten als K-means clustering zonder voorafgaande PCA. Kernel K-means en spectrale clustering daarentegen kunnen wat betreft de rekencomplexiteit efficiënter omgaan met de hoogdimensionale microroosterexperimenten aangezien deze technieken gebruik maken van de kerneltruuk, welke toelaat impliciet in de kenmerkenruimte te werken. Merk ook op dat kernel K-means clustering met een lineaire kernel vergelijkbaar is met de klassieke K-means clustering.

We toonden aan hoe verschillende interne clustervalidatiecriteria die gewoonlijk toegepast worden in de dataruimte uitgebreid kunnen worden voor toepassingen in de kenmerkenruimte. Merk op dat klassieke K-means clustering en kernel K-means clustering met een lineaire kernel optimalisatie vereisen van het aantal clusters en de random initialisatie. Kernel K-means met een RBF kernel en spectrale clustering vereisen bovendien ook de optimalisatie van de kernelparameter σ. Het tunen van deze parameters gebeurt gewoonlijk aan de hand van interne validatiecriteria. Aangezien bij kernelclusteringmethoden ook het tunen van de kernelparameters vereist wordt, zal in de toekomst hieromtrent verder onderzoek moeten verricht worden.

De verschillende clusteringtechnieken werden vervolgens getest op enkele van de reeds gebruikte datasets, namelijk data van tumoren van het colon (Alon *et al*., 1999) en data van acute leukemie (Golub *et al*., 1999).

Conclusie

Kernelclusteringmethoden zoals kernel K-means en spectrale clustering werden speciaal ontworpen voor het ontdekken van niet-lineaire relaties in de data. Bovendien werken deze technieken efficiënter op hoogdimensionale data in termen van de rekencomplexiteit. In dit hoofdstuk toonden we aan dat deze eigenschappen maken dat

kernelclusteringmethoden interessant zouden kunnen zijn voor toepassingen met microroosters (al dan niet voorafgegaan door selectie van genen voor het filteren van de data). We vergeleken klassieke en kernelclusteringalgoritmen op verschillende datasets gegenereerd met microroosters. Hieruit besloten we dat zeer goede resultaten bekomen kunnen worden met spectrale clustering in termen van interne validatiecriteria. Bij de realisatie hiervan toonden we aan hoe deze interne validatiematen uitgebreid kunnen worden naar de kenmerkenruimte. In de toekomst zou het gebruik van deze technieken kunnen leiden tot de ontdekking van nieuwe klinisch relevante subgroepen.

Hoofdstuk 7: Conclusies en toekomstig onderzoek

In dit hoofdstuk vatten we de voornaamste bevindingen van deze thesis samen. Verder stellen we kort enkele specifieke klinische onderzoeksprojecten voor waarin we in de toekomst toe willen bijdragen. Tot slot beschrijven we methodologische onderzoeksvraagstellen die we in de toekomst willen behandelen.

In deze thesis onderzochten we hoe microroostergegevens optimaal gebruikt kunnen worden bij klinische beleidsbeslissingen in de oncologie. Hiertoe maakten we gebruik van leeralgoritmen zoals LS-SVMs en kernelmethoden, beschreven in Hoofdstuk 2, welke in staat zijn hoogdimensionaliteit te hanteren alsook niet-lineaire relaties in de data te ontdekken. Concrete bijdragen van deze thesis kunnen als volgt samengevat worden:

1. Deze leeralgoritmen en kernelmethoden werden bestudeerd en verfijnd om ze meer geschikt te maken voor microroostergegevens bij problemen uit de klinische oncologie. In Hoofdstuk 3 voerden we een systematische benchmarkingstudie uit om de invloed van regularisatie, niet-lineariteit en dimensionaliteitsreductie op de performantie van klinische voorspellingen te onderzoeken. We besloten dat regularisatie of dimensionaliteitsreductie vereist is voor de classificatie van microroosterexperimenten. Bovendien geeft een niet-lineair LS-SVM model met een RBF kernel over het algemeen de beste resultaten voor de classificatie van microroosterexperimenten.

2. In Hoofdstuk 4 werden deze methoden opgenomen in een interface genaamd M@CBETH (a MicroArray Classification BEnchmarking Tool on a Host server: http://www.esat.kuleuven.be/MACBETH/) die vrij beschikbaar is en die eenvoudig gebruikt kan worden door clinici voor het maken van optimale voorspellingen. Deze webservice vindt het beste voorspellingsmodel uit verschillende classificatiemethoden door gebruik te maken van randomisaties van een dataset die voor benchmarking doeleinden gebruikt wordt.

3. We pasten deze technieken in Hoofdstuk 5 toe op een verzameling van genexpressiepatronen afkomstig van ovariale tumoren om in deze context verschillende diagnostische problemen op te lossen. We pasten een brede waaier aan klassieke en lineaire technieken toe op deze experimenten, gevolgd door de meer geavanceerde niet-lineaire technieken beschikbaar in M@CBETH. Dit toonde aan dat het mogelijk is om onderscheid te maken tussen stadium I ovariumtumoren zonder herval, platinumsensitieve en platinumresistente ovariumtumoren in een vergevorderd stadium (stadium III/IV).

4. Tenslotte vergeleken we klassieke en kernelclusteringalgoritmen op verschillende datasets gegenereerd met microroosters in Hoofdstuk 6. Hieruit besloten we dat zeer goede resultaten bekomen kunnen worden met spectrale clustering in termen van interne validatiecriteria. Bij de realisatie hiervan toonden we aan hoe deze interne validatiematen uitgebreid kunnen worden naar de kenmerkenruimte.

Toekomstig onderzoek: klinische toepassingen

1. Klinisch beleid bij ovariale tumoren

Dit project loopt in samenwerking met Prof. I. Vergote en Prof. D. Timmerman. In een eerste fase werkten we reeds aan een pilootstudie op een verzameling van cDNA microroosterexperimenten afkomstig van 20 ovariale tumoren. Zoals beschreven in hoofdstuk 5 ontwikkelden we voorspellingsmodellen op basis van deze data om het onderscheid te kunnen maken tussen stadium I ovariumtumoren zonder herval, platinumsensitieve en platinumresistente ovariumtumoren in een vergevorderd stadium (stadium III/IV).

In een volgende fase zullen we werken op een prospectieve studie. cDNA microroosterexperimenten werden reeds gegenereerd voor 50 ovariumtumoren afkomstig van 4 klassen: stadium I ovariumtumoren zonder herval, stadium I ovariumtumoren met herval, platinumsensitieve en platinumresistente ovariumtumoren in een vergevorderd stadium (stadium III/IV). In deze prospectieve studie zullen we de modellen gegenereerd in de pilootstudie evalueren, verfijnen en mogelijk ook uitbreiden. In het geval de voorspellingsperformanties van sommige van deze modellen onvoldoende blijken, zouden de nieuwe experimenten gebruikt kunnen worden om de modellen verder te verfijnen. Bovendien zullen deze extra experimenten toelaten genen te selecteren die differentieel tot expressie komen tussen de verschillende klassen met een hogere efficiëntie. In deze prospectieve studie zullen we ook de verschillen tussen stadium I ovariumtumoren met een hoog en een laag risico op herval onderzoeken. Beide stadium I klassen zullen gesitueerd worden met betrekking tot de andere klassen gebruik makende van PCA.

Voor een deel van deze patiënten hebben we ook klinische gegevens voorhanden. Bovendien zullen we in de nabije toekomst ook proteoomexperimenten uitvoeren op een deel van deze patiënten, waarvoor we reeds fondsen verkregen binnen Biopattern. Op deze manier kunnen proteoomgegevens gecombineerd worden met microrooster- en klinische gegevens.

2. Klinisch beleid bij borsttumoren

Dit project loopt in samenwerking met Prof. P. Neven en Prof. D. Timmerman. Een databank die klinische informatie bevat van meer dan 3000 patiënten met borsttumoren (zonder metastasen bij diagnose) is reeds voorhanden (verzameld sinds 1 januari 2000). Van deze patiënten zullen ook tumorweefsels en serum verzameld worden, wat toelaat moleculaire informatie van deze patiënten te verzamelen gegenereerd met proteoom- en/of microroostertechnologie. We zullen dan trachten modellen te bouwen die in staat zijn voorspellingen te maken omtrent het herval van de tumor buiten de borst. Op deze manier zou het mogelijk zijn betere voorspellingen te maken rond adjuvante therapie bij borsttumoren. We zullen zowel de technieken beschreven in dit werk als de hierna beschreven technieken toepassen.

Bovendien wordt in de nabije toekomst binnen Biopattern een dataset beschikbaar gesteld die microrooster-, proteoom- en klinische gegevens van 150 patiënten met een borsttumor (met meer dan 10 jaar opvolging) zal bevatten.

Toekomstig onderzoek: methodologische uitdagingen

Verschillende mogelijke uitbreidingen van de webservice M@CBETH met het oog op verdere optimalisatie van de performanties van voorspellingsmodellen voor gebruik in de klinische praktijk worden hier besproken:

1. De webservice M@CBETH zou eenvoudig uitgebreid kunnen worden om de huidige functionaliteit verder te verbeteren gebaseerd op enkel microroostergegevens:

    - Ten eerste zou het interessant zijn naar de klinische praktijk toe om ook schattingen van de voorspellingsprobabiliteiten te integreren om zo een idee te hebben van de betrouwbaarheid van de voorspellingen gemaakt voor individuele patiënten.
    - Ten tweede zou het nuttig kunnen zijn ook voorspellingen te maken voor classificatieproblemen die meerdere (meer dan 2) klassen omvatten om zo de nood aan het herleiden van elk classificatieprobleem tot een tweeklassen probleem te omzeilen.
    - Verder zouden nieuwe classificatietechnieken, nadat ze in benchmarkingstudies gelijkaardig aan deze in hoofdstuk 3 bewezen hebben tot goede performanties te leiden, toegevoegd kunnen worden aan deze webservice.
    - Een ander interessant aspect is meer genselectiemethoden te integreren. Het belang hiervan is voornamelijk gesitueerd in de

klinische praktijk, onder meer voor het maken van klinische voorspellingen of voor het ontdekken van tumormerkers en doelwitten voor geneesmiddelen. Twee belangrijke strategieën zouden onderzocht kunnen worden. De eerste strategie omvat het uitvoeren van kenmerkenselectiealgoritmen voor het ontwikkelen van classificatiemodellen. We gebruikten reeds klassieke, lineaire en kernel PCA met gesuperviseerde en niet-gesuperviseerde manieren om principale componenten te selecteren. Nieuwe opportuniteiten kunnen gevonden worden in het gebruik van gesuperviseerde en niet-gesuperviseerde methoden voor de selectie van individuele genen (univariaatanalyse) of methoden voor de selectie van groepen van genen (multivariaatanalyse). De tweede strategie omvat classificatietechnieken die spaarsheid creëren in de genen terwijl de classificatieperformantie geoptimaliseerd wordt. Op deze manier kunnen relevante onderliggende structuren in de genen onmiddellijk gedetecteerd worden bij het ontwikkelen van het voorspellingsmodel. Methoden ontwikkeld met dit doel zouden eerst bestudeerd en aangepast moeten worden voor toepassingen met hoogdimensionale datasets. Vervolgens kunnen al deze technieken geïntegreerd worden in het algoritme van M@CBETH. Voor beide strategieën is het echter aan te raden om ook procedures te integreren om zo automatisch de biologische relevantie en de functie van de genen die geselecteerd worden voor het bouwen van het model te controleren. Belangrijke databanken die beschouwd kunnen worden voor classificatieproblemen in de oncologie zijn onder meer: National Center of Biotechnology Information (NCBI) (http://www.ncbi.nih.gov/), Cancer Profiling Database Oncomine (http://www.oncomine.org/), Tumor Gene Database (http://www.tumor-gene.org/).

2. Verschillende databronnen zoals microrooster-, proteoom- en klinische gegevens zouden complementaire informatie kunnen bevatten met betrekking tot het klinisch gedrag. Daarom zou het interessant kunnen zijn om deze heterogene databronnen eerst afzonderlijk te bestuderen en deze dan te combineren in een (hybride) procedure om de data te analyseren. Het hoofddoel zou zijn te onderzoeken hoe gegevens afkomstig van het transcriptoom, het proteoom en klinische gegevens het best gebruikt en gecombineerd worden om het beleid bij verschillende soorten tumoren te ondersteunen.

    - In deze context zou een eerste stap zijn te focussen op klinische gegevens afzonderlijk. Het idee om het beste classificatiemodel te ontwikkelen voor elke dataset afkomstig van microroosters zou eenvoudig uitgebreid kunnen worden voor klinische datasets. In de studie op klinische gegevens van endometriumcarcinomen in De

Smet *et al.* (2006) leidden sommige van de classificatiemethoden op basis van leeralgoritmen verwerkt in M@CBETH ook tot betere performanties dan de meer traditionele technieken. Daarom zou het interessant zijn een aangepaste versie van M@CBETH te creëren specifiek voor toepassingen met klinische gegevens.

- Zoals eerder vermeld zou het mogelijk kunnen zijn extra informatie over de moleculaire biologie van weefsels en serumstalen te bekomen door studie van het proteoom. Merk op dat de methodologie vereist om deze proteoomdata momenteel nog minder ontwikkeld is wegens de significante recente ontwikkelingen binnen deze technologie. De hoogdimensionaliteit van deze data suggereert echter dat methoden toegepast op microroostergegevens ook op deze data nuttig zouden kunnen zijn. Wanneer de analyse van proteoomgegevens meer gestandaardiseerd zal zijn, zou het interessant zijn een andere aangepaste versie van M@CBETH te creëren met het oog op het genereren van goede voorspellingsmodellen op basis van proteoomgegevens.
- Uiteindelijk zal de ultieme doelstelling zijn deze 3 heterogene databronnen te integreren in een grote data analyse procedure. Een interessante onderzoeksvraagstelling zou zijn te onderzoeken of het mogelijk is de huidige voorspellingen (en mogelijk tegelijkertijd ook het ontdekken van de onderliggende structuren) op basis van microroostergegevens verder te optimaliseren door de combinatie ervan met proteoom- en klinische gegevens. Integratie van deze heterogene databronnen in een model zou uitgevoerd kunnen worden door combinatie van de datavectoren, de kernelfuncties of de modellen zelf, bijvoorbeeld door het trainen van een extra laag in het model (Suykens *et al.*, 2002) of met Bayesiaanse Netwerken (Gevaert *et al.*, 2006). Dit zou kunnen leiden tot een finale en optimaal aangepaste versie van M@CBETH.

Tot slot vatten we samen dat we naar de toekomst toe streven naar het gebruik en de ontwikkeling van de hierboven vermelde technieken voor de analyse van transcriptoom-, proteoom- en klinische gegevens van patiënten en de integratie van deze resultaten teneinde een verbeterde versie van de webservice M@CBETH te creëren voor gebruik in de klinische praktijk web.

List of acronyms

ALL	Acute lymphoblastic leukemia
AML	Acute myeloid leukaemia
ACC	Accuracy
AUC	Area under the ROC curve
BMI	Body mass index
BRCA	Breast cancer gene
cDNA	Complementary DNA
DNA	Deoxyribonucleic acid
ER	Estrogen receptor
ESI	Electrospray ionization
FDA	Fisher discriminant analysis
FIGO	International federation of gynaecology and obstetrics
FN	False negative
FP	False positive
Her-2/Neu	Human epidermal growth factor receptor
SVD	Singular value decomposition
LOO	Leave-one-out
LOO-CV	LOO cross-validation
LS-SVM	Least squares SVM
M@CBETH	Microarray classification benchmarking tool on host server
MALDI	Matrix-assisted laser desorption ionisation
MALDI-TOF	MALDI time-of-flight
mRNA	Messenger RNA
PCA	Principal component analysis
PCR	Polymerase chain reaction
PR	Progesterone receptor
RBF	Radial basis function
RNA	Ribonucleic acid
ROC	Receiver Operating Characteristic
RT-PCR	Reverse transcription-coupled PCR
SELDI	Surface-enhanced laser desorption ionisation

SELDI-TOF	SELDI ionisation time-of-flight
SOM	Self-organizing maps
SVM	Support vector machines
TN	True negative
TP	True positive

Contents

Chapter 1

Introduction

1.1 Motivation

In her article (Stikeman, 2002) entitled "The State of Biomedicine: Medical treatment will be tailored to your genetic profile" published in *Technology Review* in June 2002, Alexandra Stikeman cited: "*"It's going to totally transform medicine, there's no question about it," says Susan Lindquist, director of MIT's Whitehead Institute for Biomedical Research. "And it's going to be happening soon." Mark Levin, CEO of Cambridge, MA-based Millennium Pharmaceuticals, offers one vision of what personalized medicine might mean for a patient: "When we walk into the doctor's office 10 years from now, we'll have our genome on a chip." Using that chip, Levin says, a doctor will be able to determine what diseases a patient is predisposed to and what medicines will provide the most benefit with the fewest side effects. Even the way we think about disease will be different, says Jeffrey Augen, director of life sciences strategy at IBM, because doctors will make diagnoses based on genes and proteins rather than on symptoms or the subjective analysis of tissue samples under a microscope. "So instead of a person having chronic inflammation or cancer, he or she will have a cox-2 enzyme disorder or a specific set of genetic mutations," Augen predicted at a recent conference in Boston.*".

The recently developed microarray technology has the potential to realize these aspirations of making predictions for individual patients in clinical practice. Technological advances in molecular biology have led to the development of these microarrays. These are capable of determining the expression levels of thousands of genes simultaneously.

One important application area of this technology is clinical oncology (Friend, 1999; Nasir, 2001; Patten-Hitt, 2001; Ahr *et al.*, 2001; Ahr *et al.*, 2002; Nielsen *et al.*, 2002; Perou *et al.*, 2000; Sørlie *et al.*, 2001). As the disordered expression of genes is pivotal in the behavior of tumors,

its measurement can be very helpful to model or to predict the clinical behavior of malignant processes. By these means, the fundamental processes underlying carcinogenesis can be involved in the clinical decision process.

The development of microarray technology has led to the generation and analysis of huge amounts of data. The extraction of clinically relevant information from these data requires specific statistical procedures. This information can typically be extracted in three different ways: by making clinical predictions (Furey *et al.*, 2000), by discovering diagnostic classes (Ben-Dor *et al.*, 2001), and by selecting relevant genes or groups of genes to distinguish the different groups of tumors (Guyon *et al.*, 2002). This creates an interesting point of view for a relatively new science called bioinformatics.

For clinical applications, microarray data can be represented by an expression matrix whose rows represent the gene expression profiles and columns the expression patterns of the patients. Microarray data sets are characterized by a large number of gene expression levels for each patient and a relatively small number of patients (comprising different classes of tumors). The large number of gene expression levels for each patient is a problem for most classical statistical methods. Therefore, dimensionality reduction is often applied before using these methods (Alter *et al.*, 2000; Guyon *et al.*, 2002; Müller *et al.*, 2001). Support Vector Machines (SVM) on the contrary seem to learn and to generalize these data well despite the high dimensionality, owing to the regularization principle this method is based on (Burges, 1998; Cristianini and Shawe-Taylor, 2000; Müller *et al.*, 2001; Marron and Todd, 2002; Schölkopf *et al.*, 1999; Schölkopf *et al.*, 2001; Schölkopf and Smola, 2002; Suykens *et al.*, 2002; Van Gestel *et al.*, 2004; Vapnik, 1998). Furthermore, in the future the number of microarray experiments will continue to increase, which has consequences for data analysis since complex relationships may become apparent in microarray data sets. Unfortunately, most traditional statistical methods use linear functions to model the relationships in the data. These functions, however, are not capable of discovering complex nonlinear relationships in microarray data. This can be solved by using more complex nonlinear kernel functions, which allow for better modeling of the data.

The general objective in this work is to investigate how microarray data can be optimally used in the clinical management of neoplastic disorders (tumors), hereby emphasizing a good mathematical foundation. More specifically, the aim is to use machine learning methods like Least Squares Support Vector Machines (LS-SVM) and kernel methods, capable of both handling the high dimensionality (for both linear and nonlinear modeling) and discovering nonlinear relationships in the data, to optimize clinical microarray data analysis. In this work, the behavior of these methods is studied and fine-tuned to make them more suitable for microarray data in

clinical decision-making problems. Furthermore, these methods are incorporated into an interface that is freely available and can easily be used by clinicians. Finally, these techniques are also applied to solve several diagnostic problems.

1.2 Supporting clinical management of cancer

The classical approach to cancer management is usually based on clinical information from patient history, clinical and histopathological examinations, and on the experience of the clinician. However, the fundamental mechanisms determining the diagnostic categories, prognosis, and therapeutic choice are often ignored. Taking these mechanisms into account will be of major importance in making the right management decisions. For this purpose, one may rely on data originating from recently developed high-throughput technologies like microarrays and proteomics technologies. Since these data sources may contain complementary information about the clinical behavior to that of traditional clinical data sets, the main objective of this work is to investigate how microarray data can be optimally used in clinical management of tumors.

Microarrays allow determining the expression levels of thousands of genes simultaneously. As the dysregulated expression of genes that represent these fundamental mechanisms, lies at the origin of the tumor phenotype, its measurement can be helpful to model or to predict the clinical behavior of malignancies. By these means, the fundamental processes underlying carcinogenesis can be integrated into the clinical decision making. Using microarray data allows for optimized predictions for an individual patient (e.g., predictions about therapy response, prognosis and metastatic phenotype).

An example of making predictions about therapy response can be found in Iizuka *et al.* (2003). Hepatocellular carcinoma (malignant tumor derived from epithelial tissue of the liver) has a poor prognosis because of the high intrahepatic recurrence rate. Intrahepatic recurrence limits the potential of surgery as a cure for hepatocellular carcinoma. The current pathological prediction systems clinically applied to patients are inadequate for predicting recurrence in individuals who undergo hepatic resection. In this case, it would be useful to predict therapy response to select the patients who would benefit from surgical treatment.

Nutt *et al.* (2003) present a nice example of predicting the prognosis, as can be seen in Figure 1.1. Among high-grade gliomas (a tumor springing from the neuroglia or connective tissue of the brain), anaplastic oligodendrogliomas have a more favorable prognosis than glioblastomas. Moreover, although glioblastomas are resistant to most available therapies,

anaplastic oligodendrogliomas are often chemosensitive. Unfortunately, both subtypes cannot be distinguished from each other based on only the histopathology (microscopic examination of tissue samples). However, by predicting the prognosis it is possible to fine-tune treatment.

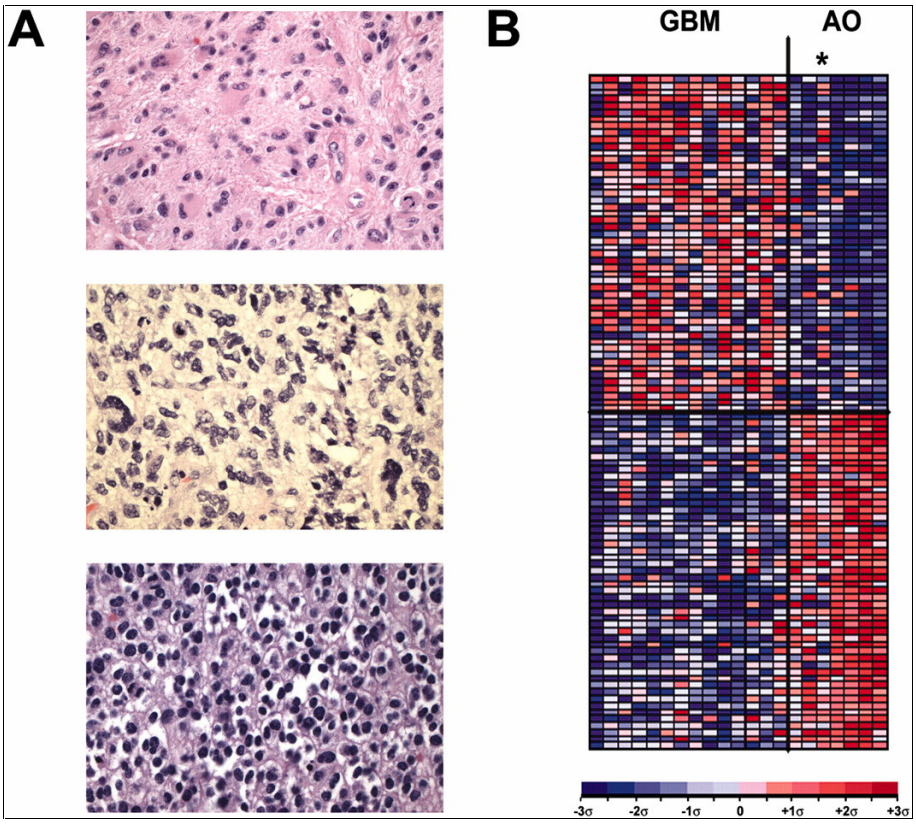


Figure 1.1 : *Characterization of classic high-grade gliomas from Nutt et al. (2003). A. Histopathological features of the different types of classic high-grade gliomas: (top) classic glioblastoma featuring cells; (middle) classic glioblastoma; (bottom) classic anaplastic oligodendroglioma. B. Gene expression profiles of high-grade gliomas. Genes were ranked according to their correlation with the classic glioblastoma (GBM) versus classic anaplastic oligodendroglioma (AO) distinction. Results are shown for the top 50 genes of each distinction. Each column represents a single glioma sample, and each row represents a single gene. For each gene, red indicates a high level of expression relative to the mean; blue indicates a low level of expression relative to the mean. The standard deviation from the mean is indicated.*

An example of predicting the metastatic phenotype (transmission of pathogenic microorganisms or cancerous cells from an original site to one or more sites elsewhere in the body, usually by way of the blood vessels or lymphatics) is presented in van 't Veer *et al.* (2002). For breast cancer

patients without tumor cells in local lymph nodes at diagnosis (lymph node negative), it is useful to predict the presence of distant subclinical metastases (poor prognosis) based on the primary tumor. Predicting the metastatic phenotype allows selecting patients who would benefit from adjuvant therapy as well as selecting patients for whom this adjuvant therapy would mean unnecessary toxicity.

The next section elaborates on the details of high-throughput technologies like microarrays used in previous examples and more recent proteomics technology. A description of the characteristics of the resulting data sets elucidates the requirement of machine learning methods to analyze these data.

1.3 High-throughput technologies

Recent technological advances have enabled molecular biologists to study the transcriptome and proteome. The transcriptome is the collection of mRNA or genes that are expressed in tissues and can be measured by the microarray technology. The proteome is the collection of all proteins present in tissues or samples and can be measured for example by means of mass spectrometry.

1.3.1 Microarrays

Microarrays are a recent technology to determine the expression levels of thousands of genes simultaneously in tissues. For clinical purposes, microarrays have mainly been used in oncology, as previously stated. Parallel measurements of these expression levels result in data vectors that contain thousands of values, which are called expression patterns. A microarray consists of a reproducible grid of several DNA-probes attached to a small solid support. Labeled cDNA prepared from extracted mRNA, is hybridized with the complementary DNA-probes attached to the microarray. The hybridizations are measured by means of a laser scanner and transformed quantitatively. Two important types of microarrays are cDNA microarrays and oligonucleotide arrays. cDNA microarrays consist of about ten thousands of known cDNAs (obtained after PCR amplification) that are spotted in an ordered matrix on a glass slide. Oligonucleotide arrays (or DNA chips) are constructed by the synthesis of oligonucleotides on silicium chips. Figure 1.2 gives a schematic overview of an experiment with the cDNA technology. Both technologies have specific characteristics that will not be discussed here.

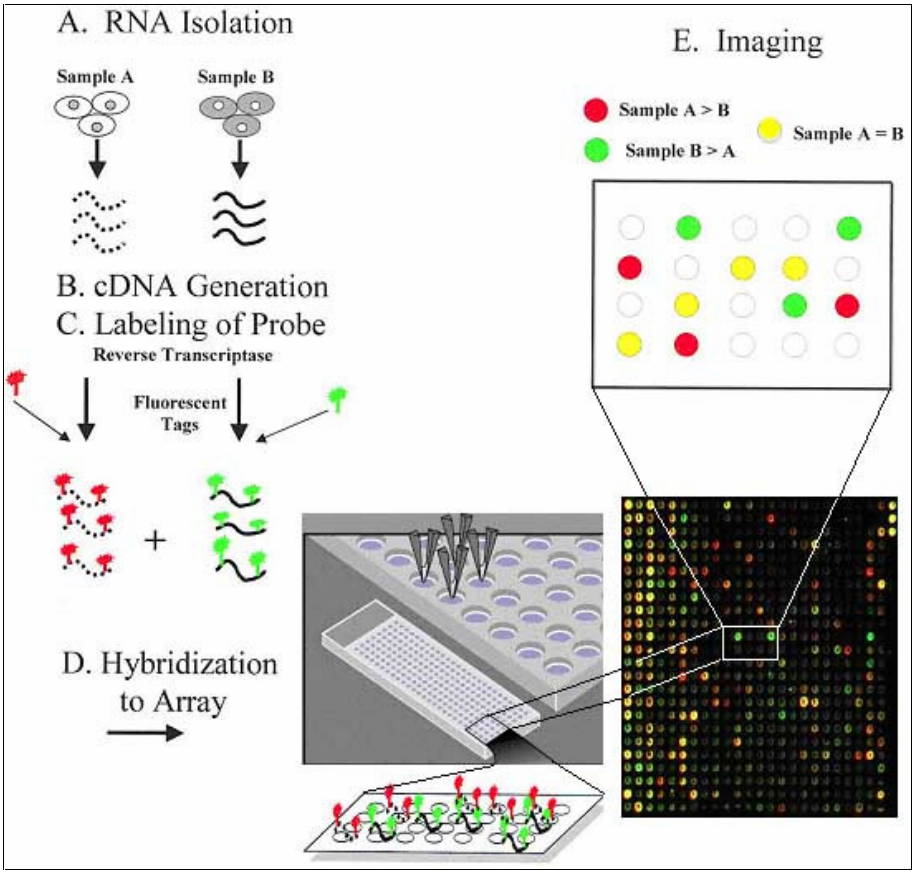


Figure 1.2 : *Schematic overview of an experiment with a cDNA-microarray from Albelda and Sheppard (2000). A. The total mRNA of the test sample (tumor - red) and reference sample (green) are isolated. B. The mRNA from each sample is treated with reverse transcriptase. C. The mRNA of each sample is labeled with a distinct fluorescent tag: the test sample red and the reference sample green. D. Spotting of the presynthesized DNA-probes (derived from the genes to be studied) on the glass slide. These probes are the purified products from PCR amplification of the associated DNA clones. The two pools of labeled RNA are hybridized to separate arrays and washed. E. The arrays are imaged using a specialized fluorimeter. The red and green intensities (measure for the hybridization by the test and reference sample) of each spot are determined. The relative expression levels (intensity in the red channel / intensity in the green channel) are calculated. E. In this example, the two pools are mixed before hybridization to one array. The genes only expressed in Sample A would be red in color, genes only expressed in Sample B would be green, and those genes expressed equally in both samples would be yellow.*

1.3.2 Proteomics

Microarrays do not capture all relevant phenomena in a cell on a molecular level because of posttranscriptional modifications and regulation of biologically active molecules. Studying the proteome (the protein complement to the transcriptome) makes it possible to obtain additional information about the molecular biology of tissues and samples and most likely the pathomechanism behind tumors. Recently developed technology like the ProteinChip technology developed by Ciphergen Biosystems (http://www.ciphergen.com/), enables to quantify the presence of a large subset of proteins in a sample. ProteinChip technology has already been applied to some selected cases in oncology (Kozak *et al.*, 2003; Petricoin *et al.*, 2002a; Petricoin *et al.*, 2002b; Wilson *et al.*, 2004). In essence, the samples undergo a basic mass analysis using a MALDI mass spectrometer (together with ESI mass spectrometers). However, in this technology there is an added discriminatory dimension added by applying different chemical companion matrices to the wells where the samples reside. Combined, this approach is known as SELDI mass spectrometry, as illustrated in Figure 1.3. Qualitatively, these technologies provide mass spectra that contain thousands of discrete peak intensity values, each associated with a mass/charge value, which in its turn is associated with a (unknown) protein or a fragment of a protein. Therefore, these spectra are characteristic for the proteins and peptides or a subset of proteins and peptides present in a sample. The output consists (similarly to microarray data) of huge data vectors where each component is representative for the amount of an unspecified protein (or its characteristic fragments) that is present in the sample at hand. The methodology specifically designed to analyze these proteomic data is currently less developed because of the significant recent developments in this technology. Therefore, these data would form a prime candidate for the application of methods that are used for microarray data analysis.

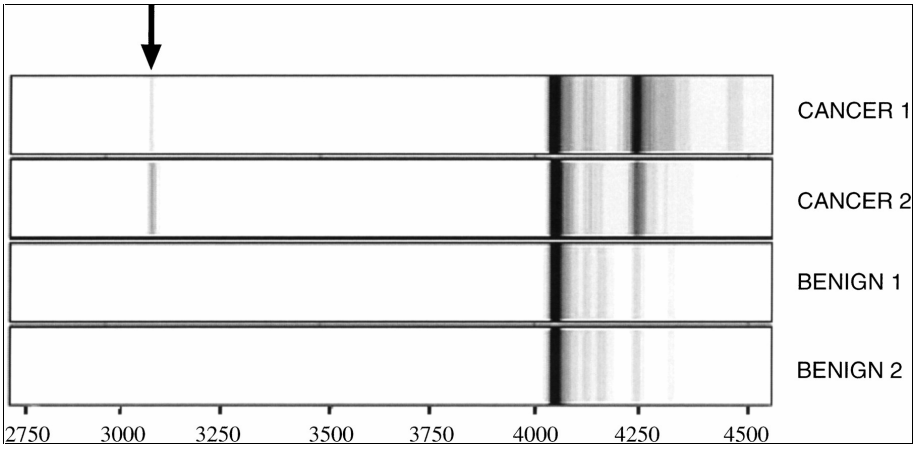


Figure 1.3 : *Proteomic spectra from Petricoin et al. (2002a). Representative surface-enhanced laser desorption ionization time-of-flight (SELDI-TOF) mass spectra (gel view) from sera of two prostate cancer patients (top) are compared with spectra from a set of sera from two men with benign conditions (bottom). A serum protein spectrum in our study is composed of mass/charge (m/z) amplitudes over a range of 0–20 000. A small window (m/z values from 2750 to 4500) surrounding one of the seven discriminatory m/z values (3080) within the pattern is shown. As indicated by the arrow, the peak ion signature value appears distinctly different between sera from men with benign conditions and cancer.*

1.4 Machine learning techniques and statistical techniques

Clinical microarray data can be analyzed from different viewpoints. The three main perspectives are: (1) making clinical predictions (classification), (2) discovering diagnostic classes (clustering experiments), and (3) selecting relevant genes (or groups of genes) or dimensionality reduction (feature extraction). Table 1.1 presents an overview of these objectives and the respective traditional statistical and machine learning methods. Each of these objectives will be discussed below. Although feature extraction is usually the first step in microarray data analysis, thereby preceding classification and clustering, we discuss classification and clustering first because the emphasis of this thesis lies on the latter techniques.

Clinical objectives	Traditional statistical techniques	Support Vector Machines and kernel methods
Making clinical predictions **(Classification)**	Logistic regression Fisher Discriminant Analysis (FDA)	Support vector Machines (SVM) or Least Squares SVM (LS-SVM) with linear or nonlinear kernels
Discovering diagnostic classes **(Clustering)**	Hierarchical clustering K-means clustering	Spectral clustering Kernel K-means clustering
Selecting relevant genes or groups of genes **Dimensionality reduction** **(Feature extraction)**	Univariate analysis: · Statistical methods (Hypothesis tests) Multivariate Analysis: · Linear Principal Component Analysis (PCA) · Stepwise logistic regression	Multivariate analysis: · Kernel PCA

Table 1.1 : *Overview of the three main viewpoints with examples of corresponding statistical algorithms and machine learning methods considered during this research.*

1.4.1 Classification

In clinical practice, it would be valuable to make predictions about for example prognosis, therapy response, stage, and histopathological diagnosis, based on measurements done with microarrays (possibly combined with other clinical information). This is realized with statistical models that classify tumors, often based on selected features. The parameters of the model are determined based on a collection of patients for whom it is already known to which class they belong. These are the patients for whom for example stage, histopathological diagnosis, prognosis, and therapy response, are already known. This collection of patients is referred to as the training set, which is supposed to be used for training the model. In practice, this corresponds to determining the parameters of the model. The trained

model can then be applied afterwards to make predictions for previously unseen patients, which form the test set.

Microarray data sets are characterized by high dimensionality in the sense of a small number of patients and a large number of gene expression levels for each patient. Most classification methods from statistics, like for example logistic regression and Fisher Discriminant Analysis (FDA), have problems with the high dimensional nature of microarray data and require dimensionality reduction first. On the contrary, machine learning techniques like Support Vector Machines (SVMs) are capable of learning and generalizing these data well (Mukherjee *et al.*, 1999; Furey *et al.*, 2000). Most classification methods from statistics also rely on linear functions and are unable to discover nonlinear relationships in microarray data, if any. By using nonlinear kernel functions when applying SVMs, one aims at a better understanding of these data (Brown *et al.*, 2000), especially when more patient data may become available in the future. A more detailed comparison between classification methods from statistics and machine learning will be described below.

Moreover, these machine learning methods may also be useful in case of analyzing classical clinical data, as is shown in our study on the usage of new models to predict depth of infiltration in endometrial carcinoma based on transvaginal sonography[1]. This study uses standard logistic regression models and compares these models to the Least Squares Support Vector Machine (LS-SVM) models (Suykens *et al.*, 2002) with linear and Radial Basis Function (RBF) kernels. Although logistic regression analysis is one of the standard statistical methods used for clinical classification problems (Epstein *et al.*, 2002; Friedland *et al.*, 2005; Timmerman *et al.*, 2005), these LS-SVM might still improve the performance, which will not further be discussed[2].

Statistical methods such as multivariate logistic regression and FDA intend to build a classification model that fits a set of patients ('training' set) optimally. Unfortunately, this strategy may easily result in a model that fits these training patients too well and is therefore not capable of making good predictions for previously unknown patients ('independent', 'prospective' or 'test' set). This problem is often referred to as overfitting the training patients, and leads to poor generalization towards previously unknown patients. SVMs are a relatively new method based on the principle of

---

[1] This study is accepted for publication in the journal *Ultrasound in Obstetrics and Gynecology* (De Smet *et al.*, 2006b).

[2] This opinion paper is accepted for publication in the journal *Ultrasound in Obstetrics and Gynecology* (Pochet and Suykens, 2006b).

statistical learning theory (Vapnik, 1998) to solve classification and regression problems. This method tries to learn and generalize well when building a model using a given set of patients. This way, SVMs perform reasonably well on a training set, but not at the expense of the performance when making predictions for previously unseen patients.

FDA and logistic regression try to fit a model as good as possible on the patients of the training set. Even with samples that do not follow the general underlying distribution in the case of outliers, these methods fit the training set too well, leading to a substantial number of misclassified patients when applied prospectively. SVMs try to generalize well when building a model using the given set of patients. In SVMs, optimization of the generalization performance is achieved by controlling two terms (i.e., by minimizing the classification error on the training set together with minimizing the complexity of the model). This trade-off is represented by a regularization parameter in the LS-SVM formulation.

Another disadvantage of FDA and logistic regression is that these techniques cannot look for possible nonlinear structures in a set of patients. When nonlinear relationships are present, a nonlinear decision boundary may result in an overall better performance. Unlike these methods, SVMs are designed to generate more complex decision boundaries. Using an LS-SVM with a simple linear kernel function corresponds to a linear decision boundary. Instead of a linear kernel, more complex kernel functions like for example the commonly used RBF kernel can be chosen. An RBF kernel requires optimization of the kernel parameter, which controls the curvature of the decision boundary. Figure 1.4 shows an example where using an LS-SVM with an RBF kernel is more appropriate than an LS-SVM with a simple linear kernel. Using this more complex decision boundary it is possible to better describe the nonlinearity in this set of patients than with a linear decision boundary.

LS-SVMs are reformulations and qualitatively similar to standard SVMs and they have already extensively been used for various classification problems, including medical ones (Van Gestel *et al.*, 2004; Timmerman *et al.*, 2005). LS-SVM models can easily be trained using LS-SVMlab version 1.5 (http://www.esat.kuleuven.be/sista/lssvmlab/) for MATLAB, as will be explained in the next chapter.

We summarize the two advantages of the recent SVMs over FDA and logistic regression. Unlike the latter methods, SVMs have means to prevent the model to be sensitive to outliers in the data, resulting in a model that is capable of making good predictions for prospective analyses. Moreover, SVMs can cope with nonlinearity in the data by using more complex nonlinear kernel functions instead of a simple linear kernel.

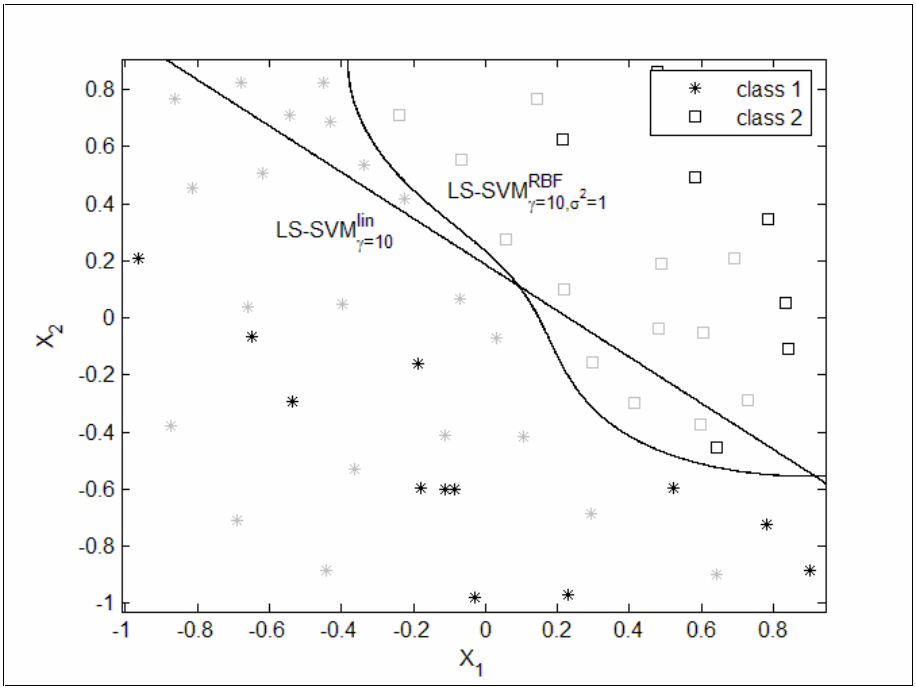


Figure 1.4 : *Using LS-SVM with a more complex decision boundary (represented by an RBF kernel) is more appropriate than with a simple linear decision boundary (using a linear kernel) in case of nonlinear structures in a set of patients. Training samples are represented in black, prospective samples in grey. Class 1 (*) and Class 2 (□) represent for example a set of diseased and a set of disease-free patients respectively. The information known for each of these patients is represented by the two variables X1 and X2, which constitute the axes of this plot.*

1.4.2 Clustering

Clustering techniques are generally applied to microarray data for the identification of clinical classes, which could allow refining clinical management. Cluster analysis of entire microarray experiments (expression patterns from patients or tissues) allows for the discovery of possibly unknown diagnostic categories without knowing the properties of these classes in advance. These clusters could form the basis of new diagnostic schemes in which the different categories contain patients with less clinical variability.

Clustering of microarray experiments has already shown to be useful in a large number of cancer studies. Alon *et al.* (1999), for example, separated cancerous colon tissues from non-cancerous colon tissues by

applying two-way clustering. The distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) has been rediscovered by using self-organizing maps (SOM) by Golub *et al.* (1999). By using hierarchical clustering, Van 't Veer *et al.* (2002) were able to distinguish between the presence (poor prognosis) and the absence (good prognosis) of distant subclinical metastases in breast cancer patients where the histopathological examination did not show tumor cells in local lymph nodes at diagnosis (lymph node negative).

For this purpose, methods such as the classical K-means clustering and hierarchical clustering are commonly used (Handl *et al.*, 2005; Bolshakova *et al.*, 2005). These methods are based on simple distance or similarity measures (e.g., the Euclidean distance). Therefore, only linear distance measures can be applied to the data using these techniques. Validation techniques are used to assess and compare the performance of different clustering methods. These methods can also be employed for tuning the cluster settings. Internal validation can be used to assess the quality of a clustering result based on statistical properties. External validation reflects the level of agreement of a clustering result with an external partition, e.g., existing diagnostic classes generally used by experts in clinical practice.

Recently, methods have emerged for clustering data where the clusters are not linearly separable. Two important methods are kernel K-means clustering (Dhillon *et al.*, 2004a; Dhillon *et al.*, 2004b; Zhang and Rudnicky, 2002) and the related spectral clustering (Cristianini *et al.*, 2002; Ng *et al.*, 2001). Introducing these techniques in microarray data analysis could allow for dealing with nonlinear relationships in the data and improving the computational complexity caused by high dimensional data. Since these techniques are based on kernel functions, this will also have an impact on the usage and development of the validation techniques.

1.4.3 Feature extraction

For the selection of relevant genes and groups of genes, the correlations between gene expression patterns and class labels are investigated. Not all gene expressions are correlated with the different diagnostic classes. The goal is to find gene expressions or groups of gene expressions that are correlated with the different diagnostic classes. Methods for selection of features or gene expressions will allow a deeper insight into the molecular biology of tumors. This can open up new perspectives for finding drug targets and tumor markers.

The simplest way to realize this is to select individual genes (univariate analysis) (Golub *et al.*, 1999; Rickman *et al.*, 2001; Sørlie *et al.*, 2001; Sotiriou *et al.*, 2002; Van 't Veer *et al.*, 2002; Welsh *et al.*, 2001;

Wigle *et al.*, 2002). This can be done using statistical methods (method proposed by Golub *et al.* (1999), *t*-test, Wilcoxon rank-sum test,...) that reflect the correlation between individual genes and a class difference. Another approach is to select groups of genes (multivariate analysis) (Alter *et al.*, 2000; Alter *et al.*, 2003; Guyon *et al.*, 2002). The classical Principal Components Analysis (PCA), which is an unsupervised method that looks for the most informative combination of gene expressions for a set of patients, is a suitable procedure to realize this. These methods are also capable of reducing the dimensionality of the genes as a preliminary step to the previous two objectives (classification and clustering). Note that PCA is not very suitable for dimensionality reduction towards classification because the former is unsupervised and the latter supervised. In this thesis, we therefore consider supervised as well as unsupervised selection of principal components. Furthermore, in this work we will focus on a more complex approach consisting of the kernel version of the classical PCA, namely the kernel PCA, which allows for using nonlinear kernel functions like RBF kernels.

1.5 Main contributions of this thesis

In this thesis, several research topics have been investigated. In this section, we give a general outline of the research subjects that are discussed in this dissertation, followed by other research items not explicitly described in this text.

In general, we realized the following challenges:

1. We studied and tested the behavior of statistical and machine learning techniques facing the specific problems induced by microarray data;
2. We optimized and fine-tuned these methods for biomedical applications;
3. We incorporated these methods into an interface that can easily be used by clinicians;
4. We applied these methods to solve several diagnostic problems.

More specifically, we focused on developing classification models for making optimal clinical predictions in oncology based on microarray data. For this purpose, we used classification and dimensionality reduction methods mentioned in Sections 1.4.1 and 1.4.3 and generally described in Chapter 2. In Chapter 3, we first performed a benchmarking study to assess the role of nonlinearity, dimensionality reduction and regularization together with microarray data. This resulted in several conclusions described in Chapter 3, which we also published in the journal *Bioinformatics* (Pochet *et al.*, 2004). Since it is important to carefully develop an optimal classifier for each cancer classification problem based on microarray data, we developed

an algorithm implemented in the freely available web service M@CBETH (http://www.esat.kuleuven.be/MACBETH/). This is described in Chapter 4 and also published in the journal *Bioinformatics* (Pochet *et al.*, 2005). Finally, in Chapter 5 we applied the findings of both Chapters 3 and 4 on an ovarian cancer microarray data set generated in a project we collaborated in (see Section 1.7). This resulted in a publication in *International Journal of Gynecological Cancer* (De Smet *et al.*, 2006a), a conference contribution at the *European Society of Gynaecological Oncology Conference (ESGO2005)* in Istanbul together with an abstract published in the *International Journal of Gynecological Cancer* (Van Gorp *et al.*, 2005), and a conference contribution in the *Computational Systems Bioinformatics Conference (CSB2005)* at Stanford University, California (Pochet *et al.*, 2005). In the same chapter, we also commented on two other recently published studies that have shortcomings in the techniques used to develop prediction models. We wrote two letters as a response to these publications, which were recently published in the journals *Clinical Cancer Research* (De Smet *et al.*, 2005) and *International Journal of Cancer* (Gevaert *et al.*, 2006).

A second main research topic in this dissertation is the usage of clustering algorithms for discovering diagnostic classes based on microarray data. In this context, we studied the advantages of using the nonlinear clustering techniques described in Section 1.4.2 and generally described in Chapter 2. In Chapter 6, we formulated extensions in the feature space for commonly used validation criteria. The work presented in this chapter has also been accepted for publication as a chapter "*Kernel clustering for knowledge discovery in clinical microarray data analysis*" (Pochet *et al.*, 2006a) in the book "*Kernel methods in bioengineering, communications and image processing*". Both main research subjects of this dissertation are discussed in more detail in the next section.

During the last few years, we have also contributed to other research not explicitly presented here. This includes our work on the analysis of classical clinical data in several specific applications. We generated models to predict depth of infiltration in endometrial carcinoma based on transvaginal sonography using traditional statistical and machine learning methods. A full paper on this subject together with an opinion paper on the advantages of machine learning methods compared to the traditional methods in clinical decision making have been accepted for publication in the journal *Ultrasound in Obstetrics and Gynecology*. Furthermore, we investigated several clinical questions based on a database containing classical clinical information from about 3,000 breast cancer patients in a project we collaborated in (see Section 1.7). In a first study, we investigated the usefulness of a negative progesterone receptor status (PR) as a predictor for a positive lymph node status in estrogen receptor (ER) positive breast tumors taking into account tumor size and tumor grade. For this purpose, we

used multivariate logistic regression analysis using stepwise selection to analyze clinical information of 1,472 women. We concluded that when analyzing the prognostic effect of a negative PR and lymph node status in women with an ER positive breast cancer, premenopausal (50 years or younger) and postmenopausal (older than 50 years) women should be considered separately. This study resulted in a letter that is accepted for publication in *Journal of Clinical Oncology* (Neven *et al.*, 2006a). Furthermore, we also investigated the usefulness of HER-2/neu as a predictor for axillary lymph node invasion in operable breast cancer. Univariate analysis (using the Fisher's Exact test) and multivariate analysis (using stepwise logistic regression) on 1741 breast cancer patients showed that HER-2/neu predicts lymph node invasion in ER positive PR positive breast cancer. A revised version of a full paper is submitted to *Journal of Clinical Oncology* (Neven *et al.*, 2006b). In another study, we investigated the correlation between the Body Mass Index (BMI) and HER-2 in postmenopausal breast cancer patients (older than 50 years) using univariate analysis techniques on clinical information of 539 women. This study revealed that in postmenopausal women with an operable breast cancer, there is an inverse association between BMI and HER-2 overexpression, which is most pronounced in ER positive breast cancer. A revised version of a full paper describing this finding is submitted to *Journal of the National Cancer Institute* (Van Mieghem *et al.*, 2006d). Some additional studies are accepted for publication as abstracts in the *International Journal of Gynecological Cancer*, among others on the prognostic value of ER in PR positive breast cancer (De Maeyer *et al.*, 2006), the effect of the BMI on the PR status in postmenopausal women with an ER positive breast cancer (Leunen *et al.*, 2006), how weight and BMI affect HER-2 expression in postmenopausal women with breast cancer (Van Mieghem *et al.,* 2006a), and the influence of body composition and hormone replacement therapy (HRT) on PR expression in postmenopausal women with an ER positive breast cancer (Van Mieghem *et al.*, 2006b). These studies were also presented on the *Flemish Gynecology Oncologic Group Conference (FGOG2006)*. A final study on the subject of the influence of weight and BMI on HER-2 expression in postmenopausal women with breast cancer is accepted for publication as an abstract in the *European Journal of Cancer* (Van Mieghem *et al*., 2006c), also presented at the *European Breast Cancer Conference (EBCC2006)*.

1.6 Chapter-by-chapter overview

In this dissertation, several research topics are described in different chapters. In this section, a summary is given of each chapter, elaborating on own contributions. The relations between the different chapters are

visualized in Figure 1.5. An overview of the methods, software and data sets used for the analyses in this work is presented in Figure 1.6. A list of own publications can be found in the beginning of this text.

Chapter 2: Least Squares Support Vector Machines and kernel methods

This chapter is devoted to a detailed description of the existing traditional statistical and machine learning techniques that are used further in this dissertation. These methods comprise classification, clustering, and dimensionality reduction techniques. The classical methods are described first, followed by their kernel version counterparts.

Chapter 3: Prediction models: benchmarking of microarray data classification

This chapter focuses on applying classification methods to microarray data, which can be useful to support clinical management decisions for individual patients, for example in oncology. The aim here is to systematically benchmark the role of nonlinear versus linear techniques and dimensionality reduction methods. We performed a systematic benchmarking study by comparing linear versions of standard classification and dimensionality reduction techniques with their nonlinear version counterparts based on nonlinear kernel functions with a radial basis function (RBF) kernel. A total of 9 binary cancer classification problems, derived from 7 publicly available microarray data sets (see Appendix A), and 20 randomizations of each problem are examined.

Chapter 4: M@CBETH web service: microarray classification tool

This chapter further elaborates on the clinical classification of microarray data by generating a web service, since comparing classifiers and selecting the best for each microarray data set can be a tedious and non-straightforward task. We developed the M@CBETH (a MicroArray Classification BEnchmarking Tool on a Host server) web service to offer the microarray community a simple tool for making optimal two-class predictions. M@CBETH aims at finding the best prediction among different classification methods by using randomizations of the benchmarking data set. This way, the M@CBETH web service intends to introduce an optimal use of clinical microarray data classification.

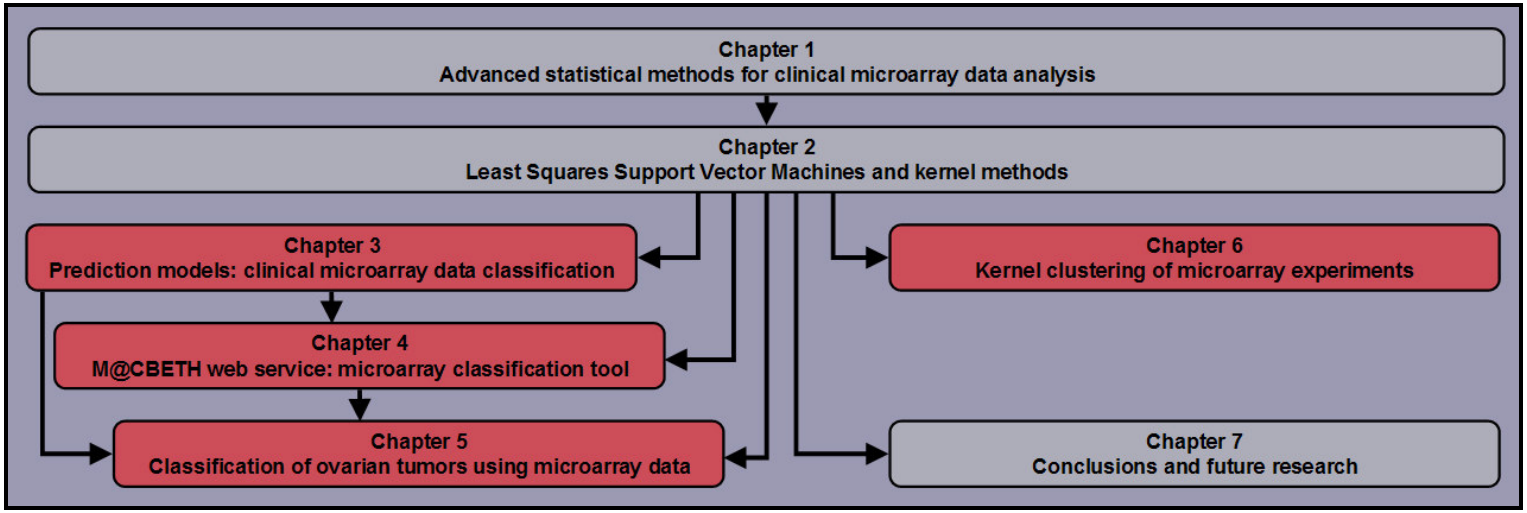


Figure 1.5 : *Relationships between the chapters of this dissertation. For clarity reasons, the arrows that connect every chapter with Chapter 7 (conclusions) are not indicated. The chapters that contain own research contributions are highlighted in red. Chapter 1 sheds light on our motivation to optimize clinical microarray data analysis by using machine learning methods. Existing machine learning techniques like Least Squares Support Vector Machines (LS-SVM) and kernel methods that are used throughout the rest of this dissertation are described in Chapter 2. In Chapter 3, we will investigate the influence of using classification techniques from machine learning when developing prediction models based on clinical microarray data. These findings result in the development of the M@CBETH web service in Chapter 4. Experience gained from Chapters 3 and 4 will be exploited to analyze microarray data from ovarian cancer within a project we collaborated in. In contrast with the previous chapters, Chapter 6 is devoted to investigating the usefulness of kernel methods for clustering microarray experiments. Finally, specific clinical cases comprising micorarray, proteomics and clinical data will be discussed in Chapter 7 (future research), followed by a description of new research directions that could lead to an improved version of the M@CBETH web service.*

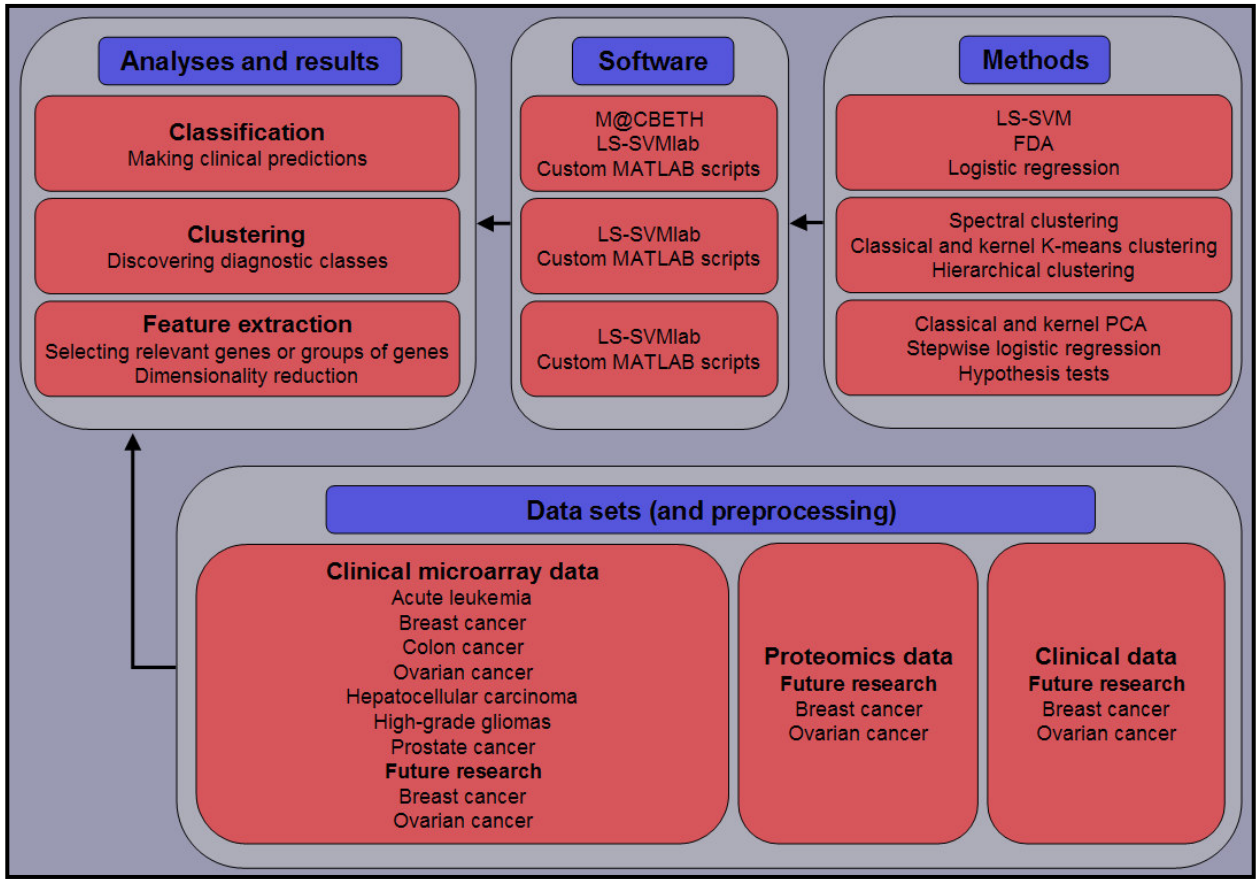


Figure 1.6 : *Overview of the three main perspectives for analyzing clinical microarray data: 1. making clinical predictions (classification), 2. discovering diagnostic classes (clustering experiments), and 3. selecting relevant genes (or groups of genes) or dimensionality reduction (feature extraction). The methods (together with the required software) and the data sets used in this work are represented. In the future, microarray da-ta will be combined with proteomics and clinical data for two specific clinical cases.*

Chapter 5: Classification of ovarian tumors using microarray data

This chapter is completely devoted to the analysis of gene expression data from ovarian tumors. We first applied the findings of both previous chapters on cDNA microarrays of ovarian tumors we generated in the context of a project (see below for collaborations). In a pilot study, we investigated whether prognostic information is reflected in the expression patterns of ovarian carcinoma samples, only using linear techniques. We then investigated whether the results of this pilot study can be confirmed and perhaps even further optimized by using the linear and nonlinear machine learning techniques incorporated in the M@CBETH web service. In both studies, prediction models are generated that will be used in a prospective study in the future. In this chapter, we also studied two other recently published studies that have shortcomings in the techniques used to develop prediction models and for which we formulated suggestions for improvements.

Chapter 6: Knowledge discovery: clustering of microarray data

This chapter considers classical K-means, kernel K-means and spectral clustering algorithms and discusses their advantages and disadvantages in the context of clinical microarray data analysis. Since classical K-means clustering experiences time complexity inconveniences when dealing with the high dimensional microarray experiments, Principal Component Analysis (PCA) is used as a preceding dimensionality reduction step. Kernel K-means and spectral clustering are capable of dealing with these data in a computationally more efficient way since these make use of the kernel trick, which allows them to work implicitly in the feature space. We described several internal and external cluster validation criteria commonly used in the input data space and we extended these for usage in the feature space. The advantages of nonlinear clustering techniques in case of clinical microarray data analysis are further demonstrated by means of the clustering results on several microarray data sets related to cancer.

Chapter 7: Conclusions and future research

This chapter first highlights our main accomplishments and then gives directions for future research on the shorter term and formulates future prospects on the longer term. Concerning future research on the short term, two specific clinical projects are considered that have already started or are planned, including a project on ovarian cancer and a project on breast cancer. In both projects, microarray, proteomics and clinical data will become available. We intend to apply the techniques that are described in this dissertation on these heterogeneous data sources. Concerning future

prospects on the longer term, we describe several interesting and possible extensions that may result in an improved version of the M@CBETH web service for usage in clinical practice.

1.7 Cooperations

The research described in this dissertation was realized in cooperation with clinicians at the University Hospitals Leuven, specifically on ovarian cancer and on endometrial cancer with Prof. I. Vergote and Prof. D. Timmerman of the Department of Obstetrics and Gynaecology and Gynaecologic Oncology of the University Hospitals, Leuven. Microarray experiments are generated in close collaboration with the Microarray Facility of the VIB (Flanders Interuniversity Institute for Biotechnology) located at the K.U.Leuven.

The research topics on breast cancer[3] mentioned in Section 1.5 that are not described in this dissertation were realized in cooperation with Prof. P. Neven at the University Hospitals, Leuven. We also started investigating the analysis of proteomics data obtained by mass spectrometry in the context of endometriosis in cooperation with Prof. T. D'Hooghe at the University Hospitals, Leuven. There also exist close collaborations with ProMeta (Interfacultary Center for Proteomics and Metabolomics) at the K.U.Leuven for data analysis.

[3] This resulted in a letter that is accepted for publication in *Journal of Clinical Oncology* (Neven *et al.*, 2006a) and the submission of a revised version of a full paper to the same journal (Neven *et al.*, 2006b). Another revised version of a full paper is submitted to *Journal of the National Cancer Institute* (Van Mieghem *et al.*, 2006d). Moreover, four abstracts are accepted for publication in *International Journal of Gynecological Cancer* (De Maeyer *et al.*, 2006; Leunen *et al.*, 2006; Van Mieghem *et al*., 2006a; Van Mieghem *et al*., 2006b) and one abstract is accepted for publication in *European Journal of Cancer* (Van Mieghem *et al*., 2006c).

Chapter 2

Least Squares Support Vector Machines and kernel methods

2.1 Introduction

As highlighted in Table 1.6 in previous chapter, microarray data can be analyzed from different perspectives in clinical applications: (1) making clinical predictions (classification), (2) discovering diagnostic classes (clustering experiments), and (3) selecting relevant genes (or groups of genes) or performing dimensionality reduction (feature extraction). The traditional statistical and machine learning methods that are considered for realizing these objectives in this work are described in this chapter, as can be seen in Table 1.5 in previous chapter. Although feature extraction is usually the first step in microarray data analysis, thereby preceding classification and clustering, we discuss classification and clustering first because the emphasis of this thesis lies on the latter techniques. Classification methods are discussed in Section 2.2. Section 2.3 elaborates on clustering algorithms. Finally, dimensionality reduction methods that can be used as a preprocessing step before performing classification or clustering are described in Section 2.4.

2.2 Classification methods

One of the most frequently used classification methods is Fisher Discriminant Analysis (FDA). This technique, however, is not suitable for handling high dimensional data. Therefore, dimensionality reduction techniques are required for preprocessing the data. In this work, we consider the FDA algorithm as the standard traditional statistical method to which the machine learning methods can be compared. The kernel version of FDA can be viewed as a special case of Least Squares Support Vector Machines (LS-

SVM). Owing to the regularization principle they are based on, such methods have already shown to be capable of directly dealing with high dimensional data in a number of other application area like text mining (Joachims *et al.*, 2002) and image analysis (Gupta *et al.*, 2002). This machine learning method will be proved to have a better behavior on microarray data further in this work. Moreover, since LS-SVM can be regarded as a kernel method, this technique can be used for developing linear as well as nonlinear classification models. In this section, the FDA algorithm is first explained, followed by the LS-SVM algorithm.

2.2.1 Fisher Discriminant Analysis

Fisher Discriminant Analysis (FDA) (Bishop, 1995) projects the microarray experiments $\mathbf{x}_i \in \Re^d$ from the original input space to a one-dimensional variable $z_i \in \Re$ and makes a discrimination based on this projected variable. In this one-dimensional space one tries to achieve a high discriminatory power by maximizing the between-class variances and to minimize the within-class variances for the two classes. The gene expression patterns are projected as follows

$$z = f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b,$$

with $f(\cdot) : \Re^d \rightarrow \Re$. One is interested then in finding a line such that the following objective of a Rayleigh quotient is maximized:

$$\max_{\mathbf{w}} J_{FD}(\mathbf{w}) = \frac{\mathbf{w}^T \sum_B \mathbf{w}}{\mathbf{w}^T \sum_W \mathbf{w}}.$$

The means of the expression patterns for class C_1 and class C_2 are $\varepsilon[\mathbf{x}^{(1)}] = \mu^{(1)}, \varepsilon[\mathbf{x}^{(2)}] = \mu^{(2)}$. The between and within covariance matrices related to class C_1 and class C_2 are $\sum_B = [\mu^{(1)} - \mu^{(2)}][\mu^{(1)} - \mu^{(2)}]^T$, $\sum_W = \varepsilon\{[\mathbf{x} - \mu^{(1)}][\mathbf{x} - \mu^{(1)}]^T\} + \varepsilon\{[\mathbf{x} - \mu^{(2)}][\mathbf{x} - \mu^{(2)}]^T\}$, where the latter is the sum of the two covariance matrices $\sum_{W_1}$, $\sum_{W_2}$ for the two classes. Note that the Rayleigh quotient is independent of the bias term b.

By choosing a threshold z_0, it is possible to classify a new microarray experiment as belonging to class C_1, if $z(\mathbf{x}) \geq z_0$, and classify it as belonging to class C_2, otherwise. Assuming that the projected data is the sum of a set of random variables allows invoking the central limit theorem

and modeling the class-conditional density functions $p(z \mid C_1)$ and $p(z \mid C_2)$ using normal distributions. After using Bayes' theorem to calculate the posterior probabilities $p(C_1 \mid z)$ and $p(C_2 \mid z)$, the threshold b follows from solving

$$P(C_1 \mid b) = P(C_2 \mid b).$$

In practice we do not solve this, but we evaluate the posterior probabilities using the test sample and assign it to C_1 if $p(C_1 \mid z) > p(C_2 \mid z)$ and to C_2 if $p(C_1 \mid z) < p(C_2 \mid z)$. Note that a new gene expression pattern can be classified with higher confidence if the difference between z and b is larger (and therefore also the difference between $p(C_1 \mid z)$ and $p(C_2 \mid z)$).

2.2.2 Least Squares Support Vector Machine classifiers

Least Squares Support Vector Machines (LS-SVM) (Suykens *et al.*, 2002; Suykens and Vandewalle, 1999; Van Gestel *et al.*, 2002; Pelckmans *et al.*, 2002) are a modified version of Support Vector Machines (Vapnik, 1998; Schölkopf *et al.*, 1999; Cristianini and Shawe-Taylor, 2000; Schölkopf *et al.*, 2001; Schölkopf and Smola, 2002). Since this work focuses on using the former, the latter is not further discussed. LS-SVM comprises a class of kernel machines with primal-dual interpretations related to kernel FDA, kernel PCA, kernel PLS (kernel Partial Least Squares), kernel CCA (kernel Canonical Correlation Analysis), recurrent networks and others. For classification this modification leads to solving a linear system instead of a quadratic programming problem, which makes LS-SVM easier and much faster than SVM on for example microarray data sets. The benchmarking study of (Van Gestel *et al.*, 2004) on 20 UCI data sets revealed that the results of LS-SVM are similar to those of SVM.

Given is a training set $\{\mathbf{x}_i, y_i\}_{i=1}^N$ with gene expression patterns $\mathbf{x}_i \in \Re^d$ and corresponding binary class labels $y_i \in \{-1,+1\}$. Vapnik's SVM classifier formulation was modified in (Suykens and Vandewalle, 1999) into the following LS-SVM formulation:

$$\min_{\mathbf{w},b,e} J_P(\mathbf{w},e) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + \gamma\frac{1}{2}\sum_{i=1}^N e_i^2,$$

such that

$$y_i[\mathbf{w}^T\varphi(\mathbf{x}_i) + b] = 1 - e_i, \qquad i = 1,...,N,$$

for a classifier in the primal space that takes the form

$$y(\mathbf{x}) = \text{sign}[\mathbf{w}^T \varphi(\mathbf{x}) + b],$$

where $\varphi(\cdot) : \Re^d \to \Re^{d_i}$ is the mapping to the high dimensional feature space and the regularization parameter. In the case of a linear classifier one could easily solve this primal problem, but in general $\mathbf{w}$ might be infinite dimensional (in case the feature map φ is infinite dimensional). For this nonlinear classifier formulation, the Lagrangian is solved, which results in the following dual problem to be solved in α, b:

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

where the kernel trick $K(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}_j)$ can be applied within the $\mathbf{\Omega}$ matrix

$$\mathbf{\Omega}_{ij} = y_i y_j \varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}_j) = y_i y_j K(\mathbf{x}_i, \mathbf{x}_j), \qquad i, j = 1, ..., N.$$

The classifier in the dual space takes the form

$$y(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b.$$

The chosen kernel function should be positive definite and satisfy the Mercer condition. The kernel functions used in this work are the linear kernel $K(\mathbf{x}, \mathbf{x}_i) = \mathbf{x}_i^T \mathbf{x}$ and the Radial Basis Function (RBF) kernel $K(\mathbf{x}, \mathbf{x}_i) = \exp\{-\|\mathbf{x} - \mathbf{x}_i\|_2^2 / \sigma^2\}$. The polynomial kernel of degree p, $K(\mathbf{x}, \mathbf{x}_i) = (\tau + \mathbf{x}_i^T \mathbf{x})^p$, with $\tau > 0$, is another commonly used kernel function.

Note that using LS-SVM with a linear kernel without regularization ($\gamma \to \infty$) is in fact the counterpart of classical linear FDA, but the latter needs dimensionality reduction while the former can solve the problem without dimensionality reduction in the dual form as the size of the linear system to be solved is $(N+1) \times (N+1)$ and is not determined by the number of gene expression levels. Hence, the advantage of using kernel methods like SVM or LS-SVM is that they can be used without performing dimensionality reduction first, which is not the case for the classical linear regression method FDA.

Practical note

LS-SVM models can easily be trained using LS-SVMlab version 1.5 for MATLAB, available at http://www.esat.kuleuven.be/sista/lssvmlab/. When training LS-SVM models with a linear and an RBF kernel, it is necessary to optimize the regularization parameter γ. Using an RBF kernel also requires tuning of the kernel parameter σ. These parameter(s) can be tuned in LS-SVMlab with the *tunelssvm* function using a *linesearch* approach for the LS-SVM with a linear kernel (tuning of γ) and a *gridsearch* approach for the LS-SVM with an RBF kernel (tuning of σ and γ) when optimizing the *(leave-one-out) cross-validation* performance on the training set. These parameter settings can subsequently be used when training the definitive model with the *trainlssvm* function. The *simlssvm* function allows making predictions for new patients using the previously built model. Since regularization is performed in LS-SVM models, the generalization of this technique on an independent set of patients can be expected to be better than standard FDA or logistic regression.

2.3 Clustering

Because of their conceptual simplicity and their wide availability in standard software packages, traditional clustering techniques such as K-means (Tavazoie *et al.*, 1999; Rosen *et al.*, 2005) and hierarchical clustering algorithms (Eisen *et al.*, 1998) are the predominant clustering methods in a wide range of applications. Therefore, this work focuses on a class of linear and nonlinear kernel clustering techniques based on the traditional K-means clustering. Kernel clustering methods have already shown to be useful in text mining applications (De Bie *et al.*, 2004; Dhillon *et al.*, 2004) and image data analysis (Zhang and Rudnicky, 2002) among others. These kernel clustering methods have recently emerged for clustering data in which the clusters are not linearly separable and to find nonlinear relationships in the data. Moreover, these techniques allow for dealing with high dimensional data, which makes it specifically interesting for application on microarray data. The kernel K-means and spectral clustering algorithms are considered in this work for this purpose. In this section, the classical K-means clustering algorithm is first explained, followed by the kernel K-means and the spectral clustering algorithms.

2.3.1 K-means clustering

K-means clustering aims at partitioning the data set $\{\mathbf{x}_i\}_{i=1}^N$ with expression patterns $\mathbf{x}_i \in \Re^d$, into G clusters $C_1,...,C_G$ such that the expression patterns in each cluster are more similar to each other than to the expression patterns in other clusters (Dubes and Jain, 1988). The centers or centroids (i.e., prototypes) of all clusters $\mathbf{m}_1,...,\mathbf{m}_G$ are returned as representatives of the clusters, together with the cluster assignments of all expression patterns. K-means clustering is based on the mean squared error criterion. The general objective is to obtain a partition that minimizes the mean squared error for a fixed number of clusters, where the mean squared error is the sum of the Euclidean distances between each expression pattern and its cluster center.

Suppose a set of expression patterns $\{\mathbf{x}_i\}_{i=1}^N$. The objective function (i.e., the mean squared error criterion) is then defined as

$$\text{se} = \sum_{k=1}^{G}\sum_{i=1}^{N} z_{C_k,\mathbf{x}_i} \left\|\mathbf{x}_i - \mathbf{m}_k\right\|^2,$$

where $z_{C_k,\mathbf{x}_i}$ is an indicator function defined as

$$z_{C_k,\mathbf{x}_i} = \begin{cases} 1 & \text{if } \mathbf{x}_i \in C_k; \\ 0 & \text{otherwise,} \end{cases}$$

with

$$\sum_{k=1}^{G} z_{C_k,\mathbf{x}_i} = 1, \qquad \forall i,$$

and $\mathbf{m}_k$ is the center of cluster C_k defined as

$$\mathbf{m}_k = \frac{1}{n_k}\sum_{i=1}^{N} z_{C_k\mathbf{x}_i}\mathbf{x}_i \ .$$

The Euclidean distance is often used as the dissimilarity function $D(\mathbf{x}_i, \mathbf{m}_k)$. The iterative K-means clustering algorithm first proposed by MacQueen (1967) optimizes this non-convex objective function as described in Table 2.1.

This algorithm can easily be implemented and works very well for compact and hyper-spherically shaped clusters. Although convergence is always reached, K-means does not necessarily find the best clustering (i.e.,

the global minimum for the objective function). The result of the algorithm is highly dependent on the number of clusters G and the initial selection of the G cluster centroids. Cluster validation criteria are required to choose the optimal settings for G and the initialization.

Finally, note that the classical K-means clustering experiences time complexity inconveniences when dealing with the high dimensional data. Using PCA (without selection of principal components) as a preceding dimensionality reduction step results in an improved computational complexity while the actual clustering results remain similar.

K-means clustering algorithm *(parameters:* $\{\mathbf{x}_i\}_{i=1}^N$, G *)*

1. Select G initial centroids $\mathbf{m}_1,...,\mathbf{m}_G$.

2. Assign each expression pattern $\mathbf{x}_i, 1 \le i \le N,$ to cluster C_k based on the indicator function:

$$z_{C_k,\mathbf{x}_i} = \begin{cases} 1 & D(\mathbf{x}_i, \mathbf{m}_k) < D(\mathbf{x}_i, \mathbf{m}_h), \forall h \neq k,\ k,\ h = 1,...,G; \\ 0 & \text{otherwise.} \end{cases}$$

3. Calculate the new centers $\mathbf{m}_k$ of all clusters C_k, with n_k the number of expression patterns in each cluster C_k, as

$$\mathbf{m}_k = \frac{1}{n_k} \sum_{i=1}^{N} z_{C_k,\mathbf{x}_i} \mathbf{x}_i.$$

4. Repeat steps (2) and (3) until convergence (no change).

5. Return $\mathbf{m}_k, \qquad k = 1,...,G.$

Table 2.1 : *K-means clustering algorithm.*

2.3.2 Kernel K-means clustering

Kernel K-means clustering is an extension of the linear K-means clustering algorithm in order to find nonlinear structures in the data. Consider a nonlinear mapping $\varphi(\cdot)$ from the input space to the feature space. No explicit construction of the nonlinear mapping $\varphi(\cdot)$ is required, since in this feature space inner products can easily be computed by using the kernel trick $K(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}_j)$. The choice of a positive definite kernel K

satisfying the Mercer condition (Vapnik, 1998) results in a kernel matrix $\mathbf{K}$ of size $N \times N$, which is symmetric and positive definite. The kernel matrix $\mathbf{K}$ holds all pairwise inner products of the input data $\mathbf{K}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j), \forall i, j = 1,...,N.$ This kernel trick can be applied to each algorithm that can be expressed in terms of inner products.

The objective function of kernel K-means clustering is exactly the same as the objective function of the classical K-means clustering stated earlier, except for the fact that it is now rewritten in terms of inner products that can be replaced by a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ (Dhillon *et al.*, 2004; Zhang *et al.*, 2002). By introducing the feature map $\varphi(\cdot)$, the mean squared error function can be expressed in the feature space by

$$\mathrm{se}^{\varphi} = \sum_{k=1}^{G} \sum_{i=1}^{N} z_{C_k, \mathbf{x}_i} \left\| \varphi(\mathbf{x}_i) - \mathbf{m}_k^{\varphi} \right\|^2,$$

with $\mathbf{m}_k^{\varphi}$ the cluster center of cluster C_k in the feature space, defined by

$$\mathbf{m}_{\mathrm{k}}^{\varphi} = \frac{1}{n_k} \sum_{j=1}^{N} z_{C_k, \mathbf{x}_j} \varphi(\mathbf{x}_j).$$

The Euclidean distance between $\varphi(\mathbf{x}_i)$ and $\varphi(\mathbf{x}_j)$ can be written as

$$\begin{aligned} D^2(\varphi(\mathbf{x}_i), \varphi(\mathbf{x}_j)) &= \left\| \varphi(\mathbf{x}_i) - \varphi(\mathbf{x}_j) \right\|^2 \\ &= \varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}_i) - 2\varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}_j) + \varphi(\mathbf{x}_j)^T \varphi(\mathbf{x}_j) \\ &= K(\mathbf{x}_i, \mathbf{x}_i) - 2K(\mathbf{x}_i, \mathbf{x}_j) + K(\mathbf{x}_j, \mathbf{x}_j). \end{aligned}$$

The computation of distances in this feature space can then be carried out by

$$\begin{aligned} D^2(\varphi(\mathbf{x}_i), \mathbf{m}_k^{\varphi}) &= \left\| \varphi(\mathbf{x}_i) - \mathbf{m}_k^{\varphi} \right\|^2 \\ &= \left\| \varphi(\mathbf{x}_i) - \frac{1}{n_k} \sum_{j=1}^{N} z_{C_k, \mathbf{x}_j} \varphi(\mathbf{x}_j) \right\|^2 \\ &= \varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}_i) - \frac{2}{n_k} \sum_{j=1}^{N} z_{C_k, \mathbf{x}_j} \varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}_j) \\ &\quad + \frac{1}{n_k^2} \sum_{j=1}^{N} \sum_{l=1}^{N} z_{C_k, \mathbf{x}_j} z_{C_k, \mathbf{x}_l} \varphi(\mathbf{x}_j)^T \varphi(\mathbf{x}_l). \end{aligned}$$

Application of the kernel trick results in

$$D^2\left(\varphi(\mathbf{x}_i), \mathbf{m}_k^{\varphi}\right) = \mathbf{K}_{ii} + f\left(C_k, \mathbf{x}_i\right) + g\left(C_k\right),$$

with

$$f\left(C_k, \mathbf{x}_i\right) = -\frac{2}{n_k} \sum_{j=1}^{N} z_{C_k, \mathbf{x}_j} \mathbf{K}_{ij},$$

and

$$g\left(C_k\right) = \frac{1}{n_k^2} \sum_{j=1}^{N} \sum_{l=1}^{N} z_{C_k, \mathbf{x}_j} z_{C_k, \mathbf{x}_l} \mathbf{K}_{jl}.$$

This gives the following formulation of the mean squared error criterion in the feature space:

$$\mathrm{se}^{\varphi} = \sum_{k=1}^{G} \sum_{i=1}^{N} z_{C_k, \mathbf{x}_i} \left(\mathbf{K}_{ii} + f\left(C_k, \mathbf{x}_i\right) + g\left(C_k\right)\right).$$

The kernel-based K-means algorithm solving the non-convex optimization problem is then as described in Table 2.2.

Note also that in this algorithm, the factor $\mathbf{K}_{ii}$ is ignored because it does not contribute to determine the closest cluster. Remark that the term $g\left(C_k\right)$ needs to be computed only once for each cluster in each iteration, while the term $f\left(C_k, \mathbf{x}_i\right)$ is calculated once per data point.

The objective function of the kernel K-means algorithm monotonically decreases in each iteration, which also holds for the classical K-means algorithm. The general structure of the traditional algorithm is thus preserved in its nonlinear version. Nevertheless, there are two main differences between both algorithms, namely the nonlinear mapping via the kernel trick and the lack of an explicit centroid in the feature space.

This algorithm, unfortunately, is also prone to local minima since the optimization problem is not convex. Considerable effort has been devoted to finding good initial guesses or inserting additional constraints to limit the effect of this fact on the quality of the solution obtained. The spectral clustering algorithm is a relaxation of this problem for which it is possible to find the global solution.

Finally, note that the kernel K-means algorithm almost degenerates to the classical K-means algorithm when using a linear kernel in the former. However, since kernel K-means makes use of the kernel trick and works implicitly in the feature space, this method is able to deal with high dimensional data in a computationally more efficient way.

Kernel K-means clustering algorithm *(parameters:* $\{\mathbf{x}_i\}_{i=1}^N$, G, $K(\cdot,\cdot)$*)*

1. Assign an initial clustering value to each sample, $z_{C_k,\mathbf{x}_i}, 1 \le k \le G, 1 \le \mathrm{i} \le N$, forming G initial clusters $C_1,\ldots,C_G$.

2. For each cluster C_k, compute n_k and $g(C_k)$.

3. For each training sample $\mathbf{x}_i$ and cluster C_k, compute $f(C_k,\mathbf{x}_i)$.

4. Assign $\mathbf{x}_i$ to the closest cluster by computing the value of the indicator function

$$z_{C_k,\mathbf{x}_i} = \begin{cases} 1 & f(C_k,\mathbf{x}_i)+g(C_k) < f(C_h,\mathbf{x}_i)+g(C_h), \\ & \qquad \forall h \neq k, k, h = 1,\ldots,G; \\ 0 & \text{otherwise.} \end{cases}$$

5. Repeat steps (2), (3) and (4) until convergence is reached.

6. For each cluster C_k, select the sample that is closest to the center as the representative centroid of cluster C_k, by computing

$$\mathbf{m}_k = \min_{\mathbf{x}_i \text{ s.t. } z_{C_k,\mathbf{x}_i}=1} D(\varphi(\mathbf{x}_i), \mathbf{m}_k^{\varphi}), \qquad k = 1,\ldots,G.$$

Table 2.2 : *Kernel K-means clustering algorithm.*

2.3.3 Spectral clustering

Spectral clustering is another technique that allows nonlinear separation of clusters by using graph partitioning (Cristianini *et al.*, 2002; De Bie *et al.*, 2004; Ng *et al.*, 2001). In this approach, the cluster problem is relaxed or restated, leading to efficient algorithms requiring no more than solving an eigenvalue problem. Unlike standard clustering methods, spectral clustering does not make assumptions on Gaussian class distributions. This is realized by avoiding the use of the Euclidean distance (as dissimilarity measure) or the inner product (as similarity or affinity measure). Instead, often an affinity measure based on an RBF kernel is used:

$$\mathbf{A}(\mathbf{x}_i, \mathbf{y}_j) = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right),$$

The width σ^2 of the kernel function determines how rapidly the affinity matrix $\mathbf{A}(\mathbf{x}_i, \mathbf{y}_j) = \mathbf{A}_{ij}$ decreases with the distance between expression patterns $\mathbf{x}_i$ and $\mathbf{x}_j$. In fact, the positive definiteness of the RBF kernel is not a requirement, although an RBF kernel is often used for spectral clustering. On the other hand, as is the case for the RBF kernel, affinities should also be symmetric and all entries must be positive. The matrix containing the affinities between all pairs of microarray experiments should therefore be referred to as the affinity matrix in a spectral clustering context and not as the kernel matrix.

Before describing the actual spectral clustering algorithm used in this work, some more background information is given. Spectral clustering methods can be regarded as relaxations of graph cut problems on a fully connected graph. In this graph each node represents a gene expression pattern, and the edges between the expression patterns are assigned weights that are equal to the affinities. Clustering then corresponds to partitioning the nodes (expression patterns) in the graph into groups (clusters). Such a division of the graph nodes into two disjoint sets is called a graph cut. To achieve a good clustering, one can see that it is undesirable to separate two nodes into different clusters if they are connected by an edge with a large weight, which corresponds to a large affinity. To translate this into an optimization problem, several graph cut cost functions for clustering have been proposed in literature, among which the cut cost, the average cut cost and the normalized cut cost (Shi and Malik, 2000). The cut cost is immediately computationally tractable (Blum and Chawla, 2001), but it often leads to degenerate results (where all but one of the clusters are trivially small, see Joachims (2003)). This problem can largely be solved by using the average or normalized cut cost functions, of which the average cut cost seems to be more vulnerable to outliers (distant microarray experiments, meaning that they have low affinity to all other gene expression patterns). Unfortunately, both optimizing the average and normalized cut costs are NP-complete problems. Therefore, spectral relaxations of these optimization problems have been proposed (Shi and Malik, 2000; Ng *et al.*, 2002; Cristianini *et al.*, 2002). These spectral relaxations are known as spectral clustering algorithms.

Suppose a microarray data set $\{\mathbf{x}_i\}_{i=1}^N$ with expression patterns $\mathbf{x}_i \in \Re^d$. A well-known instance of spectral clustering, proposed by Ng *et al.* (2001), finds G clusters in the data as described in Table 2.3. Conditions in which the algorithm is expected to do well are described by Ng *et al.* (2002). Once the samples are represented by rows of $\mathbf{V}$ (with dimension G), tight clusters are formed.

Spectral clustering algorithm *(parameters:* $\{\mathbf{x}_i\}_{i=1}^{N}$, G, $K(\cdot,\cdot)$*)*

1. Form the affinity matrix $\mathbf{A}$ (with dimensions $N \times N$) defined by $\mathbf{A}_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2}\right)$, if $i \neq j$, and $\mathbf{A}_{ii} = 0$.

2. Define $\mathbf{D}$ to be the diagonal matrix of which the element (i,i) is the sum of row i of $\mathbf{A}$, and construct the matrix $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$.

3. Find $\mathbf{x}_1,...,\mathbf{x}_G$ the G largest eigenvectors of $\mathbf{L}$ (chosen to be orthogonal to each other in the case of repeated eigenvalues), and form the matrix $\mathbf{U} = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad ... \quad \mathbf{u}_G]$ (with dimensions $N \times G$) by stacking the eigenvectors in columns.

4. Form the matrix $\mathbf{V}$ from $\mathbf{U}$ by renormalizing each row of $\mathbf{U}$ to have unit length: $\mathbf{V}_{ij} = \mathbf{U}_{ij} \Big/ \left(\sum_j \mathbf{U}_{ij}^2\right)^{1/2}$.

5. Treating each row of $\mathbf{V}$ as a sample with dimension G, cluster these into G clusters via K-means or any other algorithm (that attempts to optimize an internal validation criterion).

6. Finally, assign the original sample $\mathbf{x}_i$ to cluster C_j if and only if row i of the matrix $\mathbf{V}$ is assigned to cluster C_j.

Table 2.3 : *Spectral clustering algorithm.*

2.4 Dimensionality reduction

Dimensionality reduction is often performed in an unsupervised multivariate way by using Principal Component Analysis (PCA). This technique as well as its kernel version are applied in this work and described in the next sections.

2.4.1 Principal Component Analysis

Principal Component Analysis (PCA) looks for linear combinations of gene expression levels to obtain a maximal variance over a set of patients. In fact, those combinations are most informative for this set of patients and are called the principal components. One formulation to characterize PCA

problems is to consider a given set of centered (zero mean) microarray experiments $\{\mathbf{x}_i\}_{i=1}^N$ as a cloud of points for which one tries to find projected variables $\mathbf{w}^T\mathbf{x}$ with maximal variance. This means,

$$\max_{\mathbf{w}} \mathrm{Var}(\mathbf{w}^T\mathbf{x}) = \mathbf{w}^T C\mathbf{w},$$

where the covariance matrix $\mathbf{C}$ is estimated as $\mathbf{C} = (1/(N-1))\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^T$. One optimizes this objective function under the constraint that $\mathbf{w}^T\mathbf{w} = 1$. Solving the constrained optimization problem gives the eigenvalue problem

$$\mathbf{C}\mathbf{w} = \lambda\mathbf{w}.$$

The matrix $\mathbf{C}$ is symmetric and positive semi-definite. The eigenvector $\mathbf{w}$ corresponding to the largest eigenvalue determines the projected variable having maximal variance.

2.4.2 Kernel Principal Component Analysis

Kernel Principal Component Analysis (kernel PCA) (Schölkopf *et al.*, 1998) has the same goal as classical PCA, but is capable of looking for nonlinear combinations too. The objective of kernel PCA can be formulated as

$$\max_{\mathbf{w}} \sum_{i=1}^N [\mathbf{w}^T(\varphi(\mathbf{x}_i) - \mu_\varphi)]^2,$$

with notation $\mu_\varphi = (1/N)\sum_{i=1}^N \varphi(\mathbf{x}_i)$ used for centering the microarray data in the feature space, where $\varphi(\cdot): \Re^d \to \Re^{d_i}$ is the mapping to a high dimensional feature space, which might be infinite dimensional. This can be interpreted as first mapping the gene expression patterns to a high dimensional feature space and next to projected variables. The following optimization problem is formulated in the primal weight space (Suykens *et al.*, 2003)

$$\max_{\mathbf{w},e} J_P(\mathbf{w}, e) = \gamma\frac{1}{2}\sum_{i=1}^N e_i^2 - \frac{1}{2}\mathbf{w}^T\mathbf{w},$$

such that

$$e_i = \mathbf{w}^T[\varphi(\mathbf{x}_i) - \mu_\varphi], \qquad i = 1,...,N.$$

This formulation states that the variance of the projected variables is maximized for the given N data points while keeping the norm of $\mathbf{w}$ small by the regularization term. By taking the conditions for optimality from the Lagrangian related to this constrained optimization problem, such as $\mathbf{w} = \sum_{i=1}^{N} \alpha_i [\varphi(\mathbf{x}_i) - \mu_\varphi]$ among others, and defining $\lambda = 1/\gamma$, one obtains the eigenvalue problem

$$\mathbf{\Omega}_c \alpha = \lambda \alpha,$$

with

$$\mathbf{\Omega}_{c,ij} = [\varphi(\mathbf{x}_i) - \mu_\varphi]^T [\varphi(\mathbf{x}_j) - \mu_\varphi], \qquad i, j = 1,..., N,$$

the elements for the centered kernel matrix $\mathbf{\Omega}_c$. Since the kernel trick $K(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}_j)$ can be applied to the centered kernel matrix, one may choose any positive definite kernel satisfying the Mercer condition. The centered kernel matrix can be computed as $\mathbf{\Omega}_c = \mathbf{M}_c \mathbf{\Omega} \mathbf{M}_c$ with $\mathbf{\Omega}_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{M}_c = \mathbf{I} - (1/N) 1_N 1_N^T$ the centering matrix where $\mathbf{I}$ denotes the identity matrix and 1_N is a vector of length N containing all ones. Further dimensionality reduction is done by selecting the eigenvectors corresponding to the largest eigenvalues.

2.5 Conclusion

In this chapter, a description is given of the existing traditional statistical and machine learning methods that are considered for realizing the objectives in this work, which can be seen in Table 1.5 in the previous chapter.

Classification methods used in this dissertation are FDA and LS-SVM. FDA, however, is not suitable for handling high dimensional data. Therefore, dimensionality reduction techniques are required for preprocessing the data. The kernel version of FDA can be viewed as a special case of LS-SVM.

K-means, kernel K-means and spectral clustering are the clustering algorithms considered here. Classical K-means clustering, however, experiences time complexity inconveniences when dealing with the high dimensional data. Using PCA (without selection of principal components) as a preceding dimensionality reduction step results in an improved computational complexity while the actual clustering results remain similar. The kernel K-means algorithm almost degenerates to the classical K-means

algorithm when using a linear kernel in the former. However, since kernel K-means and spectral clustering make use of the kernel trick and work implicitly in the feature space, these methods are able to deal with high dimensional data in a computationally more efficient way.

PCA and kernel PCA are the dimensionality reduction methods used as a preprocessing step before some of the classification and clustering methods. Note that kernel PCA with a linear kernel degenerates to classical PCA and therefore gives similar results, although the former is computationally more efficient on high dimensional data because of the kernel trick.

Chapter 3

Prediction models: clinical microarray data classification

3.1 Introduction

The main goal of this thesis is to support clinical management decisions in the future by using prediction models based on genomic information. Therefore, developing optimal models for each cancer classification problem is crucial and will constitute the main research topic in this chapter. In the first chapter, the characteristics of microarray data have already been highlighted. However, in the context of developing classification models, the most important characteristics of these data are repeated below before studying how to build such models in an optimal way. The statistical principles needed to develop these models have been described in the previous chapter. In this chapter, we will study how to obtain an optimal performance with prediction models based on microarray data[1].

The dimensions of microarray data sets are a crucial factor when determining which methods can or cannot be applied for making predictions. At this moment, generating microarray data sets is expensive because this technology is relatively new and still experimental. Therefore, the number of experiments that is feasible in an economic sense is limited. Microarray data sets are typically characterized by a high dimensionality in the sense of a small number of patients and a large number of gene expression levels for each patient. Most classification methods have problems with the high dimensional nature of microarray data and require dimensionality reduction first. On the contrary, SVMs are capable of learning and generalizing these

[1] The study described in this chapter has been published in the journal *Bioinformatics* (Pochet *et al.*, 2004).

data well because these are based on regularization principles (Mukherjee *et al*., 1999; Furey *et al*., 2000). Towards the future, it can be expected that the number of patients will increase when this technology becomes less expensive. Moreover, most classification methods like for example FDA rely on linear functions and are unable to discover nonlinear relationships in microarray data, if any. By using kernel functions, one aims at better understanding of these data (Brown *et al*., 2000), especially when more patient data may become available in the future.

To find an optimal strategy for making clinical predictions, we will perform a systematic benchmarking study to compare linear versions of the standard techniques applied to microarray data with their kernel version counterparts (using linear and RBF kernels) in this chapter. Note that even with a linear kernel, LS-SVM techniques could be more suitable as they contain regularization and do not require dimensionality reduction as applied in the dual space. However, using more complex kernel functions could be useful for generating prediction models on the larger microarray data sets, which would be interesting towards the future since data sets are expected to contain more microarray experiments. In the next section, we will therefore systematically investigate the influence of regularization, dimensionality reduction and nonlinearity on a wide variety of microarray data sets. The results on one specific partitioning of training, validation and test set (as often reported in literature) could easily lead to misleading results, especially in the case of a small number of patient data. Instead of doing this in an *ad hoc* manner, randomizations on all data sets are carried out in order to get a more reliable idea of the expected performance and the variation on it.

In this chapter, we will show that regularization or dimensionality reduction is required for the classification of microarray experiments. Furthermore, we will demonstrate that a nonlinear LS-SVM model with an RBF kernel is a first choice for the classification of microarray experiments.

3.2 Materials and methods

This section describes the setup of our benchmarking study. In Section 3.2.1, a discussion on the data sets is given first. Section 3.2.2 gives more details on the general preprocessing step used for all data sets. Next, Section 3.2.3 shows an inventory of the methods that are considered in the numerical experiments. These experiments are described in detail in Section 3.2.4.

3.2.1 Data sets

This study considers nine cancer classification problems, all comprising two classes. For this purpose, seven publicly available microarray data sets are used: colon cancer data (Alon *et al*., 1999), acute leukemia data (Golub *et al*., 1999), breast cancer data (Hedenfalk *et al*., 2001), hepatocellular carcinoma data (Iizuka *et al*., 2003), high-grade glioma data (Nutt *et al*., 2003), prostate cancer data (Singh *et al*., 2002) and breast cancer data (van 't Veer *et al*., 2002). More detailed information on these seven data sets is presented in Appendix A, together with guidelines for preprocessing procedures. Since the data set in Hedenfalk *et al*. (2001) contains three classes, three binary classification problems and corresponding data sets can be defined by considering each class versus the rest. In most of the data sets, all data samples have already been assigned to a training set or test set. In the cases of data sets for which a training set and test set have not been defined yet, two-third of the data samples of each class are assigned to the training set and the rest to the test set. A summary of the characteristics of all data sets can be found in Table 3.1.

Systematic benchmarking studies are important for obtaining reliable results allowing comparability and repeatability of the different numerical experiments. For this purpose, this study not only uses the original division of each data set in training and test set, but also reshuffles (randomizes) all data sets. Consequently, all numerical experiments are performed with 20 randomizations of the 9 original data sets as well. These randomizations are the same for all 9 numerical experiments on one data set (in MATLAB with the same seed for the random generator). They are also stratified, which means that each randomized training and test set contains the same number of samples of each class compared to the original training and test set. The results of all numerical experiments in the tables represent the mean and standard deviation of the results on each original data set and 20 randomizations.

3.2.2 Preprocessing: standardization

Biologists are mainly interested in the relative behavior instead of the absolute behavior of genes. Genes that are up- and downregulated together should have the same weights in subsequent algorithms. Applying standardization or rescaling (sometimes also called normalization) to the gene expression profiles can largely achieve this (Quackenbush, 2001).

data sets	class 1	class 2	training set	training set class 1	training set class 2	test set	test set class 1	test set class 2	gene expression levels	microarray technology
Alon et al., 1999	normal colon tissues	tumor colon tissues	40	14	26	22	8	14	2000	oligonucleotide
Golub et al., 1999	acute myeloid leukemia (AML) patients	acute lymphoblastic leukemia (ALL) patients	38	11	27	34	14	20	7129	oligonucleotide
Hedenfalk et al., 2001 (1)	breast cancer carriers of BRCA1 mutation	breast cancer carriers of BRCA2 or sporadic mutations	14	4	10	8	3	5	3226	cDNA
Hedenfalk et al., 2001 (2)	breast cancer carriers of BRCA2 mutation	breast cancer carriers of BRCA1 or sporadic mutations	14	5	9	8	3	5	3226	cDNA
Hedenfalk et al., 2001 (3)	breast cancer carriers of sporadic mutations	breast cancer carriers of BRCA1 or BRCA2 mutations	14	4	10	8	3	5	3226	cDNA
Lizuka et al., 2003	hepatocellular carcinoma patients with intrahepatic recurrence	hepatocellular carcinoma patients with intrahepatic non-recurrence	33	12	21	27	8	19	7129	oligonucleotide
Nutt et al., 2003	glioblastomas (high-grade gliomas)	anaplastic oligodendrogliomas (high-grade gliomas)	21	14	7	29	14	15	12625	oligonucleotide
Singh et al., 2002	tumor prostate tissues	normal prostate tissues	102	52	50	34	25	9	12600	oligonucleotide
Van 't Veer et al., 2002	breast cancer patients who developed distant metastasis within 5 years	breast cancer patients remain disease-free within 5 years	78	34	44	19	12	7	24188	cDNA

Table 3.1 : *Summary of the nine binary cancer classification problems data sets reflecting the dimensions of each data set, the size of the corresponding training and test sets, and the microarray technology of each data set. Data sets used are: colon cancer data of Alon et al. (1999); acute leukemia data of Golub et al. (1999); breast cancer data of Hedenfalk et al. (2001) taking the BRCA1 mutations versus the rest (1); breast cancer data of Hedenfalk et al. (2001) taking the BRCA2 mutations versus the rest (2); breast cancer data of Hedenfalk et al. (2001) taking the sporadic mutations versus the rest (3); hepatocellular carcinoma data of Iizuka et al. (2003); high-grade glioma data of Nutt et al. (2003); prostate cancer data of Singh et al. (2002); and breast cancer data of van 't Veer et al. (2002).*

Consider a gene expression profile, denoted by the vector $\mathbf{g} = [g^1, g^2, .., g^i, ..., g^N]$, measured for N experiments. Rescaling is commonly done by replacing every expression level g^i in $\mathbf{g}$ by

$$\frac{g^i - \mu}{\sigma},$$

where μ is the average expression level of the gene expression profile and is given by

$$\mu = \frac{\sum_{i=1}^{N} g^i}{N},$$

and σ is the standard deviation given by

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(g^i - \mu\right)^2}.$$

This is repeated for every gene expression profile in the data set and results in a collection of expression profiles all having average zero and standard deviation one (i.e., the absolute differences in expression behavior have been largely removed). The division by the standard deviation is sometimes omitted (rescaling is then called mean centering).

3.2.3 Methods

The methods used to set up the numerical experiments can be subdivided in two categories: dimensionality reduction and classification. For dimensionality reduction, classical PCA and kernel PCA are used. FDA and LS-SVM (the kernel version of FDA can be viewed as a special case of LS-SVM) are used for classification. All these methods are described in more technical detail in the previous chapter.

3.2.4 Numerical experiments

As discussed above, nine classification problems are considered in this study. The numerical experiments applied to all these problems can be divided into two subgroups, depending on the required parameter optimization procedure. First, three kinds of experiments, all without dimensionality reduction, are performed to all nine classification problems. These are LS-SVM with linear kernel, LS-SVM with RBF kernel and LS-

SVM with linear kernel and infinite regularization parameter ($\gamma \to \infty$). Next, six kinds of experiments, all using dimensionality reduction, are performed on all nine classification problems. The first two of these are based on classical PCA followed by FDA for building the classifier. Selection of the principal components is done both in an unsupervised and a supervised way. The same strategy is used in the last four of these, but kernel PCA with linear kernel as well as RBF kernel are used instead of classical linear PCA.

Since building a prediction model requires good generalization towards making predictions for previously unseen test samples, tuning the parameters is an important issue. The small sample size characterizing microarray data restricts the choice of an estimator for the generalization performance. The optimization criterion used in this study is the leave-one-out cross-validation (LOO-CV) performance. In each LOO-CV iteration (number of iterations equals the sample size), one sample is left out of the data, a classification model is trained on the rest of the data and this model is then evaluated on the left out data point. As an evaluation measure, the LOO-CV performance [*(*No. of correctly classified samples*)/(*No. of samples in the data*)* · 100]% is used.

All numerical experiments are implemented in MATLAB by using the LS-SVM and kernel PCA implementations of the LS-SVMlab toolbox (http://www.esat.kuleuven.be/sista/lssvmlab/).

The optimization procedures that are conducted to tune the parameters are discussed in the next two sections, distinguishing between the cases without and with dimensionality reduction. This is followed by a section describing how to select the principal components in an unsupervised and a supervised way. Finally, a section is devoted to measuring and comparing the performances of all numerical experiments.

Tuning parameter optimization for the case without dimensionality reduction

When using LS-SVM with a linear kernel, only the regularization constant needs to be further optimized. The value of the regularization parameter corresponding to the largest LOO-CV performance is then selected as the optimal value. Using an RBF kernel instead requires optimization of the regularization parameter γ as well as the kernel parameter σ. This is done by searching a two dimensional grid of different values for both parameters. Using LS-SVM with a linear kernel and infinite regularization parameter, which corresponds to FDA, requires no tuning parameter optimization.

After the preprocessing steps that are specifically required for each data set (as discussed in Appendix A), standardization (which is discussed above) is always performed on all the data sets to preprocess them before

usage for classification purposes. This is done by standardizing each gene expression of the data to have zero mean and unit standard deviation. Standardization of training sets as well as test sets is done by using the mean and standard deviation of each gene expression profile of the training sets.

Tuning parameter optimization for the case with dimensionality reduction

When reducing the dimensionality of the expression patterns of the patients with classical PCA and next building a prediction model by means of FDA, the number of principal components needs to be optimized first. This is realized by performing LOO-CV on the training set. An outline of the complete algorithm is shown in Table 3.2.

For each possible number of principal components (ranging between 1 and $N-2$, with N the number of training samples), the LOO-CV performance is computed. The number of principal components with best LOO-CV performance is then selected as the optimal one. If there exist different numbers of principal components with the same best LOO-CV performance, the smallest number of principal components is selected. This choice can be interpreted as minimizing the complexity of the model. If kernel PCA with a linear kernel is used instead of the classical PCA, the same method is used. Using kernel PCA with an RBF kernel not only requires optimization of the number of principal components, but also the kernel parameter σ needs to be tuned. A broad outline of the optimization procedure is described below. For several possible values of the kernel parameter, the LOO-CV performance is computed for each possible number of principal components.

The optimal number of principal components with the best LOO-CV performance is then selected for each value of the kernel parameter. If there are several optimal numbers of principal components, the smallest number of principal components is selected, again to obtain minimal model complexity. To find the optimal value for the kernel parameter, the value of the kernel parameter with best LOO-CV performance is selected. If there are several possible optimal values for the kernel parameter, also the optimal number of principal components belonging to these optimal kernel parameter values need to be considered. From these values, the optimal kernel parameter value with the smallest number of principal components is chosen. If there are still several possible optimal kernel parameter values, the smallest value of these is selected as the optimal one. Note the complexity of this optimization procedure because both the kernel parameter and the number of principal components of the kernel PCA with RBF kernel need to be optimized in the sense of the LOO-CV performance of the FDA classification.

Optimization algorithm:
kernel PCA with RBF kernel followed by FDA

(1) Generation of parameter grid

for each kernel parameter value within selected range
for each possible # principal components
for each LOO-CV iteration
• leave one sample out
• standardization
• dimensionality reduction (kernel PCA)
• selection of the principal components (unsupervised or supervised)
• classification (FDA)
• test sample left out
end

calculate LOO-CV performance
end
end

(2) Optimization of parameters

for each kernel parameter value out of a range
optimal # principal components:
1. best LOO-CV performance
2. smallest # principal components *
end

optimal kernel parameter value:
1. best LOO-CV performance
2. smallest # principal components *
3. smallest kernel parameter value *

* if more than one

Table 3.2 : *Optimization algorithm for tuning the parameters in case of kernel PCA with RBF kernel followed by FDA.*

Standardization of the samples left out in each LOO-CV iteration also needs to be done based on the mean and standard deviation of each gene expression profile of each accompanying LOO-CV training set (of all

samples except the left out samples). Concerning dimensionality reduction, it should be noted that this is also done based on the training set. First, PCA is applied to the training set, which results in eigenvalues and eigenvectors going from 1 till $N-1$. The training and test set are then projected onto those eigenvectors. As the data are centered, the last eigenvalue is equal to zero. Therefore, the last principal component is left out, which results in the number of principal components going from 1 till $N-2$. In fact, this corresponds to obtaining a low-rank approximation starting from a full rank matrix.

Supervised versus unsupervised selection of principal components

Concerning the experiments with dimensionality reduction, two ways of selecting the principal components are used. The first one simply looks at the eigenvalues of the principal components, originating from PCA. Since this method does not take into account the class labels, it is unsupervised. The other one is based on the absolute value of the score introduced by Golub *et al*. (1999), as also used in Furey *et al*. (2000):

$$F(\mathbf{x}_i) = \left| \frac{\mu_i^1 - \mu_i^2}{\sigma_i^1 - \sigma_i^2} \right|$$

This method allows finding individual gene expression profiles that help discriminating between two classes by calculating for each gene expression profile $\mathbf{x}_i$ a score based on the mean μ_i^1 (respectively μ_i^2) and the standard deviation σ_i^1 (respectively σ_i^2) of each class of samples. In our experiments, this method is applied onto the principal components instead of applying it directly to the gene expression profiles. This method takes into account the class labels and is therefore called supervised. The *n* most important principal components now correspond to the *n* principal components with either the highest eigenvalues or the highest absolute value of the score introduced by Golub.

Measuring and comparing the performance of the numerical experiments

For the results, three kinds of measures are used. The first one is the LOO-CV performance. This is estimated based on only using the training data sets for tuning the parameters.

The second measure is the accuracy (ACC), which gives an idea of the classification performance by reflecting the percentage of samples classified correctly. When measured on independent test sets, this gives an idea of the generalization performance. When measured on the training set as well, one can get an idea of the degree of overfitting by comparing both

performances. Overfitting can then be concluded in case of a high training set performance and a low test set performance.

The third measure is the Area Under the Receiver Operating Characteristic (ROC) Curve performance (AUC) (Hanley and McNeil, 1982). An ROC curve shows the separation abilities of a binary classifier: by setting different possible classifier thresholds, sensitivity versus (1 – specificity) are plotted resulting in the ROC curve. Sensitivity can be calculated as

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}},$$

and the specificity as

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}.$$

If the AUC performance equals 100% on a data set, a perfectly separating classifier is found on that particular data set, if the area equals 50%, the classifier has no discriminative power at all. This measure can be evaluated on an independent test set or training set.

Statistical significance tests are performed to allow a correct interpretation of the results, taking into account all randomizations. A non-parametric paired test, the Wilcoxon signed rank test (signrank in MATLAB) (Dawson-Saunders and Trapp, 1994), has been used to make general conclusions. A threshold of 0.05 is respected, which means that two results are statistically significantly different if the value of the Wilcoxon signed rank test applied to both of them is lower than 0.05.

3.3 Results

This section first reveals some general comments based on issues that are common among all nine classification problems. This is followed by a discussion of the most prominent results on four specific cases, which are summarized in Table 3.3. Detailed results (with statistical significance tests) on all cancer classification problems can be found in Appendix B.

For each classification problem, the results in the tables represent the statistical summary (mean and standard deviation) of the numerical experiments (expressed in terms of LOO-CV performances, training and test set accuracies, and training and test ROC performances) on the original data set and 20 randomizations of it. Since the randomizations (training and test set splits) are not disjoint, the results as well as the statistical significance

tests given in the tables are not unbiased and can in general also be too optimistic.

The results are also represented by means of boxplots (function boxplot in MATLAB), as shown in Figure 3.1. Such a boxplot consists of a

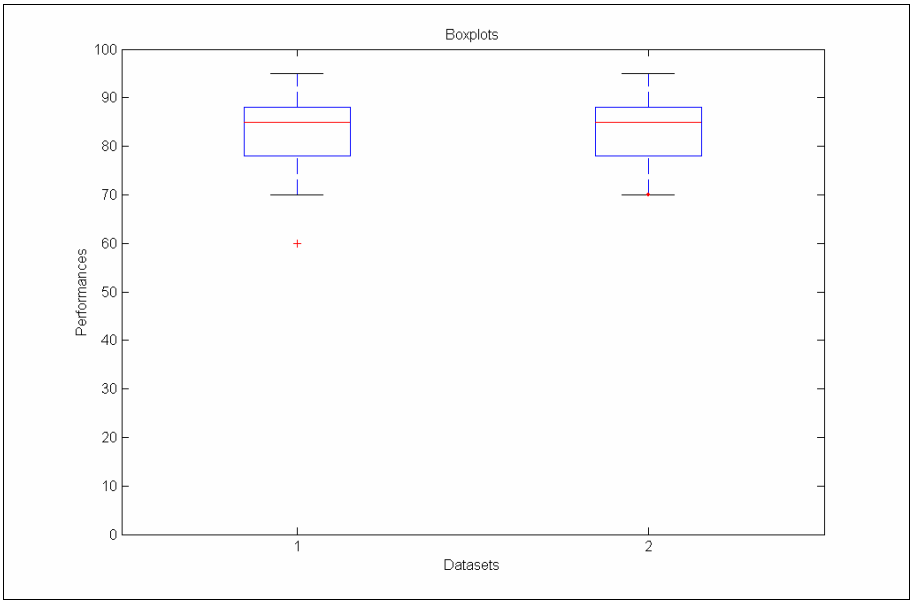


Figure 3.1 : *Illustration of some examples showing boxplots generated with MATLAB. This boxplot representation allows visualization of the lines at the lower quartile, median (visualized in red), and upper quartile values and whiskers extending from the box out to the most extreme data value within 1.5*IQR: (1) for data of which the performances contain outliers (represented by the symbol '+' visualized in red); (2) for data of which the performances do not contain outliers (represented by the red dot on the lower whisker).*

rectangular box and whisker plot for each column of the data provided. In our case, boxplots are generated to visualize the LOO-CV performances, the training and test set accuracies, and the training and test ROC performances. The box has lines at the lower quartile, median (visualized in red), and upper quartile values. The whiskers are lines extending from each end of the box to show the extent of the rest of the data. More specifically, whiskers extend from the box out to the most extreme data value within 1.5*IQR, where IQR is the interquartile range of the sample. The IQR can be computed as the difference between the 75th and the 25th percentiles of the sample. The IQR is a robust estimate of the spread of the data, since changes in the upper and lower 25% of the data do not affect it. Outliers are data with values beyond the ends of the whiskers and are represented by the symbol '+' visualized in red. If there is no data outside the whisker, a red dot is placed at the bottom whisker. This is also explained in the MATLAB Help documentation.

3.3.1 General findings

One general remark is that constructing the randomizations in a stratified way already seems to result in a large variance (it would have been even larger if constructed in a non-stratified way).

Another remark is that the LOO-CV performance is not a good indicator for the accuracy or the area under the ROC curve of the test set. This raises the question whether or not this LOO-CV performance is a good method for tuning the parameters. Since microarray data are characterized by a small sample size, LOO-CV has to be applied with care as one may easily overfit in this case.

For all data sets except the one containing the acute leukemia data (Golub *et al*., 1999), the LOO-CV performance, the test set accuracy and also the area under the ROC curve of the test set of the experiment based on LS-SVM with linear kernel and $\gamma \to \infty$ (i.e., no regularization) is significantly worse than all other experiments. This clearly indicates that regularization is very important when performing classification without previous dimensionality reduction, even for linear models. In the further discussion treating the individual data sets, this experiment will be left out.

The acute leukemia data (Golub *et al*., 1999) clearly comprises an easy classification problem, since the variances on the results caused by the randomizations are quite small compared to the other data sets. All experiments on this data set also seem to end up in quite similar results, so in fact it hardly matters which classification method is applied on this data set.

Observing the optimal values for the tuning parameters leads to the following remarks. When LS-SVM with a linear kernel is applied, typical values for the mean regularization parameter γ on each data set are ranging between $1e^{-3}$ and $1e^{+3}$. When using LS-SVM with an RBF kernel, typical values for the mean regularization parameter γ as well as the mean kernel parameter σ^2 on each data set both are ranging between $1e^{+10}$ and $1e^{+15}$. Optimal values for the kernel parameter σ^2 are quite large because they are scaled with the large input dimensionality of microarray data. Using kernel PCA with an RBF kernel before classification often results in test set performances that are worse than when using kernel PCA with a linear kernel, which means that overfitting occurs. Typical values for the mean kernel parameter σ^2 of the kernel PCA with RBF kernel on each data set highly depend on the way the principal components are selected. When using the unsupervised way for selecting the principal components, the mean of kernel parameter values σ^2 tends to go to $1e^{+20}$. Using the supervised way for selecting the principal components, $1e^{+0}$ is often selected as the optimal value for the kernel parameter σ^2, which leads to bad test set performances compared to the other experiments (seriously overfitting).

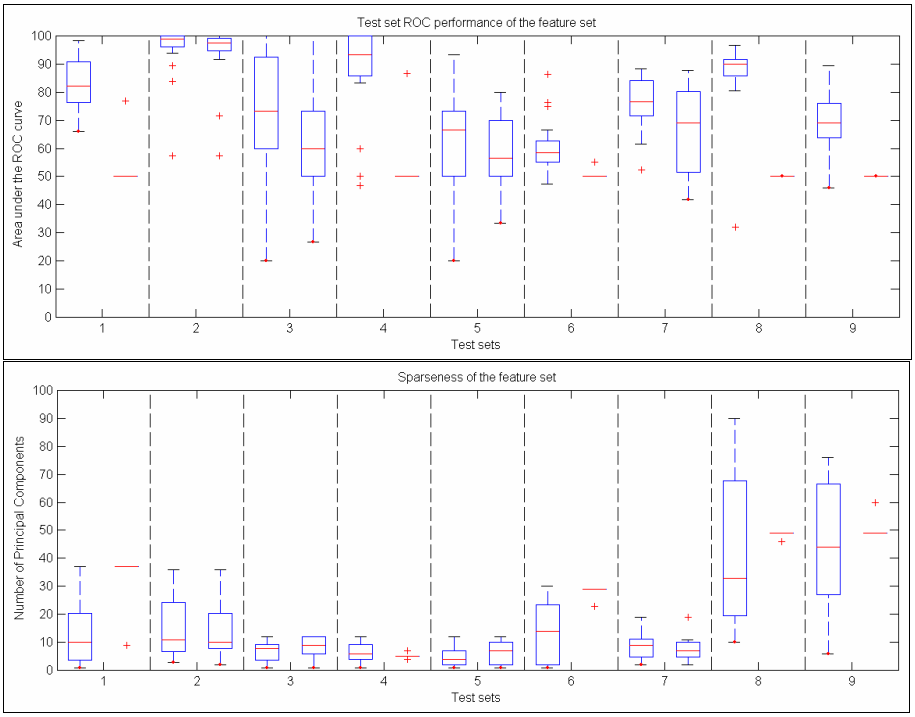


Figure 3.2 : *Illustration of the test set ROC performance (upper part) and the sparseness (lower part) of the optimally selected feature set based on boxplots of the areas under the ROC curve of the test set and boxplots of the optimal number of principal components respectively of all nine cancer classification problems. For each data set, the areas under the ROC curve of the test set and the optimal number of principal components of kernel PCA with a linear kernel (selecting the principal components in a supervised way) followed by FDA are represented on the left, the areas under the ROC curve of the test set and the optimal number of principal components of kernel PCA with an RBF kernel (selecting the principal components in a supervised way) followed by FDA on the right. Concerning the data sets, the order of Table 3.1 is respected. It has been observed in this study that* ***an optimal selection - in the sense of Table 3.2 - of a large number of features is often an indication for overfitting in case of kernel PCA with RBF kernel (supervised feature selection) followed by FDA****.*

In the context of parameter optimization, it is also important to address the number of selected features and in particular the sparseness of the projections obtained by traditional and kernel PCA. Figure 3.2 represents the test set ROC performance together with the sparseness when using a linear and an RBF kernel for kernel PCA. It has been noticed in this study that classical PCA leads to approximately the same results as kernel PCA with linear kernel and is therefore not represented separately. Selection of the principal components is done in a supervised way based on the LOO-CV performance criterion. Two observations can be stated when comparing the

results of these two experiments. First, when the optimal number of principal components is relatively low when using a linear kernel and much larger when using an RBF kernel. This is an indication of overfitting when using an RBF kernel. The colon cancer data set of (Alon *et al*., 1999) (1) and the hepatocellular carcinoma data set of (Iizuka *et al*., 2003) (6) are examples of this observation. Second, when the optimal number of principal components is very large both when using a linear kernel and when using an RBF kernel, this is an indication of overfitting too in both cases. The prostate cancer data set of (Singh *et al*., 2002) (8) and the breast cancer data set of (van 't Veer *et al*., 2002) (9) are illustrations of this observation.

3.3.2 Most prominent results on four specific cases

This section contains a discussion of the most prominent results on four specific cases, also summarized in Table 3.3.

Breast cancer data set (Hedenfalk *et al.*, 2001): BRCA1 mutations versus the rest

Concerning the test set accuracies, LS-SVM with RBF kernel obviously performs better than all other methods. Using an RBF kernel when doing kernel PCA by contrast, clearly performs worse when the eigenvalues are used for selection of the principal components. The results of the area under the ROC curve of the test set show that using LS-SVM results in much better performances than all other experiments, even when using a linear kernel. Both methods for selecting the principal components seem to perform very similarly, but in some cases using the absolute value of the Golub score tends to perform slightly better. Remarkable in this case is that the test set accuracy of LS-SVM with RBF kernel is much better than LS-SVM with linear kernel, although the area under the ROC curve of both experiments is practically equal. This is also an indication of how important it is to find a good decision threshold value, which corresponds to an operating point on the ROC curve.

High-grade glioma data set (Nutt *et al.*, 2003)

Concerning the test set performances, the experiment using LS-SVM with RBF kernel is significantly better than using LS-SVM with linear kernel. For this data set both methods for selection of the principal components give similar results.

Table 3.3 : *Summary of the results of the numerical experiments on four binary cancer classification problems, comprising the LOO-CV performance, the accuracy (ACC) on training and test set, and the area under the ROC curve (AUC) on training and test set. The results visualized in bold followed by (+) are statistically significantly better than the other results. The results in bold followed by (−) are statistically significantly worse than the other results.*

Hedenfalk et al., 2001: BRCA1 mutations	LOO-CV performance	ACC training set	ACC test set	AUC training set	AUC test set
LS-SVM linear kernel	78.23±7.13	87.76±14.14	64.29±6.99	100.00±0.00	**81.90±18.19** (+)
LS-SVM RBF kernel	82.65±8.12	98.64±6.08	**75.00±12.20** (+)	100.00±0.00	**82.22±17.38** (+)
LS-SVM linear kernel (no regularization)	46.94±21.21	47.62±9.94	**52.98±19.25** (−)	47.14±14.38	**52.70±24.16** (−)
PCA + FDA (unsupervised PC selection)	81.63±7.17	95.24±7.09	64.29±12.96	93.93±12.67	67.62±21.83
PCA + FDA (supervised PC selection)	84.01±9.58	97.96±4.49	68.45±15.25	97.86±5.25	71.75±21.12
kPCA lin + FDA (unsupervised PC selection)	81.29±7.13	95.24±6.73	63.10±13.07	96.55±5.64	66.35±20.23
kPCA lin + FDA (supervised PC selection)	84.35±8.99	98.30±4.36	67.86±15.70	98.45±4.12	72.38±22.23
kPCA RBF + FDA (unsupervised PC selection)	91.16±7.28	94.90±6.29	**54.17±11.79** (−)	95.36±7.98	60.63±16.25
kPCA RBF + FDA (supervised PC selection)	92.52±5.16	98.30±5.36	63.69±10.85	97.68±7.72	64.13±18.54

Nutt et al., 2003	LOO-CV performance	ACC training set	ACC test set	AUC training set	AUC test set
LS-SVM linear kernel	75.74±8.93	90.02±14.16	61.25±11.75	99.47±1.03	79.25±6.06
LS-SVM RBF kernel	78.23±7.99	98.41±7.10	**69.95±8.59** (+)	100.00±0.00	**81.04±6.64** (+)
LS-SVM linear kernel (no regularization)	50.79±16.65	50.79±12.75	**48.93±10.88** (−)	50.63±16.40	**50.68±15.15** (−)
PCA + FDA (unsupervised PC selection)	80.95±7.49	92.29±7.12	67.82±7.24	97.72±2.80	77.48±10.50
PCA + FDA (supervised PC selection)	81.41±7.19	92.97±10.14	65.52±11.01	96.65±5.69	77.37±9.04
kPCA lin + FDA (unsupervised PC selection)	80.73±7.12	92.52±6.98	68.31±6.78	97.91±2.74	77.98±10.43
kPCA lin + FDA (supervised PC selection)	81.86±6.67	95.24±8.57	67.32±11.04	98.15±4.02	76.53±8.96
kPCA RBF + FDA (unsupervised PC selection)	86.62±5.99	94.78±9.05	**64.20±11.19** (−)	97.30±6.60	**70.80±15.44** (−)
kPCA RBF + FDA (supervised PC selection)	85.94±5.78	96.15±7.29	**58.13±12.24** (−)	98.25±3.78	**66.33±15.48** (−)

Singh et al., 2002	LOO-CV performance	ACC training set	ACC test set	AUC training set	AUC test set
LS-SVM linear kernel	90.10±1.42	100.00±0.00	84.31±13.66	100.00±0.00	**91.28±5.20** (+)
LS-SVM RBF kernel	91.22±1.19	99.95±0.21	**88.10±4.93** (+)	100.00±0.00	**92.04±5.03** (+)
LS-SVM linear kernel (no regularization)	50.33±0.92	51.45±7.03	**48.18±10.25** (−)	51.10±8.27	**50.98±12.38** (−)
PCA + FDA (unsupervised PC selection)	90.38±1.83	97.62±1.95	83.89±13.63	99.67±0.38	88.93±11.39
PCA + FDA (supervised PC selection)	90.57±1.53	97.57±3.34	82.49±13.35	99.40±0.99	86.74±12.95
kPCA lin + FDA (unsupervised PC selection)	90.34±1.75	97.57±1.90	85.01±9.07	99.67±0.38	89.98±7.30
kPCA lin + FDA (supervised PC selection)	90.57±1.53	97.57±3.34	82.49±13.35	99.40±0.99	86.73±12.96
kPCA RBF + FDA (unsupervised PC selection)	91.60±1.50	98.97±1.75	85.01±11.00	99.84±0.32	89.90±9.64
kPCA RBF + FDA (supervised PC selection)	100.00±0.00	100.00±0.00	**28.71±10.02** (−)	100.00±0.00	**50.00±0.00** (−)

Van 't Veer et al., 2002	LOO-CV performance	ACC training set	ACC test set	AUC training set	AUC test set
LS-SVM linear kernel	68.99±4.22	100.00±0.00	**67.92±8.58** (+)	100.00±0.00	**73.30±11.01** (+)
LS-SVM RBF kernel	69.05±3.55	100.00±0.00	**68.42±7.62** (+)	100.00±0.00	**73.98±10.69** (+)
LS-SVM linear kernel (no regularization)	52.14±6.04	74.66±24.04	**57.14±9.08** (−)	74.73±25.26	**64.60±13.18** (−)
PCA + FDA (unsupervised PC selection)	71.31±3.57	91.27±10.04	57.39±15.57	94.61±6.80	65.16±12.30
PCA + FDA (supervised PC selection)	73.44±3.19	97.31±5.62	**66.92±9.90** (+)	98.77±3.16	67.91±12.64
kPCA lin + FDA (unsupervised PC selection)	71.18±3.62	91.21±10.33	60.90±14.49	94.46±7.22	66.01±13.45
kPCA lin + FDA (supervised PC selection)	73.63±3.89	97.13±6.63	**65.41±7.54** (+)	98.54±3.98	69.22±11.01
kPCA RBF + FDA (unsupervised PC selection)	74.91±6.54	90.66±11.08	51.38±15.91	93.77±8.75	60.26±16.57
kPCA RBF + FDA (supervised PC selection)	100.00±0.00	100.00±0.00	**36.84±0.00** (−)	100.00±0.00	**50.00±0.00** (−)

Prostate cancer data set (Singh *et al.*, 2002)

The test set performances show that the experiment using kernel PCA with RBF kernel and selecting the principal components by means of the supervised method clearly gives bad results. Using the eigenvalues for selection of the principal components seems to give better results than using the supervised method. According to the test set accuracy, the experiment applying LS-SVM with RBF kernel even performs slightly better than those experiments using the eigenvalues for selection of the principal components. When looking at the area under the ROC curve of the test set, both experiments applying LS-SVM perform slightly better than those experiments using the eigenvalues for selection of the principal components.

Breast cancer data set (van 't Veer *et al.*, 2002)

When looking at the test set performances, it is obvious that the experiment using kernel PCA with RBF kernel and selecting the principal components by means of the supervised method leads to bad results. Using LS-SVM gives better results than performing dimensionality reduction combined with an unsupervised way for the selection of the principal components. According to the area under the ROC curve of the test set, using LS-SVM gives better results than all experiments performing dimensionality reduction. Both methods for selecting the principal components seem to perform similarly, but in some cases using the absolute value of the Golub score tends to perform slightly better.

3.4 Discussion

In this section, a discussion is given on the three main conclusions that can be derived from this study. The first conclusion can be formulated when assessing the role of nonlinearity for the case without dimensionality reduction, the second can be derived by studying the importance of generalization, and the third conclusion can be inferred by assessing the role of nonlinearity for the case with dimensionality reduction.

3.4.1 Assessing the role of nonlinearity for the case without dimensionality reduction

When considering only the experiments without dimensionality reduction (i.e., LS-SVM with linear kernel and LS-SVM with RBF kernel) using a well-tuned RBF kernel never resulted in overfitting on all tried data sets. The test set performances obtained when using an RBF kernel often

appear to be similar to those obtained when using a linear kernel, but in some cases an RBF kernel ends up in even better classification performances. This is illustrated in Figure 3.3. The fact that using LS-SVM with an RBF kernel does not result in overfitting, even for simple classification problems, can be explained by looking to the optimal values of the kernel parameter. When optimizing the kernel parameter of the RBF kernel for such a problem, the obtained value seems to be large. Using an RBF kernel with the kernel parameter σ set to infinity corresponds to using a linear kernel, aside from a scale factor (Suykens *et al.*, 2002). Until now, most microarray data sets are quite small and they may represent quite easily separable classification problems. It can be expected that those data sets will become larger or perhaps represent more complex classification problems in the future. In this case, the use of nonlinear kernels as the commonly used RBF kernel becomes important. Considering this, it may be useful to explore the effect of using other kernel functions.

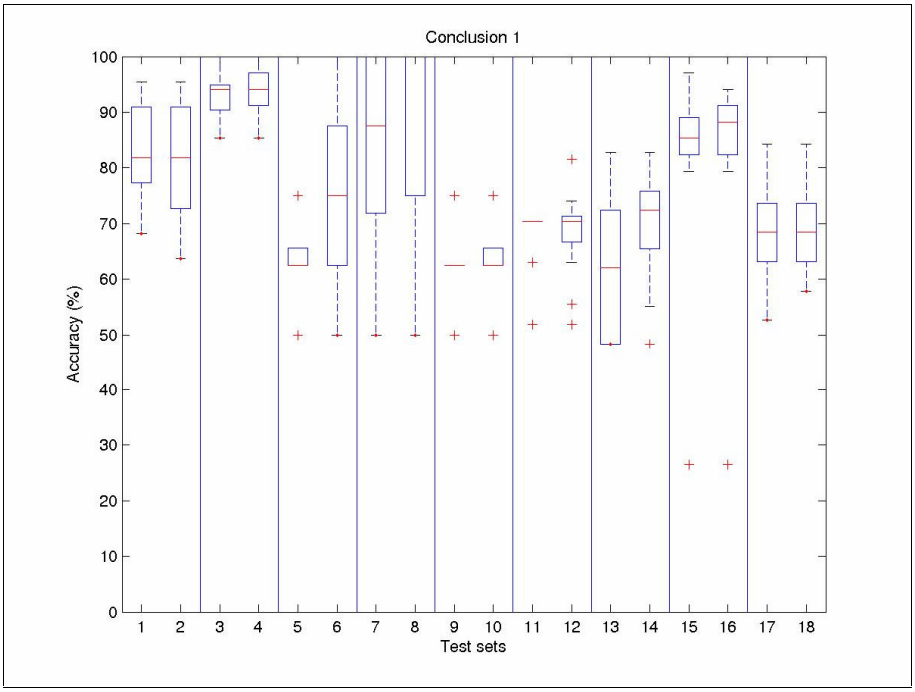


Figure 3.3 : *Illustration of the first main conclusion based on boxplots (boxplot in MATLAB) of the test set accuracies of all nine binary cancer classification problems: When performing classification with LS-SVM (without dimensionality reduction),* ***using well-tuned RBF kernels can be applied without risking overfitting****. The results obtained with well-tuned RBF kernels are never worse and sometimes even statistically significantly better compared with using a linear kernel. For each data set, the test set accuracies of LS-SVM with a linear kernel are represented on the left, the test set accuracies of LS-SVM with an RBF kernel on the right. Concerning the data sets, the order of Table 3.1 is respected.*

When comparing the experiments with and without dimensionality reduction, an important issue is that LS-SVM with RBF kernel (experiment without dimensionality reduction) never performs worse than all other methods.

3.4.2 The importance of regularization

When looking at the experiment using LS-SVM with linear kernel and the regularization parameter γ set to infinity (i.e., without regularization), the following issue can be seen. Using LS-SVM without regularization corresponds to FDA (Suykens *et al.*, 2002). Figure 3.4 shows that this experiment hardly performs better than random classification on all data sets, except on the acute leukemia data set of (Golub *et al.*, 1999), which represents an easily separable classification problem. Regularization appears to be important when applying classification methods onto microarray data without doing a dimensionality reduction step first.

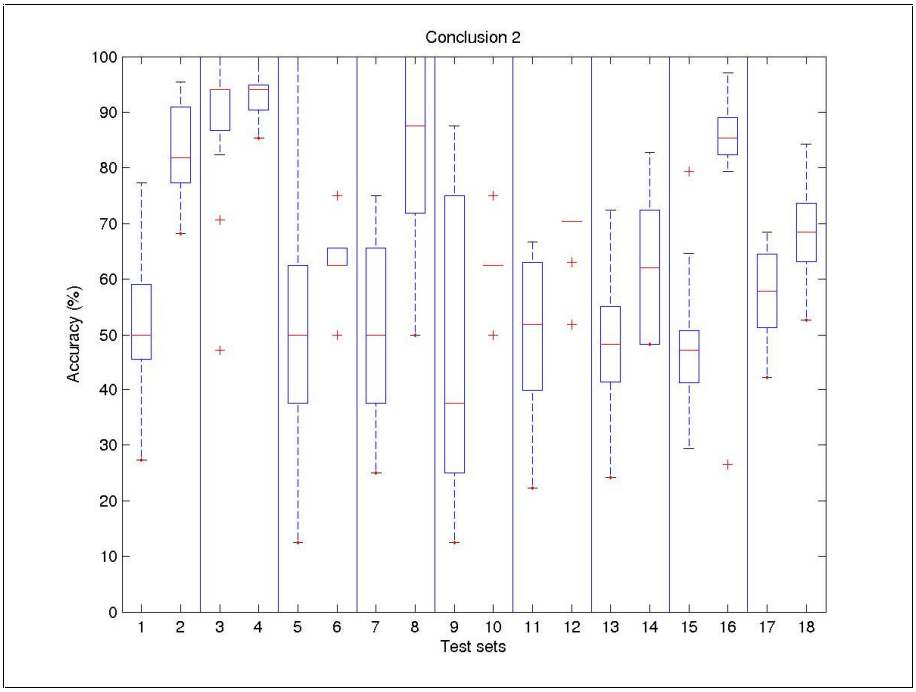


Figure 3.4 : *Illustration of the second main conclusion based on boxplots of the test set accuracies of all nine cancer classification problems: Even for classification with linear classifiers like LS-SVM with linear kernel,* ***performing regularization is important****. For each data set, the test set accuracies of LS-SVM with a linear kernel without regularization are represented on the left, the test set accuracies of LS-SVM with a linear kernel with regularization on the right. The latter shows much better performance. Concerning the data sets, the order of Table 3.1 is respected.*

3.4.3 Assessing the role of nonlinearity in case of dimensionality reduction

When considering only the experiments using dimensionality reduction, another important issue becomes clear. Comparing the results of using an RBF kernel with those of using a linear kernel when applying kernel PCA before classification, reveals that using an RBF kernel easily results in overfitting. This is represented by Figure 3.5. The best results are obtained by simply using a linear kernel when doing kernel PCA, which are similar to those when using classical PCA. (Gupta *et al.*, 2002) states a similar conclusion for face recognition based on image data. When comparing both methods for selection of the principal components, namely the unsupervised way based on the eigenvalues with the supervised way based on the absolute value of the score introduced by (Golub *et al.*, 1999),

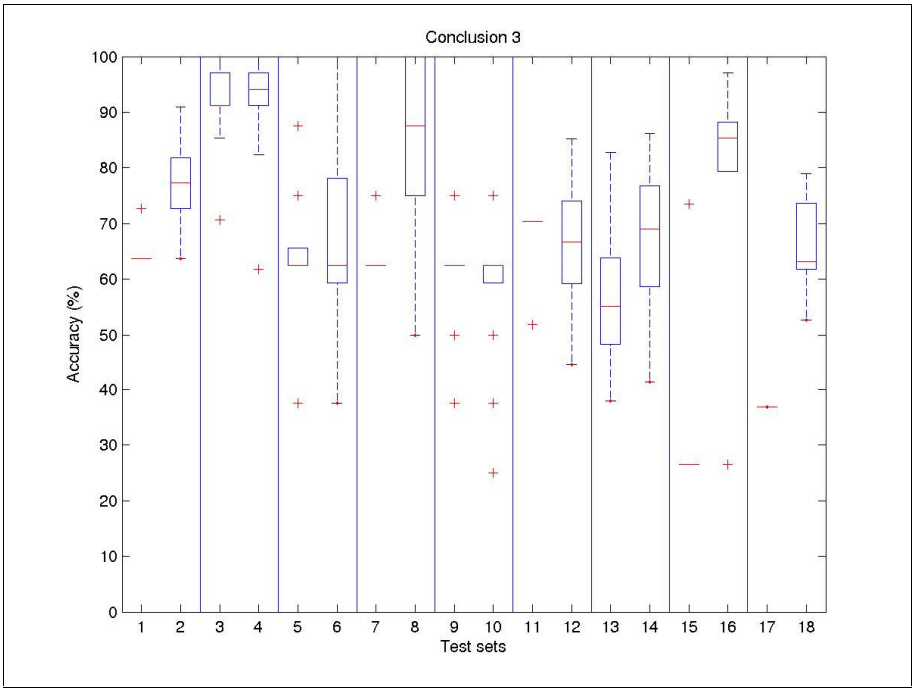


Figure 3.5 : *Illustration of the third main conclusion based on boxplots of the test set accuracies of all nine cancer classification problems: When performing kernel PCA before classification,* ***using an RBF kernel for kernel PCA tends to result in overfitting***. ***Kernel PCA with linear kernel gives better results***. *For each data set, the test set accuracies of kernel PCA with an RBF kernel (selecting the principal components in a supervised way) followed by FDA are represented on the left, the test set accuracies of kernel PCA with a linear kernel (selecting the principal components in a supervised way) followed by FDA on the right. Concerning the data sets, the order of Table 3.1 is respected.*

no general conclusions can be made. It depends on the data set whether one method is better than the other or not. The combination of using kernel PCA with RBF kernel and supervised selection of the principal components tends to result in overfitting. All this can be explained by ignoring relevant principal components (Bishop, 1995) since a linear selection of features followed by building nonlinear classifiers may be not ideal.

In the context of feature selection, some interesting issues become clear when studying the ROC performance and the sparseness of the classical and kernel PCA projections. When comparing the results of using a linear kernel with those of using an RBF kernel for kernel PCA when selection of the principal components is done in a supervised way as shown in Figure 3.2, two situations indicating overfitting can be recognized. First, overfitting occurs in case of using kernel PCA with an RBF kernel followed by supervised selection of principal components when the optimal number of principal components is relatively low in case of using a linear kernel for kernel PCA and much larger in case of using an RBF kernel. Second, overfitting also occurs when using kernel PCA with both a linear and an RBF kernel followed by supervised selection of principal components when the optimal number of principal components is large both when using a linear kernel for kernel PCA and when using an RBF kernel.

When comparing the experiments with and without dimensionality reduction, also worth mentioning is the fact that performing dimensionality reduction requires optimization of the number of principal components. This parameter, belonging to the unsupervised PCA, needs to be optimized in the sense of the subsequent supervised FDA (see outline of the optimization algorithm in the section on numerical experiments). In practice, this appears to be quite time consuming, especially in combination with other parameters that need to be optimized (e.g. kernel parameter of kernel PCA with RBF kernel).

3.5 Conclusion

In the past, using classification methods in combination with microarrays has shown to be promising for guiding clinical management in oncology. In this study, several important issues have been formulated to optimize the performance of clinical predictions based on microarray data. Those formulations are based on nonlinear techniques and dimensionality reduction methods as well as on regularization techniques, taking into consideration the probability of increasing size and complexity of microarray data sets in the future.

A first important conclusion from benchmarking nine microarray data set problems is that when performing classification with LS-SVM

(without dimensionality reduction), using an RBF kernel can be applied without risking overfitting on all data sets studied. The results obtained with an RBF kernel are never worse and sometimes even better than when using a linear kernel. A second conclusion is that using LS-SVM without regularization (without dimensionality reduction) gives bad results, which stresses the importance of applying regularization even in the linear case. A final important conclusion is that when performing kernel PCA before classification, using an RBF kernel for kernel PCA tends to lead to overfitting, especially when using supervised feature selection. It has been observed that an optimal selection of a large number of features is often an indication for overfitting. Kernel PCA with linear kernel gives better results.

Nevertheless, although it was possible to derive some important general conclusions out of this study, the best classification method to build the most optimal prediction model may differ for each cancer classification problem. Since it is obvious that building an optimal prediction model is of major importance with respect to using such models in clinical practice in the future, finding the best classification method in each specific case is an indispensable issue. Concluding with the idea that it remains essential to carefully consider each cancer classification problem individually, we can proceed to the next chapter.

Chapter 4

M@CBETH web service: microarray classification tool

4.1 Introduction

In the previous chapter, we concluded with the observation that the best classification method to build the most optimal prediction model may differ for classifying each cancer microarray data set. Therefore, it is essential to develop the best classifier for each microarray data set on an individual basis. This not only includes finding the best classification method for each data set, but also fine-tuning of all parameters (e.g., regularization parameter, kernel parameter, number of principal components), which is important in the model building process. However, exploring all combinations to find the most optimal classifier can be complicated. In this context, we first state the relevant issues for this chapter. In the next section, a web service is developed as a solution to the remaining unsolved issues[1].

As mentioned before, using microarray data allows making predictions on, for example, therapy response, prognosis, and metastatic phenotype of an individual patient. Microarray technology is useful in supporting clinical management decisions for individual patients (e.g., breast cancer (van 't Veer *et al*., 2002), acute myeloid leukemia (Valk *et al*., 2004), and ovarian cancer (De Smet *et al*., 2006a)) in combination with classification methods (Furey *et al*., 2000). Finding the best classifier for each data set can be a tedious and non-straightforward task for users not familiar with these classification techniques. In this chapter, a web service is presented that compares, for each microarray data set introduced to this

[1] The web service M@CBETH that is presented in this chapter has been published in the journal *Bioinformatics* (Pochet *et al.*, 2005).

service, different classifiers and selects the best in terms of randomized independent test set performances.

Systematic benchmarking of microarray data classification revealed that either regularization or dimensionality reduction is required to obtain good independent test set performances, as discussed in Chapter 3. Regularization - as is performed in SVM (Cristianini and Shawe-Taylor, 2000) - already led to the Gist web service, which offers SVM classification on the web (Pavlidis *et al*., 2004). The web service presented in this chapter allows comparing different classification methods. By exploring different combinations of nonlinearity and dimensionality reduction, the benchmarking study presented in the previous chapter showed that the optimal classifier can differ for each data set. Also important, but often underestimated in the model building process, is the fine-tuning of all parameters (e.g. regularization parameter, kernel parameter and number of principal components). Exploring all combinations to find the optimal classifier for each data set can be complicated. Therefore, we intend to design a web service to offer the microarray community a simple tool for making optimal predictions.

In this chapter, we will incorporate the methods used in the benchmarking study in the previous chapter into an interface called M@CBETH (a MicroArray Classification BEnchmarking Tool on a Host server) that is freely available and can easily be used by clinicians for making optimal two-class predictions. We will clarify how this web service finds the best prediction among different classification methods by using randomizations of a benchmarking dataset.

4.2 M@CBETH web site

The M@CBETH web site can be found on http://www.esat.kuleuven.be/MACBETH/ and it offers two services: benchmarking and prediction. After registration and logging on to the web service, users can request benchmarking or prediction analyses. Users are notified by email about the status of their analyses running on the host server. They can also check this on the analysis results page, which gives an overview of all analyses and contains links to corresponding results pages.

Benchmarking, the main service on the M@CBETH web site, involves selection and training of an optimal model based on the submitted benchmarking data set and corresponding class labels. This model is then stored for immediate or later use on prospective data. Benchmarking results in a table showing summary statistics [leave-one-out cross-validation (LOO-CV), training set accuracy (ACC) and area under the receiver operating characteristic curve (AUC) performance, test set ACC and AUC] for all

selected classification methods, highlighting the best method. Prospective data can also be submitted and evaluated immediately during the same benchmarking analysis.

By using the prediction service, the M@CBETH web site offers a way for later evaluation of prospective data by reusing an existing optimal prediction model (built up in a previous benchmarking analysis by the same user). For both services, if the corresponding prospective labels are submitted, the prospective accuracy is calculated. Otherwise, labels are predicted for all prospective samples. This latter application is useful for classifying new unseen patients in clinical practice.

The M@CBETH web service is intended for the classification of patient samples, supposing microarray data are represented by an expression matrix characterized by high dimensionality in the sense of a small number of patients and a large number of gene expression levels for each patient. Two kinds of data formats are accepted: spreadsheet-like tab-delimited text files and matrix-like MATLAB files. Data sets are not allowed to contain missing values. Class labels are restricted to '+1' or '−1'. Several publicly available microarray data sets are present on the web site in correct data format as examples.

Users can select the classification methods that will be compared (default selection set to the best overall and most efficient methods from the benchmarking study), change the number of randomizations (default 20, while keeping in mind that results are more reliable when the number of randomizations is large) and switch off normalization.

4.3 Algorithm

An overview of the algorithm behind this web service is presented in Figure 4.1. The algorithm is as follows. The benchmarking data set is reshuffled until the number of requested randomizations is reached. All randomizations are split (two-third of the samples for training, the rest as test set) in a stratified way (class labels are equally distributed over the training–test split). Iteratively, all selected classification methods are applied to all randomizations. In each iteration, selection of the parameters is first performed by means of LOO-CV, then the model is trained based on the training set and finally, this model is then applied onto the test set resulting in a test set ACC. The mean randomized test set ACC is calculated for each classification method. The best generalizing method - with best test set ACC - is then used for building the optimal classifier onto the complete benchmarking data set, which is stored for application onto prospective data sets.

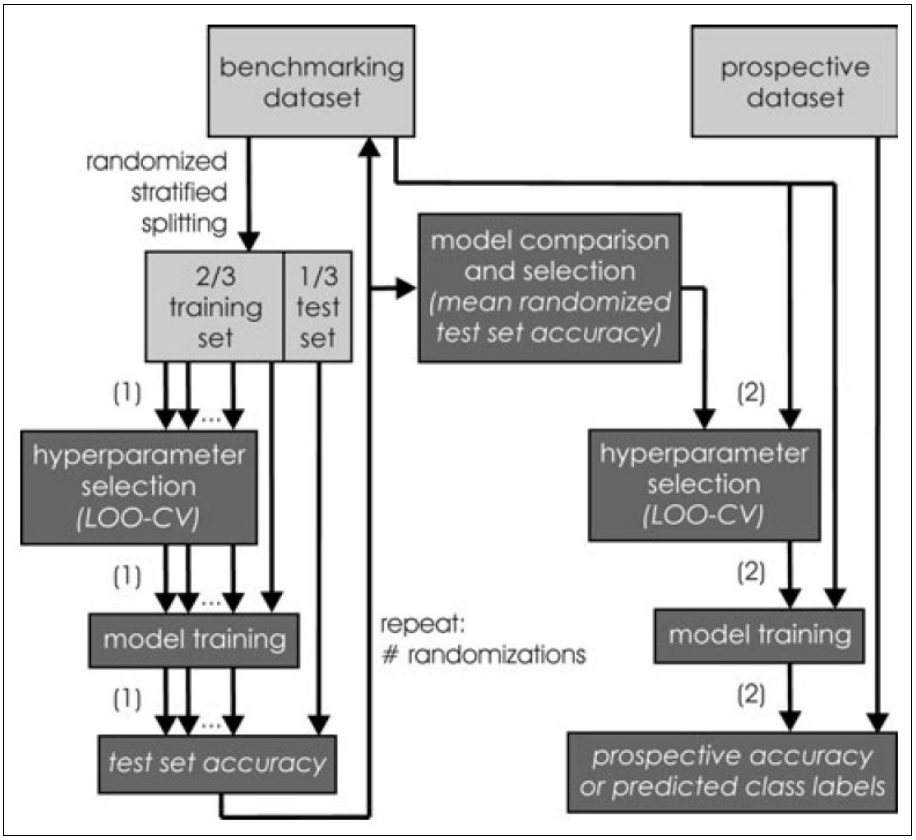


Figure 4.1 : *Overview of the algorithm behind the M@CBETH web service. The benchmarking data set is reshuffled until the number of requested randomizations is reached. Iteratively, all selected classification methods (1) are applied to all randomizations. In each iteration, selection of the parameters is first performed by means of LOO-CV, then the model is trained based on the training set and finally, this model is then applied onto the test set resulting in a test set ACC. The mean randomized test set ACC is calculated for each classification method. The best generalizing method (2) - with best test set ACC - is then used for building the optimal classifier onto the complete benchmarking data set, which is stored for application onto prospective data sets.*

Nine different classification methods - based on LS-SVM (with linear and RBF kernels), FDA, PCA and kernel PCA (with linear and RBF kernels) - are considered:

1. LS-SVM with linear kernel;
2. LS-SVM with RBF kernel;
3. LS-SVM with linear kernel without regularization;
4. PCA (unsupervised selection of principal components) followed by FDA;

5. PCA (supervised selection of principal components) followed by FDA;
6. Kernel PCA with linear kernel (unsupervised selection of principal components) followed by FDA;
7. Kernel PCA with linear kernel (supervised selection of principal components) followed by FDA;
8. Kernel PCA with RBF kernel (unsupervised selection of principal components) followed by FDA;
9. Kernel PCA with RBF kernel (supervised selection of principal components) followed by FDA.

More detailed information on these methods can be found in the previous two chapters.

4.4 Practical issues

In this section, more detailed information is given on how to use this M@CBETH web service. As previously mentioned, users can request benchmarking or prediction analyses. They are notified by email about the status of their analyses running on the host server and they can check the status and the results of their analyses on the analysis results page.

To guarantee an efficient use of this web service, a maximum of 5 benchmarking analyses is allowed to run simultaneously on the host server, the rest will be scheduled. Since prediction analyses are fast compared to benchmarking analyses, they are always allowed to start immediately.

4.4.1 Benchmarking service

Remember that benchmarking, the main service on the M@CBETH web site, involves selection and training of an optimal model based on the submitted benchmarking data set and corresponding class labels. This model is then stored for immediate or later use on prospective data. Using the benchmarking service results in a table showing summary statistics (leave-one-out cross-validation (LOO-CV), training set accuracy (ACC) and area under the Receiver Operating Characteristic curve performance (AUC), test set ACC and AUC) for all selected classification methods, highlighting the best method in red. Prospective data can be submitted and evaluated immediately during the same benchmarking analysis. If the corresponding prospective labels are submitted, the prospective accuracy is calculated. Otherwise, labels are predicted for all prospective samples. This latter application is useful for classifying new unseen patients in clinical practice.

Data set and class label files

As stated before, the web service is intended for classification of patient samples, supposing microarray data is represented by an expression matrix characterized by high dimensionality in the sense of a small number of patients and a large number of gene expression levels for each patient. Two kinds of data formats are accepted. First, spreadsheet-like tab-delimited (comma or space-delimited is also possible) text files (see Figures 4.2 and 4.3) with extension '.txt' are allowed. Furthermore, also matrix-like MATLAB files (see Figures 4.4 and 4.5) with extension '.mat' are accepted. More specific information on the data format:

- Data sets are not allowed to contain missing values.
- Class labels are restricted to '+1' (or just '1') and '-1'.
- All data must be numeric data (numbers may contain points, but no commas). All gene expression and sample descriptors must therefore be removed.
- The number of gene expression levels in a particular prediction data set must be the same as the number of gene expression levels in the corresponding benchmarking data sets.
- The number of samples in a particular data set must be the same as the number of corresponding class labels.
- The number of gene expression levels is supposed to be much larger than the number of samples. This allows that rows as well as columns are allowed to represent gene expression levels as well as samples.
- The size of the files is limited to 40 Mbytes in order to secure the server.

-214	-139	-76	-135	-106	-138	-72	-413	5	-88
-153	-73	-49	-114	-125	-85	-144	-260	-127	-105
-58	-1	-307	265	-76	215	238	7	106	42
88	283	309	12	168	71	55	-2	268	219
-295	-264	-376	-419	-230	-272	-399	-541	-210	-178
-558	-400	-650	-585	-284	-558	-551	-790	-535	-246
199	-330	33	158	4	67	131	-275	0	328
-176	-168	-367	-253	-122	-186	-179	-463	-174	-148
252	101	206	49	70	87	126	70	24	177
206	74	-215	31	252	193	-20	-169	506	183

Figure 4.2 : *Illustration of an extract of spreadsheet-like tab-delimited text data set file that can be submitted to M@CBETH.*

-1	-1	-1	-1	-1	-1	1	1	1	1

Figure 4.3 : *Illustration of an extract of spreadsheet-like tab-delimited text class labels file that can be submitted to M@CBETH.*

8589.4	9164.3	3825.7	6246.4	3230.3	2510.3	7126.6	4028.7	9330.7	5271.5
5468.2	6719.5	6970.4	7823.5	3694.4	1960.7	3779.1	3156.2	7017.2	4740.8
4263.4	4883.4	5370	5955.8	3400.7	1566.3	3705.6	2870.3	4723.8	3318.5
4064.9	3718.2	4705.6	3975.6	3463.6	3072.8	6594.5	4417.6	9491.5	6792.3
1997.9	2015.2	1166.6	2002.6	2181.4	1810.2	2460.9	1854.1	5346.5	2632.9
5282.3	5569.9	1572.2	2130.5	2922.8	1673.6	3775.7	2828.3	1557.1	5449.2
2169.7	3849.1	1325.4	1531.1	2069.2	1290.4	2621.4	1427.5	1969.1	4623.2
2773.4	2793.4	1472.3	1714.6	2948.6	2465.8	2047.3	3390.7	2295.4	3277.4
7526.4	7017.7	3297	3869.8	3303.4	1675.5	6411.3	4373	6880.3	4488.1
4607.7	4802.3	2786.6	4989.4	3109.4	1312.8	3857.1	3080.5	6162.9	3343.8

Figure 4.4 : *Illustration of an extract of matrix-like MATLAB data set file that can be submitted to M@CBETH.*

-1	1	-1	1	-1	1	-1	1	-1	1

Figure 4.5 : *Illustration of an extract of matrix-like MATLAB class labels file that can be submitted to M@CBETH.*

Several publicly available microarray data sets (preprocessing and missing value estimation as proposed in the original publications) are present on the example data page (see Figure 4.6) in correct data format as an example. Remark that users are responsible for preprocessing and missing value estimation of their own data before submitting it to the M@CBETH web service.

Classification methods

Users can select the classification methods that will be compared. The default selection is set to the best overall and most efficient methods from the benchmarking study. This includes (1) LS-SVM with linear kernel, (2) LS-SVM with RBF kernel, (6) kernel PCA with linear kernel (unsupervised selection of principal components) followed by FDA, and (7) kernel PCA with linear kernel (supervised selection of principal components) followed by FDA. Note that PCA and kernel PCA are based on centered expression and kernel matrices respectively. More detailed information on these methods can be found in the previous chapter.

Figure 4.6 : *Download page containing several publicly available microarray data sets in correct data format as an example.*

Randomizations

It is possible to change the number of randomizations. The default value is 20. Users should keep in mind that results are more reliable when the number of randomizations is large (preferably at least 20). The maximum number of randomizations allowed is 100 to limit the time for computation.

To get an idea how long a specific analysis will take on your data set, it is recommended to first perform the analysis you prefer with zero randomizations (this means that only a first data set split is used). The time needed for this relatively short analysis can then be multiplied by the number of requested randomizations to have an idea of the duration of the complete analysis. Of course, because of other users using the web service, this is only a rough estimation.

Normalization

Users can switch off normalization, although performing normalization is better from a statistical viewpoint. Normalization is done by standardizing each gene expression of the data to have zero mean and unit standard deviation. Normalization of training sets as well as test sets is done by using the mean and standard deviation of each gene expression profile of the training sets.

4.4.2 Prediction service

Via the prediction service, users can evaluate prospective data in a later stage by reusing an existing optimal prediction model built in one of their previous benchmarking analyses. If the user submits the corresponding prospective labels, the prospective accuracy is calculated. Otherwise, labels are predicted for all prospective samples.

Data set and class label files

This was already discussed in Section 4.2.1.

Optimal prediction models

An existing optimal prediction model (built in a previous benchmarking analysis) needs to be selected for evaluation of prospective data. Remark that by viewing the results of an existing model, this model will be automatically selected for prediction.

4.5 Examples

The M@CBETH web service can be used for 5 different possible applications:

1. Benchmarking analysis;
2. Benchmarking analysis with immediate evaluation of prospective data (calculation of prospective accuracy);
3. Benchmarking analysis with immediate evaluation of prospective data (prediction of prospective samples);
4. Delayed prediction analysis of prospective data (calculation of prospective accuracy);
5. Delayed prediction analysis of prospective data (prediction of prospective samples).

This section shows an example of a simple benchmarking analysis, a benchmarking analysis with immediate evaluation of prospective data (prediction of prospective samples), and a delayed prediction analysis of prospective data (calculation of prospective accuracy).

4.5.1 Example 1: Benchmarking analysis

A benchmarking data set with corresponding class labels is submitted to the benchmarking service. All classification methods are selected. The number of randomizations is 20 and normalization is switched off. Figure 4.7 shows the completed benchmarking page based on the prostate cancer data of Singh *et al.* (2002).

The results page of this benchmarking analysis is presented in Figure 4.8. This consists of a table comparing all selected classification methods and highlighting the best in red.

4.5.2 Example 2: Benchmarking analysis with immediate evaluation of class labels for prospective data (prediction of prospective samples)

A benchmarking data set with corresponding class labels is submitted to the benchmarking service, as well as a prospective data set (without corresponding class labels). The default selection of classification methods is preserved. The number of randomizations is 20 and

normalization is switched off. Figure 4.9 shows the completed benchmarking page based on the high-grade glioma data of Nutt *et al.* (2003).

The results page of this benchmarking analysis is presented in Figure 4.10. This starts with a table comparing all selected classification methods and highlighting the best in red. This is followed by the evaluation of the best model onto the prospective data, given by the predicted class labels for all prospective samples.

4.5.3 Example 3: Prediction analysis for delayed evaluation of prospective data (calculation of prospective accuracy)

A prospective data set with corresponding class labels is submitted to the prediction service. An existing optimal model is selected. Figure 4.11 shows the completed prediction page based on the high-grade glioma data of Nutt *et al.* (2003).

The results page of this prediction analysis is presented in Figure 4.12. This consists of the evaluation of the selected optimal model onto the prospective data, given by the prospective accuracy and a description of all misclassified samples.

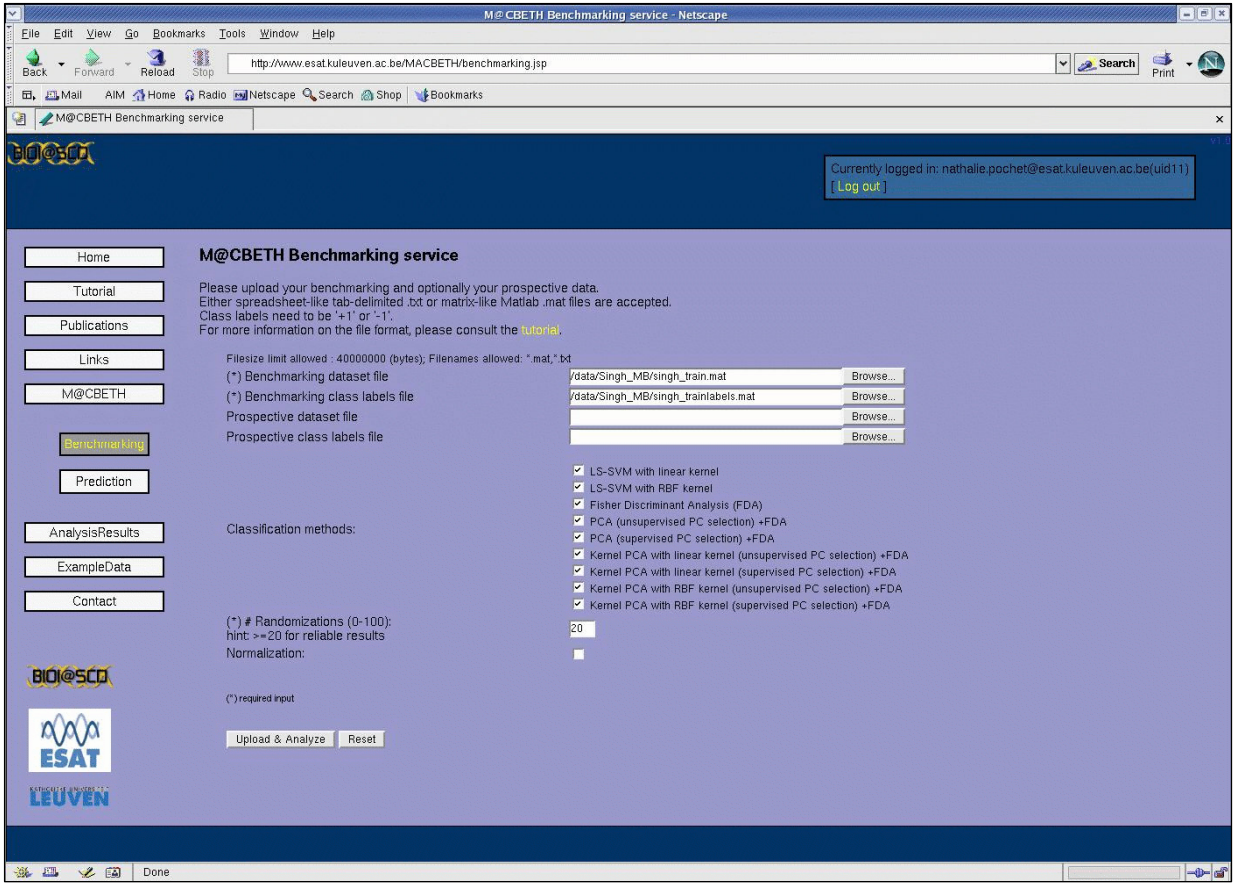


Figure 4.7 : *Completed benchmarking page for the performance of a benchmarking analysis. Benchmarking data and class labels are submitted to the benchmarking service, all classification methods are selected, the number of randomizations is set to 20 and normalization is switched off. Starting the benchmarking analysis will result in selection of the best classification method for this data set and building of the optimal model using this method. This model is then stored for delayed prospective analyses.*

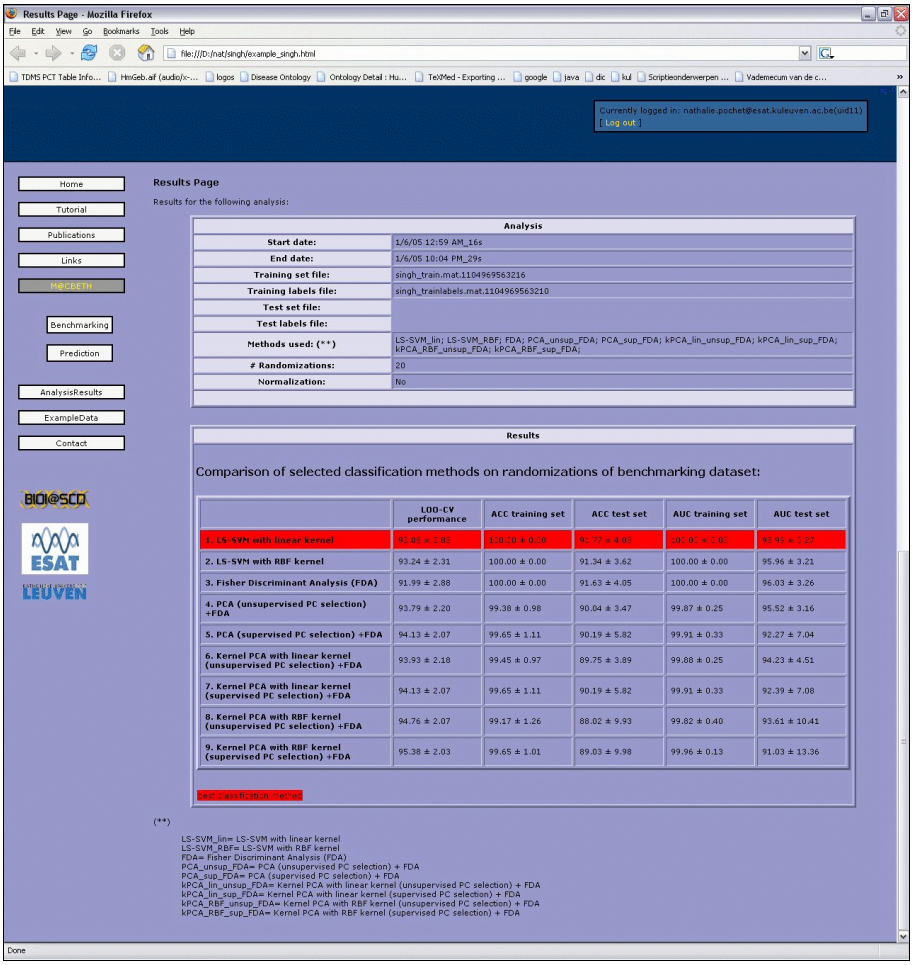


Figure 4.8 : *Results page for performance of a benchmarking analysis. The table shows the summary statistics for all selected classification methods, highlighting the best method in red.*

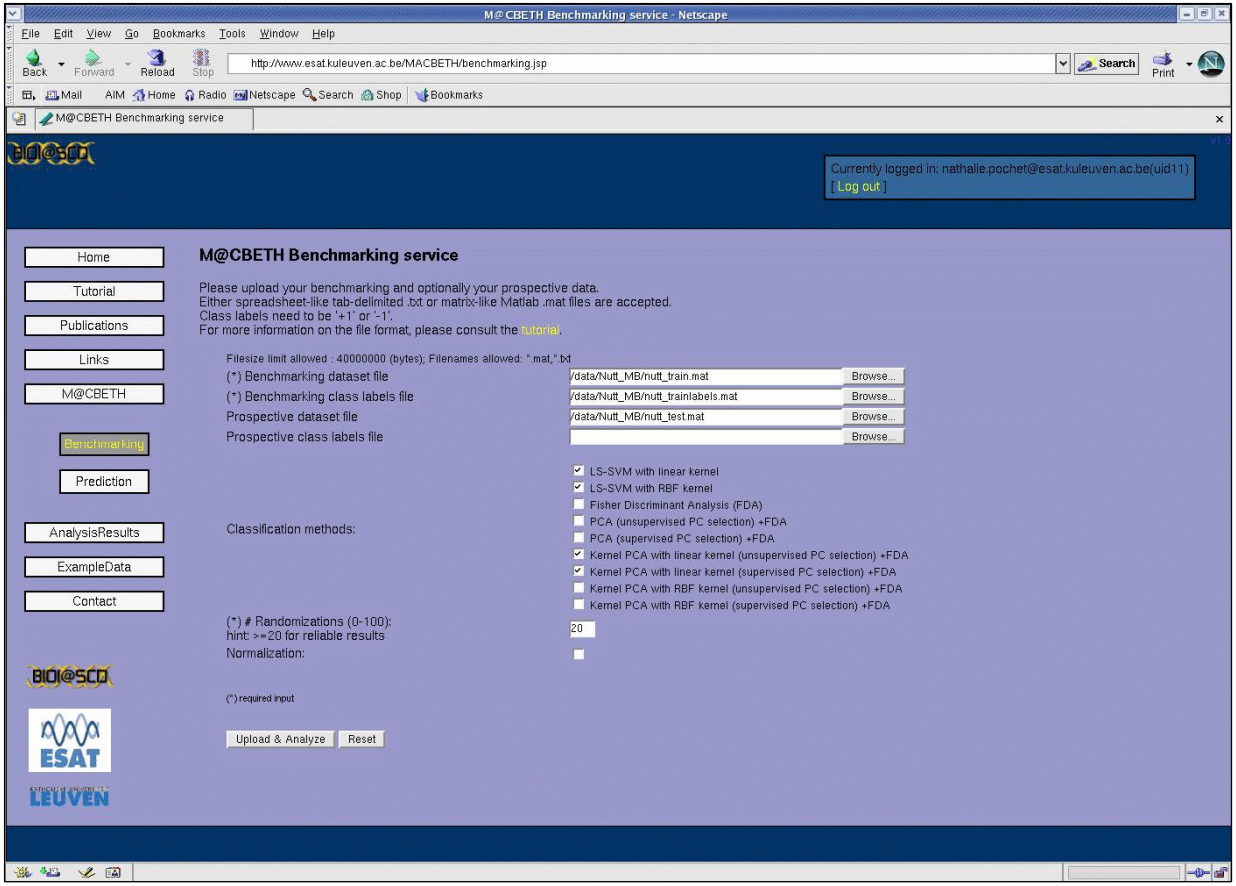


Figure 4.9 : *Completed benchmarking page for the performance of a benchmarking analysis with immediate evaluation of prospective data (prediction of prospective samples). In contrast to previous benchmarking analysis, prospective data is immediately submitted and the default selection of classification methods is preserved. In addition to the generation and storage of the optimal model, this also results in the prediction of the prospective class labels. Note that in case prospective class labels are also submitted, the prospective accuracy is calculated.*

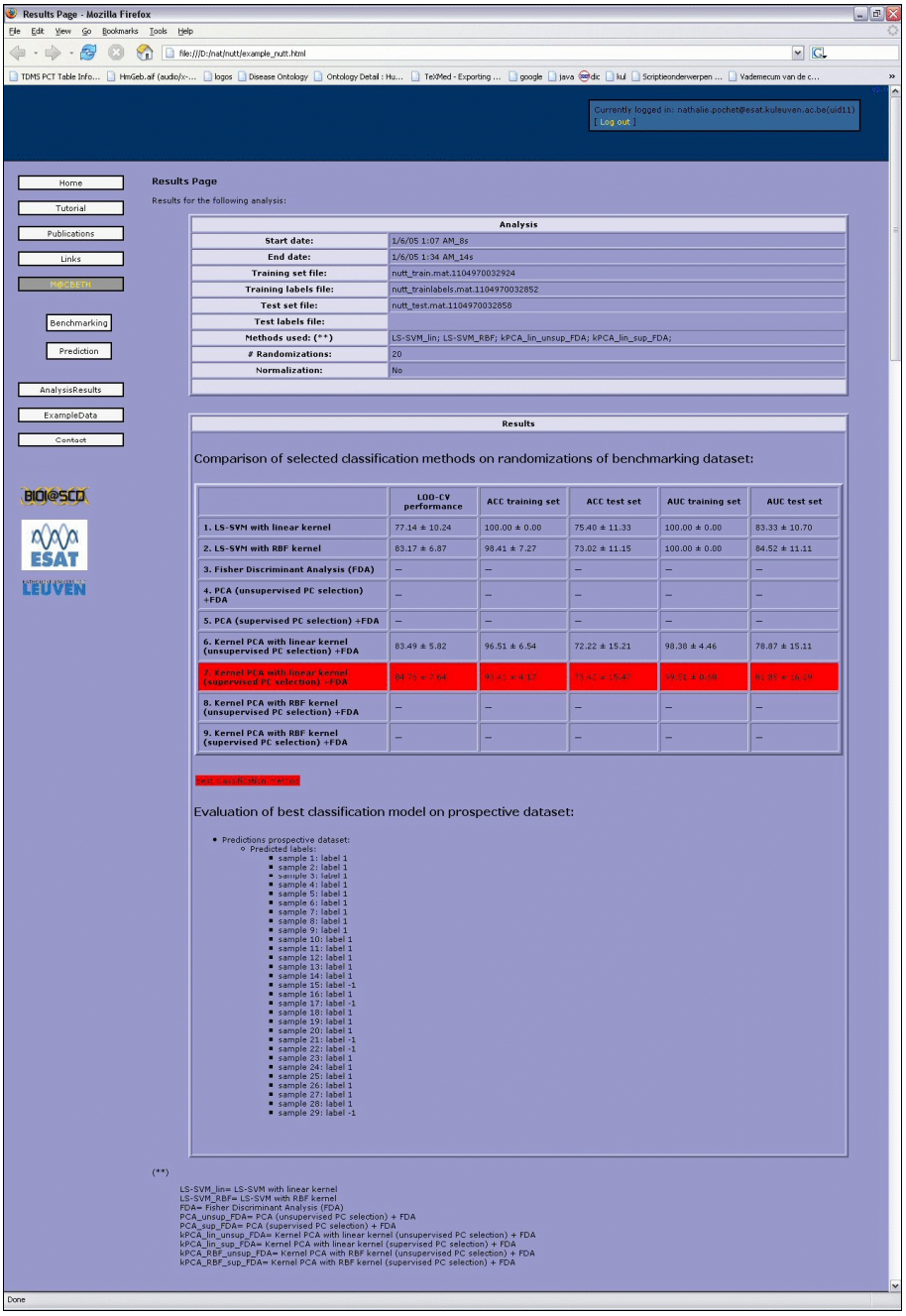


Figure 4.10 : *Results page for performance of a benchmarking analysis with immediate evaluation of prospective data (prediction of prospective samples). The table shows the summary statistics for all selected classification methods, highlighting the best method in red. This is followed by the prediction of the prospective class labels.*

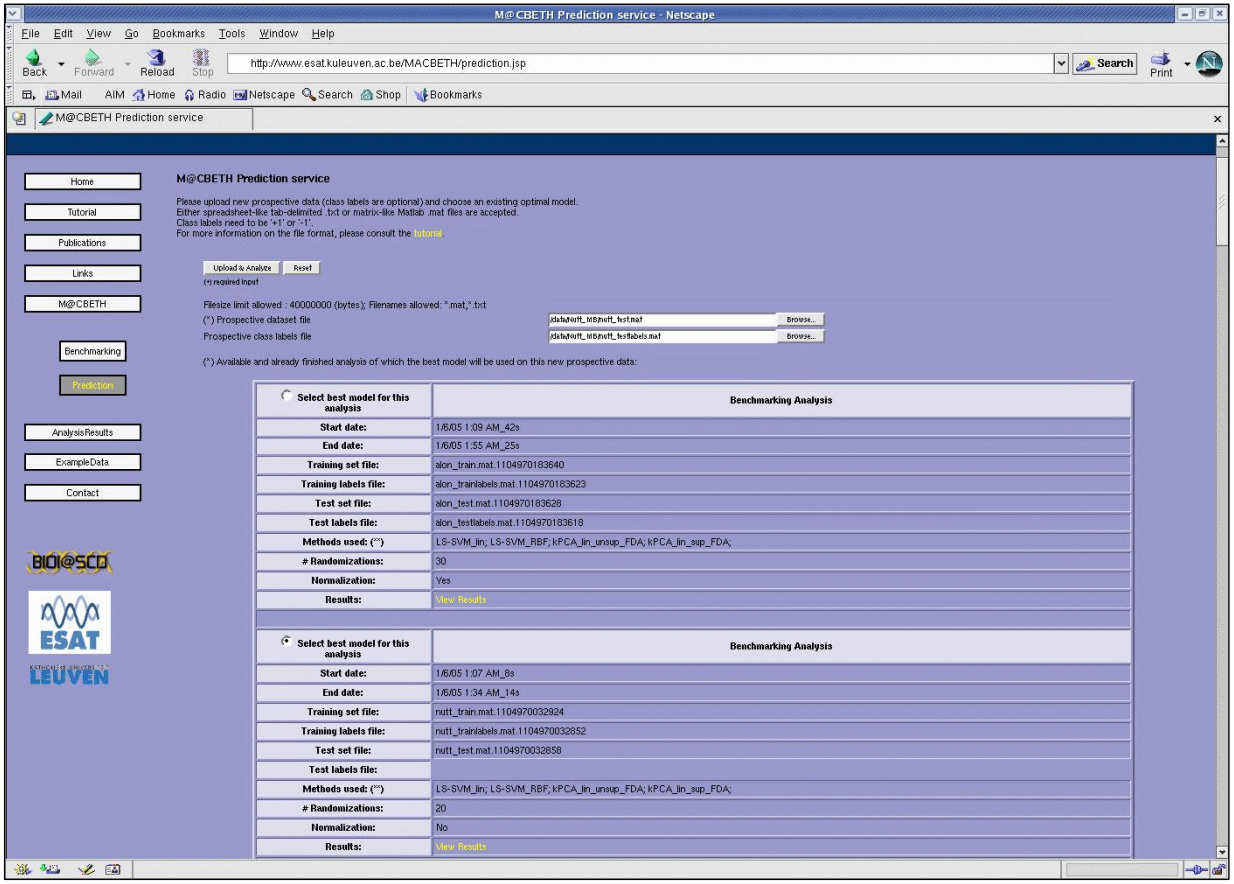


Figure 4.11: *Completed prediction page for the performance of a prediction analysis for delayed evaluation of prospective data (calculation of prospective accuracy). Prospective data and class labels are submitted to the prediction service and an existing optimal model for this data set is selected. This analysis will result in the calculation of the prospective accuracy. Note that in case prospective class labels were not submitted, this would result in the prediction of the prospective class labels.*

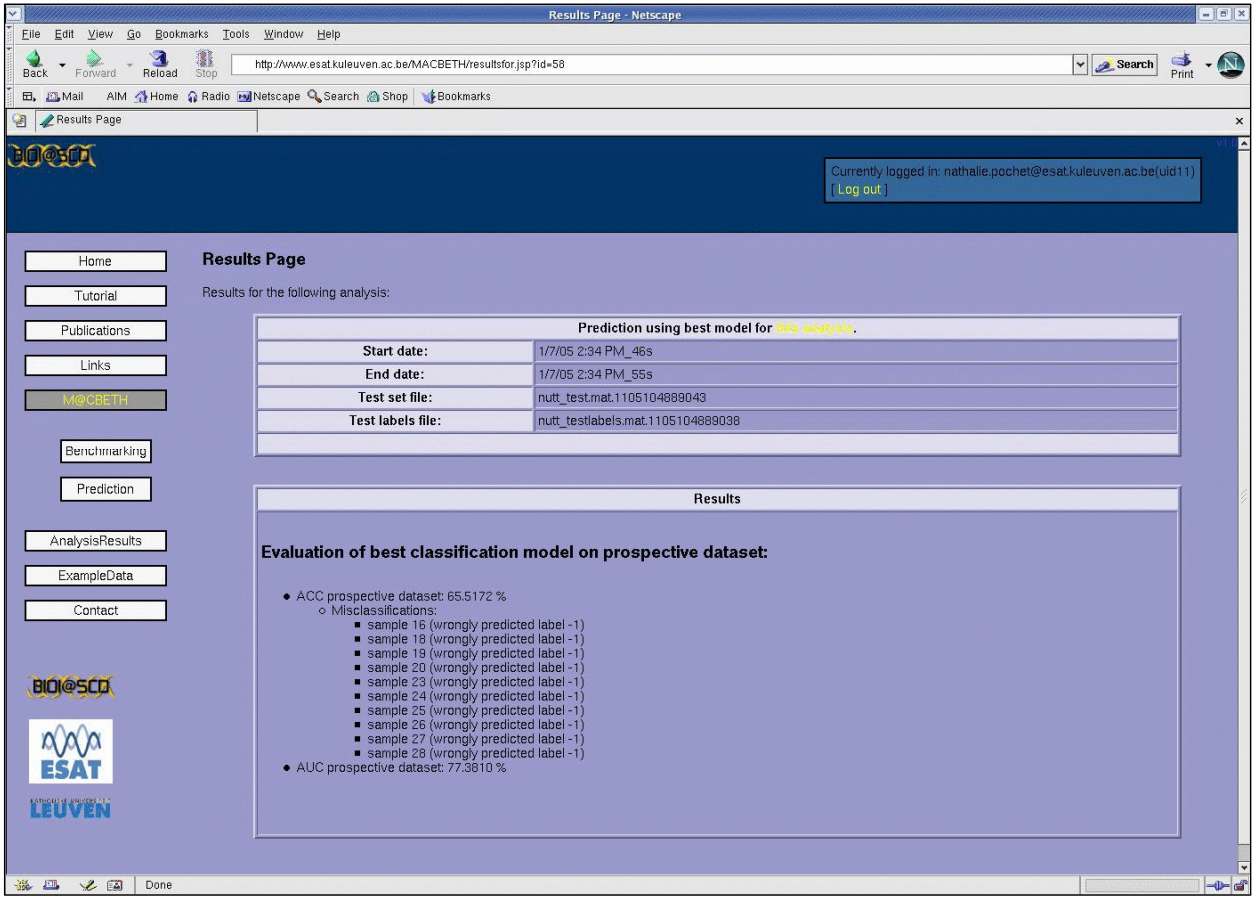


Figure 4.12 : *Results page for the performance of a prediction analysis for delayed evaluation of prospective data (calculation of prospective accuracy). Since both prospective data and class labels were submitted to M@CBETH, this analysis results in the calculation of the prospective accuracy, followed by a list of the wrongly predicted samples.*

4.6 International impact

Despite the recent introduction of M@CBETH, people from all over the world have already shown interest in this tool. Since the online publication of M@CBETH (hit statistics have been measured since the 24th of May, 2005 until the 15th of May, 2006), the web site has been visited 608 times, including 265 visitors from abroad (see Figures 4.17, 4.18, and 4.19). Apart from individual users, the web site was visited by three main groups of users: biomedical companies, research groups (universities), and medical centers and hospitals. Table 4.1 shows a selection of these visitors, spread out over these three groups.

Figure 4.17 : *Summary of the hit statistics of M@CBETH. The web site has already been visited 608 times, including 265 visitors from abroad, since measuring the hit statistics shortly after the online publication of M@CBETH.*

Page views per month - Oldest on top

May 2006	45
April 2006	42
March 2006	37
February 2006	26
January 2006	34
December 2005	24
November 2005	28
October 2005	36
September 2005	37
August 2005	57
July 2005	95
June 2005	131
May 2005	16
April 2005	0
March 2005	0
February 2005	0
January 2005	0
December 2004	0
November 2004	0
October 2004	0
September 2004	0
August 2004	0
July 2004	0
June 2004	0
May 2004	0
April 2004	0
March 2004	0
February 2004	0
Total	**608**

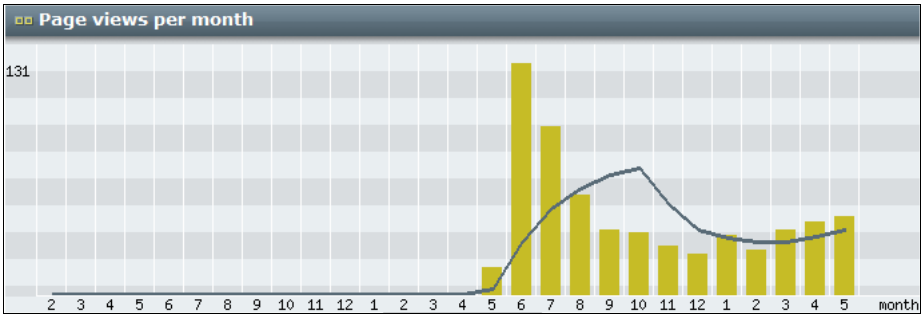
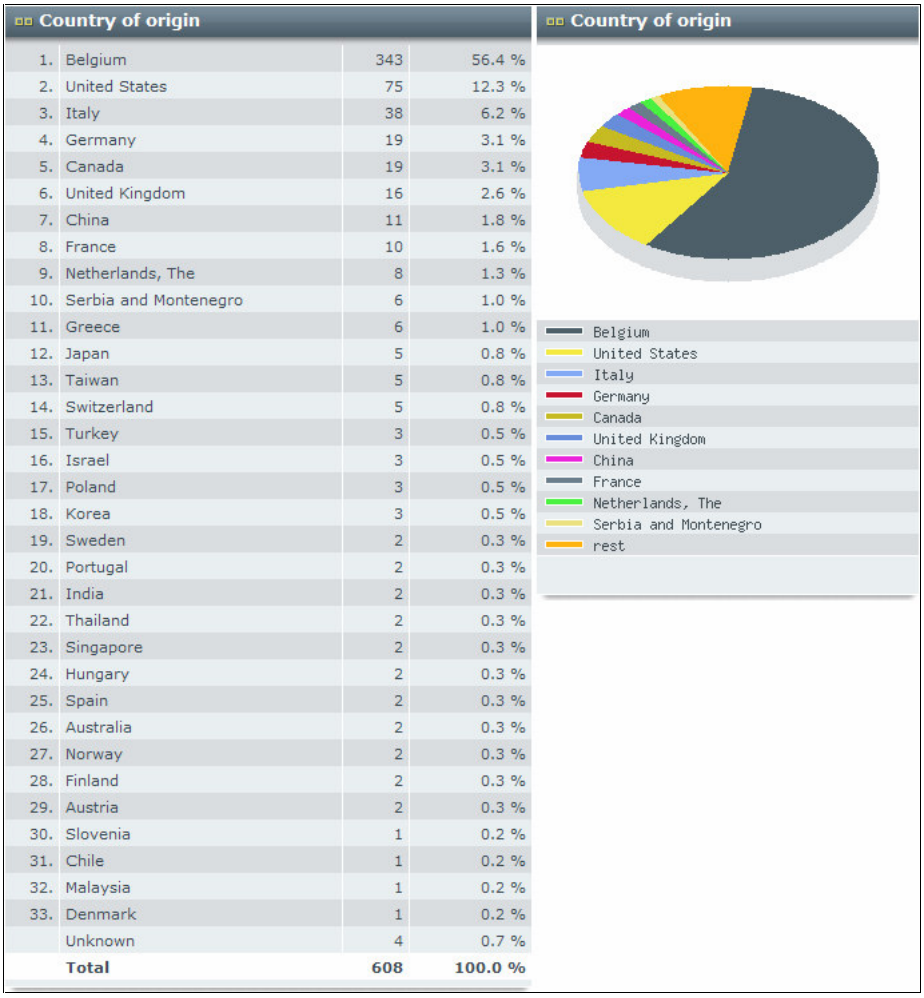


Figure 4.18 : *Page views per month for M@CBETH.*



Country of origin

1.	Belgium	343	56.4 %
2.	United States	75	12.3 %
3.	Italy	38	6.2 %
4.	Germany	19	3.1 %
5.	Canada	19	3.1 %
6.	United Kingdom	16	2.6 %
7.	China	11	1.8 %
8.	France	10	1.6 %
9.	Netherlands, The	8	1.3 %
10.	Serbia and Montenegro	6	1.0 %
11.	Greece	6	1.0 %
12.	Japan	5	0.8 %
13.	Taiwan	5	0.8 %
14.	Switzerland	5	0.8 %
15.	Turkey	3	0.5 %
16.	Israel	3	0.5 %
17.	Poland	3	0.5 %
18.	Korea	3	0.5 %
19.	Sweden	2	0.3 %
20.	Portugal	2	0.3 %
21.	India	2	0.3 %
22.	Thailand	2	0.3 %
23.	Singapore	2	0.3 %
24.	Hungary	2	0.3 %
25.	Spain	2	0.3 %
26.	Australia	2	0.3 %
27.	Norway	2	0.3 %
28.	Finland	2	0.3 %
29.	Austria	2	0.3 %
30.	Slovenia	1	0.2 %
31.	Chile	1	0.2 %
32.	Malaysia	1	0.2 %
33.	Denmark	1	0.2 %
	Unknown	4	0.7 %
	Total	**608**	**100.0 %**

Figure 4.19 : *Countries of origin of people visiting M@CBETH.*

Biomedical companies
Abbott Laboratories (United States)
AstraZeneca (United Kingdom)
Johnson & Johnson (United States)
Novartis AG (Basel, Switzerland)
Sigma-Aldrich Company (Germany)
Research groups (Universities)
Centre National de la Recherche Sientifique (Gif-sur-Yvette, France)
Centro Nacional de Investigaciones Oncológicas (Madrid, Spain)
Clemson University (Clemson, United States)
Copenhagen University (Copenhagen, Denmark)
Eidgenössische Technische Hochschule Zürich (Zürich, Switzerland)
Ernst-Moritz-Arndt-Universität Greifswald (Greifswald, Germany)
Florida State University (Tallahassee, United States)
Greek Research and Technology Network (Greece)
Harvard University (Cambridge, United States)
Leibniz-Rechenzentrum der Bayerischen Akademie (München, Germany)
Medical Research Council (Harwell, United Kingdom)
National Centre for Software Technology (India)
National Institute of Advanced Industrial Science and Technology (Japan)
Salk Institute for Biological Studies (United States)
Stanford University (Stanford, United States)
Taiwan Academic Network (Chiayi, Taiwan)
Tampere University of Technology (Tampere, Finland)
The National Research Council (Napoli, Italy)
The University of California (San Francisco, United States)
The University of New Mexico (Albuquerque, United States)
Universidade do Porto (Porto, Portugal)
Università di Cagliari (Cagliari, Italy)
Università di Milano (Milano, Italy)
Universita di Pisa (Pisa, Italy)
Universität Tübingen (Tübingen, Germany)
Universität Ulm (Ulm, Germany)
Universität Würzburg (Würzburg, Germany)
University of British Columbia (Canada)
University of California (San Diego, United States)
University of Cambridge (Cambridge, United Kingdom)
University of Colorado (Healt, Denver, United States)
University of Florida (Gainesville, United States)
University of Louisville (Louisville, United States)
University of Massachusetts (Dorchester, United States)
University of Michigan (Ann Arbor, United States)
University of Pennsylvania (Philadelphia, United States)
University of Science Technology (Trondheim, Norway)
University of Toronto (Canada)
University of Wisconsin (Milwaukee, United States)
Veterinärmedizinische Universität (Vienna, Austria)
Washington University (Saint Louis, United States)
Medical centers and Hospitals
Deutsches Krebsforschungszentrum (Heidelberg, Germany)
Forschungszentrum (Karlsruhe, Germany)
National Institutes of Health (Bethesda, United States)

Table 4.1 : *Selection of visitors from biomedical companies, research groups (universities), and medical centers and hospitals.*

4.7 Conclusion

Because comparing classifiers and selecting the best for each microarray data set is a tedious and non-straightforward task, a web service has been developed in this chapter. The M@CBETH (a MicroArray Classification BEnchmarking Tool on a Host server) web service offers the microarray community a simple tool for making optimal two-class predictions. This web service aims at finding the best prediction among different classification methods by using randomizations of the benchmarking data set. This way, the M@CBETH web service enables an optimal use of clinical microarray data classification. The M@CBETH web site is freely available at http://www.esat.kuleuven.be/MACBETH/ and has already shown initial international impact after its recent introduction.

In the next chapter, this web service as well as other methods based on the previous chapter will be applied to microarray data from an ovarian cancer project in cooperation with Prof. I. Vergote and Prof. D. Timmerman of the Department of Obstetrics and Gynaecology and Gynaecologic Oncology of the University Hospitals, Leuven. The microarray experiments have been generated in collaboration with the MicroArray Facility (VIB) at the K.U.Leuven.

Chapter 5

Classification of ovarian tumors using microarray data

5.1 Introduction

In the previous chapters we studied several ways of developing prediction models using microarray data and presented a web service that allows users to easily generate models in a statistically sound way. In this chapter, the general principles described before will be applied to microarray data of ovarian tumors generated in a project we collaborated in.

In the beginning of this chapter we will elaborate on our ovarian cancer project[1] with Prof. I. Vergote and Prof. D. Timmerman of the Department of Obstetrics and Gynaecology and Gynaecologic Oncology of the University Hospitals, Leuven. In this context, we first study the experiments using a broad range of classical linear techniques in Section 5.2. We apply the M@CBETH web service on these experiments in Section 5.3 using also nonlinear machine learning techniques to develop prediction models. The microarray experiments were generated in close collaboration with the MicroArray Facility (VIB) and are also described and indicated in Section 5.2, together with the methodology to analyze them.

Further on, the remainder of this chapter will be dedicated to two other recently published studies that have shortcomings in the techniques used to develop prediction models.

In this chapter, will show that it is possible to distinguish between stage I without recurrence, platin-sensitive advanced-stage and platin-

[1] In this context, we published a full manuscript (De Smet *et al.*, 2006a) and an abstract (Van Gorp *et al.*, 2005) in the journal *International Journal of Gynecological Cancer*.

resistant advanced-stage ovarian tumors based on a set of gene expression patterns. This will be proved by applying a broad range of classical and linear techniques on these experiments, followed by the set of more advanced nonlinear techniques incorporated in M@CBETH.

5.2 Expression patterns of ovarian tumors: general analysis

Epithelial ovarian cancer is the leading cause of death from gynecological malignancies in women. Stage at diagnosis is one of the most important prognostic factors but only partially explains the heterogeneous behavior of these patients. For example, 10-50% of patients with early-stage disease will recur after initial surgery (Trimbos *et al.*, 2003) and a subset of patients with stage III or IV disease will prove to be resistant to platin-based chemotherapy (platin-resistance is defined as originally proposed by Markman *et al.*, 1991). However, at this moment no clinical or pathological parameters are available that can predict these events with sufficient accuracy.

Recently, several studies have been published investigating whether various aspects of ovarian cancer affect the global expression behavior of these malignancies captured by microarrays (Berchuck *et al.*, 2004; Schwartz *et al.*, 2002; Lancaster *et al.*, 2004; Welsh *et al.*, 2001; Jazaeri *et al.*, 2002; Lu *et al.*, 2004).

Within our project, we formulated the following two important clinical cases in ovarian cancer that we want to investigate and for which we aim to develop appropriate prediction models based on microarray data to predict whether:

1. A patient with stage III or IV (FIGO) ovarian tumor will relapse within 6 months after the last therapeutic intervention. Since standard chemotherapy for advanced ovarian cancer is usually platinum based (e.g., carboplatinum + paclitaxel), this model will be able to predict platinum resistance (or chemosensitivity of the tumor). This has mainly prognostic significance but might allow for developing new therapeutic strategies in the future for tumors that are predicted not to respond adequately to the standard chemotherapeutic regimen.
2. A patient with stage I ovarian tumor will have a recurrence after initial surgery. The subset of women with early-stage disease and, according to our model, with a high probability of recurrence are suitable candidates that might maximally benefit from adjuvant treatment (chemotherapy and/or lymphadenectomy) while the women with early-stage disease and

a low probability of recurrence might be spared the side-effects of adjuvant therapy.

Hereto, we already carried out a pilot study and are working on a prospective study at this moment. In the future, we will finish the prospective study and start with new studies, as will be described in Chapter 7.

The pilot study focused on the first clinical case described above, and also looked at the differences between early-stage and advanced-stage disease. For this purpose, we generated a set of cDNA microarray experiments comprising 20 ovarian tumors. More specifically, we wanted to investigate whether the differences between stage I and advanced-stage (Fédération Internationale de Gynécologie et Obstétrie (FIGO) stage III-IV) ovarian cancer, and between platin-sensitive and platin-resistant advanced-stage disease are reflected in the expression patterns. Furthermore, in this section we also develop predictions models based on the 20 ovarian tumors of the pilot study. We will evaluate these models on cDNA microarray experiments comprising 50 ovarian tumors of the prospective study in the nearby future. Moreover, in the prospective study, we also aim to focus on the second clinical case described above, as will be further discussed in Chapter 7.

5.2.1 Materials and methods

This section first describes the tumor characteristics of the samples included in the pilot study. Next, details are given on the procedures that were followed to generate the expression patterns.

Tumor characteristics

Tissue collection and analysis was approved by the ethics committee of the Faculty of Medicine of the K.U.Leuven. Patients signed an informed consent form. Tumor biopsies were sampled during primary or interval surgery and were taken from three groups of patients: 7 from patients with stage I without recurrence, 7 from patients with advanced-stage platin-sensitive (all with a platin-free interval of at least twelve months after first-line platin-based chemotherapy), and 6 from patients with advanced-stage platin-resistant (all with progression during or recurrence within six months after first-line platin-based chemotherapy) disease. In this chapter we will refer to these three classes with the following notation: I, A_s and A_r, respectively. Patient and tumor characteristics are summarized in Table 5.1. It should be noted that all advanced-stage ovarian carcinoma were serous while the stage I patients had different histopathological types. All patients in class A_s and A_r and none of the patients in class I received chemotherapy. Chemotherapy was platin-based and contained paclitaxel as the second drug

	Class A_r n = 6	Class A_s n = 7	Class I n = 7	Reference pool n = 21
Mean age (range), years	61.2 (54–72)	59.1 (47–75)	51.7 (19–86)	58.2 (19-86)
Histologic type, n				
Serous carcinoma	6	7	2	16
Endometrioid carcinoma	-	-	3	3
Mucinous carcinoma	-	-	1	1
Mixed carcinoma*	-	-	1	1
FIGO Stage, n				
I	-	-	7	8
III	5	6	-	10
IV	1	1	-	3
Differentiation grade, n				
Borderline	-	-	2	2
Grade 1	-	-	3	3
Grade 2	4	2	1	5
Grade 3	2	5	2	12
Operation, n				
Primary surgery.	3	7	7	20
Interval surgery after 3 courses of chemotherapy.	2	-	-	1
Diagnostic biopsy, no surgery.	1	-	-	-
Residual tumorload after surgery, n				
0 cm	3	7	7	20
0-1 cm	-	-	-	-
1-2 cm	1	-	-	1
> 2cm	1	-	-	-
Time to progression, n				
< 6 months after first-line chemotherapy	6	-	-	3
> 12 months after fist-line chemotherapy	-	5	-	3
no recurrence	-	2	7	15
Current status, n				
No evidence of disease	-	5	7	16
Alive with evidence of disease	-	1	-	2
Died of disease	6	1	-	4
Median follow-up, months**	14.5	43.0	48.0	40.0

* One sample of class I and one sample of the reference pool were of a mixed histology with a clear cell and an endometrioid component.
** In patients dying of disease the follow-up was ended at the time of death.

Table 5.1 : *Clinical information on tumor samples in class A_r (stage III-IV ovarian tumors with platinum resistance), A_s (stage III-IV ovarian tumors without platinum resistance), I (stage I tumors) and the common reference pool consisting of 21 ovarian carcinomas.*

in 5 out of 6 cases in class A_r and 6 out of 7 cases in class A_s. The remaining two patients received platin and cyclophosphamide.

Microarray procedures

Each tumor was hybridized twice (with dye-swap) against a common reference pool on an array containing 21,372 probes and specifically enriched for genes related to ovarian cancer (i.e., after an extensive review of literature and publicly available databases, we included 3,855 genes possibly involved in ovarian cancer). From each patient, mRNA was amplified and labeled with Cy3 and Cy5 according to Puskas *et al.* (2002). All protocols can be downloaded from ArrayExpress (http://www.ebi.ac.uk/arrayexpress; P-MEXP-580 and P-MEXP-581). The reference pool was obtained by pooling amplified RNA derived from 4, 6, and 7 samples included in class A_r, A_s, and I, respectively, and 4 samples of other ovarian tumors, not further included in the study (also see Table 5.1). Microarray data and information recommended by the MIAME (Minimum Information About a Microarray Experiment) (Brazma *et al.*, 2001) guidelines can be found on the following web site (http://www.esat.kuleuven.be/~fdesmet/ovarian/).

Data analysis

Gene expression data were analyzed using MATLAB 6.5. Background corrected intensities were log-transformed, and subsequently normalized using the intensity-dependent Lowess fit (Yang *et al.*, 2002) procedure. The mean of the replicate and normalized log-ratios (i.e., patient over reference) was used as a measure for expression.

To rank the genes with respect to their differential expression between two of the three classes of ovarian tumors, we calculated p-values for each gene using the Wilcoxon rank sum test (Dawson-Saunders and Trapp, 1994; Troyanskaya *et al.*, 2002; De Smet *et al.*, 2004). We used the method proposed by Storey and Tibshirani (2003) to estimate, independently of the choice of a p-value threshold, the number of truly alternative genes (i.e., these are the genes whose expression is really affected by the difference between the classes and whose degree of differential expression is therefore not expected to be accidental – these genes are said to be actually differentially expressed) (De Smet *et al.*, 2004).

We applied classical PCA to the 21,372-dimensional expression patterns and to the 3,000-dimensional expression patterns obtained after selection of the 3,000 genes with the largest amount of differential expression between the three classes (p-values obtained with the Kruskal-Wallis test) (Dawson-Saunders and Trapp, 1994).

We carried out supervised class prediction using LS-SVM and estimated the test set performance of this classification technique using a leave-one-out (LOO) approach. This approach is similar to the optimization procedures discussed in Chapter 3, except for the fact that the number of genes is not optimized but previously defined to be 3,000. More specifically, the approach used in this study is as follows. Iteratively, the expression pattern from one sample was removed from the data. The remaining samples were used to select the 3,000 genes with the largest degree of differential expression and, based on this selection, to build an LS-SVM model with a linear kernel. Note that we intended to use linear techniques in this study (nonlinear machine learning techniques will be applied in the next study using the M@CBETH web service). This model was finally used to classify the sample left out. The proportion of left-out samples correctly classified in this procedure is an estimate of the actual predictive performance on an independent test set.

Finally, we built two prediction models using the experiments of all 20 patients of this pilot study for later evaluation in the prospective study. The 3,000 genes with the largest amount of differential expression between the three classes that were earlier selected using the Kruskal-Wallis test, are used to train two LS-SVM models with a linear kernel.

5.2.2 Results and discussion

In this section, we describe the results obtained from analyzing expression patterns of 7 stage I without recurrence (class I), 7 stage III/IV platin-sensitive (class A_s) and 6 stage III/IV platin-resistant (class A_r) ovarian carcinomas.

First, we quantified the degree of differential expression of each gene between class I and A_r, class I and A_s, and class A_s and A_r using the Wilcoxon rank sum test. The three respective lists with the 500 best scoring genes (lowest p-value) can be found on the web site (http://www.esat.kuleuven.be/~fdesmet/ovarian/). These lists however, contain false positives and miss false negatives. Although the p-values cannot be used to identify the individual false positive and false negative genes, procedures are available to estimate their proportions (De Smet *et al.*, 2004) and to calculate the number of genes expected to be *actually* differentially expressed (i.e., the sum of the true positives and false negatives). Using the method proposed by Storey and Tibshirani (2003), the number of genes actually differentially expressed was estimated to be 7059 between class I and A_r, 4943 between class I and A_s, and 2028 between class A_s and A_r. These numbers suggest that the expression patterns indeed reflect the differences between the classes under investigation and that the amount

of differential expression is larger between class I and A_r than between class I and A_s. Note however that – since not all members of class I were serous carcinomas (unlike the samples in class A_s an A_r) – we cannot exclude that some of the differential expression between class I and A_s and between class I and A_r is caused by the difference in histopathological types.

The 20 expression patterns were also analyzed with PCA. The three principal component directions explaining the largest variation in the data (i.e., that are associated with the largest eigenvalues) were selected and each of the 20 expression patterns were projected onto these three vectors (Figure 5.1). This analysis shows a clear separation between patients from class I and class A_r with class A_s lying in between, which indicates the order of transition between the different tumor types.

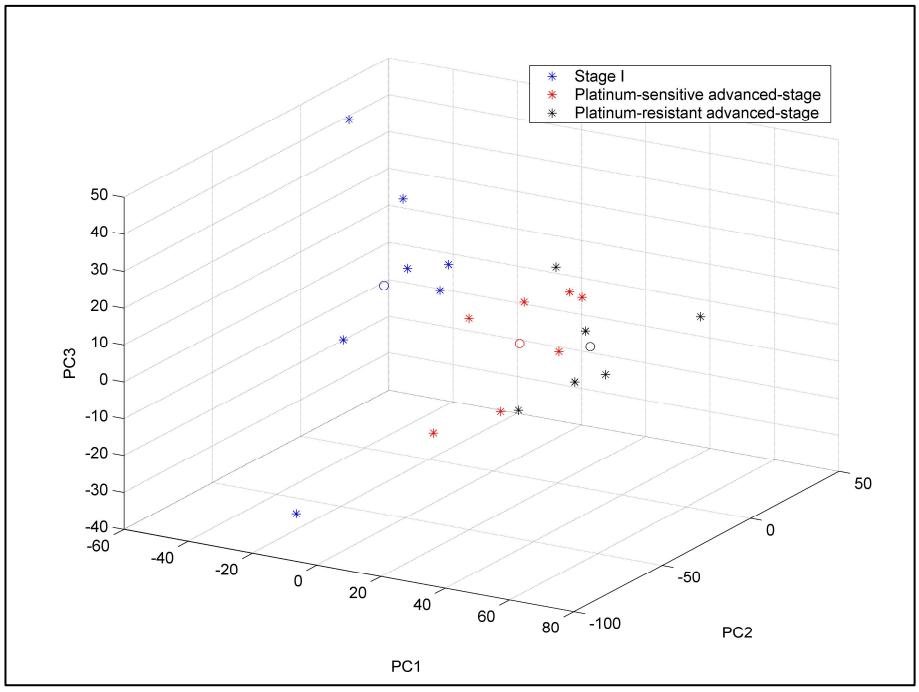


Figure 5.1 : *PCA of the expression patterns of the 20 ovarian tumor samples. Projection of the 21372-dimensional expression patterns on the three first principal component directions (PC1, PC2, and PC3) associated with the largest eigenvalues. * = individual sample; O = mean projected expression pattern in each class; blue = stage I; red = platin-sensitive advanced-stage (III or IV); black = platin-resistant advanced-stage.*

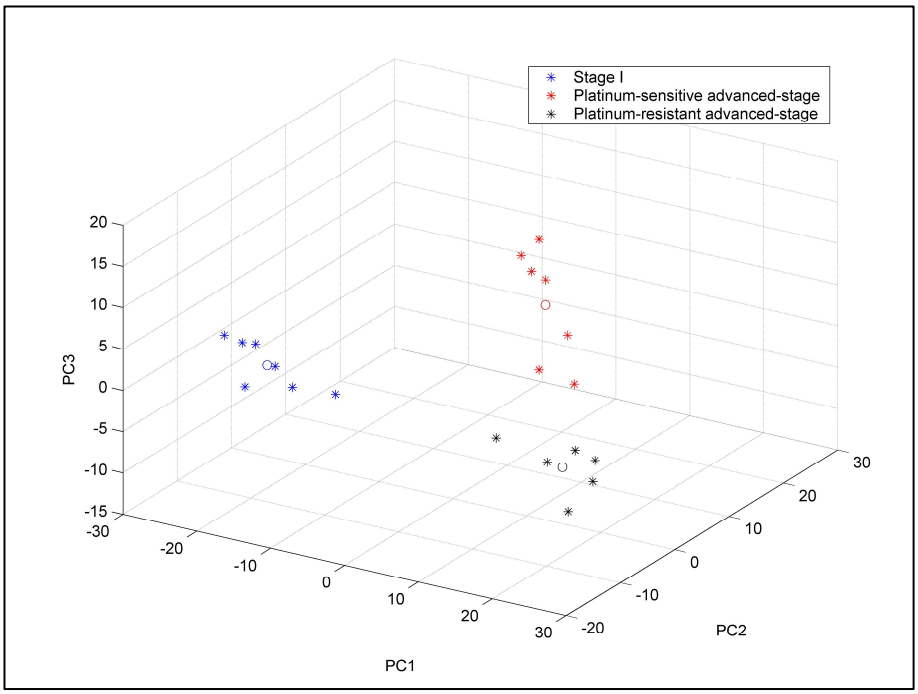


Figure 5.2 : *PCA of the expression patterns of the 20 ovarian tumor samples. PCA after selection of the 3000 genes with the largest degree of differential expression between stage I, platin-sensitive advanced-stage and platin-resistant advanced-stage disease (i.e., 3,000 genes with the smallest p-value obtained using the Kruskal-Wallis test).*

To enhance the separation between the different tumor types, we repeated PCA after selection of the 3,000 genes with the largest amount of differential expression (determined by the Kruskal-Wallis test) between the three classes (Figure 5.2). This results into three well-separated clusters that almost perfectly coincide with the known classes. This observation suggests that the three different types of ovarian tumors can be accurately identified using their expression patterns.

The almost perfect segregation between the three classes obtained by supervised gene selection prior to PCA, however, could be caused by random effects (the 3,000 selected genes could possibly contain a high number of false positives) that might not be confirmed by new experiments. To determine whether it is possible to assign independent ovarian tumor samples to the correct class, we applied LS-SVM to the expression data. This resulted in an estimated LOO classification accuracy of 100% for the distinction between stage I and advanced-stage disease and 76.92% for the distinction between class A_s and A_r (2 samples from class A_r and 1 sample from class A_s were misclassified). Moreover, if one single model was trained

(using all 13 high-stage samples) to distinguish between class A_s and A_r and subsequently applied to the 7 stage I patients, these were all assigned to class A_s. This again indicates that stage I disease is more similar to platin-sensitive than platin-resistant advanced-stage disease.

The data presented here suggest that gene expression patterns can indeed be used to discriminate, with reasonable accuracy, between stage I, platin-sensitive advanced-stage and platin-resistant advanced-stage ovarian tumors. Thus microarray technology might, for example, be useful to help clinicians to select stage I patients with a very low risk of recurrence making adjuvant chemotherapy unnecessary, or for patients with advanced ovarian cancer to predict platin resistance. Note that the former corresponds to the second clinical case formulated earlier in this chapter and that we will investigate this in the prospective study in the future: for example, by looking where stage I patients with a high risk of recurrence lie with respect to the other classes of ovarian cancer patients in a PCA plot.

5.3 Expression patterns of ovarian tumors: M@CBETH

In the previous chapter, the M@CBETH web service was presented that compares, for each microarray data set introduced to this service, different classifiers and selects the best in terms of randomized independent test set performances. As a test case, we applied our tool on the ovarian cancer microarray data set described in this chapter[2]. In the previous section, we only used classical and linear techniques to study these data. In this section, however, we will investigate the influence of using nonlinear machine learning techniques included in M@CBETH. This way, the prediction models developed in the previous section may be even further optimized.

Application of the M@CBETH benchmarking service on both cases discussed in the pilot study automatically results in storage of an optimal model for each case. This allows a delayed evaluation of both models on the patients from the prospective study in the future. Note that no gene selection is performed when using M@CBETH, which means that all genes are included in these prediction models.

---

[2] In this context, we contributed in the *Computational Systems Bioinformatics Conference (CSB2005)* at Stanford University, California (Pochet *et al.*, 2005).

5.3.1 Results

We first discuss the case where we tried to discriminate the early-stage from the advanced-stage ovarian tumor samples. Next, the case of discriminating the platin-sensitive from the platin-resistant advanced-stage ovarian tumor samples is considered.

Discrimination between early-stage and advanced-stage ovarian tumor samples

Figure 5.3 shows the results of submitting all 20 samples to the benchmarking service, considering the 7 early-stage as one class and the 13 advanced-stage samples as the other class. LS-SVM with an RBF kernel is selected as the best classification method for this classification problem with an average test set accuracy (ACC) of about 92% and an average test set Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) of about 99%.

	LOO-CV	training ACC	test ACC	training AUC	test AUC
1.	90.48 ± 5.22	100.00 ± 0.00	91.27 ± 8.53	100.00 ± 0.00	99.40 ± 2.73
2.	92.18 ± 3.85	98.30 ± 7.79	92.06 ± 10.03	100.00 ± 0.00	98.81 ± 5.46
3.	49.32 ± 14.62	50.68 ± 8.72	51.59 ± 19.65	50.05 ± 5.82	49.40 ± 23.54
4.	94.22 ± 5.35	99.32 ± 2.15	83.33 ± 14.91	99.89 ± 0.48	93.45 ± 9.37
5.	95.58 ± 4.78	99.66 ± 1.56	85.71 ± 16.90	100.00 ± 0.00	93.45 ± 12.88
6.	93.88 ± 5.19	98.64 ± 3.66	83.33 ± 14.91	99.68 ± 1.06	95.24 ± 8.36
7.	94.90 ± 5.12	99.66 ± 1.56	86.51 ± 17.17	100.00 ± 0.00	93.45 ± 12.88
8.	97.28 ± 4.21	99.66 ± 1.56	82.54 ± 16.22	99.89 ± 0.48	89.88 ± 14.04
9.	97.96 ± 3.31	99.66 ± 1.56	82.54 ± 14.41	100.00 ± 0.00	90.18 ± 15.13

Figure 5.3 : *Results of discriminating early-stage from advanced stage ovarian tumor samples. Methods are: 1. LS-SVM with linear kernel, 2. LS-SVM with Radial Basis Function (RBF) kernel, 3. LS-SVM with linear kernel without regularization, 4. PCA (unsupervised principal component (PC) selection) followed by FDA, 5. PCA (supervised PC selection) followed by FDA, 6. Kernel PCA with linear kernel (unsupervised PC selection) followed by FDA, 7. Kernel PCA with linear kernel (supervised PC selection) followed by FDA, 8. Kernel PCA with RBF kernel (unsupervised PC selection) followed by FDA, 9. Kernel PCA with RBF kernel (supervised PC selection) followed by FDA. The best classification method is highlighted.*

Discrimination between platin-sensitive and platin-resistant advanced-stage ovarian tumor samples

Figure 5.4 shows the results of submitting 13 samples to the benchmarking service, taking the 7 platin-sensitive advanced-stage samples as one class and the 6 platin-resistant advanced-stage samples as the other class. LSSVM with a linear kernel is selected as the best classification method for this classification problem with an average test set ACC of about 75% and an average test set AUC of about 82%.

	LOO-CV	training ACC	test ACC	training AUC	test AUC
1.	56.08 ± 8.22	100.00 ± 0.00	75.00 ± 17.68	100.00 ± 0.00	82.14 ± 22.56
2.	63.49 ± 7.97	85.71 ± 20.83	60.71 ± 16.90	100.00 ± 0.00	82.74 ± 22.53
3.	50.26 ± 5.69	51.32 ± 8.94	52.38 ± 24.88	53.33 ± 7.80	46.43 ± 38.96
4.	64.55 ± 14.32	91.53 ± 12.12	61.90 ± 15.04	96.43 ± 6.92	80.95 ± 23.59
5.	67.20 ± 10.23	100.00 ± 0.00	64.29 ± 20.27	100.00 ± 0.00	78.57 ± 19.82
6.	66.67 ± 14.49	92.06 ± 12.74	59.52 ± 14.74	97.14 ± 6.04	76.19 ± 23.02
7.	65.08 ± 13.28	97.88 ± 5.69	58.33 ± 21.41	97.86 ± 7.68	73.81 ± 21.62
8.	79.89 ± 9.04	96.30 ± 9.51	53.57 ± 8.96	97.62 ± 7.18	64.29 ± 26.89
9.	83.60 ± 10.32	100.00 ± 0.00	57.14 ± 11.57	100.00 ± 0.00	65.48 ± 20.12

Figure 5.4 : *Results of discriminating platin-sensitive from platin-resistant advanced-stage ovarian tumor samples. Methods: see Figure 5.3.*

5.3.2 Discussion

By applying the M@CBETH benchmarking service on these two binary cancer classification problems in ovarian cancer, we confirmed our previous findings that the differences between stage I and advanced-stage ovarian cancer, and between platin-sensitive and platin-resistant advanced-stage disease are reflected in the expression patterns.

From a methodological point of view, it can be concluded that it is important to choose an optimal classification method for each microarray data set.

The study in this section indicates that further optimization of the prediction models developed in the context of the pilot study, may be possible by also considering nonlinear techniques. However, we would like to stress that direct comparison of both studies is not allowed as such. Below, we will explain the differences between both studies to support this statement. To end this discussion, a solution is proposed to easily make comparisons possible. The actual implementation and the determination of the most optimal classifier for each of both cases, however, will be left for future research within the prospective study.

In the pilot study, we selected only 3,000 genes (using the Kruskal-Wallis test) for inclusion in the model. Using the M@CBETH web service, we did not perform gene selection as such. This also implies that the final models generated in both studies differ with respect to the number of genes these are based on.

Moreover, the test set performance of the model generated in the pilot study was estimated using a LOO approach with removal of the left-out sample *before* gene selection. On the other hand, the test set performance (ACC) calculated in the M@CBETH web service is an averaged test performance of the 20 test set splits ($1/3^{rd}$ of the samples in each reshuffling/randomization of the data set). It is also not allowed to make comparisons with the LOO-CV performance calculated by M@CBETH, since this is an averaged LOO-CV performance of the 20 training set splits ($2/3^{rd}$ of the samples in each reshuffling/randomization of the data set), which is in fact used for optimizing the parameters and is certainly not a test performance.

An easy and straightforward solution to circumvent these inconveniences when comparing different classifiers, is to implement the approach followed in the pilot study into the M@CBETH web service. Note, however, that the fixed number of genes (3,000) that are selected for model building, may be not optimal for general usage on other microarray data sets. More ideal would be to also optimize the number of genes that are selected, but this may become time consuming.

5.4 Problems with other studies investigating expression patterns of ovarian tumors

This section presents and discusses two recently published studies that have shortcomings in the techniques used to develop prediction models.

5.4.1 Importance of the independency of test samples

In this section, we take a closer look at the article by Hartmann *et al.* (2005) investigating whether it is possible to apply gene expression patterns to discriminate between ovarian tumors with early and late relapse after platinum-paclitaxel combination chemotherapy[3]. Among others, the authors claim to have derived a 14-gene predictive model with an independent test set accuracy of 86% and a positive predictive value of 95%.

After examination of the data analysis strategy of Hartmann *et al.*, we noticed that the test set has been used to perform prior model selection and therefore cannot be called independent. Summarized and after data preprocessing, the authors constructed 100 SVM models each based on a set of genes with the highest signal-to-noise ratio derived from a random selection (70% of 51 training samples) of the training set. Subsequently, these 100 models were all tested on the (wrongfully called independent) test set (28 samples) and the top model with the fewest prediction errors was selected and reported.

Unfortunately, this selection implies that information from the test set was used to choose a model that optimally fits this particular test set but might perform worse on another and independently chosen test set. As a consequence the reported performance indices might be too optimistic and will probably be impossible to reproduce on new prospective data. In our experience and due to the high dimensional nature of microarray data, even the slightest use of a so called independent test set (or the use of the left-out samples in cross-validation studies (Simon *et al.*, 2003)) within the model building process will result in an overly optimistic performance of a classifier based on expression patterns. After model selection and to obtain a realistic estimate of the true performance, it is therefore important to test a new model on completely independent and prospective data.

The authors have the choice between three options to develop and to assess their prediction models in a sound way. A first possibility is to use the complete data set for building the model, thus performing gene selection on all microarray experiments (which is optional) and then applying a classification method to actually generate the model. In this case, the model can then be assessed by estimating the (leave-one-out) cross-validation performance. This is in fact the strategy that we followed in our pilot study described above because of the limited number of microarray experiments available in our study. Remember that if cross-validation is used to estimate

[3] A letter we wrote as a response to this recently published study has been published in the journal *Clinical Cancer Research* (De Smet *et al.*, 2005).

the prediction accuracy, then the entire model-building process, including the selection of informative genes, should be repeated for each cross-validation training set. Note that when prospective data becomes available in a later stage, it is possible to also assess the model by calculating the independent test set performance on these data.

A second possibility is to exploit the availability of a training and a test set in the case of Hartmann *et al.* In this case, the model can be generated based on the training data, thus building the classifier based on these data (and possibly performing gene selection first). Assessing this model can then easily be done by calculating the independent test set performance based on the test set (using the same subset of relevant genes when gene selection was performed). Note that if a separate data set is used for validation, this should be sufficiently large to provide a meaningful and reliable estimate for the prediction accuracy. Important to mention is also that an independent test set accuracy can easily be calculated by submitting a benchmarking (training) data set and a prospective (test) data set to the M@CBETH web service. This can be done in one step with a benchmarking analysis, or by using benchmarking and prediction analyses when prospective data becomes available in a later stage.

A third possibility is to again use the complete data set for building the model, but this time assessing the model by calculating the mean randomized test set performance over a number of reshufflings of the data set, as is done in M@CBETH (see Chapter 4 for details on the algorithm). Remember that we also followed this strategy by applying the M@CBETH web service on our data set, as described above. Note that M@CBETH automatically generates and stores an optimal prediction model, which can then be used for later evaluation of prospective data.

To substantiate our objections with respect of the strategy followed by Hartmann *et al.*, we implemented a similar data analysis scheme in MATLAB based on 14-gene LS-SVM models from LS-SVMlab. We subsequently applied our script on 10 randomly generated data sets each subdivided in a training and test set (expression levels uniformly and independently drawn between 0 and 1) with the same dimensions and composition as reported by Hartmann *et al.* For a true independent test set and since the random data does not contain any information about the process under study, one could expect an accuracy around 50%. However, the 10 test set accuracies returned by our MATLAB script (one for each training + test set) ranged between 71.43% and 82.14% and were significantly ($P = 0.002$; sign test) different from 50%. Therefore, these results indicate that the procedure described in Hartmann *et al.* strongly overestimates the accuracy that can be expected on independent data. Also noteworthy was the observation that the accuracies on the (in this case truly independent) test set indeed varied around a mean of about 50% if the model

selection step was omitted. In the latter case we considered all 1,000 models (100 models for each random data set) and not only the 10 models selected for their optimal performance on the test set.

Finally, we want to mention that Hartmann *et al.* stated that the reported accuracy of 86% was unlikely to occur by chance alone. This was – similarly as above – assessed by comparing this result with a series of test set accuracies obtained through random models (in this case generated by randomly permuting the outcome labels of the training set). However, this assessment only indicates that the reported accuracy is relatively better than the test set accuracies of the random models. Since our simulation showed that these values themselves are overestimated, this evaluation does not say anything about the validity of the absolute value of the reported accuracy of 86%. Nevertheless, this assessment seems to indicate that the expression patterns indeed contain information about the time of relapse after chemotherapy.

5.4.2 Problems with clinical model assessment

In this section, we analyze the article by Helleman *et al.* (2005) investigating whether a gene set identified using microarrays could be used to predict platin resistance in ovarian cancer[4]. The authors studied a training set obtained from 24 tumors that were analyzed using cDNA microarrays. This set contained 5 women that were platin-resistant (the non-responders) and 19 women that were platin-sensitive (the responders). The authors concluded that 69 genes were differentially expressed between the responders and the non-responders. An algorithm based on clustering was used to identify the most predictive genes among these 69 genes in the training set. This resulted in 9 genes (the differential expression of these genes was later confirmed with qRT-PCR) that could significantly discriminate between the responders and the non-responders in the training set. Subsequently, this 9-gene set was used to predict platin resistance in an independent test set of 72 tumors (9 non-responders and 63 responders) using expression levels measured with qRT-PCR. This resulted in a sensitivity of 89% and a specificity of 59%.

Remember the definitions of sensitivity and specificity in the context of ROC Curves from Chapter 3. In the case of Helleman *et al.*, the model is developed in order to predict platin resistance (positives). This means that

[4] A letter we wrote as a response to this recently published study has been accepted for publication in the journal *International Journal of Cancer* (Gevaert *et al.*, 2006).

the sensitivity of 89% reflects the potential of the model to predict platin resistance (true positives) for all patients who actually are platin-resistant (true positives and false negatives). The specificity of 59% on the other hand, reflects the potential of the model to predict platin-sensitivity (true negatives) for all patients who actually are platin-sensitive (true negatives and false positives). The model proposed by Helleman *et al.* is therefore appropriate to detect the patients with platin-resistance (positives).

However, the approach described by Helleman *et al.* is not optimally tuned for implementation in clinical practice. For women that are platin-sensitive (the responders – negatives with respect to the model), the non-platinum containing regimen strategies remain suboptimal (Thigpen *et al.*, 2005). Therefore, it is important to accurately identify patients that will respond to platin-based chemotherapy. In practice, this amounts to developing a new model that considers the patients with platin-sensitivity as positives and the platin-resistant patients as negatives, in contrary to the current model. Because the specificity of the model of Helleman *et al.* is only 59%, 41% of the responders will be predicted to have platin-resistance and will therefore be wrongfully assigned to the group of patients where other management options are recommended. Although 89 % (value of the sensitivity) of the women with platin-resistance are correctly classified by the model of Helleman *et al.*, this is less critical in a clinical setting since these patients have worse prognosis which can, at this moment, only be minimally improved by different treatment strategies.

5.5 Conclusion

This chapter concentrated on the analysis of gene expressions coming from ovarian tumors. We first presented our ovarian cancer microarray experiments that were generated within our project. In a pilot study, we studied these experiments using a broad range of classical linear techniques. Next, we applied the M@CBETH web service on these experiments using also nonlinear machine learning techniques to develop prediction models.

Both studies investigated whether it is possible to distinguish between stage I without recurrence, platin-sensitive advanced-stage and platin-resistant advanced-stage ovarian tumors. In the pilot study, results were obtained by studying the number of genes that exhibit non-accidental differential expression between the different tumor classes, by performing unsupervised PCA, and by using a LOO approach together with LS-SVM for developing classifiers. These results indicated that gene expression patterns could be useful in clinical management of ovarian cancer. These findings were confirmed by applying the M@CBETH benchmarking service.

Furthermore, from a methodological point of view it can be concluded that it is important to choose an optimal classification method for each microarray data set. This study also indicates that further optimization of the prediction models developed in the context of the pilot study, may be possible by also considering nonlinear techniques.

The remainder of this chapter was dedicated to other recently published studies that have shortcomings in the techniques used to develop prediction models.

Chapter 6

Kernel clustering of microarray experiments

6.1 Introduction

In contrast to all previous chapters, which were concentrated on developing prediction models for clinical applications, this chapter is devoted to discovering diagnostic classes[1]. Clustering techniques are generally applied to microarray data for the identification of clinical classes, which could allow refining clinical management. Cluster analysis of entire microarray experiments (expression patterns from patients or tissues) allows for the discovery of possibly unknown diagnostic categories without knowing the properties of these classes in advance. These clusters could form the basis of new diagnostic schemes in which the different categories contain patients with less clinical variability.

Clustering microarray experiments has already shown to be useful in a large number of cancer studies. Alon *et al.* (1999), for example, separated cancerous colon tissues from non-cancerous colon tissues by applying two-way clustering. The distinction between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) has been rediscovered by using self-organizing maps (SOM) by Golub *et al.* (1999). By using hierarchical clustering, Van 't Veer *et al.* (2002) were able to distinguish between the presence (poor prognosis) and the absence (good prognosis) of distant sub-clinical metastases in breast cancer patients where the histopathological

[1] The content of this chapter has been accepted for publication as the chapter "*Kernel clustering for knowledge discovery in clinical microarray data analysis*" in the book "*Kernel methods in bioengineering, communications and image processing*" (Pochet *et al.*, 2006a).

examination did not show tumor cells in local lymph nodes at diagnosis (lymph node negative).

For this purpose, methods such as the classical K-means clustering and hierarchical clustering are commonly used (Handl *et al.*, 2005; Bolshakova *et al.*, 2005). These methods are based on simple distance or similarity measures (e.g., the Euclidean distance). Therefore, only linear distance measures can be applied to the data using these techniques. Recently, methods have emerged for clustering data of which the clusters are not linearly separable. Two important methods are kernel K-means clustering (Dhillon *et al.*, 2004a; Dhillon *et al.*, 2004b; Zhang and Rudnicky, 2002) and the related spectral clustering (Cristianini *et al.*, 2002; Ng *et al.*, 2001). Introducing these techniques in microarray data analysis would allow for dealing with nonlinear relationships in the data and improving the computational complexity caused by high dimensional data.

Validation techniques are used to assess and compare the performance of different clustering methods. These methods can also be employed for tuning the cluster settings (e.g., optimizing the number of clusters and tuning the kernel parameters). A recent review of Handl *et al.* (2005) presents the state-of-the-art in cluster validation on high dimensional data, among others on microarray data, referring to some previous important manuscripts in the field (Bolshakova and Azuaje, 2003; Halkidi *et al.*, 2001). Two main kinds of validation techniques are internal and external validation. Internal validation assesses the quality of a clustering result based on statistical properties (e.g., assessing the compactness of a cluster, or maximizing the inter-cluster distances while minimizing the intra-cluster distances). External validation reflects the level of agreement of a clustering result with an external partition (e.g., existing diagnostic classes generally used by experts in clinical practice). The Global Silhouette index, the Distortion score and the Calinski-Harabasz index (F-statistic) are commonly used for internal validation, the Rand index and Adjusted Rand index for external validation.

This chapter uses the classical K-means, kernel K-means and spectral clustering algorithms discussed in Chapter 2 and discusses their advantages and disadvantages in the context of clinical microarray data analysis. Since classical K-means clustering experiences time complexity inconveniences when dealing with the high dimensional microarray experiments, PCA is used as a preceding dimensionality reduction step. Kernel K-means and spectral clustering are capable of dealing with the high dimensional microarray experiments in a computationally more efficient way since these make use of the kernel trick, which allows them to work implicitly in the feature space. Several internal and external cluster validation criteria commonly used in the input data space are described and extended for usage in the feature space. The advantages of nonlinear

clustering techniques in case of clinical microarray data analysis are further demonstrated by means of the clustering results on several microarray data sets related to cancer.

In this chapter, we will demonstrate that very good results can be obtained with spectral clustering in terms of internal validation criteria. To realize this, we will show how these internal validation measures can be extended in feature space.

6.2 Preprocessing

This chapter uses standardization, which is discussed in Chapter 3, as a preceding preprocessing step for all clustering methods. Since time complexity inconveniences can be seen when applying classical K-means clustering to the high dimensional microarray experiments, standardization is followed by PCA to obtain a representation of the data with a reduced dimensionality. Note, however, that the results remain similar to those obtained without performing PCA. It is possible to further preprocess the data by selection of principal components. This, however, is not done in this work. For a detailed description of the PCA algorithm, we refer to the second chapter. Although kernel clustering techniques are capable of handling high dimensional data in a computationally more efficient way, one should not forget the possible benefits of performing preprocessing steps that (possibly) remove noise before using any clustering technique. In this section, we therefore describe filtering as another unsupervised preprocessing step, which is also commonly used for that purpose. However, this technique is not used in this work.

Filtering

A set of microarray experiments, generating gene expression profiles (measurements of a single gene under several conditions), frequently contains a considerable number of genes that do not really contribute to the clinical process that is being studied. The expression values of these profiles often show little variation over the different experiments (they are called “constitutive” with respect to the clinical process studied). Moreover, these constitutive genes will have seemingly random and meaningless profiles after standardization (division by a small standard deviation resulting in noise inflation), which is a very common preprocessing step. Another problem with microarray data sets is the fact that these regularly contain highly unreliable expression profiles with a considerable number of missing values. Due to their number, replacing these missing values in these expression profiles is not possible within the desired degree of accuracy.

If these data sets were passed to the clustering algorithms as such, the quality of the clustering results could significantly degrade. A simple solution (that can also be used in combination with other preprocessing steps), is to remove at least a fraction of the undesired genes from the data. This procedure is in general called filtering (Eisen *et al.*, 1998). Filtering involves removing gene expression profiles from the data set that do not satisfy one or possibly more criteria. Commonly used criteria include a minimum threshold for the standard deviation of the expression values in a profile (removal of constitutive genes) and a threshold on the maximum percentage of missing values. Another similar method for filtering takes a fixed number or fraction of genes best satisfying one criterion (like the criteria stated above).

6.3 Classical clustering methods

In a recent review, Handl *et al.* (2005) state that although there have recently been numerous advances in the development of improved clustering techniques for (biological and clinical) microarray data analysis (e.g., biclustering techniques (Madeira and Oliveira, 2004; Sheng *et al.*, 2003) adaptive quality-based clustering (De Smet *et al.*, 2002) and gene shaving (Hastie *et al.*, 2000)), traditional clustering techniques such as K-means (Tavazoie *et al.*, 1999; Rosen *et al.*, 2005) and hierarchical clustering algorithms (Eisen *et al.*, 1998) remain as the predominant methods. According to this review, this fact is arguably more owing to their conceptual simplicity and their wide availability in standard software packages than to their intrinsic merits. In this context, this chapter focuses on a class of linear and nonlinear clustering techniques based on the traditional K-means clustering. A more detailed description of the K-means clustering algorithm can be found in Chapter 2.

6.4 Kernel clustering methods

In microarray data analysis, some kernel clustering methods have already shown to be useful. For example, Qin *et al.* (2003) proposed a kernel hierarchical clustering algorithm on microarray data for identifying groups of genes that share similar expression profiles. Support Vector clustering (Ben-Hur *et al.*, 2001) is another clustering method based on the approach of Support Vector Machines. These kernel clustering methods have recently emerged for clustering data in which the clusters are not linearly separable in order to find nonlinear relationships in the data. Moreover, these techniques allow for dealing with high dimensional data in a computationally more efficient way, which makes it specifically interesting for application on

microarray data. Kernel K-means and spectral clustering were already presented in Chapter 2.

6.5 Cluster validation methods and their kernel versions

Validation of the clustering results can be done internally, for example, by assessing the quality of a clustering result based on statistical properties, and externally, for example, by reflecting the level of agreement of a clustering result with an external partition (e.g., existing diagnostic classes generally used in clinical practice (Handl *et al.*, 2005; Halkidi and Vazirgiannis, 2005; Bolshakova *et al.*, 2005; Jain and Dubes, 1988; Milligan and Cooper, 1985)). Moreover, internal cluster validation techniques can also be used for selecting the best clustering result when comparing different clustering methods, several random initializations, different number of clusters, a range of kernel parameters (e.g., the width σ of the RBF kernel), and so on. In this section, a formulation of three well-known internal validation methods in the input space (Global Silhouette index, Calinski-Harabasz index (F statistic), and Distortion score) and two external validation methods (Rand index and Adjusted Rand index) are given first (applied in the input space) for reason of completeness. However, to be useful for kernel K-means clustering (and eventually other kernel clustering methods as well), we also derive the internal validation criteria for usage in the feature space.

6.5.1 Internal validation

Global Silhouette index

An expression pattern from a patient can be considered to be well clustered if its distance to the other expression patterns of the same cluster is small and the distance to the expression patterns of other clusters is larger. This criterion can be formalized by using the Silhouette index (Kaufman and Rousseeuw, 1990), for example, for testing the cluster coherence.

Suppose $\mathbf{x}_i$ is an expression pattern that belongs to cluster C_k. Call $v(\mathbf{x}_i)$ (also called the within-cluster dissimilarity) the average distance of $\mathbf{x}_i$ to all other expression patterns from C_k. Suppose C_h is a cluster different from C_k. Define $w(\mathbf{x}_i)$ (also called the between-cluster dissimilarity) as the minimum over all clusters C_h different from C_k of the

average distance from $\mathbf{x}_i$ to all expression patterns of C_h. The Silhouette width $s(\mathbf{x}_i)$ of expression patterns $\mathbf{x}_i$ is now defined as follows:

$$s(\mathbf{x}_i) = \frac{w(\mathbf{x}_i) - v(\mathbf{x}_i)}{\max\left(v(\mathbf{x}_i), w(\mathbf{x}_i)\right)},$$

with $\mathbf{x}_i \in C_k$, and

$$v(\mathbf{x}_i) = \frac{1}{n_k - 1} \sum_{\substack{\mathbf{x}_j \in C_k \\ \mathbf{x}_j \neq \mathbf{x}_i}} \left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2,$$

and

$$w(\mathbf{x}_i) = \min_{\substack{h \neq k \\ h,k=1,\ldots,G}} \left\{ \frac{1}{n_h} \sum_{\mathbf{x}_j \in C_h} \left\| \mathbf{x}_i - \mathbf{x}_j \right\|^2 \right\}.$$

Note that $-1 \leq s(\mathbf{x}_i) \leq 1$. Consider two extreme situations now. Firstly, suppose that the within dissimilarity $v(\mathbf{x}_i)$ is significantly smaller than the between dissimilarity $w(\mathbf{x}_i)$. This is the ideal case and $s(\mathbf{x}_i)$ will be approximately equal to 1. This occurs when $\mathbf{x}_i$ is 'well clustered' and there is little doubt that $\mathbf{x}_i$ is assigned to an appropriate cluster. Secondly, suppose that $v(\mathbf{x}_i)$ is significantly larger than $w(\mathbf{x}_i)$. Now $s(\mathbf{x}_i)$ will be approximately -1 and $\mathbf{x}_i$ has in fact been assigned to the wrong cluster (worst case scenario).

Two other measures can now be defined: the average Silhouette width of a cluster and the average Silhouette width of the entire data set. The first is defined as the average of $s(\mathbf{x}_i)$ for all expression patterns of a cluster and the second is defined as the average of $s(\mathbf{x}_i)$ for all expression patterns in the data set. This last value can be used to compare different cluster results and can be used as an inherent part of clustering algorithms, if its value is optimized (maximized) during the clustering process.

When applying this validation measure in combination with kernel clustering methods (performing the actual clustering in the feature space), using this definition of the Silhouette index leads to wrong results since the distances between the expression patterns are computed in the input space. We therefore derive the definition of the Silhouette index for computation in the feature space.

By introducing the feature map $\varphi(\cdot)$, $v(\mathbf{x}_i)$ and $w(\mathbf{x}_i)$ can be expressed in the feature space as

$$v^{\varphi}(\mathbf{x}_i) = \frac{1}{n_k - 1} \sum_{\substack{\mathbf{x}_j \in C_k \\ \mathbf{x}_j \neq \mathbf{x}_i}} \left\| \varphi(\mathbf{x}_i) - \varphi(\mathbf{x}_j) \right\|^2,$$

and

$$w^{\varphi}(\mathbf{x}_i) = \min_{\substack{h \neq k \\ h,k=1,\ldots,G}} \left(\frac{1}{n_h} \sum_{\mathbf{x}_j \in C_h} \left\| \varphi(\mathbf{x}_i) - \varphi(\mathbf{x}_j) \right\|^2 \right),$$

for $\mathbf{x}_i \in C_k$.

Replacing all the dot products by a kernel function $K(\cdot,\cdot)$ results in

$$v^{\varphi}(\mathbf{x}_i) = \frac{1}{n_k - 1} K(\mathbf{x}_i, \mathbf{x}_i) - \frac{2}{n_k - 1} \sum_{\substack{\mathbf{x}_j \in C_k \\ \mathbf{x}_j \neq \mathbf{x}_i}} K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{n_k - 1} \sum_{\substack{\mathbf{x}_j \in C_k \\ \mathbf{x}_j \neq \mathbf{x}_i}} K(\mathbf{x}_j, \mathbf{x}_j),$$

and

$$w^{\varphi}(\mathbf{x}_i) = \min_{\substack{h \neq k \\ h,k=1,\ldots,G}} \left(\frac{1}{n_h} K(\mathbf{x}_i, \mathbf{x}_i) - \frac{2}{n_h} \sum_{\mathbf{x}_j \in C_h} K(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{n_h} \sum_{\mathbf{x}_j \in C_h} K(\mathbf{x}_j, \mathbf{x}_j) \right),$$

for $\mathbf{x}_i \in C_k$.

Consequently, the Silhouette index can be computed in the feature space as

$$s^{\varphi}(\mathbf{x}_i) = \frac{w^{\varphi}(\mathbf{x}_i) - v^{\varphi}(\mathbf{x}_i)}{\max\left(v^{\varphi}(\mathbf{x}_i), w^{\varphi}(\mathbf{x}_i)\right)}.$$

Calinski-Harabasz Index

The Calinski-Harabasz index (Calinski and Harabasz, 1974; Milligan and Cooper, 1985), also called F statistic, is also a measure of inter-cluster dissimilarity (nominator) over intra-cluster dissimilarity (denominator). For n expression patterns and G clusters, the Calinski-Harabasz index CH is defined as

$$CH = \frac{\sum_{k=1}^{G} n_k \left\| \mathbf{m}_k - \mathbf{m} \right\|^2 \Big/ (G-1)}{\sum_{k=1}^{G} \sum_{\mathbf{x}_i \in C_k} \left\| \mathbf{x}_i - \mathbf{m}_k \right\|^2 \Big/ (n-G)}.$$

A larger value for CH indicates a better clustering, since the between cluster dissimilarity is then supposed to be large, while the within cluster dissimilarity is then supposed to be small. Maximum values of the CH index are often used to indicate the correct number of partitions in the data. The nominator can partially be rewritten using

$$\sum_{k=1}^{G} n_k \left\| \mathbf{m}_k - \mathbf{m} \right\|^2 = \sum_{k=1}^{G} n_k \left\| \frac{1}{n_k} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i - \frac{1}{n} \sum_{\mathbf{x}_j \in D} \mathbf{x}_j \right\|^2,$$

where n_k denotes of gene expression patterns in cluster C_k with centroid $\mathbf{m}_k$, and $\mathbf{m}$ the centroid of the entire data set D. The denominator can partially be rewritten using

$$\sum_{k=1}^{G} \sum_{\mathbf{x}_i \in C_k} \left\| \mathbf{x}_i - \mathbf{m}_k \right\|^2 = \sum_{k=1}^{G} \sum_{\mathbf{x}_i \in C_k} \left\| \mathbf{x}_i - \frac{1}{n_k} \sum_{\mathbf{x}_l \in C_k} \mathbf{x}_l \right\|^2.$$

Therefore, the CH index is can be written as

$$CH = \frac{\sum_{k=1}^{G} n_k \left\| \frac{1}{n_k} \sum_{\mathbf{x}_i \in C_k} \mathbf{x}_i - \frac{1}{n} \sum_{\mathbf{x}_j \in D} \mathbf{x}_j \right\|^2 \Big/ (G-1)}{\sum_{k=1}^{G} \sum_{\mathbf{x}_i \in C_k} \left\| \mathbf{x}_i - \frac{1}{n_k} \sum_{\mathbf{x}_l \in C_k} \mathbf{x}_l \right\|^2 \Big/ (n-G)}.$$

By introducing the feature map $\varphi(\cdot)$, the CH index can be expressed as

$$CH^{\varphi} = \frac{\sum_{k=1}^{G} n_k \left\| \frac{1}{n_k} \sum_{\mathbf{x}_i \in C_k} \varphi(\mathbf{x}_i) - \frac{1}{n} \sum_{\mathbf{x}_j \in D} \varphi(\mathbf{x}_j) \right\|^2 \Big/ (G-1)}{\sum_{k=1}^{G} \sum_{\mathbf{x}_i \in C_k} \left\| \varphi(\mathbf{x}_i) - \frac{1}{n_k} \sum_{\mathbf{x}_l \in C_k} \varphi(\mathbf{x}_l) \right\|^2 \Big/ (n-G)}.$$

Similarly as done for the Global Silhouette index, the kernel trick can then be applied when calculating the CH index in feature space.

Distortion Score

The mean squared error criterion, which is the objective function in both classical and kernel K-means clustering, can be used for internal validation. In this context, the mean squared error criterion is called the Distortion score.

For a data set $\{\mathbf{x}_i\}_{i=1}^N$ with expression patterns $\mathbf{x}_i \in \Re^d$, the Distortion score is formulated as

$$\text{se} = \sum_{k=1}^{G}\sum_{i=1}^{N} z_{C_k,\mathbf{x}_i} \left\|\mathbf{x}_i - \mathbf{m}_k\right\|^2,$$

with the indicator function $z_{C_k,\mathbf{x}_i}$ defined as

$$z_{C_k,\mathbf{x}_i} = \begin{cases} 1 & \text{if } \mathbf{x}_i \in C_k; \\ 0 & \text{otherwise,} \end{cases}$$

with

$$\sum_{k=1}^{G} z_{C_k,\mathbf{x}_i} = 1 \qquad \forall i,$$

and the centroid (or prototype) $\mathbf{m}_k$ of cluster C_k defined as

$$\mathbf{m}_k = \frac{1}{n_k}\sum_{j=1}^{N} z_{C_k\mathbf{x}_j}\mathbf{x}_j \ .$$

In the feature space, the Distortion score can be expressed by

$$\begin{aligned} \text{se}^{\varphi} &= \sum_{k=1}^{G}\sum_{i=1}^{N} z_{C_k,\mathbf{x}_j} \left\|\varphi(\mathbf{x}_i) - \mathbf{m}_k^{\varphi}\right\|^2 \\ &= \sum_{k=1}^{G}\sum_{i=1}^{N} z_{C_k,\mathbf{x}_i} \left(\mathbf{K}_{ii} + f(C_k,\mathbf{x}_i) + g(C_k)\right), \end{aligned}$$

with $f(C_k,\mathbf{x}_i)$, $g(C_k)$ and $\mathbf{m}_k^{\varphi}$ defined as in the kernel K-means algorithm in Chapter 2.

6.5.2 External validation

Rand Index

The Rand index (Rand, 1971; Yeung *et al.*, 2001a; Yeung *et al.*, 2001b) is a measure that reflects the level of agreement of a cluster result with an external partition, for example, an existing partition of a known cluster structure of the data. This external criterion could for example be the existing diagnostic classes generally used by experts in clinical practice (e.g., groups of patients with a similar type of cancer, groups of patients responding to therapy in a similar way, or groups of patients with a similar kind of diagnosis), a predefined cluster structure if one is clustering synthetic data where the clusters are known in advance, or another cluster result obtained using other parameter settings for a specific clustering algorithm or obtained using other clustering algorithms. Note that the latter could be used to investigate how sensitive a cluster result is to the choice of the algorithm or parameter setting. If this result proves to be relatively stable, one could assume that pronounced structures are present in the data possibly reflecting subcategories that are clinically relevant.

Suppose one wants to compare two partitions (the cluster result at hand and the external criterion) of a set of N expression patterns. Suppose that s is the number of expression pattern pairs that are placed in the same subset (or cluster) in both partitions. Suppose that v is the number of expression pattern pairs that are placed in different subsets in both partitions. The Rand index is then defined as the fraction of agreement between both partitions

$$r = \frac{s+v}{M},$$

with M the maximum number of all expression pattern pairs in the data set, namely $M = N(N-1)/2$. This can also be rewritten by $M = s+t+u+v$, with t the number of expression pattern pairs that are placed in the same cluster according to the external criterion but in different clusters according to the cluster result, and u the number of expression pattern pairs that are placed in the same cluster according to the cluster result but in different clusters according to the external criterion. The Rand index lies between 0 and 1 (1 if both partitions are identical) and can be viewed as the proportion of agreeing expression pattern pairs between two partitions.

Adjusted Rand Index

One disadvantage of the Rand index is that the expected value of two random partitions is not a constant value and depends on the number of clusters G (Yeung *et al.*, 2001a). In order to compare clustering results with different numbers of clusters G, the Adjusted Rand index is proposed by Hubert and Arabie (1985).

This index assumes that the u and v partitions are picked at random such that the number of objects in the classes and clusters are fixed. Let n_{ij} be the number of expression patterns that are in both cluster u_i (according to the external criterion) and cluster v_j (according to the cluster result). Let $n_{i\cdot}$ and $n_{\cdot j}$ be the number of expression patterns in cluster u_i and cluster v_j, respectively. The general form of an index with a constant expected value is

$$\frac{\text{index} - \text{expected index}}{\text{maximum index} - \text{expected index}},$$

which is bounded above by 1 and below by -1, and has an expected value of 0 for random clustering. Under the assumptions, it can be shown (Hubert and Arabie, 1985) that

$$E\left[\sum_{i,j}\binom{n_{ij}}{2}\right] = \left(\sum_{i}\binom{n_{i\cdot}}{2}\sum_{j}\binom{n_{\cdot j}}{2}\right)\Big/\binom{n}{2}.$$

The expression $s + v$ can be simplified to a linear transformation

$$\sum_{i,j}\binom{n_{ij}}{2},$$

According to Hubert and Arabie (1985), the Adjusted Rand index ar can then be simplified by

$$ar = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \left(\sum_{i}\binom{n_{i\cdot}}{2}\sum_{j}\binom{n_{\cdot j}}{2}\right)\Big/\binom{n}{2}}{\frac{1}{2}\left(\sum_{i}\binom{n_{i\cdot}}{2} + \sum_{j}\binom{n_{\cdot j}}{2}\right) - \left(\sum_{i}\binom{n_{i\cdot}}{2}\sum_{j}\binom{n_{\cdot j}}{2}\right)\Big/\binom{n}{2}}.$$

6.6 Experiments

In this section, the clustering and cluster validation methods described are demonstrated on acute leukemia data (Golub *et al.*, 1999) and colon cancer data (Alon *et al.*, 1999), which are described in more detail in Appendix A. Specific preprocessing steps for both data sets are also discussed in Appendix A. Further preprocessing of both data sets is done by standardization, as discussed in Chapter 3. For classical K-means, this is followed by PCA (without selection of principal components), also discussed in Chapter 2.

Tuning of the parameters (number of clusters, random initialization, and so on) in the context of the classical clustering techniques has already been discussed previously in a large number of publications. We therefore only refer to some of these studies (Halkidi and Vazirgiannis, 2005; Handl *et al.*, 2005). However, since kernel clustering methods also require tuning of the kernel parameters, some research effort still needs to be performed on this subject. Note that classical K-means clustering and kernel K-means clustering with a linear kernel require optimization of number of clusters and the random initialization. Kernel K-means with an RBF kernel and spectral clustering, however, also require the additional optimization of the kernel parameter σ. Tuning these parameters needs to be performed based on internal validation criteria. In these experiments, we choose the value for σ that corresponds to the maximum of the Global Silhouette index.

Since both data sets contain two given diagnostic categories and since we only want to illustrate these techniques, we restrict the number of clusters G to be equal to two, although in general these methods can be applied in a straightforward way for finding more than two clusters as well. Tuning of the number of clusters G has already been studied a lot, therefore we only want to focus on the optimization of the kernel parameter σ. However, future research needs to be done on more complex datasets (containing more that two clusters). The initialization is optimized by repeating each K-means or kernel K-means algorithm a hundred times, selecting the best result based on the Distortion score within these algorithms (note that this is done for each value of σ). Optimization of this kernel parameter σ is done using the Global Silhouette index. Note that only intervals for σ with meaningful cluster results are considered. For the optimal value of σ, both external validation indices (i.e., the Rand and Adjusted Rand index) are reported as well.

6.7 Results and discussion

Tuning curves (tuning the kernel parameter σ based on the Global Silhouette index) followed by a table presenting Global Silhouette, Rand and Adjusted Rand indices for the optimal kernel parameter σ, are first shown for the acute leukemia data and then for the colon cancer data.

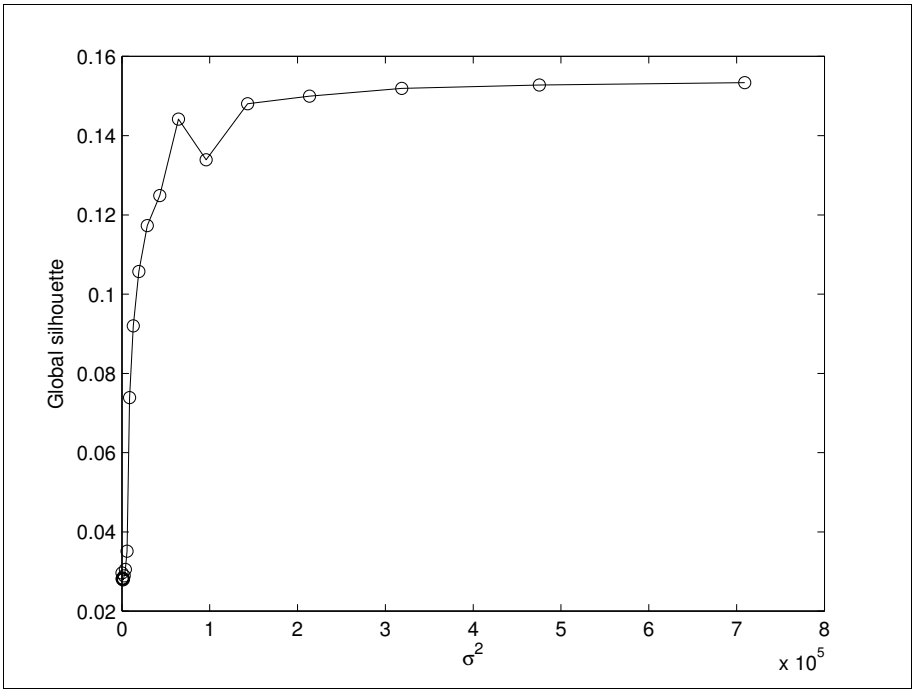


Figure 6.1 : *Tuning curve of kernel K-means clustering on the acute leukemia data. The tuning curve shows the Global Silhouette index (y -axis) for a range of values for kernel parameter σ (x -axis). Note that after a significant increase, the Global Silhouette index only slightly increases with increasing values for σ, although the distribution of the samples to the cluster partitions is stabilized. We therefore choose a very large value as the optimal value for σ^2.*

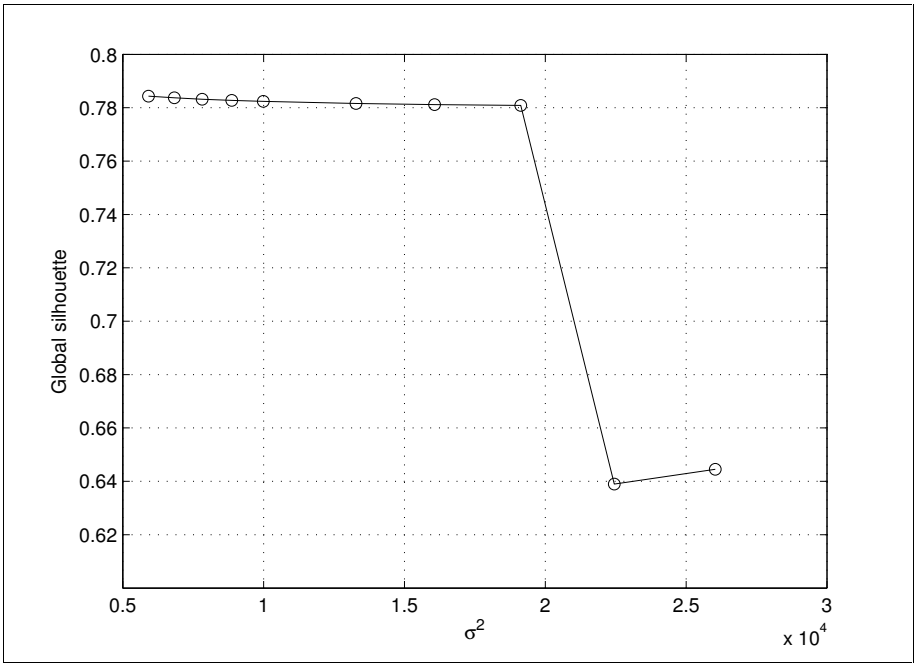


Figure 6.2 : *Tuning curve of spectral clustering on the acute leukemia data. See Figure 6.1 for more detailed information on the tuning curve. Note that the Global Silhouette index clearly shows a maximum for a* σ^2 *value of 5913.0.*

	Kernel parameter σ^2	(Kernel) Global Silhouette index	Adjusted Rand index	Rand index
K-means clustering	-	0.12988	-0.021418	0.49335
Kernel K-means clustering with linear kernel	-	0.15456	-0.017564	0.49452
Kernel K-means clustering with RBF kernel	709220.0	0.15337	-0.017564	0.49452
Spectral clustering	5913.0	0.78436	0.00258	0.49656

Table 6.1 : *Global Silhouette, Rand and Adjusted Rand indices for the optimal kernel parameter* σ *are given for all clustering methods on the acute leukemia data.*

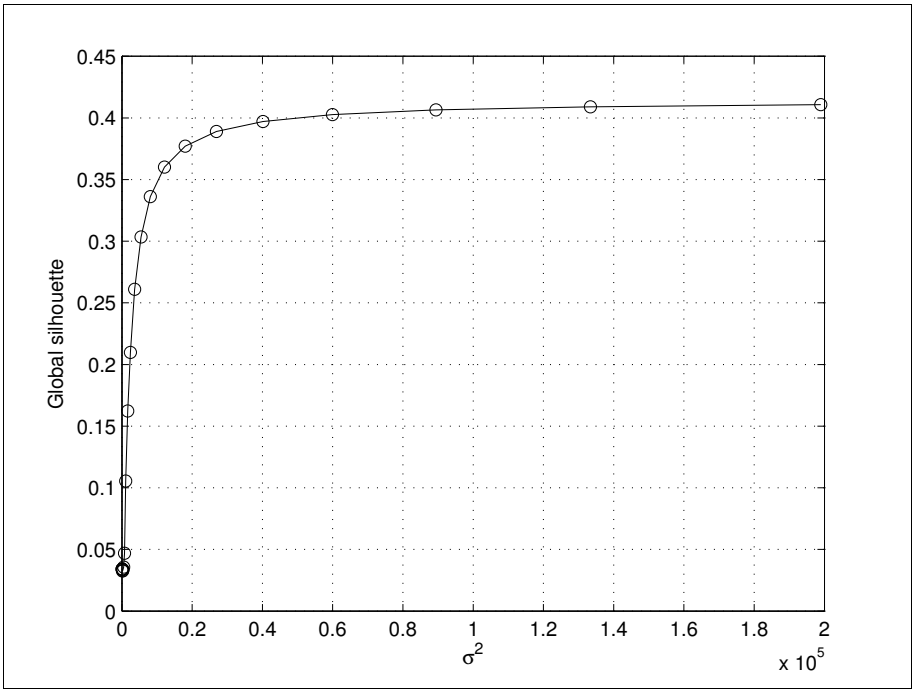


Figure 6.3 : *Tuning curve of kernel K-means clustering on the colon cancer data. See Figure 6.1 for more detailed information on the tuning curve. Note that after a significant increase, the Global Silhouette index only slightly increases with increasing values for* σ*, although the distribution of the samples to the cluster partitions is stabilized. We therefore choose a very large value as the optimal value for* σ^2.

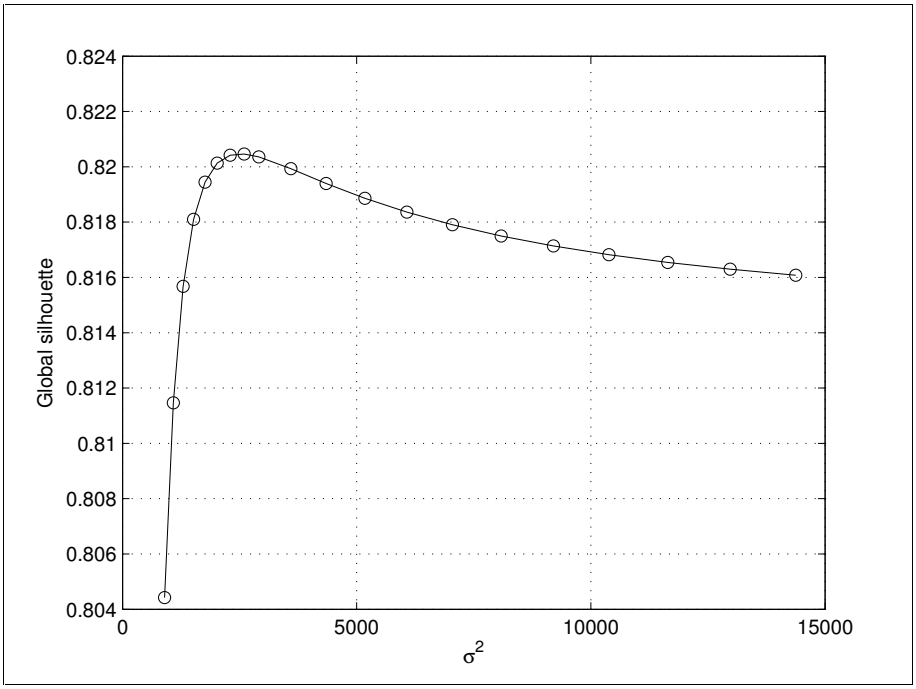


Figure 6.4 : *Tuning curve of spectral clustering on the colon cancer data. See Figure 6.1 for more detailed information on the tuning curve. Note that the Global Silhouette index clearly shows a maximum for a* σ^2 *value of 2596.4.*

	Kernel parameter σ^2	(Kernel) Global Silhouette index	Adjusted Rand index	Rand index
K-means clustering	-	0.3948	-0.0058061	0.49656
Kernel K-means clustering with linear kernel	-	0.41423	-0.0058	0.49656
Kernel K-means clustering with RBF kernel	198970.0	0.41073	-0.0058061	0.49656
Spectral clustering	2596.4	0.82046	-0.0058	0.49656

Table 6.2 : *Global Silhouette, Rand and Adjusted Rand indices for the optimal kernel parameter* σ^2 *are given for all clustering methods on the colon cancer data.*

From these results, we can conclude that spectral clustering, unlike the other clustering algorithms, gives very good and consistent clustering results in terms of the Global Silhouette index (internal validation) or its kernel version for both data sets.

It can also be observed that the results obtained by any optimally tuned clustering algorithm (classical K-means, kernel K-means with linear or RBF kernel, and spectral clustering) are not correlated to the given diagnostic categories (external partitions). However, this does not mean that these clustering results are clinically or biologically irrelevant. These could, for example, correspond to other known or unknown diagnostic categories.

6.8 Conclusion

Kernel clustering methods like kernel K-means and spectral clustering are especially designed for clustering data that contain clusters that are not linearly separable or to handle nonlinear relationships in the data. Moreover, these techniques allow for dealing with high dimensional data in a computationally more efficient way. It was shown in this chapter that these properties make kernel clustering methods specifically interesting for application on the high dimensional microarray data (with or without preprocessing steps). Using these techniques for knowledge discovery in clinical microarray data analysis may therefore allow the discovery of new clinically relevant groups in the future.

However, model selection (tuning the kernel width σ and the number of clusters) is often skipped for both kernel K-means and spectral clustering in most related publications. In this chapter, we compared classical and kernel clustering algorithms on several microarray datasets. This revealed that very good results can be obtained with spectral clustering in terms of internal validation criteria. To realize this, we showed how these internal validation measures for tuning the parameters (in this chapter we only demonstrated this for tuning the kernel parameter σ) can be extended in feature space. Nevertheless, more research is required on the usage and the efficiency of internal validation criteria and their kernel versions for tuning the parameters, possibly leading to new and more efficient measures.

Chapter 7

Conclusions and future research

7.1 General conclusions and accomplishments

The application of support vector machines and kernel methods to microarray data in this work has lead to several tangible results and observations, which we will summarize in this section. The main contributions of this thesis in terms of journal publications, conference contributions and software tools were already described in Section 1.5 in Chapter 1. We will continue this chapter with a short description of some specific clinical problems that will be studied in the future. Finally, we will conclude this dissertation with describing some challenging topics for further methodological research and also some interesting possible extensions of the M@CBETH web service towards the future.

In the past, using classification methods for developing prediction models based on microarray data has already shown to be promising for guiding clinical management in oncology. Since three of the chapters in this dissertation are focused on generating such models in a mathematical sound way, we will first formulate some general strategies that can be followed for building and assessing classification models. Further on, we will focus on the specific conclusions for each chapter.

The most common and straightforward strategy is to assess the discriminatory power of a prediction model by calculating an independent test set performance. However, this is only possible in case a substantial microarray data set is available allowing for dividing this data set in a training set and a test set. The model can then be generated based on the training set and evaluated on the test set. The test set should be large enough to obtain a reliable value for this independent test set performance. In the future, the models we generated in Chapter 5 for two cases in ovarian cancer will be assessed on a prospective data set using this strategy. This can easily be done by using the M@CBETH web service presented in Chapter 4 (in

case the desired classification method is offered by this tool). This requires the application of the benchmarking service with submission of both benchmarking (training) data and prospective (test) data. In case prospective data is only available in a later stage, it is also possible to first apply the benchmarking service on the training data and then later evaluate the prospective data based on the stored optimal model using the prediction service. In Chapter 5, we first did the actual model building by using a gene selection method (Kruskal-Wallis test for selecting the 3000 most relevant genes) followed by a classification method (LS-SVM). Next, we also used the benchmarking service in M@CBETH to generate such models.

Another strategy to assess the generalization performance of a prediction model is to calculate the leave-one-out cross-validation (LOO-CV) performance based on the complete data set. This is especially well-suited for small microarray data sets, like the ovarian cancer data set described in Chapter 5. Remember from the benchmarking study in Chapter 3 that feature (genes or principal components) selection needs to be repeated in each LOO iteration (for each LOO training set, not using the left out sample). This strategy is also followed for both ovarian cancer cases in the pilot study described in Chapter 5 (performing selection of the 3000 relevant genes in each LOO iteration). Note that the final prediction model should be developed using the complete data set. Later evaluation of this model on prospective data is possible.

A final strategy is to assess the performance of the prediction model using a mean randomized test set performance, as is calculated by the M@CBETH web service (see Figure 4.1 and Chapter 4 for more details on the algorithm). This strategy is also especially well-suited in case of small microarray data sets. This measure can easily be calculated by using the benchmarking service of the M@CBETH web service. In Chapter 5, we also applied the benchmarking service of M@CBETH on both ovarian cancer cases. The final prediction model are automatically generated by M@CBETH using the complete data set and stored for later evaluation on prospective data in the future.

Specific conclusions and accomplishments for each chapter are summarized. In Chapter 3, we investigated the influence of regularization, nonlinearity and dimensionality reduction with the aim of optimizing the performance of clinical predictions based on microarray data, taking into consideration the probability of increasing size and complexity of microarray data sets in the future. For this purpose, we performed a systematic benchmarking study based on nonlinear techniques, dimensionality reduction methods and regularization techniques. Three main conclusions were derived from this study. A first important conclusion from benchmarking nine microarray data set problems is that when performing classification with least squares SVM (without dimensionality reduction),

using an RBF kernel can be applied without risking overfitting on all data sets studied. The results obtained with an RBF kernel are never worse and sometimes even better than when using a linear kernel. A second conclusion is that using LS-SVM without regularization (without dimensionality reduction) ends up in very bad results, which stresses the importance of applying regularization even in the linear case. A final important conclusion is that when performing kernel PCA before classification, using an RBF kernel for kernel PCA tends to lead to overfitting, especially when using supervised feature selection. It has been observed that an optimal selection of a large number of features is often an indication for overfitting. Kernel PCA with linear kernel gives better results.

Nevertheless, although it was possible to derive some important general conclusions out of this study, a good classification method to build an optimal prediction model may differ for each cancer classification problem. Since it is obvious that building an optimal prediction model is of major importance with respect to using such models in clinical practice in the future, finding the best classification method in each specific case is an indispensable issue. Therefore, it remains essential to carefully consider each cancer classification problem individually. However, comparing classifiers and selecting the best for each microarray data set has been proven in Chapter 3 to be a tedious and non-straightforward task. Therefore, a web service was developed in Chapter 4. The M@CBETH (a MicroArray Classification BEnchmarking Tool on a Host server) web service offers the microarray community a simple tool for making optimal two-class predictions. This web service aims at finding the best prediction among different classification methods by using randomizations of the benchmarking data set. This way, M@CBETH intends to introduce an optimal use of clinical microarray data classification. The M@CBETH web site is freely available at http://www.esat.kuleuven.be/MACBETH/ and has already been proven to have an international impact despite its relatively young age.

Chapter 5 concentrated on the analysis of gene expression patterns coming from ovarian tumors. In a pilot study, we studied microarray experiments from ovarian tumors generated within a project we collaborated in, using a broad range of classical and linear techniques. Next, we applied the M@CBETH web service on these experiments using also nonlinear machine learning techniques to develop prediction models. Both studies investigated whether it is possible to distinguish between stage I without recurrence, platin-sensitive advanced-stage and platin-resistant advanced-stage ovarian tumors. These results of the pilot study indicated that gene expression patterns could be useful in clinical management of ovarian cancer, which was confirmed by the second study where we applied the M@CBETH benchmarking service. Furthermore, from a methodological

point of view we concluded that it is important to optimally choose an optimal classification method for each microarray data set. This latter study also indicated that further optimization of the prediction models developed in the context of the pilot study, may be possible by also considering nonlinear techniques. The remainder of this chapter was dedicated to other recently published studies that have shortcomings in the techniques used to develop prediction models.

In Chapter 6, we indicated that kernel clustering methods like kernel K-means and spectral clustering are especially designed for clustering data that contain clusters that are not linearly separable or to handle nonlinear relationships in the data. Moreover, these techniques allow for dealing with high dimensional data in a computationally more efficient way. It was shown in this chapter that these properties make kernel clustering methods specifically interesting for application on the high dimensional microarray data (with or without preprocessing steps). Using these techniques for knowledge discovery in clinical microarray data analysis may therefore allow the discovery of new clinically relevant groups in the future. In this chapter, we compared classical and kernel clustering algorithms on several microarray datasets. This revealed that very good results can be obtained with spectral clustering in terms of internal validation criteria. To realize this, we showed how these internal validation measures for tuning the parameters (in this chapter we only demonstrated this for tuning the kernel parameter σ) can be extended in feature space. Nevertheless, more research is required on the usage and the efficiency of internal validation criteria and their kernel versions for tuning the parameters, possibly leading to new and more efficient measures.

7.2 Future research

In this section we will first discuss two specific project proposals in which we are involved. In this research we aim to apply the techniques described in this dissertation for these specific clinical problems. Both projects involve the usage of microarray, proteomic and clinical data. Therefore, more advanced strategies, as outlined in the general research prospects, will be followed apart from the methodology we already used in this work.

7.2.1 Future research: clinical applications

Clinical management of ovarian cancer

This project is in collaboration with Prof. I. Vergote and Prof. D. Timmerman. In a first stage, we already carried out a pilot study on a set of cDNA microarray experiments comprising 20 ovarian tumors. As described in Chapter 5, we developed prediction models based on these data to distinguish between early-stage and advanced-stage ovarian tumors, and between chemo-resistant and chemo-sensitive advanced-stage ovarian tumors.

In a next stage, we will work on a prospective study. cDNA microarray experiments have been generated for 50 ovarian tumors, comprising four classes: early-stage disease with recurrence, early-stage disease without recurrence, chemo-resistant advanced-stage disease and chemo-sensitive advanced-stage disease. In this prospective study, we will evaluate, refine and possibly extend the models generated in the pilot study. In case prediction performances of some of these models appear to be not sufficient, the new experiments could be used to further refine the models. Moreover, these additional experiments will allow for selecting genes that are differentially expressed between the different classes with a higher efficiency. In this prospective study, we will also investigate the difference between early-stage ovarian tumors with a high and a low risk of recurrence. Both stage I classes will be situated with respect to the other classes using PCA.

For a subset of these patients, we have clinical data available. Moreover, we also obtained funds within Biopattern[1], which is a European project aiming for integrating research on eHealth within Europe, to generate proteomics experiments on a subset of these patients in the nearby future. This way, proteomic data can be combined with microarray and clinical data. Note that we also made our microarray data from the pilot study available to the partners of Biopattern.

Clinical management of breast cancer

This project is in collaboration with Prof. P. Neven and Prof. D. Timmerman. A database containing clinical information from more than 3,000 patients with breast tumors (without metastases at diagnosis) is already available (collected since the 1st of January, 2000). From these patients

[1] FP6-NoE Biopattern: EU funded network of excellence – 6th framework programme priority 2 – Project 508803: 'Computational Intelligence for biopattern analysis in Support of eHealthcare'.

tumor tissues and serum are collected as well, which allows to collect molecular information of these patients using proteomics and/or microarray technology. In the nearby future, we will apply for funding in order to generate microarray and proteomics data for a subset of these patients. The aim here is to build models that are capable of making predictions about the recurrence of the tumor outside the breast. This way it would be possible to make better decisions about adjuvant therapy in breast cancer. We will apply the techniques described in this work, as well as techniques that will be described in the general research prospects.

Moreover, within Biopattern a data set containing microarray, proteomics and clinical data from 150 breast cancer patients (with more than 10 years of follow-up) will become available in the near future.

7.2.2 Future research: methodological challenges

Several extensions of the M@CBETH web service are possible in the future in order to further optimize the performances of prediction models for usage in clinical practice. These will be discussed below.

The M@CBETH web service could easily be extended with some extra features in order to further improve its current functionality, limited to microarray data. First, it would be interesting towards clinical practice to also integrate an estimation of the prediction probabilities in order to have an idea of the reliability of the predictions for individual patients. Second, integrating and allowing multi-class predictions could be more advantageous, this way avoiding the need to derive two-class problems from each classification problem. Furthermore, new classification techniques that have been proven to have good performances in benchmarking studies similar to the one we did in Chapter 3, could be added to this web service.

Another interesting aspect when considering the application of M@CBETH on microarray data is to study and to include more gene selection methods. The importance of this is particularly situated in clinical practice, for example to make clinical predictions or to find tumor markers and drug targets. Two main strategies could be investigated. The first strategy involves performing feature selection algorithms before developing classification models. We already used classical, linear and kernel Principal Component Analysis with supervised and unsupervised principal component selection. New opportunities could be found in using supervised and unsupervised methods for selection of individual genes (univariate analysis). Another possibility is to use other methods for selection of subsets of genes, apart from PCA, like iterative methods such as Recursive Feature Elimination (Guyon *et al.*, 2002), Automatic Relevance Determination (Van Gestel *et al.*, 2001a; Van Gestel *et al.*, 2001b; Li *et al.*, 2002), LASSO

(Tibshirani, 1996). Another possibility is using Bayesian Networks (Gevaert *et al.*, 2006) for multivariate gene selection followed by for example LS-SVM for building the actual classifier, although these Bayesian Networks can also be used to generate classifiers. The second strategy involves classification techniques that obtain sparseness in the genes while optimizing the classification performance. This way, relevant underlying structures in the genes could be immediately detected while building a prediction model. Possible methods to investigate would be linear and nonlinear Componentwise LS-SVM and Structure Determination algorithms (Pelckmans *et al.*, 2005a; Pelckmans *et al.*, 2005b; Pelckmans *et al.*, 2006) and Bayesian Networks. Note that all these techniques could be integrated in the randomization algorithm of M@CBETH. For both strategies, it would be recommendable to include procedures to automatically check the biological relevance and function of the genes that are selected for building the model. Important databases that can be considered for classification problems in oncology are, among others, the database of the National Center of Biotechnology Information (NCBI) (http://www.ncbi.nih.gov/), the Cancer Profiling Database Oncomine (http://www.oncomine.org/), the Tumor Gene Database (http://www.tumor-gene.org/).

Different data sources like microarray, proteomic and clinical data may contain complementary information with respect to clinical behavior. Therefore, it would be very interesting to first study each of these data sources separately and then to combine these heterogeneous data in one (hybrid) data analysis procedure. The main goal would be to investigate how data obtained from studying the transcriptome, proteome and clinical data could be used and combined to guide the management of specific types of cancer.

In this context, a first step would be to focus on clinical data separately. The idea of developing the best classifier for each clinical microarray data set could also easily be extended towards classical clinical data sets. In the study on clinical data from endometrial tumors in De Smet *et al.* (2006), the some of the classification methods from machine learning that we use on microarray data in M@CBETH as well, resulted also in better performances than the traditional techniques. Therefore, it would be interesting to create an adapted version of M@CBETH focused on generating prediction models with good performances for clinical data.

As said before, it would be possible to obtain additional information about the molecular biology of tissues and samples by studying the proteome. Note that the methodology specifically designed to analyze these proteomic data is currently less developed because of the significant recent developments in this technology. The high dimensional nature of these data, however, suggests that these data could form a prime candidate for the application of methods that are used for microarray data analysis. When

proteomic data analysis would become more conventional, it would be interesting to create another adapted version of M@CBETH, focusing on generating prediction models with good performances for proteomic data.

Finally, the ultimate goal would be to integrate all three heterogeneous data sources (microarray, proteomic and clinical data) in one (hybrid) data analysis procedure. An important research topic would be to investigate whether it is possible to further optimize the predictions (and possibly discovering the underlying structures simultaneously) by combining microarray and proteomic data, potentially complemented with clinical data. Integration of these heterogeneous data sources in one model could be carried out by combining the data vectors, the kernel functions or the models themselves, for example by training an additional layer (Suykens *et al.*, 2002) or by using Bayesian Networks (Gevaert *et al.*, 2006). This could lead to a final and optimally adapted version of M@CBETH.

In conclusion, the use and development of the techniques mentioned in this thesis for the analysis of patient specific transcriptomic and proteomic patterns and the integration of the results in an improved version of the M@CBETH web service for usage in clinical practice will be and remain the main focus of our research.

Appendix A

Data sets

In this appendix we will list and give an overview of the characteristics of the data sets we downloaded from the Internet and that were used in this work. Furthermore, some guidelines to the specific preprocessing procedures that are required for these data sets are stated.

A.1 Colon cancer data set (Alon *et al.*, 1999)

Colon adenocarcinoma tissues were collected from patients and from some of these patients, paired normal colon tissue also was obtained. Gene expression in 40 tumor and 22 normal colon tissue samples was analyzed with an Affymetrix oligonucleotide array complementary to more than 6500 human genes. The data set contains the expression of the 2000 genes with highest minimal intensity across the 62 tissues. Each gene intensity has been derived from the about 20 feature pairs that correspond to the gene on the chip by using a filtering process. The data is otherwise unprocessed (i.e., no standardization has been performed yet).

The training set consists of 40 colon tissues of which 14 are normal and 26 tumor samples. The test set consists of 22 tissues of which 8 are normal and 14 tumor samples. The number of gene expression levels is 2000. The goal here is to classify the tissues as being cancerous or non-cancerous.

A.2 Acute leukemia data set (Golub *et al.*, 1999)

The initial leukemia data set consisted of 38 bone marrow samples obtained from adult acute leukemia patients at the time of diagnosis, before chemotherapy. RNA prepared from bone marrow mononuclear cells was hybridized to high-density oligonucleotide microarrays, produced by

Affymetrix and containing 6817 human genes. An independent collection of 34 leukemia samples contained a broader range of samples: the specimens consisted of 24 bone marrow and 10 periferal blood samples, derived from both adults and children. This collection also contained samples from different reference laboratories that used different sample preparation protocols.

The training set consists of 38 leukemia patients of who 11 suffer from acute myeloid leukemia (AML) and 27 from acute lymphoblastic leukemia (ALL). The test set consists of 34 patients of who 14 suffer from AML and 20 from ALL. The number of gene expression levels is 7129. Separating the AML samples from the ALL samples is the issue here.

The acute leukemia data have already been used frequently in previous microarray data analysis studies. Preprocessing of this data set is done by thresholding and log-transformation, similar as in the original publication. Thresholding is achieved by restricting gene expression levels to be larger than 20, e.g. expression levels which are smaller than 20 will be set to 20. Concerning the log-transformation, the natural logarithm of the expression levels is taken.

A.3 Breast cancer data set (Hedenfalk *et al.*, 2001)

RNA from samples of primary breast tumors from 7 carriers of the BRCA1 mutation, 8 carriers of the BRCA2 mutation, and 7 patients with sporadic cases of breast cancer have been hybridized to a cDNA microarray containing 6512 complementary DNA clones of 5361 genes. The goal here is to classify the different mutations, so three combinations are possible in this case.

First, tissues with BRCA1 mutations are separated from the tissues with BRCA2 or sporadic mutations. The training set consists of 14 breast cancer tissues of which 4 have a BRCA1 mutation and 10 have not. The test set consists of 8 tissues of which 3 have a BRCA1 mutation and 5 have not. The number of gene expression levels is 3226.

Second, tissues with BRCA2 mutations are separated from the tissues with BRCA1 or sporadic mutations. The training set consists of 14 breast cancer tissues of which 5 have a BRCA1 mutation and 9 have not. The test set consists of 8 tissues of which 3 have a BRCA1 mutation and 5 have not. The number of gene expression levels is 3226.

Third, tissues with sporadic mutations are separated from the tissues with BRCA1 or BRCA2 mutations. The training set consists of 14 breast

cancer tissues of which 4 have a BRCA1 mutation and 10 have not. The test set consists of 8 tissues of which 3 have a BRCA1 mutation and 5 have not. The number of gene expression levels is 3226.

A.4 Hepatocellular carcinoma data set (Iizuka *et al.*, 2003)

mRNA expression profiles in tissue specimens from a training set comprising 33 patients with hepatocellular carcinoma have been hybridized with high-density oligonucleotide microarrays representing about 6000 genes. The same has been done for a blinded set of samples from 27 newly enrolled patients. Since hepatocellular carcinoma has a poor prognosis because of the high intrahepatic recurrence rate, the goal here is to predict early intrahepatic recurrence or non-recurrence.

The training set consists of 33 hepatocellular carcinoma tissues of which 12 suffer from early intrahepatic recurrence and 21 not. The test set consists of 27 hepatocellular carcinoma tissues of which 8 suffer from early intrahepatic recurrence and 19 not. The number of gene expression levels is 7129.

A.5 High-grade glioma data set (Nutt *et al.*, 2003)

50 high-grade glioma samples were carefully selected, 28 glioblastomas and 22 anaplastic oligodendrogliomas, all were primary tumors sampled before therapy. The classic subset of tumors were cases diagnosed similarly by all examining pathologists, and each case resembled typical depictions in standard textbooks. A total of 21 classic tumors was selected, and the remaining 29 samples were considered non-classic tumors, lesions for which diagnosis might be controversial. Affymetrix arrays are used to determine the expression of about 12000 genes. The goal here is to separate the glioblastomas from the anaplastic oligodendrogliomas, which allows appropriate therapeutic decisions and prognostic estimation.

The training set consists of 21 gliomas with classic histology of which 14 are glioblastomas and 7 anaplastic oligodendrogliomas. The test set consists of 29 gliomas with non-classic histology of which 14 are glioblastomas and 15 are anaplastic oligodendrogliomas. The number of gene expression levels is 12625.

A.6 Prostate cancer data set (Singh *et al.*, 2002)

High-quality expression profiles were successfully derived from 52 prostate tumors and 50 non-tumor prostate samples from patients undergoing surgery. Oligonucleotide microarrays were containing probes for approximately 12600 genes and ESTs. Since prostate tumors are among the most heterogeneous of cancers, both histologically and clinically, the goal here is to classify tumor and non-tumor samples.

The training set consists of 102 prostate tissues of which 50 are normal and 52 tumor samples. The test set consists of 34 tissues of which 9 are normal and 25 tumor samples. The number of gene expression levels is 12600.

A.7 Breast cancer data set (Van ’t Veer *et al.*, 2002)

78 primary breast cancers (34 from patients who developed distant metastases within 5 years and 44 from patients who continue to be disease-free after a period of at least 5 years) have been selected from patients who were lymph node negative and under 55 years of age at diagnosis. Two hybridizations were carried out for each tumor using a fluorescent dye reversal technique on microarrays containing approximately 25000 human genes synthesized by an inkjet oligonucleotide technology. The goal here is to predict the presence of subclinical metastases in order to provide a strategy to select patients who would benefit from adjuvant therapy.

The training set consists of 78 breast cancer patients of which 34 develop metastases within 5 years and 44 remain disease-free within 5 years. The test set consists of 19 patients of which 12 develop metastases within 5 years and 7 remain disease-free within 5 years. The number of gene expression levels is 24188. This data set contained some missing values. Gene expression levels lacking for all patients are left out. The rest of the missing values are estimated based on the correlations between the gene expressions.

The breast cancer data set in van ’t Veer *et al*. (2002) contains missing values. Those have been estimated based on 5% of the gene expression profiles that have the largest correlation with the gene expression profile of the missing value.

Authors	URL
Alon *et al.*	http://microarray.princeton.edu/oncology/affydata/index.html
Golub *et al.*	http://www-genome.wi.mit.edu/cancer/
Hedenfalk *et al.*	http://research.nhgri.nih.gov/microarray/NEJM_Supplement/
Iizuka *et al.*	http://surgery2.med.yamaguchi-u.ac.jp/research/DNAchip/hcc-recurrence/index.html
Nutt *et al.*	http://www.broad.mit.edu/cgi-bin/cancer/data sets.cgi
Singh *et al.*	http://www.broad.mit.edu/cgi-bin/cancer/data sets.cgi
van 't Veer *et al.*	http://www.rii.com/publications/default.htm (log ratios) http://www.nature.com (suppl. inform. - degree of diff.)

Table A.1 : *Overview of the URLs of the different microarray data sets.*

Appendix B

Detailed results

In this appendix we show and discuss the detailed results (with statistical significance results) of the numerical experiments done on all cancer classification problems of Chapter 3.

B.1 Colon cancer data set (Alon *et al.*, 1999)

When looking at the results shown in Table B.1 and visualized in Figure B.1 and the statistical significance tests in Table B.2, the following statements can be derived for this data set:

1. The LOO-CV performances of the simulations containing kernel PCA with RBF kernel are significantly better than all other LOO-CV performances. Considering both of these, the one based on the absolute value of the Golub score is also significantly better than the one using the eigenvalues for selection of the principal components. And the LOO-CV performances of LS-SVM with a linear or an RBF kernel are even slightly worse than the simulations using classical PCA or kernel PCA with a linear kernel, especially when the principal components are selected based on the eigenvalues.
2. On the other hand, when looking at the accuracy results on the test set, LS-SVM with a linear or an RBF kernel seems to perform better than all other simulations, except for the simulations using classical or kernel PCA with a linear kernel when using the eigenvalues for selection of the principal components. The test set accuracy when using kernel PCA with an RBF kernel and selection of the principal components by using the absolute value of the Golub score is clearly much worse than all other simulations.
3. For the area under the ROC curve of the test set, the simulation using kernel PCA with an RBF kernel and selection of the principal

Alon et al. (1999)

Results (LOO-CV performances, training and test set accuracies, and training and test ROC performances)

	LOO-CV performance	ACC training set	ACC test set	AUC training set	AUC test set
(1) LS-SVM linear kernel	83.33±5.47	99.64±0.87	82.03±7.49	99.99±0.06	84.78±7.39
(2) LS-SVM RBF kernel	83.69±5.38	98.33±2.36	81.39±9.19	99.80±0.38	84.91±7.72
(3) LS-SVM linear kernel (no regularization)	53.57±14.57	49.40±8.93	51.73±12.19	48.59±12.82	51.74±12.65
(4) PCA + FDA (unsupervised PC selection)	86.43±4.12	90.95±5.32	80.30±9.65	95.15±3.82	85.29±8.41
(5) PCA + FDA (supervised PC selection)	84.76±6.17	95.24±5.56	76.84±7.41	97.33±4.34	83.08±8.49
(6) kPCA lin + FDA (unsupervised PC selection)	86.43±4.12	90.95±5.32	80.30±9.65	95.15±3.82	85.29±8.41
(7) kPCA lin + FDA (supervised PC selection)	84.52±6.44	95.24±5.56	76.84±7.41	97.33±4.34	83.40±8.89
(8) kPCA RBF + FDA (unsupervised PC selection)	88.21±5.07	87.86±11.24	75.11±15.02	91.95±11.41	80.72±14.71
(9) kPCA RBF + FDA (supervised PC selection)	99.76±1.06	100.00±0.00	64.07±1.94	100.00±0.00	51.28±5.70

Table B.1 : *Colon cancer data set (Alon et al., 1999): results (LOO-CV performances, training and test set accuracies, and training and test ROC performances) of all numerical experiments.*

Alon et al. (1999)

Statistical significance tests for LOO-CV performances

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.507813	0.000058	0.001113	0.128135	0.001113	0.192185	0.000167	0.000054
(2)	0.507813	1.000000	0.000058	0.001099	0.180288	0.001099	0.283767	0.000155	0.000054
(3)	0.000058	0.000058	1.000000	0.000057	0.000058	0.000057	0.000058	0.000057	0.000036
(4)	0.001113	0.001099	0.000057	1.000000	0.056524	1.000000	0.044549	0.001953	0.000079
(5)	0.128135	0.180288	0.000058	0.056524	1.000000	0.056524	1.000000	0.000703	0.000082
(6)	0.001113	0.001099	0.000057	1.000000	0.056524	1.000000	0.044549	0.001953	0.000079
(7)	0.192185	0.283767	0.000058	0.044549	1.000000	0.044549	1.000000	0.000670	0.000082
(8)	0.000167	0.000155	0.000057	0.001953	0.000703	0.001953	0.000670	1.000000	0.000071
(9)	0.000054	0.000054	0.000036	0.000079	0.000082	0.000079	0.000082	0.000071	1.000000

Alon et al. (1999)

Statistical significance tests for test set accuracies

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.507813	0.000059	0.210449	0.010278	0.210449	0.010278	0.006836	0.000082
(2)	0.507813	1.000000	0.000059	0.484375	0.015219	0.484375	0.015219	0.039063	0.000127
(3)	0.000059	0.000059	1.000000	0.000059	0.000059	0.000059	0.000059	0.000074	0.001071
(4)	0.210449	0.484375	0.000059	1.000000	0.050996	1.000000	0.050996	0.064453	0.000149
(5)	0.010278	0.015219	0.000059	0.050996	1.000000	0.050996	1.000000	0.895670	0.000185
(6)	0.210449	0.484375	0.000059	1.000000	0.050996	1.000000	0.050996	0.064453	0.000149
(7)	0.010278	0.015219	0.000059	0.050996	1.000000	0.050996	1.000000	0.895670	0.000185
(8)	0.006836	0.039063	0.000074	0.064453	0.895670	0.064453	0.895670	1.000000	0.011523
(9)	0.000082	0.000127	0.001071	0.000149	0.000185	0.000149	0.000185	0.011523	1.000000

Alon et al. (1999)

Statistical significance tests for test set ROC performances

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.859650	0.000059	0.513192	0.135862	0.513192	0.211331	0.454919	0.000068
(2)	0.859650	1.000000	0.000059	0.726978	0.098283	0.726978	0.143702	0.395416	0.000059
(3)	0.000059	0.000059	1.000000	0.000059	0.000059	0.000059	0.000059	0.000059	0.986130
(4)	0.513192	0.726978	0.000059	1.000000	0.098547	1.000000	0.156365	0.091309	0.000069
(5)	0.135862	0.098283	0.000059	0.098547	1.000000	0.098547	1.000000	0.985101	0.000069
(6)	0.513192	0.726978	0.000059	1.000000	0.098547	1.000000	0.156365	0.091309	0.000069
(7)	0.211331	0.143702	0.000059	0.156365	1.000000	0.156365	1.000000	0.895996	0.000069
(8)	0.454919	0.395416	0.000059	0.091309	0.985101	0.091309	0.895996	1.000000	0.000140
(9)	0.000068	0.000059	0.986130	0.000069	0.000069	0.000069	0.000069	0.000140	1.000000

Table B.2 : *Colon cancer data set (Alon et al., 1999): statistical significance tests for the LOO-CV performances (upper part), the test set accuracies (middle part), and the test ROC performances (lower part) of all numerical experiments. The numerical experiments are numbered as defined in Table B.1.*

components by using the absolute value of the Golub score is clearly very bad. All other simulations perform similarly.

4. For this data set, selection of the principal components by using the eigenvalues always seems to result in better performances than when using the absolute value of the Golub score.

Remark that the LOO-CV performance is not a good indicator for the accuracy or the area under the ROC curve of the test set.

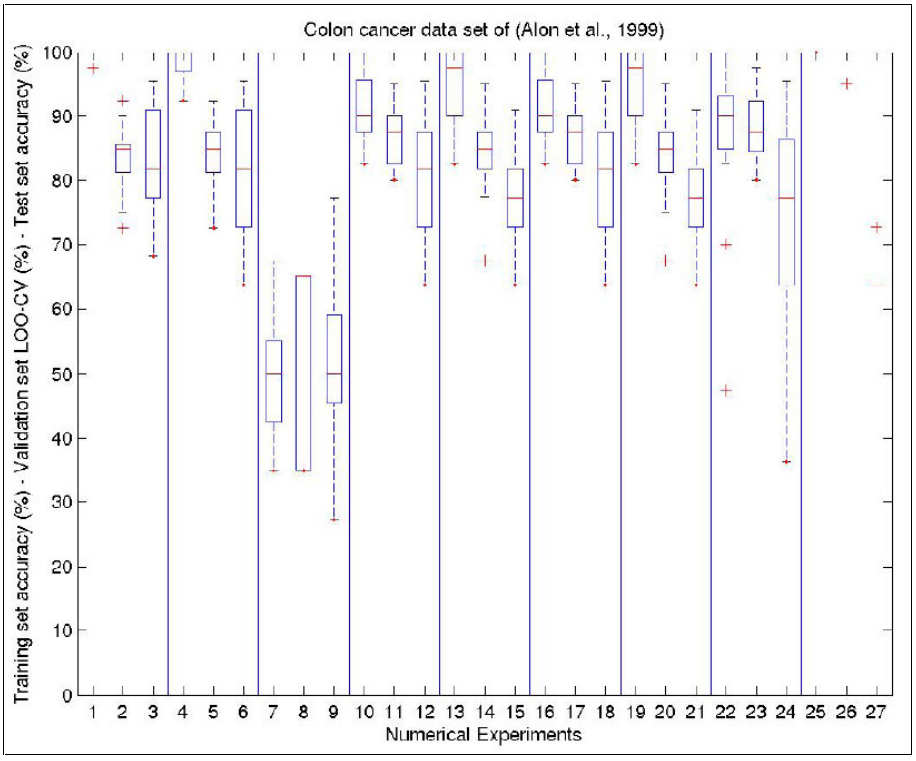


Figure B.1 : *Colon cancer data set (Alon et al., 1999): boxplots representing the training set accuracy (first), the LOO-CV performance (second) and the test set accuracy (third) of all numerical experiments.* ***Legend:*** *1,2,3 = LS-SVM with a linear kernel; 4,5,6 = LS-SVM with an RBF kernel; 7,8,9 = LS-SVM with a linear kernel without regularization; 10,11,12 = PCA + FDA (unsupervised principal component selection); 13,14,15 = performance of PCA + FDA (supervised principal component selection); 16,17,18 = kernel PCA with a linear kernel + FDA (unsupervised principal component selection); 19,20,21 = kernel PCA with a linear kernel + FDA (supervised principal component selection); 22,23,24 = kernel PCA with an RBF kernel + FDA (unsupervised principal component selection); 25,26,27 = kernel PCA with an RBF kernel + FDA (supervised principal component selection).*

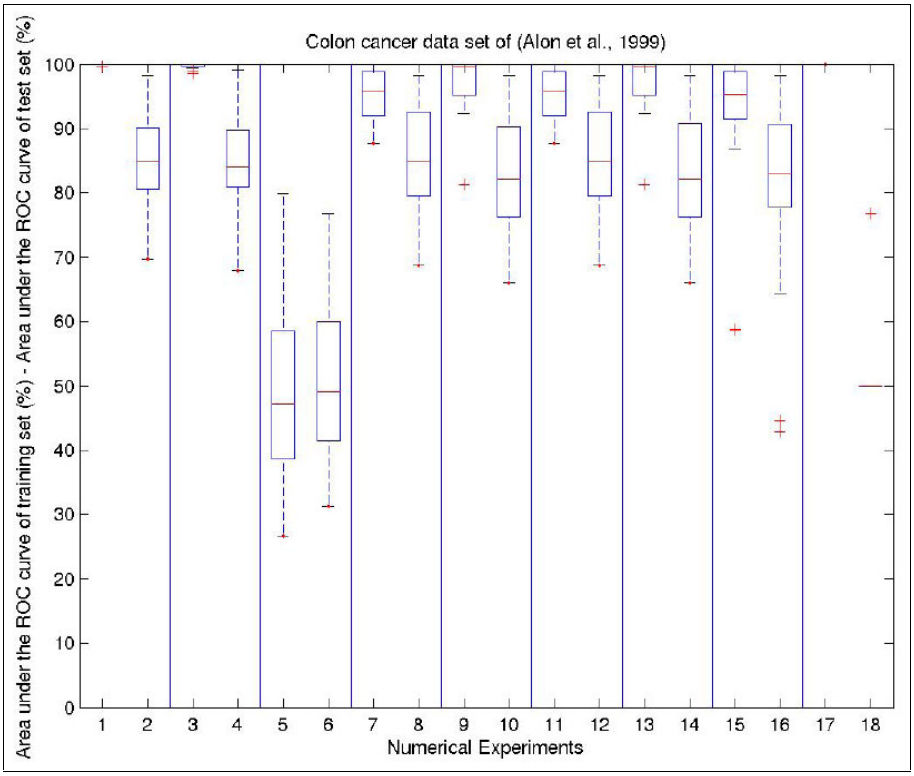


Figure B.2 : *Colon cancer data set (Alon et al., 1999): boxplots representing the training set AUC (first) and the test set AUC (second) of all numerical experiments.* ***Legend:*** *1,2 = LS-SVM with a linear kernel; 3,4 = LS-SVM with an RBF kernel; 5,6 = LS-SVM with a linear kernel without regularization; 7,8 = PCA + FDA (unsupervised principal component selection); 9,10 = performance of PCA + FDA (supervised principal component selection); 11,12 = kernel PCA with a linear kernel + FDA (unsupervised principal component selection); 13,14 = kernel PCA with a linear kernel + FDA (supervised principal component selection); 15,16 = kernel PCA with an RBF kernel + FDA (unsupervised principal component selection); 17,18 = kernel PCA with an RBF kernel + FDA (supervised principal component selection).*

B.2 Acute leukemia data set (Golub *et al.*, 1999)

The following statements can be derived for this data set when looking at the results shown in Table B.3 and visualized in Figures B.3 and B.4 and the statistical significance tests in Table B.4:

1. All simulations with dimensionality reduction have a better LOO-CV performance than the simulations using LS-SVM. On its turn, LS-SVM

Golub et al. (1999)

Results (LOO-CV performances, training and test set accuracies, and training and test ROC performances)

	LOO-CV performance	ACC training set	ACC test set	AUC training set	AUC test set
(1) LS-SVM linear kernel	94.99±2.68	100.00±0.00	92.86±4.12	100.00±0.00	98.35±1.85
(2) LS-SVM RBF kernel	95.99±2.52	100.00±0.00	93.56±4.12	100.00±0.00	98.33±1.85
(3) LS-SVM linear kernel (no regularization)	57.02±19.85	93.61±15.93	87.39±14.61	94.15±14.77	93.04±13.86
(4) PCA + FDA (unsupervised PC selection)	97.62±2.28	99.50±1.31	94.40±3.84	99.98±0.07	98.16±1.76
(5) PCA + FDA (supervised PC selection)	97.87±2.24	99.50±1.31	93.56±4.59	99.97±0.10	97.13±4.04
(6) kPCA lin + FDA (unsupervised PC selection)	97.62±2.28	99.50±1.31	94.40±3.84	99.98±0.07	98.16±1.76
(7) kPCA lin + FDA (supervised PC selection)	98.12±2.17	99.62±1.23	92.44±8.05	99.98±0.07	95.37±9.37
(8) kPCA RBF + FDA (unsupervised PC selection)	98.37±1.90	99.12±1.69	89.50±9.41	99.97±0.10	93.64±11.35
(9) kPCA RBF + FDA (supervised PC selection)	98.37±1.90	99.62±0.92	92.02±6.36	99.97±0.10	94.25±10.18

Table B.3 : *Acute leukemia data set (Golub et al., 1999): results (LOO-CV performances, training and test set accuracies, and training and test ROC performances) of all numerical experiments.*

Golub et al. (1999)

Statistical significance tests for LOO-CV performances

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.015625	0.000057	0.000061	0.000352	0.000061	0.000228	0.000175	0.000170
(2)	0.015625	1.000000	0.000056	0.001953	0.003906	0.001953	0.001953	0.000122	0.000122
(3)	0.000057	0.000056	1.000000	0.000052	0.000054	0.000052	0.000052	0.000052	0.000052
(4)	0.000061	0.001953	0.000052	1.000000	1.000000	1.000000	0.648438	0.031250	0.117188
(5)	0.000352	0.003906	0.000054	1.000000	1.000000	1.000000	1.000000	0.218750	0.125000
(6)	0.000061	0.001953	0.000052	1.000000	1.000000	1.000000	0.648438	0.031250	0.117188
(7)	0.000228	0.001953	0.000052	0.648438	1.000000	0.648438	1.000000	0.570313	0.562500
(8)	0.000175	0.000122	0.000052	0.031250	0.218750	0.031250	0.570313	1.000000	1.000000
(9)	0.000170	0.000122	0.000052	0.117188	0.125000	0.117188	0.562500	1.000000	1.000000

Golub et al. (1999)

Statistical significance tests for test set accuracies

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.437500	0.041016	0.075928	0.136108	0.075928	0.363159	0.168736	0.952148
(2)	0.437500	1.000000	0.021484	0.351196	0.552979	0.351196	0.918945	0.043019	0.433594
(3)	0.041016	0.021484	1.000000	0.003628	0.037113	0.003628	0.125500	0.775684	0.100695
(4)	0.075928	0.351196	0.003628	1.000000	0.494141	1.000000	0.451660	0.007813	0.046021
(5)	0.136108	0.552979	0.037113	0.494141	1.000000	0.494141	1.000000	0.029175	0.140625
(6)	0.075928	0.351196	0.003628	1.000000	0.494141	1.000000	0.451660	0.007813	0.046021
(7)	0.363159	0.918945	0.125500	0.451660	1.000000	0.451660	1.000000	0.131348	0.436523
(8)	0.168736	0.043019	0.775684	0.007813	0.029175	0.007813	0.131348	1.000000	0.262955
(9)	0.952148	0.433594	0.100695	0.046021	0.140625	0.046021	0.436523	0.262955	1.000000

Golub et al. (1999)

Statistical significance tests for test set ROC performances

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	1.000000	0.310303	0.509766	0.568848	0.509766	0.568912	0.049377	0.016072
(2)	1.000000	1.000000	0.310303	0.553467	0.622854	0.553467	0.622913	0.061584	0.016072
(3)	0.310303	0.310303	1.000000	0.497437	0.660175	0.497437	0.698056	0.678893	0.958726
(4)	0.509766	0.553467	0.497437	1.000000	0.319767	1.000000	0.236236	0.015625	0.020259
(5)	0.568848	0.622854	0.660175	0.319767	1.000000	0.319767	1.000000	0.586069	0.044922
(6)	0.509766	0.553467	0.497437	1.000000	0.319767	1.000000	0.236236	0.015625	0.020259
(7)	0.568912	0.622913	0.698056	0.236236	1.000000	0.236236	1.000000	0.758247	0.198242
(8)	0.049377	0.061584	0.678893	0.015625	0.586069	0.015625	0.758247	1.000000	0.371747
(9)	0.016072	0.016072	0.958726	0.020259	0.044922	0.020259	0.198242	0.371747	1.000000

Table B.4 : *Acute leukemia data set (Golub et al., 1999): statistical significance tests for the LOO-CV performances (upper part), the test set accuracies (middle part), and the test ROC performances (lower part) of all numerical experiments. The numerical experiments are numbered as defined in Table B.1.*

with an RBF kernel also has a better LOO-CV performance than when using a linear kernel.

2. The test set accuracy on the other hand is quite similar for all simulations.
3. About the area under the ROC curve of the test set, the simulations using kernel PCA with an RBF kernel seem to end up in slightly worse results than all other simulations.
4. For this data set both methods for selection of the principal components seem to give similar results.

This data set clearly comprises an easy classification problem, since the variances on the results caused by the randomizations are quite small compared to other data sets. All simulations also seem to end up in quite similar results, so in fact it hardly matters which classification method is applied on this data set.

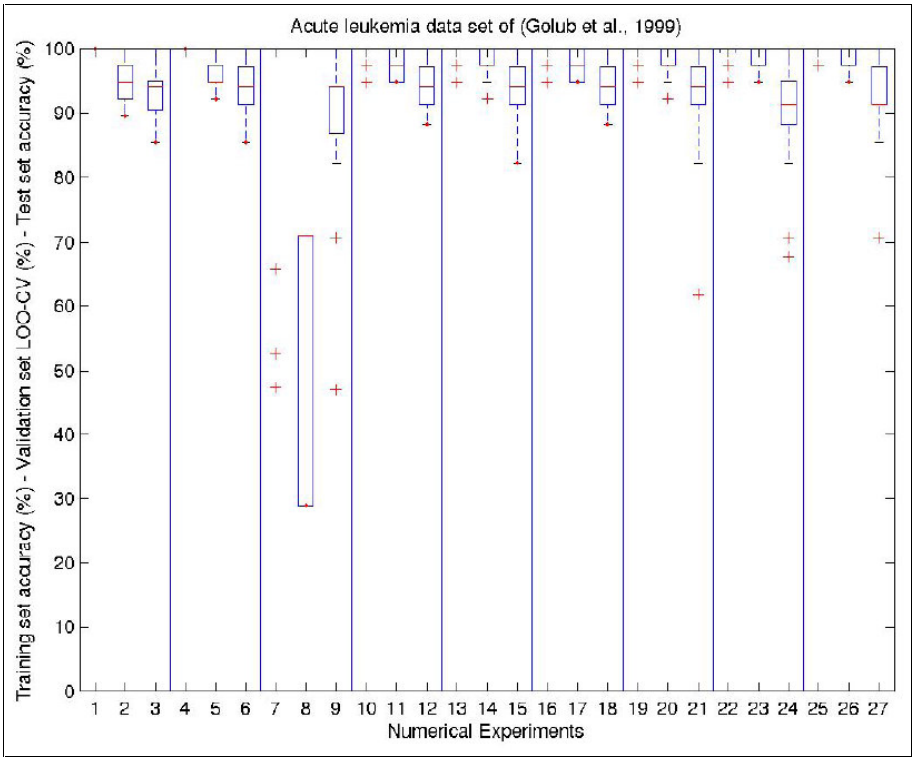


Figure B.3 : *Acute leukemia data set (Golub et al., 1999): boxplots representing the training set accuracy (first), the LOO-CV performance (second) and the test set accuracy (third) of all numerical experiments.* ***Legend:*** *See Figure B.1.*

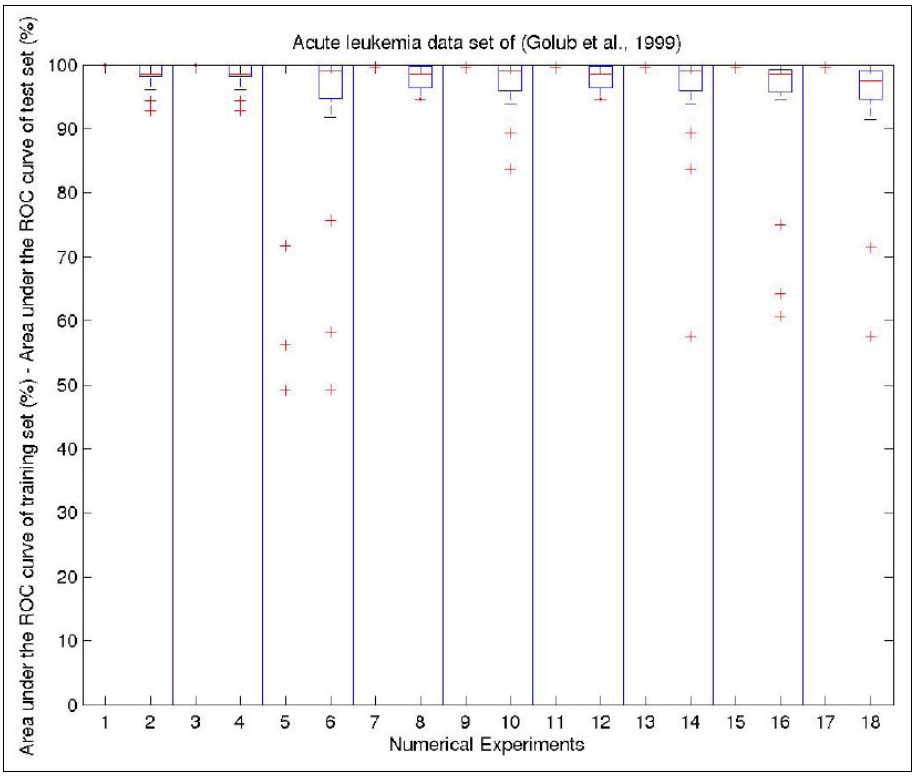


Figure B.4 : *Acute leukemia data set (Golub et al., 1999): boxplots representing the training set AUC (first) and the test set AUC (second) of all numerical experiments.* ***Legend:*** *See Figure B.2.*

B.3 Breast cancer data set (Hedenfalk *et al.*, 2001): BRCA1 mutations versus the rest

Similarly, when looking at the results shown in Table B.5 and visualized in Figures B.5 and B.6 and the statistical significance tests in Table B.6, the following statements can be derived for this data set:

1. The LOO-CV performances of the simulations containing kernel PCA with an RBF kernel are much better than all other LOO-CV performances. And the LOO-CV performance of LS-SVM with a linear kernel is also slightly worse than the rest of the simulations.
2. For the test set accuracies, LS-SVM with an RBF kernel obviously performs better than all other simulations. Using an RBF kernel when

Hedenfalk et al. (2001): BRCA1 mutations versus the rest

Results (LOO-CV performances, training and test set accuracies, and training and test ROC performances)

	LOO-CV performance	ACC training set	ACC test set	AUC training set	AUC test set
(1) LS-SVM linear kernel	78.23±7.13	87.76±14.14	64.29±6.99	100.00±0.00	81.90±18.19
(2) LS-SVM RBF kernel	82.65±8.12	98.64±6.08	75.00±12.20	100.00±0.00	82.22±17.38
(3) LS-SVM linear kernel (no regularization)	46.94±21.21	47.62±9.94	52.98±19.25	47.14±14.38	52.70±24.16
(4) PCA + FDA (unsupervised PC selection)	81.63±7.17	95.24±7.09	64.29±12.96	93.93±12.67	67.62±21.83
(5) PCA + FDA (supervised PC selection)	84.01±9.58	97.96±4.49	68.45±15.25	97.86±5.25	71.75±21.12
(6) kPCA lin + FDA (unsupervised PC selection)	81.29±7.13	95.24±6.73	63.10±13.07	96.55±5.64	66.35±20.23
(7) kPCA lin + FDA (supervised PC selection)	84.35±8.99	98.30±4.36	67.86±15.70	98.45±4.12	72.38±22.23
(8) kPCA RBF + FDA (unsupervised PC selection)	91.16±7.28	94.90±6.29	54.17±11.79	95.36±7.98	60.63±16.25
(9) kPCA RBF + FDA (supervised PC selection)	92.52±5.16	98.30±5.36	63.69±10.85	97.68±7.72	64.13±18.54

Table B.5 : *Breast cancer data set (Hedenfalk et al., 2001): BRCA1 mutations versus the rest: results (LOO-CV performances, training and test set accuracies, and training and test ROC performances) of all numerical experiments.*

Hedenfalk et al. (2001): BRCA1 mutations versus the rest

Statistical significance tests for LOO-CV performances

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.001953	0.000061	0.012207	0.012711	0.032227	0.004649	0.000051	0.000069
(2)	0.001953	1.000000	0.000186	0.503906	0.445190	0.395996	0.223877	0.000204	0.000115
(3)	0.000061	0.000186	1.000000	0.000123	0.000209	0.000124	0.000191	0.000055	0.000055
(4)	0.012207	0.503906	0.000123	1.000000	0.173828	0.984375	0.023438	0.000220	0.000097
(5)	0.012711	0.445190	0.000209	0.173828	1.000000	0.155518	1.000000	0.000549	0.000383
(6)	0.032227	0.395996	0.000124	0.984375	0.155518	1.000000	0.056396	0.000158	0.000080
(7)	0.004649	0.223877	0.000191	0.023438	1.000000	0.056396	1.000000	0.000671	0.000061
(8)	0.000051	0.000204	0.000055	0.000220	0.000549	0.000158	0.000671	1.000000	0.127930
(9)	0.000069	0.000115	0.000055	0.000097	0.000383	0.000080	0.000061	0.127930	1.000000

Hedenfalk et al. (2001): BRCA1 mutations versus the rest

Statistical significance tests for test set accuracies

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.000977	0.021033	0.862305	0.273193	0.623047	0.319422	0.007813	0.872559
(2)	0.000977	1.000000	0.000368	0.013485	0.050781	0.007346	0.033691	0.000432	0.001696
(3)	0.021033	0.000368	1.000000	0.044541	0.021836	0.057631	0.027803	0.853823	0.088921
(4)	0.862305	0.013485	0.044541	1.000000	0.288391	0.750000	0.375122	0.016968	0.905273
(5)	0.273193	0.050781	0.021836	0.288391	1.000000	0.164551	1.000000	0.007536	0.163086
(6)	0.623047	0.007346	0.057631	0.750000	0.164551	1.000000	0.210368	0.064331	0.940430
(7)	0.319422	0.033691	0.027803	0.375122	1.000000	0.210368	1.000000	0.010074	0.219727
(8)	0.007813	0.000432	0.853823	0.016968	0.007536	0.064331	0.010074	1.000000	0.033508
(9)	0.872559	0.001696	0.088921	0.905273	0.163086	0.940430	0.219727	0.033508	1.000000

Hedenfalk et al. (2001): BRCA1 mutations versus the rest

Statistical significance tests for test set ROC performances

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	1.000000	0.001035	0.014633	0.007446	0.002424	0.010010	0.000719	0.001392
(2)	1.000000	1.000000	0.000896	0.008365	0.009568	0.001837	0.011719	0.000538	0.001702
(3)	0.001035	0.000896	1.000000	0.043749	0.022721	0.075974	0.013837	0.227236	0.077732
(4)	0.014633	0.008365	0.043749	1.000000	0.445068	0.742188	0.296265	0.233930	0.420835
(5)	0.007446	0.009568	0.022721	0.445068	1.000000	0.217562	0.750000	0.098650	0.164063
(6)	0.002424	0.001837	0.075974	0.742188	0.217562	1.000000	0.156562	0.280945	0.827502
(7)	0.010010	0.011719	0.013837	0.296265	0.750000	0.156562	1.000000	0.090668	0.138672
(8)	0.000719	0.000538	0.227236	0.233930	0.098650	0.280945	0.090668	1.000000	0.270569
(9)	0.001392	0.001702	0.077732	0.420835	0.164063	0.827502	0.138672	0.270569	1.000000

Table B.6 : *Breast cancer data set (Hedenfalk et al., 2001): BRCA1 mutations versus the rest: statistical significance tests for the LOO-CV performances (upper part), the test set accuracies (middle part), and the test ROC performances (lower part) of all numerical experiments. The numerical experiments are numbered as defined in Table B.1.*

doing kernel PCA on the other hand, clearly performs worse when the eigenvalues are used for selection of the principal components.

3. The results of the area under the ROC curve of the test set show that using LS-SVM results in much better performances than all other simulations, even when using a linear kernel.
4. Both methods for selecting the principal components seem to perform very similarly, but in some cases using the absolute value of the Golub score tends to perform slightly better.

Remarkably in this case is that the test set accuracy of LS-SVM with an RBF kernel is much better than LS-SVM with a linear kernel, although the area under the ROC curve of both simulations is practically equal. This can be an indication of how important it is to find a good decision threshold value, which corresponds to an operating point on the ROC curve.

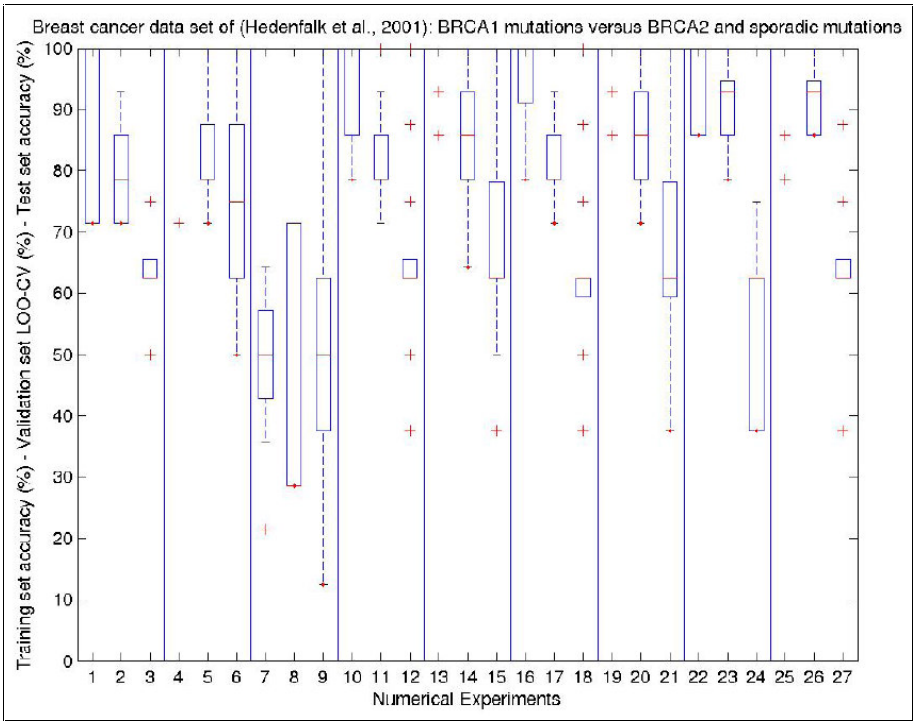


Figure B.5 : *Breast cancer data set (Hedenfalk et al., 2001): BRCA1 mutations versus the rest: boxplots representing the training set accuracy (first), the LOO-CV performance (second) and the test set accuracy (third) of all numerical experiments.* ***Legend:*** *See Figure B.1.*

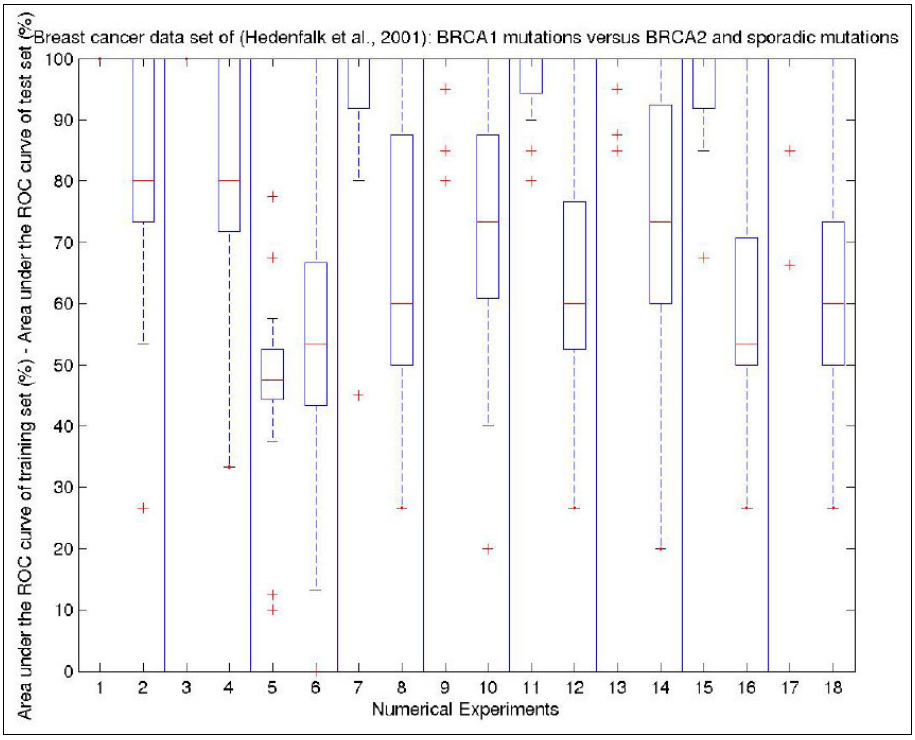


Figure B.6 : *Breast cancer data set (Hedenfalk et al., 2001): BRCA1 mutations versus the rest: boxplots representing the training set AUC (first) and the test set AUC (second) of all numerical experiments.* ***Legend:*** *See Figure B.2.*

B.4 Breast cancer data set (Hedenfalk *et al.*, 2001): BRCA2 mutations versus the rest

Similarly, the following statements can be derived for this data set when looking at the results shown in Table B.7 and visualized in Figures B.7 and B.8 and the statistical significance tests in Table B.8:

1. The LOO-CV performances of the simulations containing kernel PCA with an RBF kernel are much better than all other LOO-CV performances. And the LOO-CV performances of the simulations using LS-SVM are worse than the rest of the simulations. Using LS-SVM with a linear kernel is even worse than when using an RBF kernel.
2. On the other hand, the test set accuracies of the simulations containing kernel PCA with an RBF kernel are clearly worse than those of all other simulations.

Hedenfalk et al. (2001): BRCA2 mutations versus the rest

Results (LOO-CV performances, training and test set accuracies, and training and test ROC performances)

	LOO-CV performance	ACC training set	ACC test set	AUC training set	AUC test set
(1) LS-SVM linear kernel	78.91±10.22	94.90±12.50	84.52±16.77	100.00±0.00	95.24±9.68
(2) LS-SVM RBF kernel	84.35±7.51	100.00±0.00	88.10±15.66	100.00±0.00	94.92±10.27
(3) LS-SVM linear kernel (no regularization)	49.32±14.27	50.00±6.61	50.00±18.09	48.25±9.13	53.65±25.40
(4) PCA + FDA (unsupervised PC selection)	88.78±6.07	91.84±7.73	85.71±12.37	96.51±4.73	90.00±12.43
(5) PCA + FDA (supervised PC selection)	90.82±5.46	97.28±6.03	85.12±12.57	98.84±3.33	88.41±15.21
(6) kPCA lin + FDA (unsupervised PC selection)	89.46±6.08	92.52±7.79	84.52±12.14	96.61±4.78	89.21±12.30
(7) kPCA lin + FDA (supervised PC selection)	91.16±5.79	97.96±5.46	83.33±14.60	99.47±1.93	87.30±17.23
(8) kPCA RBF + FDA (unsupervised PC selection)	99.66±1.52	100.00±0.00	67.26±9.82	100.00±0.00	58.41±16.45
(9) kPCA RBF + FDA (supervised PC selection)	100.00±0.00	93.54±2.10	63.10±2.66	100.00±0.00	51.75±7.81

Table B.7 : *Breast cancer data set (Hedenfalk et al., 2001): BRCA2 mutations versus the rest: results (LOO-CV performances, training and test set accuracies, and training and test ROC performances) of all numerical experiments.*

Hedenfalk et al. (2001): BRCA2 mutations versus the rest

Statistical significance tests for LOO-CV performances

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.000488	0.000083	0.000408	0.000374	0.000388	0.000265	0.000053	0.000053
(2)	0.000488	1.000000	0.000055	0.006447	0.000549	0.002808	0.000122	0.000052	0.000051
(3)	0.000083	0.000055	1.000000	0.000051	0.000053	0.000054	0.000052	0.000038	0.000035
(4)	0.000408	0.006447	0.000051	1.000000	0.471558	0.500000	0.339294	0.000131	0.000106
(5)	0.000374	0.000549	0.000053	0.471558	1.000000	0.720703	1.000000	0.000070	0.000070
(6)	0.000388	0.002808	0.000054	0.500000	0.720703	1.000000	0.535339	0.000177	0.000156
(7)	0.000265	0.000122	0.000052	0.339294	1.000000	0.535339	1.000000	0.000113	0.000112
(8)	0.000053	0.000052	0.000038	0.000131	0.000070	0.000177	0.000113	1.000000	1.000000
(9)	0.000053	0.000051	0.000035	0.000106	0.000070	0.000156	0.000112	1.000000	1.000000

Hedenfalk et al. (2001): BRCA2 mutations versus the rest

Statistical significance tests for test set accuracies

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.343750	0.000140	0.957642	0.979187	0.874020	0.772949	0.001038	0.000481
(2)	0.343750	1.000000	0.000077	0.422607	0.373047	0.249146	0.252930	0.000767	0.000157
(3)	0.000140	0.000077	1.000000	0.000082	0.000050	0.000080	0.000070	0.001699	0.005326
(4)	0.957642	0.422607	0.000082	1.000000	1.000000	0.500000	0.554688	0.000582	0.000213
(5)	0.979187	0.373047	0.000050	1.000000	1.000000	0.781250	1.000000	0.000941	0.000228
(6)	0.874020	0.249146	0.000080	0.500000	0.781250	1.000000	0.863281	0.000694	0.000218
(7)	0.772949	0.252930	0.000070	0.554688	1.000000	0.863281	1.000000	0.002159	0.000380
(8)	0.001038	0.000767	0.001699	0.000582	0.000941	0.000694	0.002159	1.000000	0.125000
(9)	0.000481	0.000157	0.005326	0.000213	0.000228	0.000218	0.000380	0.125000	1.000000

Hedenfalk et al. (2001): BRCA2 mutations versus the rest

Statistical significance tests for test set ROC performances

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	1.000000	0.000194	0.033691	0.001953	0.019531	0.001953	0.000081	0.000030
(2)	1.000000	1.000000	0.000195	0.025879	0.001953	0.021729	0.001953	0.000106	0.000030
(3)	0.000194	0.000195	1.000000	0.000213	0.000253	0.000213	0.000386	0.389954	0.986101
(4)	0.033691	0.025879	0.000213	1.000000	1.000000	1.000000	1.000000	0.000296	0.000111
(5)	0.001953	0.001953	0.000253	1.000000	1.000000	0.740723	1.000000	0.000395	0.000100
(6)	0.019531	0.021729	0.000213	1.000000	0.740723	1.000000	0.799805	0.000304	0.000116
(7)	0.001953	0.001953	0.000386	1.000000	1.000000	0.799805	1.000000	0.000646	0.000118
(8)	0.000081	0.000106	0.389954	0.000296	0.000395	0.000304	0.000646	1.000000	0.125000
(9)	0.000030	0.000030	0.986101	0.000111	0.000100	0.000116	0.000118	0.125000	1.000000

Table B.8 : *Breast cancer data set (Hedenfalk et al., 2001): BRCA2 mutations versus the rest: statistical significance tests for the LOO-CV performances (upper part), the test set accuracies (middle part), and the test ROC performances (lower part) of all numerical experiments. The numerical experiments are numbered as defined in Table B.1.*

3. The same holds for the area under the ROC curve of the test set. Except that in this case using LS-SVM clearly performs even better than all other simulations.
4. Both methods for selection of the principal components seem to give similar results for this data set.

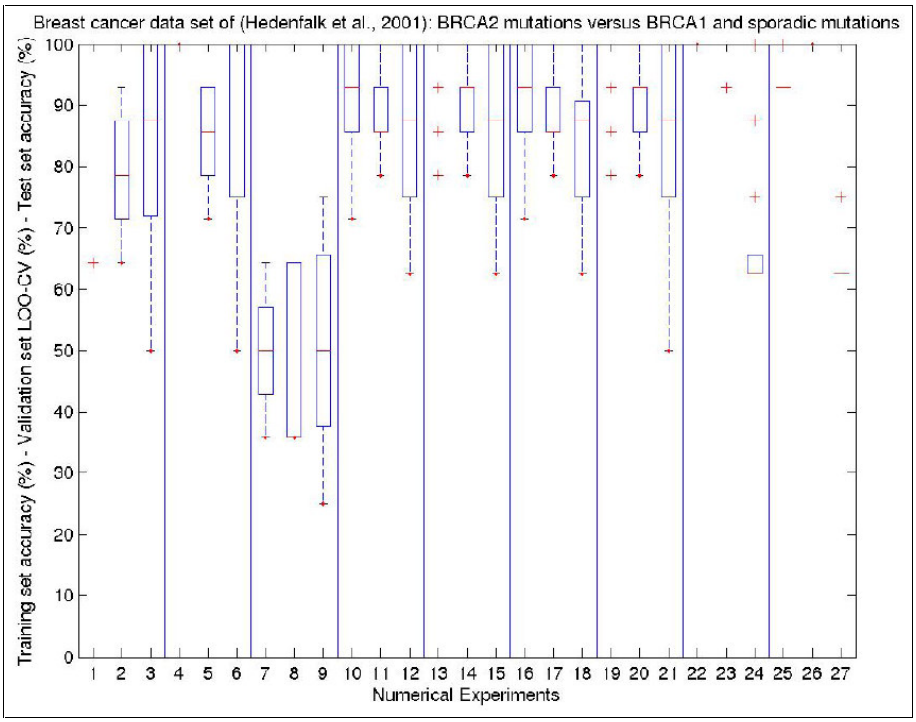


Figure B.7 : *Breast cancer data set (Hedenfalk et al., 2001): BRCA2 mutations versus the rest: boxplots representing the training set accuracy (first), the LOO-CV performance (second) and the test set accuracy (third) of all numerical experiments.* ***Legend:*** *See Figure B.1.*

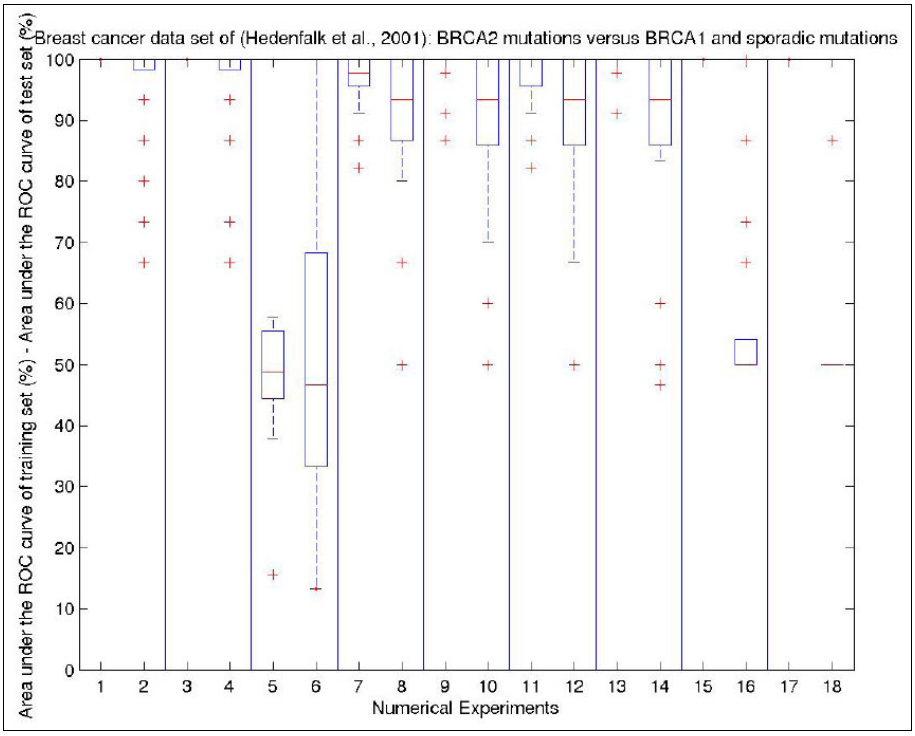


Figure B.8 : *Breast cancer data set (Hedenfalk et al., 2001): BRCA2 mutations versus the rest: boxplots representing the training set AUC (first) and the test set AUC (second) of all numerical experiments.* ***Legend:*** *See Figure B.2.*

B.5 Breast cancer data set (Hedenfalk *et al.*, 2001): sporadic mutations versus the rest

Similarly, when looking at the results shown in Table B.9 and visualized in Figures B.9 and B.10 and the statistical significance tests in Table B.10, the following statements can be derived for this data set:

1. The LOO-CV performances of the simulations containing kernel PCA with an RBF kernel are much better than all other LOO-CV performances. And the LOO-CV performance of LS-SVM with a linear kernel is also slightly worse than the rest of the simulations.
2. For the test set accuracies, all simulations seem to have similar results.
3. Using LS-SVM seems to result in much better values of the area under the ROC curve of the test set than when doing dimensionality reduction.
4. Again, both methods for selection of the principal components seem to give similar results for this data set.

Hedenfalk et al. (2001): sporadic mutations versus the rest

Results (LOO-CV performances, training and test set accuracies, and training and test ROC performances)

	LOO-CV performance	ACC training set	ACC test set	AUC training set	AUC test set
(1) LS-SVM linear kernel	75.17±6.08	80.95±13.47	63.69±5.32	100.00±0.00	80.63±18.04
(2) LS-SVM RBF kernel	76.53±7.36	91.84±12.91	64.88±6.24	100.00±0.00	83.17±18.81
(3) LS-SVM linear kernel (no regularization)	51.02±21.40	50.00±6.61	47.02±24.06	48.10±6.68	47.30±25.85
(4) PCA + FDA (unsupervised PC selection)	76.53±5.89	84.01±12.25	64.29±13.52	84.88±17.70	64.44±22.09
(5) PCA + FDA (supervised PC selection)	77.89±7.28	93.54±9.06	60.12±11.96	96.79±5.68	64.44±18.30
(6) kPCA lin + FDA (unsupervised PC selection)	77.55±7.07	85.03±12.64	63.69±13.31	85.60±17.99	61.59±21.96
(7) kPCA lin + FDA (supervised PC selection)	78.23±7.47	93.88±9.16	60.12±11.96	96.79±5.68	63.97±18.01
(8) kPCA RBF + FDA (unsupervised PC selection)	84.35±6.47	88.44±11.34	62.50±14.43	92.02±10.37	64.60±23.56
(9) kPCA RBF + FDA (supervised PC selection)	82.65±7.17	95.92±7.50	60.12±8.29	98.69±3.75	58.41±11.44

Table B.9 : *Breast cancer data set (Hedenfalk et al., 2001): sporadic mutations versus the rest: results (LOO-CV performances, training and test set accuracies, and training and test ROC performances) of all numerical experiments.*

Hedenfalk et al. (2001): sporadic mutations versus the rest

Statistical significance tests for LOO-CV performances

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.125000	0.000061	0.462891	0.078125	0.224609	0.042969	0.000095	0.000305
(2)	0.125000	1.000000	0.000357	0.574219	0.472656	0.849609	0.296875	0.000129	0.000061
(3)	0.000061	0.000357	1.000000	0.000408	0.000366	0.000414	0.000308	0.000083	0.000124
(4)	0.462891	0.574219	0.000408	1.000000	0.289063	0.500000	0.186035	0.000344	0.000864
(5)	0.078125	0.472656	0.000366	0.289063	1.000000	0.790039	1.000000	0.000061	0.001953
(6)	0.224609	0.849609	0.000414	0.500000	0.790039	1.000000	0.559570	0.000122	0.003418
(7)	0.042969	0.296875	0.000308	0.186035	1.000000	0.559570	1.000000	0.000122	0.001953
(8)	0.000095	0.000129	0.000083	0.000344	0.000061	0.000122	0.000122	1.000000	0.209961
(9)	0.000305	0.000061	0.000124	0.000864	0.001953	0.003418	0.001953	0.209961	1.000000

Hedenfalk et al. (2001): sporadic mutations versus the rest

Statistical significance tests for test set accuracies

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.625000	0.014516	0.851563	0.431641	1.000000	0.431641	0.671875	0.156250
(2)	0.625000	1.000000	0.007136	0.875000	0.119141	0.750000	0.119141	0.613281	0.062500
(3)	0.014516	0.007136	1.000000	0.019857	0.020288	0.019810	0.020288	0.022813	0.036288
(4)	0.851563	0.875000	0.019857	1.000000	0.264160	1.000000	0.264160	0.699219	0.171875
(5)	0.431641	0.119141	0.020288	0.264160	1.000000	0.338867	1.000000	0.533691	1.000000
(6)	1.000000	0.750000	0.019810	1.000000	0.338867	1.000000	0.338867	0.835938	0.246094
(7)	0.431641	0.119141	0.020288	0.264160	1.000000	0.338867	1.000000	0.533691	1.000000
(8)	0.671875	0.613281	0.022813	0.699219	0.533691	0.835938	0.533691	1.000000	0.584961
(9)	0.156250	0.062500	0.036288	0.171875	1.000000	0.246094	1.000000	0.584961	1.000000

Hedenfalk et al. (2001): sporadic mutations versus the rest

Statistical significance tests for test set ROC performances

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.187500	0.001079	0.003562	0.000762	0.002115	0.000698	0.005611	0.000664
(2)	0.187500	1.000000	0.000955	0.002123	0.000408	0.001107	0.000333	0.004018	0.000357
(3)	0.001079	0.000955	1.000000	0.054421	0.014000	0.082382	0.015660	0.031126	0.139353
(4)	0.003562	0.002123	0.054421	1.000000	0.746930	0.500000	0.680717	0.965209	0.243949
(5)	0.000762	0.000408	0.014000	0.746930	1.000000	0.827381	1.000000	0.851835	0.162395
(6)	0.002115	0.001107	0.082382	0.500000	0.827381	1.000000	0.919766	0.471663	0.627086
(7)	0.000698	0.000333	0.015660	0.680717	1.000000	0.919766	1.000000	0.765039	0.205052
(8)	0.005611	0.004018	0.031126	0.965209	0.851835	0.471663	0.765039	1.000000	0.206280
(9)	0.000664	0.000357	0.139353	0.243949	0.162395	0.627086	0.205052	0.206280	1.000000

Table B.10 : *Breast cancer data set (Hedenfalk et al., 2001): sporadic mutations versus the rest: statistical significance tests for the LOO-CV performances (upper part), the test set accuracies (middle part), and the test ROC performances (lower part) of all numerical experiments. The numerical experiments are numbered as defined in Table B.1.*

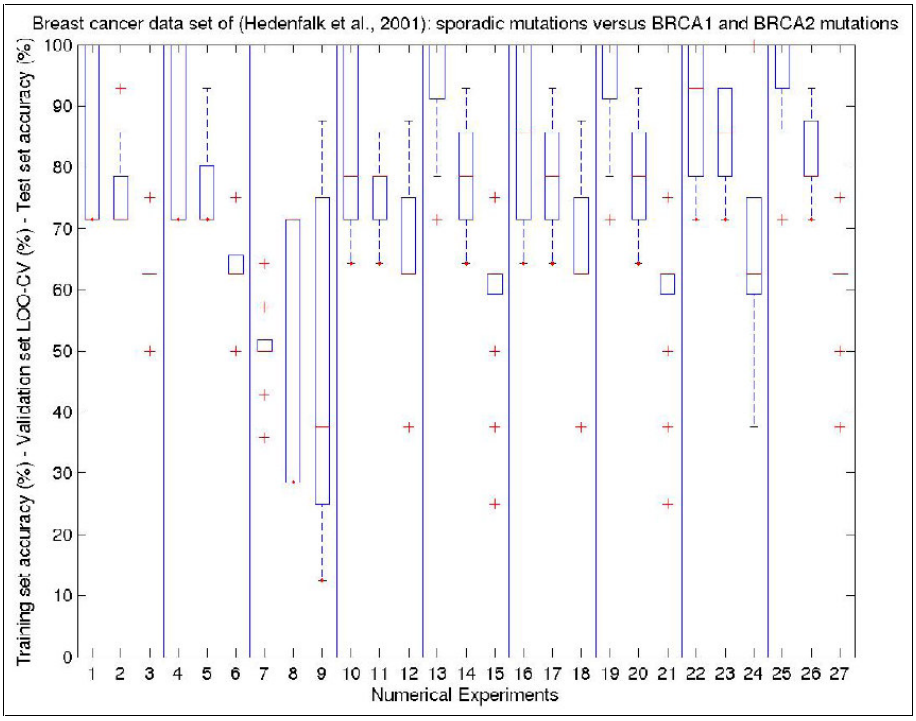


Figure B.9 : *Breast cancer data set (Hedenfalk et al., 2001): sporadic mutations versus the rest: boxplots representing the training set accuracy (first), the LOO-CV performance (second) and the test set accuracy (third) of all numerical experiments.* ***Legend:*** *See Figure B.1.*

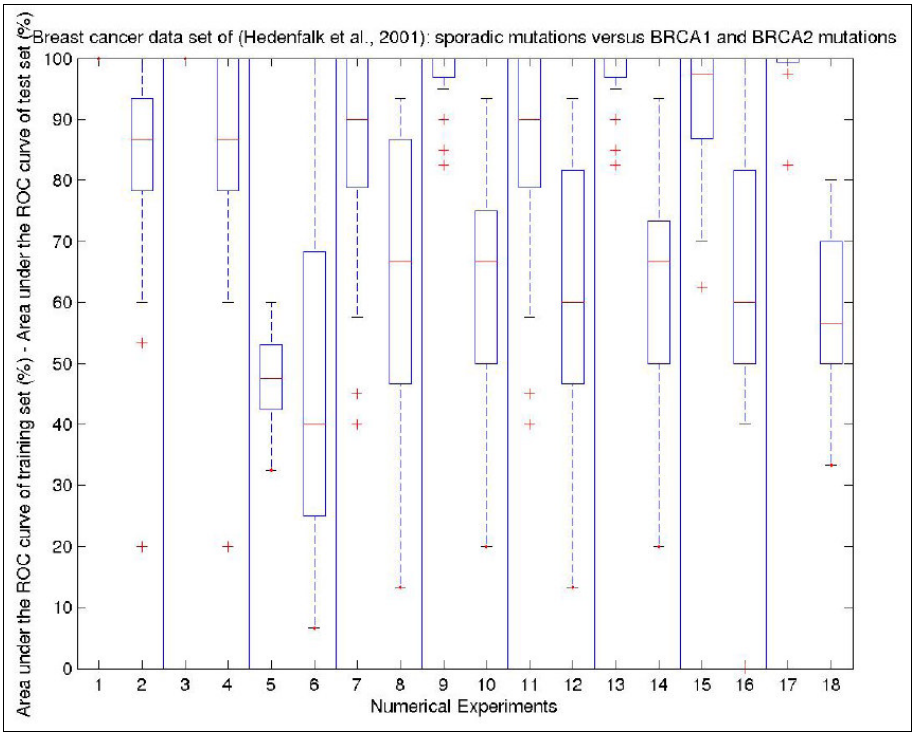


Figure B.10 : *Breast cancer data set (Hedenfalk et al., 2001): sporadic mutations versus the rest: boxplots representing the training set AUC (first) and the test set AUC (second) of all numerical experiments.* ***Legend:*** *See Figure B.2.*

B.6 Hepatocellular carcinoma data set (Iizuka *et al.*, 2003)

Similarly, when looking at the results shown in Table B.11 and visualized in Figures B.11 and B.12 and the statistical significance tests in Table B.12, the following statements can be derived for this data set:

1. The LOO-CV performances of the simulations containing kernel PCA with RBF kernel are significantly better than all other LOO-CV performances. Considering both of these, the one based on the supervised way is also significantly better than the one using the eigenvalues for selection of the principal components. The LOO-CV performances of LS-SVM with a linear or an RBF kernel are also slightly worse than the simulations using classical PCA or kernel PCA with a linear kernel. And on its turn, using a linear kernel for LS-SVM is even slightly worse than when using an RBF kernel.

Iizuka et al. (2003)

Results (LOO-CV performances, training and test set accuracies, and training and test ROC performances)

	LOO-CV performance	ACC training set	ACC test set	AUC training set	AUC test set
(1) LS-SVM linear kernel	65.80±3.76	73.88±16.21	68.43±4.52	98.54±2.20	64.60±7.28
(2) LS-SVM RBF kernel	67.53±4.09	87.16±16.73	68.61±6.32	99.32±1.28	64.04±5.98
(3) LS-SVM linear kernel (no regularization)	51.95±13.50	53.82±5.68	49.56±12.60	52.46±5.88	47.93±9.50
(4) PCA + FDA (unsupervised PC selection)	71.72±6.08	89.61±9.92	68.25±7.37	92.27±10.01	66.45±7.66
(5) PCA + FDA (supervised PC selection)	70.13±5.23	90.33±11.52	66.67±9.96	92.72±10.05	60.95±8.85
(6) kPCA lin + FDA (unsupervised PC selection)	71.72±6.08	89.61±9.92	68.25±7.37	92.27±10.01	66.45±7.66
(7) kPCA lin + FDA (supervised PC selection)	70.13±5.23	90.33±11.52	66.67±9.96	92.72±10.05	60.95±8.85
(8) kPCA RBF + FDA (unsupervised PC selection)	75.18±6.18	87.45±12.27	61.20±12.91	90.09±14.66	57.75±10.80
(9) kPCA RBF + FDA (supervised PC selection)	98.99±4.52	100.00±0.00	69.49±3.94	100.00±0.00	50.25±1.12

Table B.11 : *Hepatocellular carcinoma data set (Iizuka et al., 2003): results (LOO-CV performances, training and test set accuracies, and training and test ROC performances) of all numerical experiments.*

Iizuka et al. (2003)

Statistical significance tests for LOO-CV performances

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.015625	0.000122	0.000061	0.000502	0.000061	0.000502	0.000188	0.000027
(2)	0.015625	1.000000	0.000406	0.000244	0.000854	0.000244	0.000854	0.000193	0.000046
(3)	0.000122	0.000406	1.000000	0.000056	0.000100	0.000056	0.000100	0.000058	0.000038
(4)	0.000061	0.000244	0.000056	1.000000	0.082520	1.000000	0.082520	0.000122	0.000053
(5)	0.000502	0.000854	0.000100	0.082520	1.000000	0.082520	1.000000	0.001200	0.000082
(6)	0.000061	0.000244	0.000056	1.000000	0.082520	1.000000	0.082520	0.000122	0.000053
(7)	0.000502	0.000854	0.000100	0.082520	1.000000	0.082520	1.000000	0.001200	0.000082
(8)	0.000188	0.000193	0.000058	0.000122	0.001200	0.000122	0.001200	1.000000	0.000057
(9)	0.000027	0.000046	0.000038	0.000053	0.000082	0.000053	0.000082	0.000057	1.000000

Iizuka et al. (2003)

Statistical significance tests for test set accuracies

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.425781	0.000168	0.417969	0.721589	0.417969	0.721589	0.056702	0.250000
(2)	0.425781	1.000000	0.000288	0.774568	0.568458	0.774568	0.568458	0.055270	0.911621
(3)	0.000168	0.000288	1.000000	0.000118	0.000551	0.000118	0.000551	0.018015	0.000148
(4)	0.417969	0.774568	0.000118	1.000000	0.461723	1.000000	0.461723	0.024220	0.875549
(5)	0.721589	0.568458	0.000551	0.461723	1.000000	0.461723	1.000000	0.195585	0.378657
(6)	0.417969	0.774568	0.000118	1.000000	0.461723	1.000000	0.461723	0.024220	0.875549
(7)	0.721589	0.568458	0.000551	0.461723	1.000000	0.461723	1.000000	0.195585	0.378657
(8)	0.056702	0.055270	0.018015	0.024220	0.195585	0.024220	0.195585	1.000000	0.020264
(9)	0.250000	0.911621	0.000148	0.875549	0.378657	0.875549	0.378657	0.020264	1.000000

Iizuka et al. (2003)

Statistical significance tests for test set ROC performances

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.849775	0.000141	0.338973	0.067952	0.338973	0.067952	0.002094	0.000059
(2)	0.849775	1.000000	0.000200	0.139413	0.043710	0.139413	0.043710	0.009586	0.000059
(3)	0.000141	0.000200	1.000000	0.000122	0.001604	0.000122	0.001604	0.008685	0.190827
(4)	0.338973	0.139413	0.000122	1.000000	0.008962	1.000000	0.008962	0.002472	0.000074
(5)	0.067952	0.043710	0.001604	0.008962	1.000000	0.008962	1.000000	0.178918	0.000181
(6)	0.338973	0.139413	0.000122	1.000000	0.008962	1.000000	0.008962	0.002472	0.000074
(7)	0.067952	0.043710	0.001604	0.008962	1.000000	0.008962	1.000000	0.178918	0.000181
(8)	0.002094	0.009586	0.008685	0.002472	0.178918	0.002472	0.178918	1.000000	0.007895
(9)	0.000059	0.000059	0.190827	0.000074	0.000181	0.000074	0.000181	0.007895	1.000000

Table B.12 : *Hepatocellular carcinoma data set (Iizuka et al., 2003): statistical significance tests for the LOO-CV performances (upper part), the test set accuracies (middle part), and the test ROC performances (lower part) of all numerical experiments. The numerical experiments are numbered as defined in Table B.1.*

2. The test set accuracies show that all simulations perform similarly. Except for the simulation using kernel PCA with an RBF kernel and selecting the principal components based on the eigenvalues, which is slightly worse than some of the other simulations.
3. The area under the ROC curve of the test set reveals that the simulations using kernel PCA with an RBF kernel are clearly worse than all other simulations.
4. For this data set, using the eigenvalues for selection of the principal components often seems to give better results than when using the supervised method. Except for the test set accuracies where the simulation with kernel PCA with an RBF kernel and selection of the principal components by using the supervised way is much better than by using the eigenvalues.
5. For the simulation using kernel PCA with an RBF kernel and selection of the principal components in a supervised way, it is remarkable that although the area under the ROC curve of the test set is much worse than all other simulations, the test set accuracy is as good as the rest of the simulations. Again, this seems to indicate the importance of finding a good decision threshold value, which corresponds to an operating point on the ROC curve.

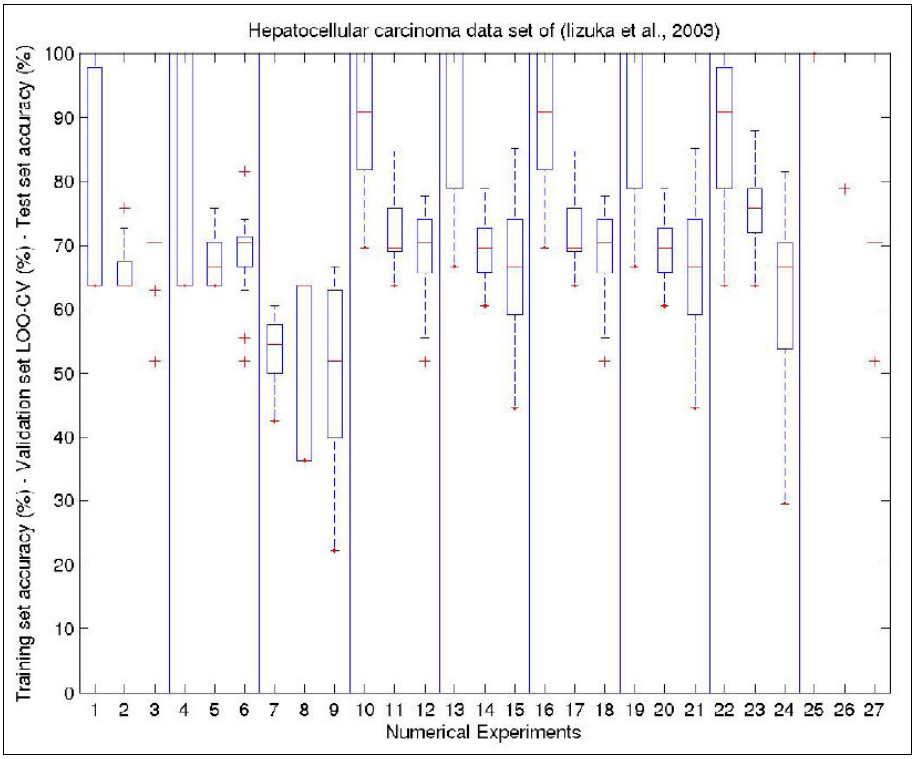


Figure B.11 : *Hepatocellular carcinoma data set (Iizuka et al., 2003): boxplots representing the training set accuracy (first), the LOO-CV performance (second) and the test set accuracy (third) of all numerical experiments.* ***Legend:*** *See Figure B.1.*

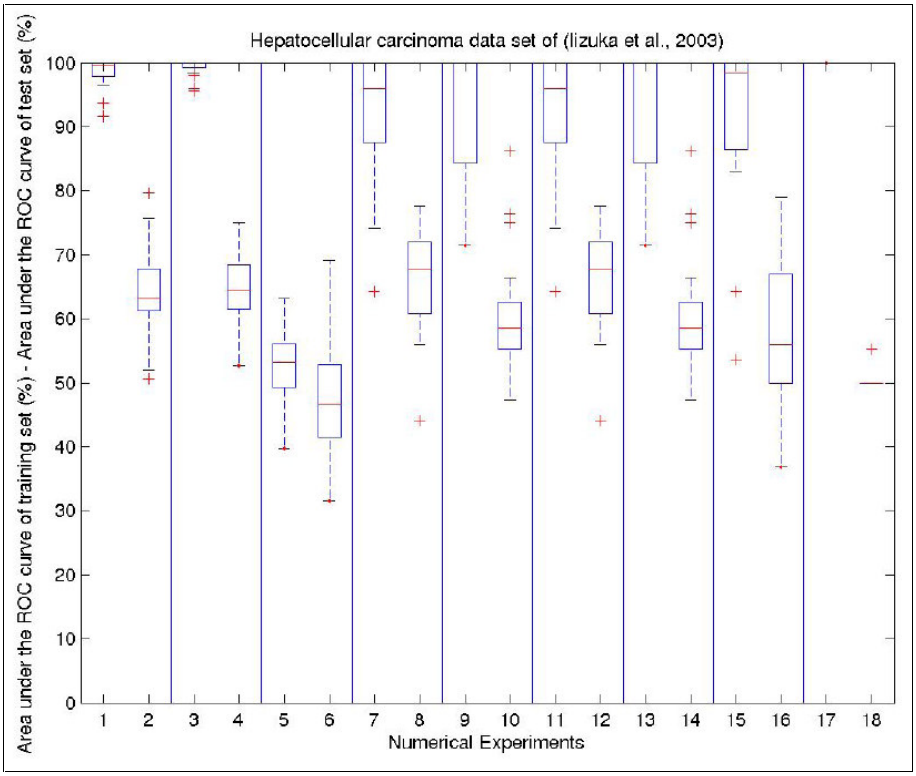


Figure B.12 : *Hepatocellular carcinoma data set (Iizuka et al., 2003): boxplots representing the training set AUC (first) and the test set AUC (second) of all numerical experiments.* ***Legend:*** *See Figure B.2.*

B.7 High-grade glioma data set (Nutt *et al.*, 2003)

Similarly, when looking at the results shown in Table B.13 and visualized in Figures B.13 and B.14 and the statistical significance tests in Table B.14, the following statements can be derived for this data set:

1. The LOO-CV performances of the simulations containing kernel PCA with RBF kernel are significantly better than all other LOO-CV performances. The LOO-CV performances of LS-SVM with a linear or an RBF kernel are also slightly worse than the simulations using classical PCA or kernel PCA with a linear kernel. And on its turn, using a linear kernel for LS-SVM is even slightly worse than when using an RBF kernel.
2. For the test set accuracies, the simulation using LS-SVM with an RBF kernel is significantly better than using LS-SVM with a linear kernel as well as using kernel PCA with an RBF kernel.

Nutt et al. (2003)

Results (LOO-CV performances, training and test set accuracies, and training and test ROC performances)

	LOO-CV performance	ACC training set	ACC test set	AUC training set	AUC test set
(1) LS-SVM linear kernel	75.74±8.93	90.02±14.16	61.25±11.75	99.47±1.03	79.25±6.06
(2) LS-SVM RBF kernel	78.23±7.99	98.41±7.10	69.95±8.59	100.00±0.00	81.04±6.64
(3) LS-SVM linear kernel (no regularization)	50.79±16.65	50.79±12.75	48.93±10.88	50.63±16.40	50.68±15.15
(4) PCA + FDA (unsupervised PC selection)	80.95±7.49	92.29±7.12	67.82±7.24	97.72±2.80	77.48±10.50
(5) PCA + FDA (supervised PC selection)	81.41±7.19	92.97±10.14	65.52±11.01	96.65±5.69	77.37±9.04
(6) kPCA lin + FDA (unsupervised PC selection)	80.73±7.12	92.52±6.98	68.31±6.78	97.91±2.74	77.98±10.43
(7) kPCA lin + FDA (supervised PC selection)	81.86±6.67	95.24±8.57	67.32±11.04	98.15±4.02	76.53±8.96
(8) kPCA RBF + FDA (unsupervised PC selection)	86.62±5.99	94.78±9.05	64.20±11.19	97.30±6.60	70.80±15.44
(9) kPCA RBF + FDA (supervised PC selection)	85.94±5.78	96.15±7.29	58.13±12.24	98.25±3.78	66.33±15.48

Table B.13 : *High-grade glioma data set (Nutt et al., 2003): results (LOO-CV performances, training and test set accuracies, and training and test ROC performances) of all numerical experiments.*

Nutt et al. (2003)

Statistical significance tests for LOO-CV performances

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.036133	0.000286	0.001380	0.000324	0.001552	0.000351	0.000189	0.000189
(2)	0.036133	1.000000	0.000192	0.050293	0.003906	0.050903	0.000977	0.000422	0.000415
(3)	0.000286	0.000192	1.000000	0.000087	0.000057	0.000087	0.000058	0.000058	0.000058
(4)	0.001380	0.050293	0.000087	1.000000	1.000000	1.000000	0.472656	0.001130	0.002197
(5)	0.000324	0.003906	0.000057	1.000000	1.000000	0.912109	0.500000	0.001684	0.000244
(6)	0.001552	0.050903	0.000087	1.000000	0.912109	1.000000	0.466797	0.000270	0.003599
(7)	0.000351	0.000977	0.000058	0.472656	0.500000	0.466797	1.000000	0.001221	0.000977
(8)	0.000189	0.000422	0.000058	0.001130	0.001684	0.000270	0.001221	1.000000	0.906250
(9)	0.000189	0.000415	0.000058	0.002197	0.000244	0.003599	0.000977	0.906250	1.000000

Nutt et al. (2003)

Statistical significance tests for test set accuracies

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.001587	0.001010	0.026721	0.197089	0.012879	0.139618	0.541289	0.164368
(2)	0.001587	1.000000	0.000088	0.139514	0.058836	0.170110	0.190063	0.024790	0.000358
(3)	0.001010	0.000088	1.000000	0.000088	0.000108	0.000088	0.000110	0.000960	0.003505
(4)	0.026721	0.139514	0.000088	1.000000	0.485510	0.500000	0.844465	0.065552	0.006562
(5)	0.197089	0.058836	0.000108	0.485510	1.000000	0.348702	0.500000	0.602308	0.002808
(6)	0.012879	0.170110	0.000088	0.500000	0.348702	1.000000	0.965222	0.034424	0.003288
(7)	0.139618	0.190063	0.000110	0.844465	0.500000	0.965222	1.000000	0.236236	0.002808
(8)	0.541289	0.024790	0.000960	0.065552	0.602308	0.034424	0.236236	1.000000	0.070409
(9)	0.164368	0.000358	0.003505	0.006562	0.002808	0.003288	0.002808	0.070409	1.000000

Nutt et al. (2003)

Statistical significance tests for test set ROC performances

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.043592	0.000060	0.347901	0.216720	0.466544	0.035355	0.006698	0.000617
(2)	0.043592	1.000000	0.000069	0.121819	0.010502	0.125760	0.001814	0.001860	0.000321
(3)	0.000060	0.000069	1.000000	0.000175	0.000141	0.000163	0.000141	0.002493	0.001155
(4)	0.347901	0.121819	0.000175	1.000000	0.972269	0.500000	0.676588	0.021621	0.018675
(5)	0.216720	0.010502	0.000141	0.972269	1.000000	0.741232	0.500000	0.073138	0.046139
(6)	0.466544	0.125760	0.000163	0.500000	0.741232	1.000000	0.476136	0.011087	0.008968
(7)	0.035355	0.001814	0.000141	0.676588	0.500000	0.476136	1.000000	0.100458	0.046139
(8)	0.006698	0.001860	0.002493	0.021621	0.073138	0.011087	0.100458	1.000000	0.281226
(9)	0.000617	0.000321	0.001155	0.018675	0.046139	0.008968	0.046139	0.281226	1.000000

Table B.14 : *High-grade glioma data set (Nutt et al., 2003): statistical significance tests for the LOO-CV performances (upper part), the test set accuracies (middle part), and the test ROC performances (lower part) of all numerical experiments. The numerical experiments are numbered as defined in Table B.1.*

3. Using LS-SVM with an RBF kernel results in a better value for the area under the ROC curve of the test set than all other simulations, except for the simulations using classical PCA or kernel PCA with a linear kernel when selecting the principal components by using the eigenvalues, which give similar results.
4. For this data set both methods for selection of the principal components seem to give similar results.

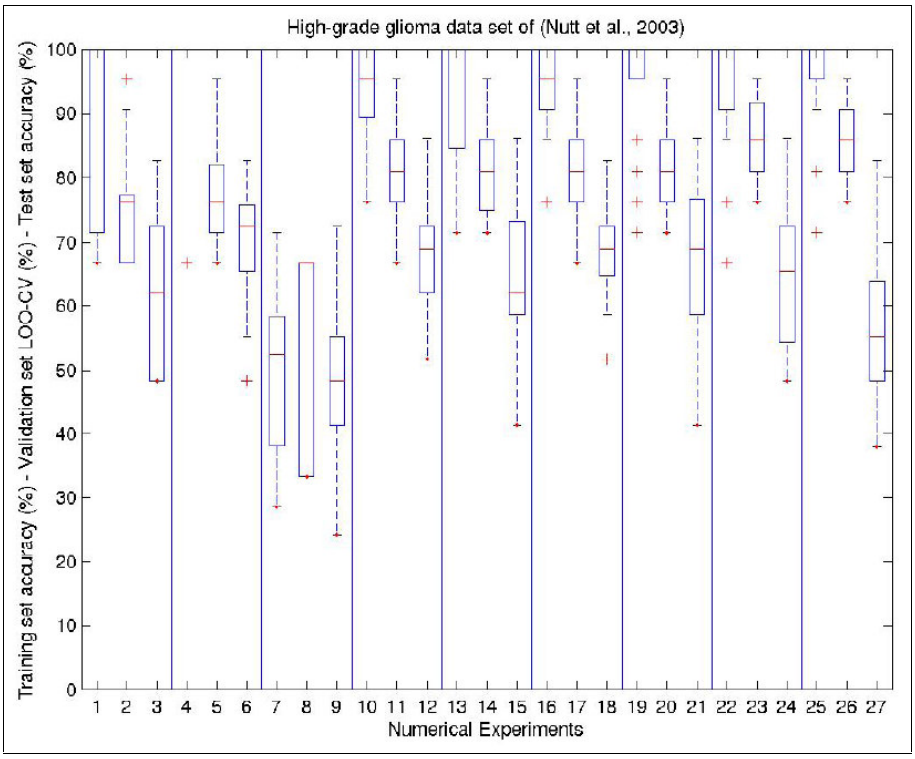


Figure B.13 : *High-grade glioma data set (Nutt et al., 2003): boxplots representing the training set accuracy (first), the LOO-CV performance (second) and the test set accuracy (third) of all numerical experiments.* ***Legend:*** *See Figure B.1.*

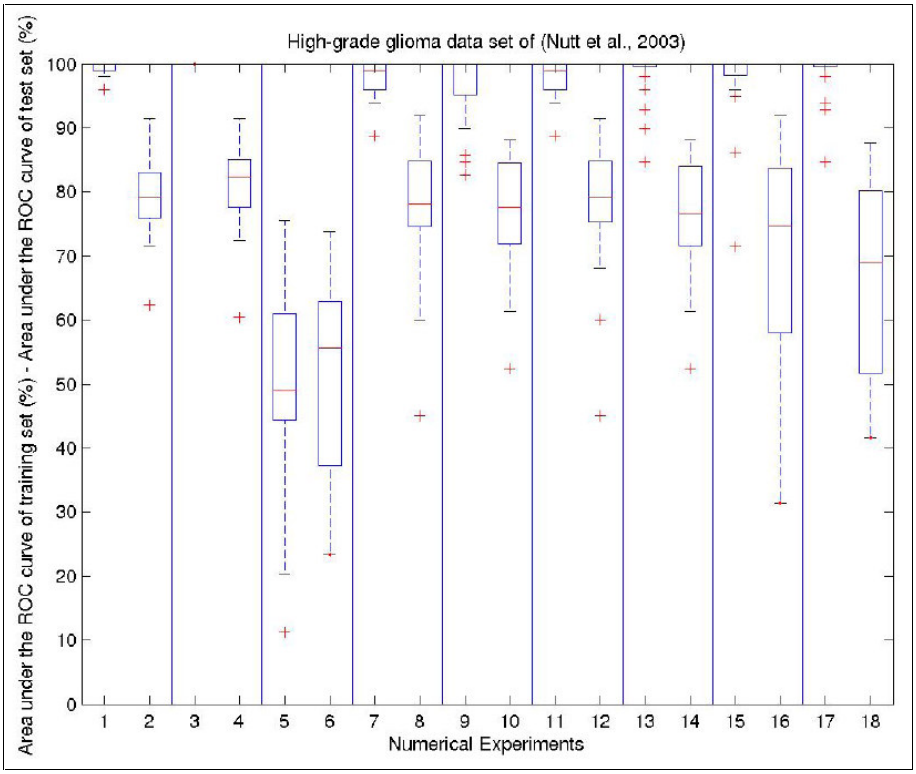


Figure B.14 : *High-grade glioma data set (Nutt et al., 2003): boxplots representing the training set AUC (first) and the test set AUC (second) of all numerical experiments.* ***Legend:*** *See Figure B.2.*

B.8 Prostate cancer data set (Singh *et al.*, 2002)

Similarly, when looking at the results shown in Table B.15 and visualized in Figures B.15 and B.16 and the statistical significance tests in Table B.16, the following statements can be derived for this data set:

1. The simulation using kernel PCA with RBF kernel and selecting the principal components by means of the supervised method clearly has the best LOO-CV performance of all simulations. And the simulation using LS-SVM with an RBF kernel as well as the simulation using kernel PCA with RBF kernel and selecting the principal components by means of the eigenvalues both perform slightly better than the rest of the simulations.
2. About the test set accuracies, the simulation using kernel PCA with RBF kernel and selecting the principal components by means of the supervised method clearly gives very bad results. Using the eigenvalues for selection of the principal components seems to give better results

Singh et al. (2002)

Results (LOO-CV performances, training and test set accuracies, and training and test ROC performances)

	LOO-CV performance	ACC training set	ACC test set	AUC training set	AUC test set
(1) LS-SVM linear kernel	90.10±1.42	100.00±0.00	84.31±13.66	100.00±0.00	91.28±5.20
(2) LS-SVM RBF kernel	90.52±1.67	100.00±0.00	85.01±13.76	100.00±0.00	92.51±4.52
(3) LS-SVM linear kernel (no regularization)	50.33±0.92	51.45±7.03	48.18±10.25	51.10±8.27	50.98±12.38
(4) PCA + FDA (unsupervised PC selection)	90.38±1.83	97.62±1.95	83.89±13.63	99.67±0.38	88.93±11.39
(5) PCA + FDA (supervised PC selection)	90.57±1.53	97.57±3.34	82.49±13.35	99.40±0.99	86.74±12.95
(6) kPCA lin + FDA (unsupervised PC selection)	90.34±1.75	97.57±1.90	85.01±9.07	99.67±0.38	89.98±7.30
(7) kPCA lin + FDA (supervised PC selection)	90.57±1.53	97.57±3.34	82.49±13.35	99.40±0.99	86.73±12.96
(8) kPCA RBF + FDA (unsupervised PC selection)	91.60±1.50	98.97±1.75	85.01±11.00	99.84±0.32	89.90±9.64
(9) kPCA RBF + FDA (supervised PC selection)	100.00±0.00	100.00±0.00	28.71±10.02	100.00±0.00	50.00±0.00

Table B.15 : *Prostate cancer data set (Singh et al., 2002): results (LOO-CV performances, training and test set accuracies, and training and test ROC performances) of all numerical experiments.*

Singh et al. (2002)

Statistical significance tests for LOO-CV performances

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.085693	0.000056	0.392437	0.090454	0.505768	0.090454	0.000276	0.000055
(2)	0.085693	1.000000	0.000051	0.506165	0.854736	0.389587	0.854736	0.012401	0.000052
(3)	0.000056	0.000051	1.000000	0.000057	0.000056	0.000057	0.000056	0.000057	0.000029
(4)	0.392437	0.506165	0.000057	1.000000	0.391805	1.000000	0.391805	0.000061	0.000054
(5)	0.090454	0.854736	0.000056	0.391805	1.000000	0.366744	1.000000	0.005493	0.000055
(6)	0.505768	0.389587	0.000057	1.000000	0.366744	1.000000	0.366744	0.000412	0.000054
(7)	0.090454	0.854736	0.000056	0.391805	1.000000	0.366744	1.000000	0.005493	0.000055
(8)	0.000276	0.012401	0.000057	0.000061	0.005493	0.000412	0.005493	1.000000	0.000055
(9)	0.000055	0.000052	0.000029	0.000054	0.000055	0.000054	0.000055	0.000055	1.000000

Singh et al. (2002)

Statistical significance tests for test set accuracies

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.611328	0.000068	0.809387	0.124512	0.855689	0.124512	0.634753	0.000065
(2)	0.611328	1.000000	0.000068	0.448242	0.034816	0.754736	0.034816	0.943057	0.000064
(3)	0.000068	0.000068	1.000000	0.000068	0.000069	0.000059	0.000069	0.000059	0.000884
(4)	0.809387	0.448242	0.000068	1.000000	0.090454	1.000000	0.090454	0.324219	0.000061
(5)	0.124512	0.034816	0.000069	0.090454	1.000000	0.046875	1.000000	0.027506	0.000064
(6)	0.855689	0.754736	0.000059	1.000000	0.046875	1.000000	0.046875	0.947266	0.000061
(7)	0.124512	0.034816	0.000069	0.090454	1.000000	0.046875	1.000000	0.027506	0.000064
(8)	0.634753	0.943057	0.000059	0.324219	0.027506	0.947266	0.027506	1.000000	0.000064
(9)	0.000065	0.000064	0.000884	0.000061	0.000064	0.000061	0.000064	0.000064	1.000000

Singh et al. (2002)

Statistical significance tests for test set ROC performances

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.117993	0.000060	0.776236	0.029394	0.776236	0.029394	0.601054	0.000059
(2)	0.117993	1.000000	0.000060	0.217717	0.002076	0.217717	0.002078	0.676450	0.000059
(3)	0.000060	0.000060	1.000000	0.000069	0.000080	0.000059	0.000080	0.000089	0.689234
(4)	0.776236	0.217717	0.000069	1.000000	0.043456	1.000000	0.047202	0.183876	0.000069
(5)	0.029394	0.002076	0.000080	0.043456	1.000000	0.043456	1.000000	0.045655	0.000069
(6)	0.776236	0.217717	0.000059	1.000000	0.043456	1.000000	0.047202	0.586003	0.000059
(7)	0.029394	0.002078	0.000080	0.047202	1.000000	0.047202	1.000000	0.045639	0.000069
(8)	0.601054	0.676450	0.000089	0.183876	0.045655	0.586003	0.045639	1.000000	0.000059
(9)	0.000059	0.000059	0.689234	0.000069	0.000069	0.000059	0.000069	0.000059	1.000000

Table B.16 : *Prostate cancer data set (Singh et al., 2002): statistical significance tests for the LOO-CV performances (upper part), the test set accuracies (middle part), and the test ROC performances (lower part) of all numerical experiments. The numerical experiments are numbered as defined in Table B.1.*

than using the supervised method. The simulation applying LS-SVM with an RBF kernel even performs slightly better than those simulations using the eigenvalues for selection of the principal components.

3. The simulation using kernel PCA with RBF kernel and selecting the principal components in a supervised way also seems to give very bad results when considering the area under the ROC curve of the test set. Again, using the eigenvalues for selection of the principal components seems to give better results than using the supervised way. The simulations applying LS-SVM even perform slightly better than those simulations using the eigenvalues for selection of the principal components.

4. Remark that the combination of using kernel PCA with RBF kernel and selecting the principal components by means of the supervised method seriously leads to overfitting.

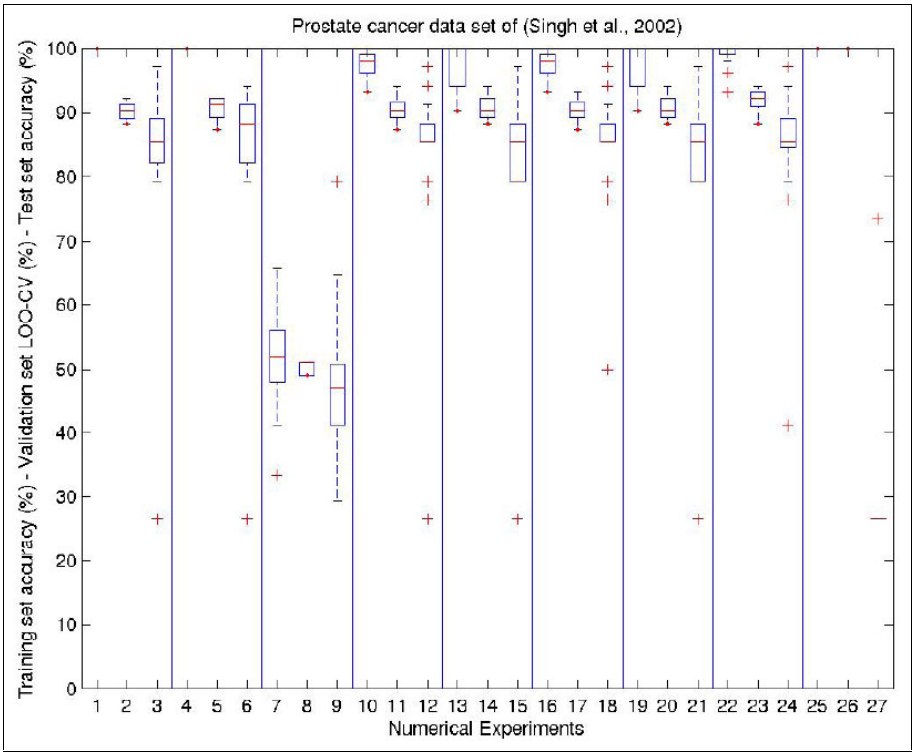


Figure B.15 : *Prostate cancer data set (Singh et al., 2002): boxplots representing the training set accuracy (first), the LOO-CV performance (second) and the test set accuracy (third) of all numerical experiments.* ***Legend:*** *See Figure B.1.*

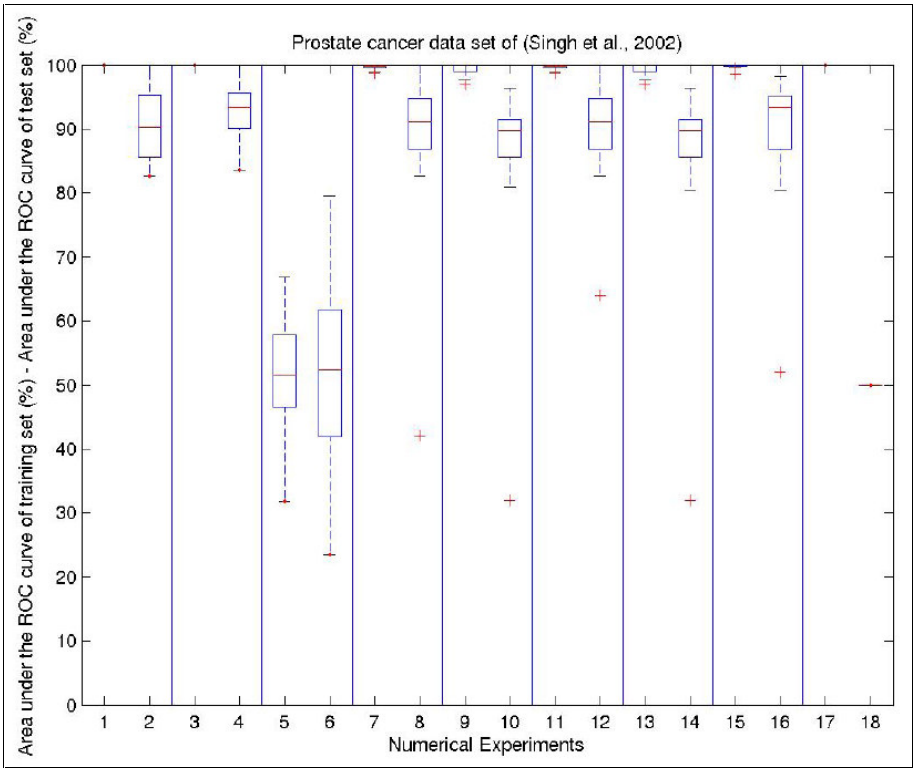


Figure B.16 : *Prostate cancer data set (Singh et al., 2002): boxplots representing the training set AUC (first) and the test set AUC (second) of all numerical experiments.* ***Legend:*** *See Figure B.2.*

B.9 Breast cancer data set (Van 't Veer *et al.*, 2002)

Similarly, when looking at the results shown in Table B.17 and visualized in Figures B.17 and B.18 and the statistical significance tests in Table B.18, the following statements can be derived for this data set:

1. The simulation using kernel PCA with RBF kernel and selecting the principal components in a supervised way clearly has the best LOO-CV performance of all simulations. And the simulations using LS-SVM both perform slightly worse than the rest of the simulations. It seems to be better to use the supervised way for selection of the principal components.
2. When looking at the test set accuracies, it is obvious that the simulation using kernel PCA with an RBF kernel and selecting the principal components by means of the supervised method is giving very bad

Van t Veer et al. (2002)

Results (LOO-CV performances, training and test set accuracies, and training and test ROC performances)

	LOO-CV performance	ACC training set	ACC test set	AUC training set	AUC test set
(1) LS-SVM linear kernel	68.99±4.22	100.00±0.00	67.92±8.58	100.00±0.00	73.30±11.01
(2) LS-SVM RBF kernel	69.05±3.55	100.00±0.00	68.42±7.62	100.00±0.00	73.98±10.69
(3) LS-SVM linear kernel (no regularization)	52.14±6.04	74.66±24.04	57.14±9.08	74.73±25.26	64.60±13.18
(4) PCA + FDA (unsupervised PC selection)	71.31±3.57	91.27±10.04	57.39±15.57	94.61±6.80	65.16±12.30
(5) PCA + FDA (supervised PC selection)	73.44±3.19	97.31±5.62	66.92±9.90	98.77±3.16	67.91±12.64
(6) kPCA lin + FDA (unsupervised PC selection)	71.18±3.62	91.21±10.33	60.90±14.49	94.46±7.22	66.01±13.45
(7) kPCA lin + FDA (supervised PC selection)	73.63±3.89	97.13±6.63	65.41±7.54	98.54±3.98	69.22±11.01
(8) kPCA RBF + FDA (unsupervised PC selection)	74.91±6.54	90.66±11.08	51.38±15.91	93.77±8.75	60.26±16.57
(9) kPCA RBF + FDA (supervised PC selection)	100.00±0.00	100.00±0.00	36.84±0.00	100.00±0.00	50.00±0.00

Table B.17 : *Breast cancer data set (Van 't Veer et al., 2002): results (LOO-CV performances, training and test set accuracies, and training and test ROC performances) of all numerical experiments.*

Van t Veer et al. (2002)

Statistical significance tests for LOO-CV performances

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.791829	0.000059	0.000477	0.000127	0.000739	0.000086	0.000086	0.000058
(2)	0.791829	1.000000	0.000058	0.000825	0.000087	0.001753	0.000058	0.000081	0.000056
(3)	0.000059	0.000058	1.000000	0.000059	0.000058	0.000059	0.000059	0.000058	0.000029
(4)	0.000477	0.000825	0.000059	1.000000	0.002225	0.750000	0.000766	0.000557	0.000057
(5)	0.000127	0.000087	0.000058	0.002225	1.000000	0.001411	0.636719	0.470021	0.000057
(6)	0.000739	0.001753	0.000059	0.750000	0.001411	1.000000	0.000563	0.000830	0.000057
(7)	0.000086	0.000058	0.000059	0.000766	0.636719	0.000563	1.000000	0.969965	0.000058
(8)	0.000086	0.000081	0.000058	0.000557	0.470021	0.000830	0.969965	1.000000	0.000082
(9)	0.000058	0.000056	0.000029	0.000057	0.000057	0.000057	0.000058	0.000082	1.000000

Van t Veer et al. (2002)

Statistical significance tests for test set accuracies

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.367188	0.000650	0.004142	0.638062	0.009748	0.295498	0.001099	0.000055
(2)	0.367188	1.000000	0.000244	0.006858	0.381369	0.022977	0.217368	0.003285	0.000055
(3)	0.000650	0.000244	1.000000	0.938114	0.001314	0.285465	0.001461	0.131122	0.000056
(4)	0.004142	0.006858	0.938114	1.000000	0.005046	0.031250	0.018422	0.246704	0.000312
(5)	0.638062	0.381369	0.001314	0.005046	1.000000	0.030904	0.562500	0.001694	0.000055
(6)	0.009748	0.022977	0.285465	0.031250	0.030904	1.000000	0.097504	0.025146	0.000141
(7)	0.295498	0.217368	0.001461	0.018422	0.562500	0.097504	1.000000	0.002325	0.000054
(8)	0.001099	0.003285	0.131122	0.246704	0.001694	0.025146	0.002325	1.000000	0.001880
(9)	0.000055	0.000055	0.000056	0.000312	0.000055	0.000141	0.000054	0.001880	1.000000

Van t Veer et al. (2002)

Statistical significance tests for test set ROC performances

experiments	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.000000	0.416992	0.001861	0.000617	0.033006	0.001472	0.016592	0.000300	0.000059
(2)	0.416992	1.000000	0.000338	0.001016	0.004005	0.001707	0.006026	0.000701	0.000059
(3)	0.001861	0.000338	1.000000	0.875710	0.489707	0.848372	0.052624	0.455194	0.000541
(4)	0.000617	0.001016	0.875710	1.000000	0.204530	0.437500	0.102315	0.196033	0.000448
(5)	0.033006	0.004005	0.489707	0.204530	1.000000	0.159193	0.820313	0.031793	0.000132
(6)	0.001472	0.001707	0.848372	0.437500	0.159193	1.000000	0.088473	0.224229	0.000389
(7)	0.016592	0.006026	0.052624	0.102315	0.820313	0.088473	1.000000	0.055863	0.000080
(8)	0.000300	0.000701	0.455194	0.196033	0.031793	0.224229	0.055863	1.000000	0.024480
(9)	0.000059	0.000059	0.000541	0.000448	0.000132	0.000389	0.000080	0.024480	1.000000

Table B.18 : *Breast cancer data set (Van 't Veer et al., 2002): statistical significance tests for the LOO-CV performances (upper part), the test set accuracies (middle part), and the test ROC performances (lower part) of all numerical experiments. The numerical experiments are numbered as defined in Table B.1.*

results. Using LS-SVM gives better results than performing dimensionality reduction combined with an unsupervised way for the selection of the principal components.

3. For the area under the ROC curve of the test set, the simulation using kernel PCA with an RBF kernel and selecting the principal components by means of the supervised method performs very badly again. Using LS-SVM gives better results than performing dimensionality reduction.

4. Both methods for selecting the principal components seem to perform very similarly, but in some cases using the absolute value of the Golub score tends to perform slightly better.

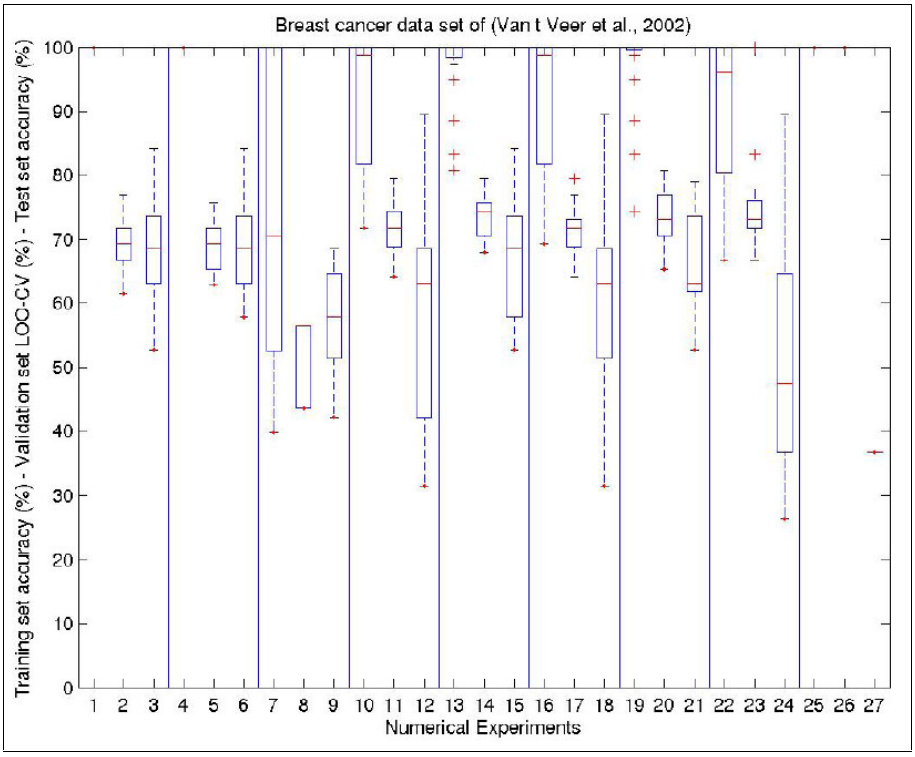


Figure B.17 : *Breast cancer data set (Van 't Veer et al., 2002): boxplots representing the training set accuracy (first), the LOO-CV performance (second) and the test set accuracy (third) of all numerical experiments.* ***Legend:*** *See Figure B.1.*

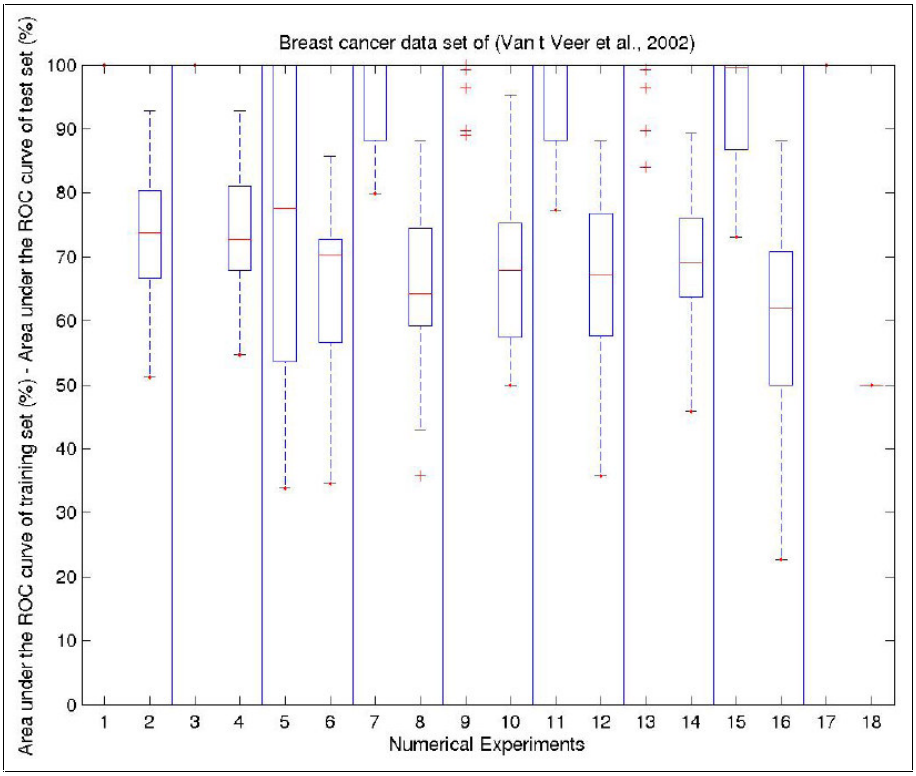


Figure B.18 : *Breast cancer data set (Van 't Veer et al., 2002): boxplots representing the training set AUC (first) and the test set AUC (second) of all numerical experiments.* ***Legend:*** *See Figure B.2.*

Bibliography

Ahr, A., Holtrich, U., Solbach, C., Scharl, A., Strebhardt, K., Karn, T., Kaufmann, M. (2001) Molecular Classification of Breast Cancer Patients by Gene Expression Profiling. *Journal of Pathology*, **195**, 312-320.

Ahr, A., Karn, T., Solbach, C., Seiter, T., Strebhardt, K., Holtrich, U., Kaufmann, M. (2002) Identification of High Risk Breast-Cancer Patients by Gene Expression Profiling. *The Lancet*, **359**, 131-132.

Albelda, S.M., Sheppard, D. (2000) Functional Genomics and Expression Profiling: Be There or Be Square. *American journal of respiratory cell and molecular biology*, **23**, 265-269.

Alon, A., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Science USA*, **96**, 6745-6750.

Alter, O., Brown, P., Botstein, D. (2000) Singular Value Decomposition for Genome-Wide Expression Data Processing And Modeling. *Proceedings of the National Academy of Science USA*, **97**, 10101-10106.

Alter, O., Brown, P.O., Botstein, D. (2003) Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *Proceedings of the National Academy of Science USA*, **100**(6), 3351-3356.

Ben-Dor, A., Friedman, N., Yakhini, Z. (2001) Class Discovery in Gene Expression Data. *Proceedings Fifth Annual International Conference on Computational Molecular Biology (RECOMB)*, 1-8.

Ben-Hur, A., Horn, D., Siegelmann, H.T., Vapnik, V. (2001) Support Vector Clustering. *Journal of Machine Learning Research*, **2**, 125-137.

Berchuck, A., Iversen, E.S., Lancaster, J.M., Dressman, H.K., West, M., Nevins, J.R. et al. (2004) Prediction of optimal versus suboptimal

cytoreduction of advanced-stage serous ovarian cancer with the use of microarrays. *American Journal of Obstetrics and Gynecology*, **190**, 910-25.

Bishop, C.M. (1995) *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, UK.

Blum, A., Chawla, S. (2001) Learning from labeled and unlabeled data using graph mincuts. *Proceedings of the 18th International Conference on Machine Learning*, 19-26.

Bolshakova, N., Azuaje, F. (2003) Cluster validation techniques for genome expression data. *Signal Processing*, **83**, 825-833.

Bolshakova, N., Azuaje, F., Cunningham, P. (2005) An integrated tool for microarray data clustering and cluster validity assessment. *Bioinformatics*, **21**(4), 451-455.

Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C. et al. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature Genetics*, **29**, 365-71.

Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M.Jr, Haussler, D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Science USA*, **97**, 262-267.

Calinski, T., Harabasz, J. (1974) A Dendrite Method for Cluster Analysis. *Communications in Statistics*, **3**(1), 1-27.

Cristianini, N., Shawe-Taylor, J. (2000) *An Introduction to Support Vector Machines (and other Kernel-Based Learning Methods)*. Cambridge University Press, Cambridge.

Cristianini, N., Shawe-Taylor, J., Kandola, J. (2002) Spectral kernel methods for clustering. In T. G. Dietterich, S. Becker, Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*. Cambridge, MA. MIT Press.

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

De Bie, T., Cristianini, N., Rosipal R. (2004) Eigenproblems in Pattern Recognition. In E. Bayro-Corrochano, editor, *Handbook of Computational Geometry for Pattern Recognition, Computer Vision, Neurocomputing and Robotics*. Springer-Verlag.

De Maeyer, L., Neven, P., Drijkoningen, M., Woestenborghs, H., Pochet, N., De Moor, B., Amant, F., Berteloot, F., Leunen, K., Van Limbergen, E., Smeets, A., Wildiers, H., Paridaens, R., Christiaens, MR., Vergote, I. (2006) The oestrogen receptor has a prognostic value in progesterone receptor

positive breast cancers. *International Journal of Gynecological Cancer (Flemish Gynecology Oncologic Group (FGOG) Conference)*, In press.

De Smet, F., Mathys, J., Marchal, K., Thijs, G., De Moor, B., Moreau, Y. (2002) Adaptive quality based clustering of gene expression profiles. *Bioinformatics*, **18**(5), 735-746.

De Smet, F., Moreau, Y., Engelen, K., Timmerman, D., Vergote, I., De Moor, B. (2004) Balancing false positives and false negatives for the detection of differential expression in malignancies. *British Journal of Cancer*, **91**, 1160-5.

De Smet, F., Pochet, N.L.M.M., De Moor, B.L.R., Van Gorp, T., Timmerman, D., Vergote, I.B., Hartmann, L.C., Damokosh, A.I., Hoersch, S. (2005) Independent test set performance in the prediction of early relapse in ovarian cancer with gene expression profiles. *Clinical Cancer Research*, **11**(21), 7958-7959.

De Smet, F., Pochet, N., Engelen, K., Van Gorp, T., Van Hummelen, P., Marchal, K., Amant, F., Timmerman, D., De Moor, B., Vergote, I. (2006a) Predicting the clinical behavior of ovarian cancer from gene expression profiles. *International Journal of Gynecological cancer*, **16**(S1), 147-151.

De Smet, F., De Brabanter, J., Konstantinovic, M.L., Pochet, N., Van den Bosch, T., Moerman, P., De Moor, B., Vergote, I., Timmerman, D. (2006b) New models to predict depth of infiltration in endometrial carcinoma based on transvaginal sonography. *Ultrasound in Obstetrics and Gynecology*, In press.

Dhillon, I.S., Guan, Y., Kulis, B. (2004a) Kernel k-means, Spectral Clustering and Normalized Cuts. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 551-556.

Dhillon, I.S., Guan, Y., Kulis, B. (2004b) A Unified View of Kernel k-means, Spectral Clustering and Graph Partitioning. *UTCS Technical Report*.

Dubes, R., Jain, A.K. (1979) Validity studies in clustering methodologies. *Pattern Recognition Letters*, **11**, 235-254.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Science USA*, **95**, 14863-14868.

Epstein, E., Skoog, L., Isberg, P., De Smet, F., De Moor, B., Olofsson, P., Gudmundsson, S., Valentin, L. (2002) An algorithm including results of gray scale and power Doppler ultrasound examination to predict endometrial malignancy in women with postmenopausal bleeding. *Ultrasound in Obstetrics and Gynecology*, **20**, 370-376.

Freedland, S.J., Isaacs, W.B., Mangold, L.A., Yiu, S.K., Grubb, K.A., Partin, A.W., Epstein, J.I., Walsh, P.C., Platz, E.A. (2005) Stronger Association between Obesity and Biochemical Progression after Radical Prostatectomy among Men Treated in the Last 10 Years. *Clinical Cancer Research*, **11**, 2883-2888.

Friend, S. H. (1999) How DNA Microarrays and Expression Profiling will Affect Clinical Practice. *British Medical Journal*, **319**, 1306-1307.

Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D. (2000) Support vector machines classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906-914.

Gevaert, O., De Smet, F., Timmerman, D., Moreau, Y., De Moor B. (2006) Integration of clinical and microarray data using Bayesian networks. In *The 14th IFAC Symposium on System Identification (SYSID2006)*.

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

Gupta, H., Agrawal, A.K., Pruthi, T., Shekhar, C., Chellappa, R. (2002) An experimental evaluation of linear and kernel-based methods for face recognition. *Workshop on the Application of Computer Vision (WACV)*, FL, USA.

Guyon, I., Weston, J., Barnhill, S., Vapnik, V. (2002) Gene Selection for Cancer Classification Using Support Vector Machines. *Machine Learning*, **46**(1-3), 389-422.

Halkidi, M., Batistakis, Y., Vazirgiannis, M. (2001) On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, **17**(2-3), 107-145.

Halkidi, M., Vazirgiannis, M. (2005) Quality Assessment Approaches in Data Mining. *The Data Mining and Knowledge Discovery Handbook*, 661-696.

Handl, J., Knowles, J., Kell, D.B. (2005) Computational cluster validation in post-genomic data analysis. *Bioinformatics*, **21**, 3201-3212.

Hanley, J.A., McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29-36.

Hartmann, L.C., Lu, K.H., Linette, G.P. et al. (2005) Gene expression profiles predict early relapse in ovarian cancer after platinum-paclitaxel chemotherapy. *Clinical Cancer Research*, **11**, 2149-55.

Hastie, T., Tibshirani, R., Eisen, M.B., Alizadeh, A., Levy, R., Staudt, L., Chan, W.C., Botstein, D., Brown, P. (2000) Gene shaving as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology*, **1**, 1-21.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M. *et al*. (2001) Gene-expression profiles in hereditary breast cancer. *New English Journal of Medicine*, **344**, 539-548.

Helleman, J., Jansen, M.P., Span, P.N., van Staveren, I.L., Massuger, L.F., Meijer-van Gelder, M.E., Sweep, F.C., Ewing, P.C., van der Burg, M.E., Stoter, G., Nooter, K., Berns, E.M. (2005) Molecular profiling of platinum resistant ovarian cancer. *International Journal of Cancer*, **118**(8), 1963-71.

Hubert, L., Arabie, P. (1985) Comparing partitions. *Journal of Classification*, 193-218.

Iizuka, N., Oka, M., Yamada-Okabe, H., Nishida, M., Maeda, Y., Mori, N., Takao, T., Tamesa, T., Tangoku, A., Tabuchi, H. *et al*. (2003) Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *The Lancet*, **361**, 923-929.

Jain, A.K., Dubes, R.C. (1988) *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, NJ.

Jazaeri, A.A., Yee, C.J., Sotiriou, C., Brantley, K.R., Boyd, J., Liu, E.T. (2002) Gene expression profiles of BRCA1-linked, BRCA2-linked, and sporadic ovarian cancers. *Journal of the National Cancer Institute*, **94**, 990-1000.

Joachims, T. (2003) Transductive learning via spectral graph partitioning. In *Proceedings of the International Conference on Machine Learning*, 290-297.

Joliffe, I.T. (1986) *Principal Component Analysis*. Springer-Verlag.

Kannan, R., Vempala, S., Vetta, A. (2004) On Clusterings: Good, Bad and Spectral. *Journal of the ACM*, **51**(3), 497-515.

Kozak, K.R., Amneus, M.W., Pusey, S.M., Su, F., Luong, M.N., Luong, S.A., Reddy, S.T., Farias-Eisner, R. (2003) Identification of biomarkers for ovarian cancer using strong anion-exchange ProteinChips: potential use in diagnosis and prognosis. *Proceedings of the National Academy of Science USA*, **100**(21), 12343-12348.

Lancaster, J.M., Dressman, H.K., Whitaker, R.S., Havrilesky, L., Gray, J., Marks, J.R. et al. (2004) Gene expression patterns that characterize advanced stage serous ovarian cancers. *The Journal of the Society for Gynecologic Investigation*, **11**, 51-9.

Leunen, K., Van Mieghem, T., Pochet, N., De Smet, F., De Moor, B., De Leyn, B., Van Limbergen, E., Amant, F., Berteloot, P., Wildiers, H., Paridaens, R., Smeets, A., Christiaens, MR., Vergote, I., Neven, P. (2006) The progesterone receptor (PR) in postmenopausal women with an oestrogen receptor (ER)-positive breast cancer: The effect of body composition? *International Journal of Gynecological Cancer (Flemish Gynecology Oncologic Group (FGOG) Conference)*, In press.

Li, Y., Campbell, C., Tipping, M. (2002) Bayesian automatic relevance determination algorithms for classifying gene expression data. *Bioinformatics*, **18**, 1332-1339.

Lu, K.H., Patterson, A.P., Wang, L., Marquez, R.T., Atkinson, E.N., Baggerly, K.A. et al. (2004) Selection of potential markers for epithelial ovarian cancer with gene expression arrays and recursive descent partition analysis. *Clinical Cancer Research*, **10**, 3291-300.

MacQueen, J. (1967) Some methods for classification and analysis of multivariate observations. In L.M. Le Cam, J. Neyman, editors, *Proceedings of the 5th Berkeley symposium on mathematical statistics and probability*, **1**, 281-297. University of California Press, Berkeley.

Madeira, S.C., Oliveira, A.L. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE Transactions on Computational Biology and Bioinformatics*, **1**, 24-45.

Markman, M., Rothman, R., Hakes, T., Reichman, B., Hoskins, W., Rubin, S. et al. (1991) Second-line platinum therapy in patients with ovarian cancer previously treated with cisplatin. *Journal of Clinical Oncology*, **9**, 389-93.

Mika, S., Schölkopf, B., Smola, A.J., Müller, K.R., Scholz, M., Rätsch, G. (1999) Kernel PCA and de-noising in feature spaces. In M. S. Kearns, S. A. Solla, D. A. Cohn, editors, *Advances in Neural Information Processing Systems 11*. MIT Press.

Milligan, G.W., Cooper, M.C. (1985) An examination of procedures for determining the number of clusters. *Psychometrika*, **50**, 159-179.

Mukherjee, S., Tamayo, P., Slonim, D., Verri, A., Golub, T., Mesirov, J.P., Poggio, T. (1999) Support vector machine classification of microarray data. *A.I. Memo 1677, Massachusetts Institute of Technology*.

Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., Schölkopf, B. (2001) An Introduction to Kernel-Based Learning Algorithms. *IEEE Transactions on Neural Networks*, **12**, 181-202.

Nasir, J. (2001) DNA Microarrays Give Breast Cancer Leads. *The Lancet Oncology*, **2**, 68.

Neven, P., Pochet, N., Drijkoningen, M., De Smet, F., De Moor, B., Amant, F., Paridaens, R., Christiaens, M.R., Vergote, I. (2006a) Estrogen receptor positive progesterone receptor negative breast cancers and other tumour characteristics. *Journal of Clinical Oncology*, In press.

Neven, P., Vanden Bempt, I., Pochet, N., Hendrickx, W., Huang, H.J., De Smet, F., De Moor, B., Drijkoningen, M., Leunen, K., Amant, F., Berteloot, P., Paridaens, R., Wildiers, H., Van Limbergen, E., Smeets, A., Christiaens, M.-R., Vergote, I. (2006b) Membrane expression of HER-2/neu as a predictor for axillary lymph node invasion in ER^+PR^+ breast cancers. Revised manuscript submitted to *Journal of Clinical Oncology*.

Ng, A., Jordan, M.I., Weiss, Y. (2002) On spectral clustering: Analysis and an algorithm. In T. G. Dietterich, S. Becker, Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA. MIT Press.

Nielsen, T. O., West, R. B., Linn, S. C., Alter, O., Knowling, M. A., O'Connell, J. X., Zhu, S., Fero, M., Sherlock, G., Pollack, J. R., Brown, P. O., Botstein, D., van de Rijn, M. (2002) Molecular Characterisation of Soft Tissue Tumours: a Gene Expression Study. *The Lancet*, **359**, 1301-1307.

Nutt, C.L., Mani, D.R., Betensky, R.A., Tamayo, P., Cairncross, J.G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M.E. *et al*. (2003) Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Research*, **63**, 1602-1607.

Patten-Hitt, E. (2001) Gene Chips Aid Cancer Diagnoses. *The Lancet Oncology*, **2**, 398.

Perou, C., Sørlie, T., Eisen, M., van de Rijn, M., Jeffrey, S., Rees, C., Pollack, J., Ross, D., Johnsen H., Akslen, L., Fluge, Ø., Pergamenschikov, A., Williams, C., Zhu, S., Lønning, P., Børresen-Dale, A.-L., Brown, P., Botstein, D. (2000) Molecular Portraits of Human Breast Tumours. *Nature*, **406**, 747-752.

Petricoin, E.F., Ornstein, D.K., Paweletz, C.P., Ardekani, A., Hackett, P.S., Hitt, B.A., Velassco, A., Trucco, C., Wiegand, L., Wood, K., Simone, C.B., Levine, P.J., Linehan, W.M., Emmert-Buck, M.R., Steinberg, S.M., Kohn, E.C., Liotta, L.A. (2002a) Serum Proteomic Patterns for Detection of Prostate Cancer. *Journal of the National Cancer Institute*, **94**, 1576-1578.

Petricoin, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, S.M., Mills, G.B., Simone, C., Fishman, D.A., Kohn, E.C., Liotta, L.A. (2002b) Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*, **359**(9306), 572-577.

Pelckmans, K., Suykens, J.A.K., Van Gestel, T., De Brabanter, J., Lukas, L., Hamers, B., De Moor, B., Vandewalle, J. (2002) LS-SVMlab: a Matlab/C

Toolbox for Least Squares Support Vector Machines. *Internal Report 02-44, ESAT-SISTA, K.U.Leuven (Leuven, Belgium)*.

Pelckmans, K., Suykens, J.A.K., De Moor, B. (2005a) Building Sparse Representations and Structure Determination on LS-SVM Substrates. *Neurocomputing*, **64**, 137-159.

Pelckmans, K., Goethals, I., De Brabanter, J., Suykens, J.A.K., De Moor, B. (2005b) Componentwise Least Squares Support Vector Machines. Wang L., ed., Chapter *Support Vector Machines: Theory and Applications*. Springer, 77-98.

Pelckmans, K., Suykens, J.A.K., De Moor, B. (2006) Additive regularization Trade-off: Fusion of Training and Validation levels in Kernel Methods. *Machine Learning*, **62**(3), 217-252.

Pochet, N., De Smet, F., Suykens, J., De Moor, B. (2004) Systematic benchmarking of micorarray data classification: assessing the role of nonlinearity and dimensionality reduction. *Bioinformatics*, **20**(17), 3185-3195.

Pochet, N.L.M.M., Janssens, F.A.L., De Smet, F., Marchal, K., Suykens, J.A.K., De Moor, B.L.R. (2005) M@CBETH: a microarray classification benchmarking tool. *Bioinformatics*, **21**(14), 3185-3186.

Pochet, N.L.M.M., Ojeda, F., De Smet, F., De Bie, T., Suykens, J.A.K., De Moor, B.L.R. (2006a) Kernel clustering for knowledge discovery in clinical microarray data analysis. Chapter 3 in *Kernel methods in bioengineering, communications and image processing*, (Camps-Valls G., Rojo-Alvarez J.L., Martinez-Ramon M., eds.), Idea Group Inc. (Hershey, Pennsylvania (US)), 2006. In press.

Pochet, N.L.M.M., Suykens, J.A.K. (2006b) Opinion. Support Vector Machines versus Logistic Regression: improving prospective performance in clinical decision making. *Ultrasound in Obstetrics and Gynecology*, In press.

Puskas, L.G., Zvara, A., Hackler, L. Jr., Van Hummelen, P. (2002) RNA amplification results in reproducible microarray data with slight ratio bias. *Biotechniques*, **32**, 1330-4, 1336, 1338, 1340.

Qin, J., Lewis, D.P., Noble, W.S. (2003) Kernel hierarchical gene clustering from microarray expression data. *Bioinformatics*, **19**, 2097-2104.

Quackenbush, J. (2001) Computational analysis of microarray data. *Nature Reviews Genetics*, 418-27.

Rand, W.M. (1971) Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, **66**, 846-850.

Rickman, D.S., Bobek, M.P., Misek, D.E., Kuick, R., Blaivas, M., Kurnit, D. M., Taylor, J., Hanash, S.M. (2001) Distinctive Molecular Profiles of High-

Grade and Low-Grade Gliomas Based on Oligonucleotide Microarray Analysis, *Cancer Research*, **61**, 6885-6891.

Rousseeuw P.J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 53-65.

Rosen, J.E., Costouros, N.G., Lorang, D., Burns, A.L., Alexander, H.R., Skarulis, M.C., Cochran, C., Pingpank, J.F., Marx, S.J., Spiegel, A.M., Libutti, S.K. (2005) Gland Size Is Associated With Changes in Gene Expression Profiles in Sporadic Parathyroid Adenomas. *Annals of Surgical Oncology*, **12**(5), 412-416.

Schölkopf, B., Smola, A.J., Müller, K.-R. (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, **10**, 1299-1319.

Schölkopf, B., Burges, C.J.C., Smola, A.J. (1999) *Advances in Kernel Methods: Support Vector Learning*. MIT Press.

Schölkopf, B., Guyon, I., Weston, J. (2001) Statistical Learning and Kernel Methods in Bioinformatics. *Proceedings NATO Advanced Studies Institute on Artificial Intelligence and Heuristics Methods for Bioinformatics*, 1-21.

Schölkopf, B., Smola, A.J. (2002) *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, UK.

Schwartz, D.R., Kardia, S.L., Shedden, K.A., Kuick, R., Michailidis, G., Taylor, J.M. et al. (2002) Gene expression in ovarian cancer reflects both morphology and biological behavior, distinguishing clear cell from other poor-prognosis ovarian carcinomas. *Cancer Research*, **62**, 4722-9.

Sheng, Q., Moreau, Y., De Moor, B. (2003) Biclustering Microarray data by Gibbs sampling. *Bioinformatics, European Conference on Computational Biology Proceedings*, **19**, ii196-ii205.

Shi, J., Malik, J. (2000) Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(8), 808-905.

Simon, R., Radmacher, M.D., Dobbin, K., McShane, L.M. (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, **95**, 14-8.

Singh, D., Febbo, P.G., Ross, K., Jackson, D.G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A.A., D'Amico, A.V., Richie, J.P. *et al*. (2002) Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, **1**, 203-209.

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist,

H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E., Børresen-Dale, A.-L. (2001) Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses with Clinical Implications. *Proceedings of the National Academy of Science USA*, **98**, 10869-10874.

Sotiriou, C., Powles, T.J., Dowsett, M., Jazaeri, A.A., Feldman, A.L., Assersohn, L., Gadisetti, C., Libutti, S.K., Liu, E.T. (2002) Gene Expression Profiles Derived from Fine Needle Aspiration Correlate with Response to Systemic Chemotherapy in Breast Cancer. *Breast Cancer Research*, **4**, R3.

Stikeman, A. (2002) The State of Biomedicine: Medical treatment will be tailored to your genetic profile. *Technology Review*.

Storey, J.D., Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Science USA*, **100**, 9440-5.

Suykens, J.A.K., Vandewalle, J. (1999) Least squares support vector machine classifiers. *Neural Processing Letters*, **9**, 293-300.

Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J. (2002) *Least Squares Support Vector Machines*. World Scientific Publishing Co., Pte, Ltd. (Singapore).

Suykens, J.A.K., Van Gestel, T., Vandewalle, J., De Moor, B. (2003) A support vector machine formulation to PCA analysis and its Kernel version. *IEEE Transactions on Neural Networks*, **14**, 447-450.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M. (1999) Systematic determination of genetic network architecture. *Nature Genetics*, **22**, 281-285.

Thigpen, T., Stuart, G., du Bois, A., Friedlander, M., Fujiwara, K., Guastalla, J.P., Kaye, S., Kitchener, H., Kristensen, G., Mannel, R., Meier, W., Miller, B., Poveda, A., Provencher, D., Stehman, F., Vergote, I. (2005) Clinical trials in ovarian carcinoma: requirements for standard approaches and regimens. *Annals of Oncology*, **16**(S8), viii13-viii19.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal Of The Royal Statistical Society Series B*, **58**(1), 267-288.

Timmerman, D., Testa, A.C., Bourne, T., Ferrazzi, E., Ameye, L., Konstantinovic, M.L., Van Calster, B., Collins, W.P., Vergote, I., Van Huffel, S., Valentin, L. (2005) Logistic Regression Model to Distinguish Between the Benign and Malignant Adnexal Mass Before Surgery: A Multicenter Study by the International Ovarian Tumor Analysis Group. *Journal of Clinical Oncology*, **23**, 8794-8801.

Trimbos, J.B., Vergote, I., Bolis, G., Vermorken, J.B., Mangioni, C., Madronal, C. et al. (2003) Impact of adjuvant chemotherapy and surgical staging in early-stage ovarian carcinoma: European Organisation for

Research and Treatment of Cancer-Adjuvant ChemoTherapy in Ovarian Neoplasm trial. *Journal of the National Cancer Institute*, **95**, 113-25.

Troyanskaya, O.G., Garber, M.E., Brown, P.O., Botstein, D., Altman, R.B. (2002) Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, **18**, 1454-61.

Van Gestel, T., Suykens, J. A.K., De Moor, B., Vandewalle, J. (2001a) Automatic relevance determination for least squares support vector machine classifiers. *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, 13-18.

Van Gestel, T., Suykens, J. A.K., De Moor, B. Vandewalle, J. (2001b) Automatic relevance determination for Least squares support vector machine regression. *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, 2416-2421.

Van Gestel, T., Suykens, J.A.K., Lanckriet, G., Lambrechts, A., De Moor, B., Vandewalle, J. (2002) Bayesian framework for least squares support vector machine classifiers, Gaussian processes and kernel Fisher discriminant analysis. *Neural Computation*, **15**, 1115-1148.

Van Gestel, T., Suykens, J.A.K., Baesens, B., Viaene, S., Vanthienen, J., Dedene, G. De Moor, B., Vandewalle, J. (2004) Benchmarking least squares support vector machine classifiers. *Machine Learning*, **54**, 5-32.

Van Gorp, T., De Smet, F., Pochet, N., Engelen, K., Van Hummelen, P., Suykens, J., Marchal, K., Amant, F., Moreau, Y., Timmerman, D., De Moor, B., Vergote, I. (2005) Predicting the clinical behavior of ovarian cancer from gene expression profiles. *International Journal of Gynecological Cancer (14th International meeting of the ESGO) (ESGO 14)*, **15**(S2), 66.

Van Mieghem, T., Leunen, K., Pochet, N., De Moor, B., Amant, F., Vanden Bempt, I., Drijkoningen, R., Christiaens, MR., Vergote, I., Neven, P. (2006a) Weight and bodymass index (BMI) affect HER-2 expression in postmenopausal breast cancer. *International Journal of Gynecological Cancer (Flemish Gynecology Oncologic Group (FGOG) Conference)*, In press.

Van Mieghem, T., Leunen, K., Pochet, N., Deleyn, A., De Moor, B., De Smet, F., Amant, F., Berteloot, P., Marquette, S., Drijkoningen, R., Van Limbergen, E., Smeets, A., Wildiers, H., Paridaens, R., Christiaens, MR., Vergote, I., Neven, P. (2006b) The progesterone receptor (PR) in postmenopausal women with an oestrogen receptor (ER)-positive breast cancer by body composition and use of hormone replacement therapy (HRT). *International Journal of Gynecological Cancer (Flemish Gynecology Oncologic Group (FGOG) Conference)*, In press.

Van Mieghem, T., Leunen, K., Pochet, N., Deleyn, A., De Moor, B., De Smet, F., Amant, F., Berteloot, P., Marquette, S., Vanden Bempt, I., Drijkoningen, R., Van Limbergen, E., Smeets, A., Wildiers, H., Paridaens, R., Christiaens, MR., Vergote, I., Neven, P. (2006c) Parameters of body composition like weight and bodymass index (BMI) affect HER-2 expression in postmenopausal women with breast cancer. *European Journal of Cancer* (*European Breast Cancer Conference (EBCC))*, In press.

Van Mieghem, T., Leunen, K., Pochet, N., De Moor, B., De Smet, F., Amant, F., Berteloot, P., Vanden Bempt, I., Drijkoningen, R., Wildiers, H., Paridaens, R., Deleyn, A., Smeets, A., Van Limbergen, E., Christiaens, M.-R., Vergote, I., Neven, P. (2006d) Body mass index and HER-2 in breast cancer patients over 50 years of age. Revised manuscript submitted to *Journal of the National Cancer Institute*.

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

Vapnik, V.N. (1998) *Statistical Learning Theory*. John Wiley and Sons, New York.

Welsh, J.B., Zarrinkar, P.P., Sapinoso, L.M., Kern, S.G., Behling, C.A., Monk, B.J. et al. (2001) Analysis of gene expression profiles in normal and neoplastic ovarian tissue samples identifies candidate molecular markers of epithelial ovarian cancer. *Proceedings of the National Academy of Science USA*, **98**, 1176-81.

Wigle, D.A., Jurisica, I., Radulovich, N., Pintilie, M., Rossant, J., Liu, N., Lu, C., Woodgett, J., Seiden, I., Johnston, M., Keshavjee, S., Darling, G., Winton, T., Breitkreutz, B.-J., Jorgenson, P., Tyers, M., Shepherd, F.A., Tsao, M.S. (2002) Molecular Profiling of Non-Small Cell Lung Cancer and Correlation with Disease-Free Survival. *Cancer Research*, **62**, 3005-3008.

Wilson, L.L., Tran, L., Morton, D.L., Hoon, D.S.B. (2004) Detection of Differentially Expressed Proteins in Early-Stage Melanoma Patients Using SELDI-TOF Mass Spectrometry. *Annals of the New York Academy of Sciences*, **1022**, 317-322.

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J. et al. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research*, **30**, e15.

Yeung, K., Haynor, D., Ruzzo, W. (2001a) Validating clustering for gene expression data. *Bioinformatics*, **17**(4), 309-318.

Yeung, K., Fraley, C., Murua, A., Raftery, A., Ruzzo, W. (2001b) Model-based clustering and data transformations for gene expression data. *Bioinformatics*, **17**(10), 977-987.

Zhang, R., Rudnicky, A.I. (2002) A Large Scale Clustering Scheme for Kernel K-Means. *Proceedings of the International Conference on Pattern Recognition*.

Publication list

Most of the work discussed in this dissertation has been published in one of the following articles, in which we contributed.

Full papers in international journals

Pochet N., De Smet F., Suykens J., De Moor B. (2004) Systematic benchmarking of micorarray data classification: assessing the role of nonlinearity and dimensionality reduction. *Bioinformatics*, **20**(17), 3185-3195. (IF 5.742)

Pochet N.L.M.M., Janssens F.A.L., De Smet F., Marchal K., Suykens J.A.K., De Moor B.L.R. (2005) M@CBETH: a microarray classification benchmarking tool. *Bioinformatics*, **21**(14), 3185-3186. (IF 5.742)

De Smet F., Pochet N.L.M.M., De Moor B.L.R., Van Gorp T., Timmerman D., Vergote I.B., Hartmann L.C., Damokosh A.I., Hoersch S. (2005) Independent test set performance in the prediction of early relapse in ovarian cancer with gene expression profiles. *Clinical Cancer Research*, **11**(21), 7958-7959. (IF 5.623)

De Smet F., Pochet N., Engelen K., Van Gorp T., Van Hummelen P., Marchal K., Amant F., Timmerman D., De Moor B., Vergote I. (2006) Predicting the clinical behavior of ovarian cancer from gene expression profiles. *International Journal of Gynecological cancer*, **16**(S1), 147-151. (IF 1.147)

Gevaert O., Pochet N., De Smet F., Van Gorp T., De Moor B., Timmerman D., Amant F., Vergote I. (2006) Molecular profiling of platinum resistant ovarian cancer: use of the model in clinical practice. *International Journal of Cancer*, In press. (IF 4.416)

Neven P., Pochet N., Drijkoningen M., De Smet F., De Moor B., Amant F., Paridaens R., Christiaens M.R., Vergote I. (2006) Estrogen receptor positive progesterone receptor negative breast cancers and other tumour characteristics. *Journal of Clinical Oncology*, In press. (IF 9.835)

De Smet F., De Brabanter J., Konstantinovic M.L., Pochet N., Van den Bosch T., Moerman P., De Moor B., Vergote I., Timmerman D. (2006) New models to predict depth of infiltration in endometrial carcinoma based on transvaginal sonography. *Ultrasound in Obstetrics and Gynecology*, In press. (IF 2.167)

Pochet N.L.M.M., Suykens J.A.K. (2006) Opinion. Support Vector Machines versus Logistic Regression: improving prospective performance in clinical decision making. *Ultrasound in Obstetrics and Gynecology*, In press. (IF 2.167)

Neven P., Vanden Bempt I., Pochet N., Hendrickx W., Huang H.J., De Smet F., De Moor B., Drijkoningen M., Leunen K., Amant F., Berteloot P., Paridaens R., Wildiers H., Van Limbergen E., Smeets A., Christiaens M.-R., Vergote I. (2006) Membrane expression of HER-2/neu as a predictor for axillary lymph node invasion in ER^+PR^+ breast cancers. Revised manuscript submitted to *Journal of Clinical Oncology*. (IF 9.835)

Van Mieghem T., Leunen K., Pochet N., De Moor B., De Smet F., Amant F., Berteloot P., Vanden Bempt I., Drijkoningen R., Wildiers H., Paridaens R., Deleyn A., Smeets A., Van Limbergen E., Christiaens M.-R., Vergote I., Neven P. (2006) Body mass index and HER-2 in breast cancer patients over 50 years of age. Revised manuscript submitted to *Journal of the National Cancer Institute*. (IF 13.856)

Abstracts in international journals presented at international conferences

Van Gorp T., De Smet F., Pochet N., Engelen K., Van Hummelen P., Suykens J., Marchal K., Amant F., Moreau Y., Timmerman D., De Moor B., Vergote I. (2005) Predicting the clinical behavior of ovarian cancer from gene expression profiles. *International Journal of Gynecological Cancer (14th International meeting of the ESGO) (ESGO 14)*, **15**(S2), 66. (IF 1.147)

Leunen K., Van Mieghem T., Pochet N., De Smet F., De Moor B., De Leyn B., Van Limbergen E., Amant F., Berteloot P., Wildiers H., Paridaens R., Smeets A., Christiaens MR., Vergote I., Neven P. (2006) The progesterone receptor (PR) in postmenopausal women with an oestrogen receptor (ER)-positive breast cancer: The effect of body composition? *International Journal of Gynecological Cancer (Flemish Gynecology Oncologic Group (FGOG) Conference)*, In press. (IF 1.147)

De Maeyer L., Neven P., Drijkoningen M., Woestenborghs H., Pochet N., De Moor B., Amant F., Berteloot F., Leunen K., Van Limbergen E., Smeets A., Wildiers H., Paridaens R., Christiaens MR., Vergote I. (2006) The oestrogen receptor has a prognostic value in progesterone receptor positive breast cancers. *International Journal of Gynecological Cancer (Flemish Gynecology Oncologic Group (FGOG) Conference)*, In press. (IF 1.147)

Van Mieghem T., Leunen K., Pochet N., De Moor B., Amant F., Vanden Bempt I., Drijkoningen R., Christiaens MR., Vergote I., Neven P. (2006) Weight and bodymass index (BMI) affect HER-2 expression in postmenopausal breast cancer. *International Journal of Gynecological Cancer (Flemish Gynecology Oncologic Group (FGOG) Conference)*, In press. (IF 1.147)

Van Mieghem T., Leunen K., Pochet N., Deleyn A., De Moor B., De Smet F., Amant F., Berteloot P., Marquette S., Drijkoningen R., Van Limbergen E., Smeets A., Wildiers H., Paridaens R., Christiaens MR., Vergote I., Neven P. (2006) The progesterone receptor (PR) in postmenopausal women with an oestrogen receptor (ER)-positive breast cancer by body composition and use of hormone replacement therapy (HRT). *International Journal of Gynecological Cancer (Flemish Gynecology Oncologic Group (FGOG) Conference)*, In press. (IF 1.147)

Van Mieghem T., Leunen K., Pochet N., Deleyn A., De Moor B., De Smet F., Amant F., Berteloot P., Marquette S., Vanden Bempt I., Drijkoningen R., Van Limbergen E., Smeets A., Wildiers H., Paridaens R., Christiaens MR., Vergote I., Neven P. (2006) Parameters of body composition like weight and bodymass index (BMI) affect HER-2 expression in postmenopausal women with breast cancer. *European Journal of Cancer* (*European Breast Cancer Conference (EBCC))*, In press. (IF 3.302)

Book chapter

Pochet N.L.M.M., Ojeda F., De Smet F., De Bie T., Suykens J.A.K., De Moor B.L.R. (2006) Kernel clustering for knowledge discovery in clinical microarray data analysis. Chapter 3 in *Kernel methods in bioengineering, communications and image processing*, (Camps-Valls G., Rojo-Alvarez J.L., and and Martinez-Ramon M., eds.), Idea Group Inc. (Hershey, Pennsylvania (US)), 2006. In press.

International conference

Pochet N.L.M.M., Janssens F.A.L., De Smet F., Marchal K., Vergote I.B., Suykens J.A.K., De Moor B.L.R. (2005) M@CBETH: optimizing clinical

microarray classification. In *Proc. of the 2005 IEEE Computational Systems Bioinformatics Conference (CSB2005)*, Stanford, California (US), pp. 89-90.

Curriculum Vitae

Personal Data

Name:	Nathalie Pochet
Born:	23/04/1977, Antwerpen, Belgium
Address:	Meuleveld 14 B-2150 Borsbeek
E-mail:	nathalie.pochet@esat.kuleuven.be nathalie.pochet@skynet.be

Education

2001-2002:	Master of Science in Bioinformatics, K.U.Leuven Cum laude Master thesis 'HIV Structural Genomics' at Tibotec-Virco, Johnson&Johnson, Mechelen, Belgium
2000-2001:	Master in Artificial Intelligence, K.U.Leuven Magna cum laude Master thesis 'Detailed 3D visualization derived from 2D images' at ESAT-ACCA, K.U.Leuven
1996-2000:	Master of Science in Industrial Engineering in Electronics (option Information and Communication Technology), De Nayer Instituut, Sint-Katelijne-Waver, Belgium Cum laude Master thesis 'Multi-resolution OpenGL 3D browser for Computational Graceful Degradation' at IMEC, Leuven

Research Experience

01/10/2002-present:

Pursuing PhD research as a Research Assistant of the IWT-Vlaanderen (Instituut voor de aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen) under the supervision of Prof. Dr. Ir. Bart De Moor, Prof. Dr. Ir. Johan Suykens, and Dr. Ir., Dr.(med) Frank De Smet in the bioinformatics group at ESAT-SCD, K.U.Leuven.

Award

01/09/2006-01/09/2007:

Pursuing postdoctoral research as a Henri Benedictus - BAEF Fellow (of the King Baudouin Foundation and the Belgian American Educational Foundation) in the lab of Prof. Dr. Ir. Kevin Verstrepen at the Bauer Center for Genomics Research, Harvard University.