TECHNICAL UNIVERSITY OF CRETE

DEPARTMENT OF ELECTRONICS AND COMPUTER

ENGINEERING

Design and Implementation of Support Vector Machines and Information Fusion Methods for Bio-medical Decision Support Systems

"Σχεδίαση και Υλοποίηση Μηχανών Διανυσμάτων Στήριξης και Μεθόδων Σύντηξης Πληροφορίας για Βιο-ιατρικά Συστήματα Υποβοήθησης Λήψης Αποφάσεων"

IOANNIS N. DIMOU

Design and Implementation of Support Vector Machines

and Information Fusion Methods

for Bio-medical Decision Support Systems

by

Ioannis N. Dimou

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy, PhD

TECHNICAL UNIVERSITY OF CRETE

DEPERTMENT OF ELECTRONIC AND COMPUTER ENGINEERING

Chairperson of the Supervisory committee: Professor Michalis Zervakis

Professor Minos Garofalakis

Professor Dionissios Hristopulos

Associate Professor Alexandros Potamianos

Associate Professor Konstantinos Balas

Associate Professor Evripidis Petrakis

Associate Professor Georgios Karystinos

Abstract

One of the key biomedical changes that took place in the last decade has been the proliferation of computerized medical diagnostic equipment and the consequent abundance of high volume of multimodal biomedical data. Such a raw information volume in principle contains a multitude of features, which can enable the construction of a highly effective representation of the patient's state and vital trends. On the other hand, the information associated with different types of data for the same subject is often controversial and overlapping. The key factor to elevating the effectiveness of diagnostic systems to the next level is to deliver the ability to summarize, mine and fuse the available information at all abstraction levels available and provide clinicians with actionable knowledge. The overall effort placed in this work focuses on this target addressed from multiple viewpoints.

From improving the data preprocessing and imputation methods, to visualizing vital statistics, to creating new kernel mappings for MRS modalities, to extending support vector machine to handle non-positive definite feature kernels, to fusing expert decisions on diagnostic outcomes, the common denominator remains the optimization of a multistep diagnostic process.

A key novelty introduced by this research is the unifying way to model, evaluate and apply standard and generic decision fusion methods. An additional contribution lies in enabling Support Vector Machine classifiers to handle domain specific (MRS) and non-positive definite kernel data, thus significantly extending their application domain.

The concepts and approaches developed in this thesis may provide applicable insight for the algorithmic tools available and hopefully help in every day clinical practice to early diagnose and improve the quality of life for several people.

Περίληψη

Μια από τις κυριότερες μεταβολές που έλαβε χώρα κατά την τελευταία δεκαετία υπήρξε η ευρεία διάδοση ψηφιακού διαγνωστικού εξοπλισμού στον τομέα της υγείας και η επακόλουθη διαθεσιμότητα μεγάλου όγκου πολυδιάστατων δεδομένων. Ο μεγάλος όγκος δεδομένων θεωρητικά περιέχει πληθώρα χαρακτηριστικών, που μπορούν να επιτρέψουν την υλοποίηση πολύ αποτελεσματικών μοντέλων της κατάστασης και των ζωτικών λειτουργιών του ασθενούς. Από την άλλη πλευρά οι διαθέσιμες πληροφορίες πολλαπλών πηγών που αφορούν τον ίδιο ασθενή είναι συχνά αντικρουόμενες και αλληλεπικαλυπτόμενες. Ο κύριος παράγοντας για την αναβάθμιση των διαγνωστικών συστημάτων στο επόμενο επίπεδο είναι η ανάπτυξη ικανότητας να κάνουν εξόρυξη, σύνοψη και συσχέτιση της πληροφορίας και να παρέχουν στο κλινικό προσωπικό εφαρμόσιμη γνώση. Η συνολική προσπάθεια που έχει καταβληθεί στην παρούσα έρευνα εστιάζεται σ’ αυτό το στόχο από πολλαπλές οπτικές γωνίες.

Από τη βελτίωση των μεθόδων προεπεξεργασίας και αναπλήρωσης δεδομένων, στην οπτικοποίηση ζωτικών στατιστικών δεικτών, στη δημιουργία νέων τελεστών για διάγνωση δομών Μαγνητικής Φασματοσκοπίας, στην επέκταση των Μηχανών Διανυσμάτων Στήριξης προκειμένου να χειριστούν μη θετικά ορισμένους πίνακες και τέλος στο συνδυασμό διαγνωστικών αποφάσεων έμπειρων συστημάτων, ο κοινός παρονομαστής παραμένει η βελτιστοποίηση μιας διαγνωστικής διαδικασίας πολλαπλών βημάτων.

Μια σημαντική συνεισφορά που εισάγεται με την παρούσα έρευνα είναι ένα ενοποιημένο πλαίσιο μοντελοποίησης, εφαρμογής και αξιολόγησης καθιερωμένων και νέων μεθόδων συνδυασμού αποφάσεων έμπειρων συστημάτων. Μια επιπλέον συνεισφορά αποτελεί και η εισαγωγή ειδικά σχεδιασμένων τελεστών Μηχανών Διανυσμάτων Στήριξης για ανάλυση δεδομένων Μαγνητικής Φασματοσκοπίας και μη θετικά ορισμένων τελεστών διευρύνοντας με τον τρόπο αυτό σημαντικά το πεδίο εφαρμογής τους.

Οι έννοιες και οι αναλυτικές προσεγγίσεις που αναπτύχτηκαν στα πλαίσια της παρούσας διατριβής θα μπορούσαν να παρέχουν έναν οδηγό για τα διαθέσιμα αναλυτικά εργαλεία και έχουν εφαρμοστεί σε πραγματικό κλινικό περιβάλλον βοηθώντας τις διαγνωστικές προσπάθειες και βελτιώνοντας την ποιότητα ζωής σημαντικού αριθμού ασθενών.

Introduction

Healthcare is a sector relying on information-intensive and sensitive procedures. The automation of diagnostic tools and the computerization of hospital operations have contributed to a steady annual growth of the volume of health related information exponentially by a factor of 2 throughout the last decade. The parallel research track of bioinformatics exhibits even steeper growth rates and affinity to data mining concepts.

The increasing availability of extensive medical datasets has triggered the development of new analytical methodologies in the context of biomedical informatics. While more than 70% of medical informatics research is targeted towards transactional clinical workflows, the area of medical decision support also shows high potential as a second stage information product. From simple summarizing statistics to state of the art pattern analysis algorithms, the underlying mechanisms that drive most medical problems show patterns that can be identified and taken into account to improve the usefulness of computerized medicine to the field-clinicians and ultimately to the patient. The aim is always to explore a problem's feature space, extract useful information and support clinicians in their time, volume and accuracy demanding decision-making tasks. Yet, interoperability at the data collection systems or protocol levels is still a highly sought goal.

The implementation of the above goals poses unique difficulties to the design of efficient next generation decision support systems based on pattern analysis. Currently, despite the revolutionary advances of many vendors in the sensors and visualization domains, the key decision support functionality is still underdeveloped. Clinicians still rely on feature-level markers with rough statistical assumptions and limited inherent predictive capability. Automation -where available- is limited to data access, visualization and marker comparison.

The cumulative research experience presented in this thesis addresses multiple aspects of the above challenge. Data and markers from Magnetic Resonance Spectroscopy, Ultrasound, genomic and blood samples are analyzed either as standalone feature sets or in the context of information fusion schemes. A common factor in the above domains is the difficulty in classifying outlier cases. Whilst many algorithms successfully diagnose mainstream pathological cases, single patients can pose a challenging problem due to inter patient differences. The treatment and side-effects cost involved in a false positive and the unquantifiable cost in loss of health resulting from a false negative diagnosis amplify the above problem. Thus, at technical level the main focus of the present thesis is to explore and enhance kernel based approaches for multiple stages of the analysis process of biomedical data (preprocessing, imputation, classification, decision fusion).

More specifically, the proposed solution approaches evolve around three axes: applying Support Vector Machine (SVM) classifiers in place of simpler legacy models, designing novel data-driven SVM kernels and improving second-order decision fusion using kernel methods. The work is organized and presented conceptually based on the above axes and enriched with and overall conclusions and insights. Each chapter is part of the author's published work, or is in the final stage of the peer review process, in international research journals in the field as shown in Appendix V.

In chapter 1, we describe the overall concept of clinical decision support systems and its current state of the art. Special emphasis is given to the practical challenges of such systems in a real operational context and to the issue of interpretability of the resulting analysis by clinical experts. Key shortcomings of the existing methodologies including missingness of data, feature noise, outcome uncertainty and quantification are analyzed as to identify important research directions. Special emphasis is given to the problem of medical data integration and the applicability of diagnostic results to wider population groups. The common algorithmic aspects converge to the fact that wider adoption of medical decision support systems depends on the maturing of algorithmic tools and supporting technologies.

In chapter 2 we explore the issue that most data mining and pattern analysis efforts face at an initial stage: i.e. the problem of sparse datasets. Delving into the various forms of data missingness and associated imputation methodologies, this part concludes by quantifying the impact of imputation on various classification schemes. The extensive evaluation of imputation algorithms in connection with strong nonlinear classifiers reveals their advantages and limitations and leads to practical model selection criteria.

In chapter 3, the novel concept of Generalized Space Support Vector Machines (GS-SVMs) is introduced as a method to circumvent the difficulties imposed by the requirement for positive definite SVM kernels. GS-SVMs are derived by extending SVMs to an additional hidden-space kernel level, which in any case assures compliance with the Mercer's conditions requirement of SVMs. Thus, arbitrary functionals can be used as kernel functions and custom kernels can be automatically defined based on the training data. This concept opens a multitude of possibilities of applying SVMs to previously intractable problems due to lack of appropriate kernel functions.

In chapter 4, the effort in application-driven kernel design is further pursued by creating a custom SVM kernel to handle a Magnetic Resonance Spectroscopy diagnostic problem. We present this case study through the description of the entire process of bringing a dataset to the right form for processing, imputing missing values, training using the custom kernel, and thoroughly testing two nonlinear diagnostic models. The overall results indicate that a limited spectral feature subset can yield comparable of better diagnostic results for brain lesion classification than standard SVM implementations.

Chapter 5 deals with higher level decision fusion in the form of classifier ensembles. Both in theory and in experiments fusion schemes are known to contribute in performance and robustness in a variety of classification problems. This work proposes the use of strong nonlinear classifiers as combiners to leverage their maturity and statistical foundations. Implementations of a wide range of established decision fusion methods are evaluated against the above combiners. More importantly, the assumptions and limitations of the various methods of aggregating outcomes are compared both in theory and practice. The findings indicate that trainable and especially discriminant hyper-classifiers provide better outlier mapping and error bounds over most single classifier and legacy combiners.

The last chapter (Chapter 6) presents overall conclusive remarks and insight based on the entirre research effort. Gaps and further interesting future research directions are identified. The author emphasizes the highly promising research opportunities from "kernels on sets" and the integration of human expert knowledge with machine learning

systems raw throughput capabilities. Additional subjects as “concept drift” and multimodal feature descriptors are also open to research in the signal processing area.

The key scientific impact of the presented work is associated with the shortcomings of the current state-of-the-art in the field. More secifically, we address:

- The need to match the appropriate data imputation algorithm to not only the missingness model but also to the classification algorithm.
- The opportunities resulting from utilizing application-specific and data-dependent kernels in biomedical problems.
- The need to leverage classifier ensembles in a combination scheme with prescribed performance levels to handle complex decision mapping requirements of difficult datasets.
- The objective quantification of classifier fusion schemes’ capability on a specific dataset via the corresponding error bounds prior to training.
- The practical needs for best practice guidance in selecting the optimal algorithms especially in the data imputation and classifier fusion stages.

The combined contributions on the above algorithmic aspects result in a sound foundation for analysis of medical diagnostic problems.

Acknowledgements

I would like to thank my supervisor Prof. M. Zervakis for his support and motivation in carrying out this project.

I would also like to thank the members of the supervising committee, Prof. Minos Garofalakis, Prof. Dionissios Hristopulos, Assoc. Prof. Alexandros Potamianos, Assoc. Prof. Konstantinos Balas, Assoc. Prof. Evripidis Petrakis, Assoc. Prof. Georgios Karystinos for their time and useful remarks.

Many thanks also due to Prof. Sabine Van Huffel, Prof. Dirk Timmerman and Ben Van Calster and the ESAT research team at Catholic University of Leuven, Belgium, for their help and guidance in significant parts of this research.

Financial support for this study was provided in part by EU: BIOPATTERN (FP6-2002-IST 508803). The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report.

Table of Contents

List of Acronyms and Abbreviations

AUC	Area Under ROC Curve
BNT	Bayesin (Belief) Network
CC	Complete Cases
CV	Cross-Validation
DA	Data Augmentation
EM	Expectation-Minimization ALgorithm
GLM	Generalized Linear Model
HD	Hot-deck imputation
IOTA	International Ovarian Tumor Association
LR	Logistic Regression
LDA	Linear Discriminant Analysis
LS-SVM	Least-Squares Support Vector Machine
MI	Multiple Imputation
MLP	Multi Layer Perceptron
RBF	Radial Basis Function
PCA	Principal Components Analysis
ROC	Receiver Operating Characteristic
RVM	Relevance Vector Machine
SVM	Support Vector Machine

List of Figures

LIST OF TABLES

Data mining has entered a golden age, whether being used to set ad prices, find new drugs more quickly or fine-tune financial models.

NY Times 06.01.2009

Equation Chapter (Next) Section 1

1 Computational Methods and Tools for Decision Support in Biomedicine

1.1 Introduction

Biomedical informatics (BMI) constitutes an expanding research field incorporating clinical and biological knowledge with information science, signal processing and intelligent systems. This chapter presents a thorough review of this field and highlights the achievements and shortcomings of each family of methods thus creating the context for the specific contributions described in the following chapters.

The main focus is in identifying the most effective methods in handling different diagnostic problems and focusing on the research gaps that need to be addressed.

1.2 Background

Due to the advances in new molecular, genomic, and biomedical techniques and applications such as genome sequencing, protein identification, medical imaging, and patient medical records, huge amounts of biomedical research data are generated on a daily basis. Originating from clinical practice, these biomedical data are available in hundreds of public and private databases, accessible via healthcare intranets or over the Internet. The digitization of critical medical information such as lab reports, patient records, research papers, and anatomic images has also resulted in large amounts of patient care data. Biomedical researchers and most importantly clinical practitioners are now facing the "info-glut" problem.

Currently, the rate of data accumulation is much faster than the rate of data interpretation. These data need to be effectively organized and analyzed in order to be useful. This fact opens new application prospects calls for more robust and practical analysis strategies.

Health informatics or medical informatics is the intersection of information science, computer science, and health care. It deals with the resources, devices, and methods required to optimize the acquisition, storage, retrieval, and use of information in health and biomedicine. Health informatics tools include not only computers but also clinical guidelines, formal medical terminologies, and information and communication systems.

Contemporary and future methods of healthcare delivery will be exploiting new technology, novel sensing devices and a plethora of modes of information generated by distributed data sources. This raw data is inevitably increasing in volume and complexity at a rate faster than the ability of primary healthcare providers to access and understand it. Several countries are currently considering issues of integrated personalized healthcare and the application of ‘intelligent’ data mining methodologies in providing medical decision support to the clinician (and the individual), using principled pattern recognition methodologies.

Within such an environment, the domain of medical imaging, with its various structural (CT, MRI, U/S) and functional (PET, fMRI) modalities, is probably on the top of the list with respect to the amount of raw data generated. Most of these modalities are explored in other chapters of this volume. Even though image inspection by human

experts enables the accurate localization of anatomic structures and/or temporal events, their systematic evaluation requires the algorithmic extraction of certain characteristic features that encode the anatomic or functional properties under scrutiny. Such imaging features, treated as markers of a disease, can subsequently be integrated with other clinical, biological and genomic markers, thus enabling more effective diagnostic, prognostic and therapeutic actions. It is the purpose of this chapter to address issues related to the decision making process, to trace developments in infrastructure and techniques, as well as to explore new frontiers in this area.

1.2.1 The medical informatics revolution

During the last decades, we are witnessing a gradual shift in the medical field. Medical professionals are increasingly being supported by advanced sensing equipment. These instruments provide objective information and assist in reducing the margin of error in diagnosis and prognosis of diseases. Detailed imaging techniques provide accurate anatomic and/or functional maps of the human body, and advanced signal processing methods performing biosignal and biochemical analyses are now largely automated, faster and increasingly accurate. In the broader medical research field, larger datasets of patients including multiple covariates are becoming available for analysis.

Figure 1.1 outlines the information flow in a medical decision support system. At an initial stage, a large amount of data is collected from various sensors and pre-processed. This data is accessibly stored in a structured format and fused with other information, such as expert knowledge. At a higher level, patterns are sought in the full dataset and translated in an intelligent way to produce meaningful and helpful reasoning. This output supports healthcare professionals during their prognostic, diagnostic and other decision making tasks. At the end of this process, feedback to the system in the form of expert evaluation or validity of analysis can be incorporated to improve performance.

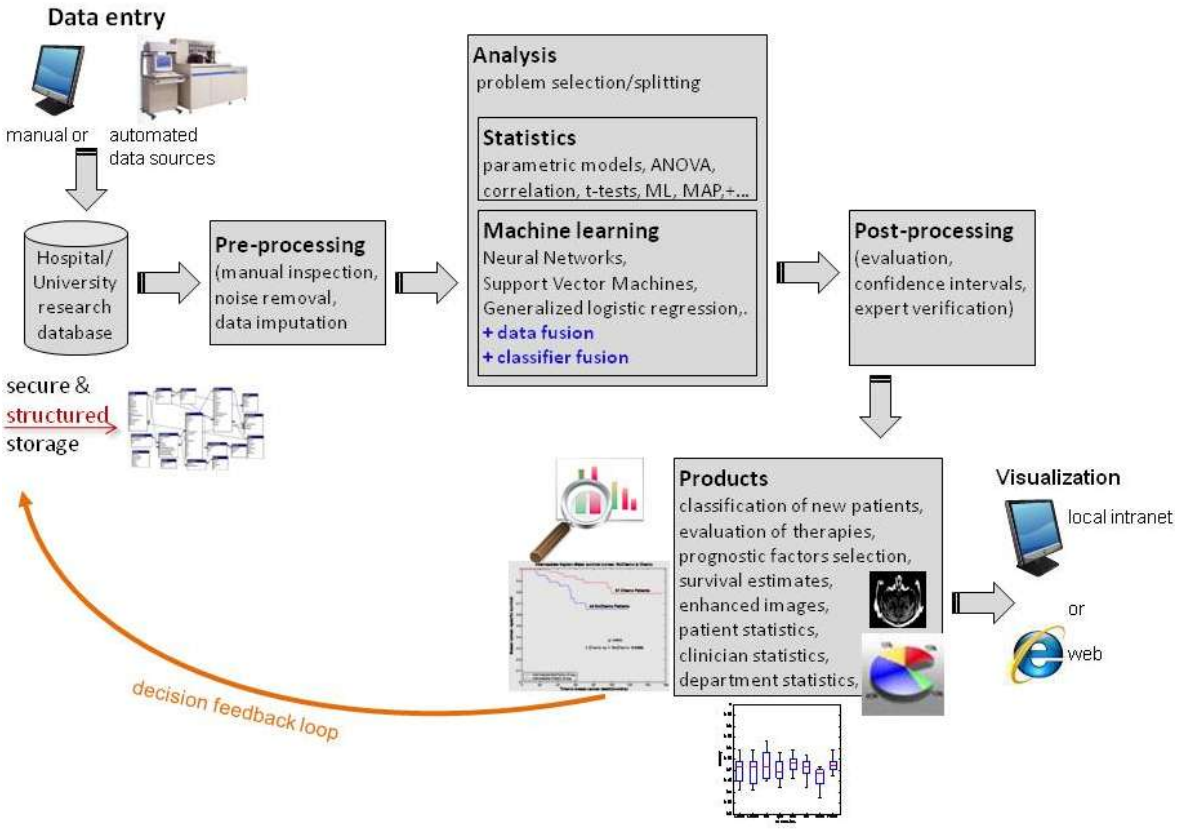


Figure 1.1 Medical informatics decision support system dataflow

This relative data abundance has resulted in a corresponding explosion of scientific papers referring to thorough statistical analysis with data mining and pattern classification techniques. New findings are more easily made available to the scientific community through the internet and cheap processing power aids the development of complex models of diseases, drugs, and effects.

In this context, the field of medical informatics emerges as the intersection of information technology with the different disciplines of medicine and health care. It deals with the resources, devices, and methods required to optimize the acquisition, storage, retrieval, analysis and use of information in health and biomedicine (Bemmel and Musen 1997). Medical informatics tools include not only computers but also clinical guidelines, formal medical terminologies, and information, communication and decision support systems. It is by now evident that medical informatics do not just provide information but also summarize it in an intelligent and comprehensive form.

1.2.2 Current State in medical informatics

According to the latest statistics, an increasing number of health care institutions are initiating the shift to informatics both for internal use and for patient accessibility. Funding for Healthcare Information Technologies (HIT) is at an all time high reaching in some countries 3.5% of their healthcare budget (Steele 2002). In most cases, such efforts are not standardized within countries and the expectations are much lower at a global scale. Although initially driven by requirements to support logistics, billing, and patient administration, today's healthcare information systems face a growing interest for higher level applications such as diagnostic decision support and treatment evaluation.

Serious efforts have been made in the last five years in creating standards for a patient's Electronic Health Record (EHR) (Katehakis, Tsiknakis et al. 2002) to facilitate data exchange not only between hospitals, labs and clinicians, but also between institutions and countries. The task of structuring an individual's medical status and history in a common format has been undertaken by a number of vendors. There are three main organizations creating standards related to EHR: HL7, CEN TC 215 and ASTM E31. HL7 operating in the United States, develops the most widely used health care-related electronic data exchange standards in North America (HL7 RIM, Clinical Document Architecture, CDA), while CEN TC 215 operating in 19 European member states, is the pre-eminent healthcare information technology standards developing organization in Europe. Both HL7 and CEN collaborate with the ASTM that operates in the United States and is mainly used by commercial laboratory vendors. ASTM's Continuity of Care Record (CCR) standard has recently been criticized for being too focused on patient file transfer in contrast to CDA's adaptability to emerging future needs and applications (Ferranti, Musser et al. 2006). In the medical imaging field DICOM is the most widely used standard (W. Dean Bidgood, S. C. Horii et al. 1997). A broader framework termed *integrating healthcare enterprises* (IHE, www.ihe.com) is an initiative by healthcare professionals and industry to improve the way computer systems in healthcare share information. IHE promotes the coordinated use of established standards such as DICOM and HL7 to address specific clinical need in support of optimal patient care. Systems developed in accordance with IHE communicate with one another better, are easier to implement, and enable care providers to use information more effectively.

The combination of EHR standards with the upward trend in funding for HIT creates good prospects in the field. However, a decisive factor in the expansion of such technologies in the applied domain is their acceptance by the healthcare personnel. There is some skepticism regarding the extensive use of computerized tools for decision support. Starting from individual data safety to reliability of automated systems, to proper training and acceptance, the attitude of clinicians is often positive only in institutions focused on research.

1.2.3 A Diversity of Applications

Within such a context and mentality, a number of key application areas can be identified in which automated data processing is already or can be effectively applied. Disease diagnosis is probably the most important application of informatics in medical decision-making. It involves the evaluation of criteria that can discriminate between different pathologies. Prognosis of the onset of pathology is also among the leading applications of pattern recognition tools in medicine (Bates 2002). Its utility ranges from preemptive advice to statistical analysis of risk for insurance companies. At subsequent stages, following a patient's initial diagnosis, treatment evaluation is a key issue, which involves monitoring of disease progress and correlation of observed results with the treatment plan; the objective being to quantify the effect of the chosen treatment on the pathology.

From a similar perspective, modeling of patient survival utilizes ideas from statistical risk and survival analysis and censoring models to estimate the probability of a health event (i.e. relapse, death) –due to a specific disease– of a patient at a specific time (Taktak, Fisher et al. 2004). It is also used for patient grouping and allocation to specialized

reference centers. Moreover, the statistical results from the survivability analysis are used to optimize follow-up monitoring of patients in a personalized manner.

At a lower level, electronic health records (EHR) are gradually becoming a reality providing near real-time data access, and thus dramatically expanding the clinician's opportunities for timely response. Although not an application in itself, these clinical information systems and the related standards elevate the functionality of upper layers of decision support systems by providing a unified structured and meaningful representation of multimodal and diverse data.

From a broader epidemiological standpoint, automated data processing has made possible the large-scale analysis of populations with the objective of identifying disease traits and epidemiological patterns that were previously beyond the reach of a single researcher.

All the above applications share common analytic aspects but at the same time pose a number of difficulties for the biomedical engineers, as they indicate a transition from disease management to personalized treatment (Shortliffe and Cimino 2006). Can we put together rational structures for the way clinical evidence is pooled, communicated, and applied to routine care? What tools and methods need to be developed to help achieve these aims in a manner that is practicable, testable, and in keeping with the fundamental goal of healthcare - the relief from disease? In the following sections, we present a perspective on the current state-of-the-art in automated data analysis and decision support.

One of the aspects that needs to be clarified relating to the use of advanced techniques for data handling in biomedicine in the actual interpretation. Analysts tend to think of any given dataset in mere numbers. The problem at hand has to be clarified to them and given in a way that lends itself to processing (classification, dimensionality reduction, preprocessing, presentation). However the engineer that creates, tests and fields a tool to perform these tasks has to keep in mind the nature and the mechanism of the problem. Not all methods are applicable to all problems. Each one has certain assumptions and limitations.

1.2.4 Need for Intelligent Knowledge extraction tools

Advanced bio-sensing methods and technologies have resulted in an explosion of information and knowledge about diseases and their treatment. As a result, our ability to characterize, understand and cope with the various forms of diseases is growing. At the same time, errors in U.S. hospitals cause from 44,000 to 98,000 deaths per year, putting medical errors, even at the more conservative estimate, above the eighth leading causes of death (Steele 2002).

It seems that difficulties and failures of medical decision-making in everyday practice are largely failures in *knowledge coupling*, due to the over-reliance on the unaided human mind to recall and organize all the relevant details (Chen, Fuller et al. 2005). They are not, specifically and essentially, failures to reason logically with the medical knowledge once it is presented completely and in a highly organized form within the framework of the patient's total and unique situation. If we are to reduce errors and provide quality of care, we must transform the current healthcare enterprise to one in which caregivers exercise their unique human capacities within supportive information systems that compensate for their inevitable human limitations.

Therefore, tools to extend the capacity of the unaided mind are required to couple the details of knowledge about a problem with the relevant knowledge from combined, evidenced and validated clinical and genomic data repositories.

1.2.5 Underlying technologies

In this direction there are a number of technologies that serve as foundations upon which the upper layer services can be built. At the data collection level, most researchers and clinicians face the need for common protocols to standardize and assist data collection, allow online incorporation of new cases and comparison of results over diverse population sets. Much like the EHR eases the transfer of health records, common data collection standards (Potamias and Moustakis 2001) are needed to facilitate interoperability of not only patient data but also analysis and presentation tools at levels ranging from institutional to international level.

The emerging GRID (Huang, Lanza et al. 2005) technologies are of great utility in the wide application field of collecting, storing, processing, and presenting the medical data in a way transparent to the end user. Utilizing bandwidth, computational power and other resources optimally, the GRID promises a cost effective solution to the unification of the medical informatics landscape. Although presently limited to research purposes, its use is undoubtedly going to extend the limits of artificial intelligence in medicine.

Moving to a more abstract (re)presentation layer, the development of medical and genomic ontologies and their alignment is a requirement for storing and handling huge datasets in a structured and logically founded way. A lot of research effort has been devoted recently towards this direction primarily in connection to DNA analysis (Alonso-Calvo, Maojo et al. 2007), as explained in section 1.3.5.

1.2.6 Specific Challenges

The breadth and depth of information already available in both medical and genomic research communities, present an enormous opportunity for improving our ability to study disease-mechanisms, reduce mortality, improve therapies and meet the demanding individualization of care needs. The inability to *share* both data and technologies developed by MI and BI research communities and by different vendors and scientific groups, is therefore severely hampering the discovery process (Martin-Sanchez, Iakovidis et al. 2004).

Among the challenges that characterize medical data pattern analysis one can identify data missingness as a key concept (Perez, Dennis et al. 2002). *Missing data* occur due to a number of reasons: inconsistent data entry, poor protocol design, death censoring, inadequate personnel familiarization with the system and inconsistent sensing equipment between collection centres. The patterns of missing data depend on the specific causes of this effect. The optimal way to handle this is in turn based upon the pattern and the covariance of the effect with other model factors. As discussed in a related review paper (Burton and Altman 2004), most often in published clinical research this phenomenon is handled inadequately. Cases or covariates with missing data are discarded or imputed in naïve ways resulting in added bias to the statistical findings. Safe imputation of medical datasets is realistically achievable up to certain missingness ratios. Above these thresholds the affected covariates have to be discarded. In particular, Expectation Maximization (EM) imputation, data augmentation and

multiple imputation are considered effective and statistically sound methods to compensate for this problem.

Apart from incomplete data, *noise* is also a crucial factor in medical informatics data. The human body has itself largely varying characteristics. Sensing equipment also introduce an additional noise component (Bruce 2001). On top of that, examiners assess the raw information in a subjective way depending on the context, their experience and other factors. The final quantifiable and electronically storable result is far from an ideal measurement. Considering this, any biomedical pattern recognition system has to be robust with respect to noise. Common practice dictates that the noise component should be removed as early as possible in the processing sequence. As an additional measure, cross validation of research results can be used in the post-processing phase to minimize output variance.

Closely related to the reliability of the input information is the concept of *uncertainty quantification*. Researchers observe very strong uncertainties in data, models, and expert opinions relating to clinical information (Ghosh 2004). Being able to quantify this factor in every step of the computational process makes it possible to infer bounds on the final decision support outcome. This is far from theoretical. The real world decisions that a healthcare professional has to make require that the information used is as concise as possible. Confidence intervals are already used in commercial imaging diagnostic support software packages. More advanced techniques as Bayesian Belief networks and Dempster-Schafer theory of evidence are still under research.

Finally, a usually overlooked part of any medical problem is that the outcome usually affects human lives. In practice this makes the *misclassification cost largely asymmetric* (Freitas, Costa-Pereira et al. 2007). A false negative patient diagnosis costs far more than a false positive one (usually not even measurable in medical expenses). Yet, in most clinical classification research papers, cost functions are assumed symmetric as a default for simplicity reasons. Another important problem is that the *models' assumptions* are not always analyzed in detail. It is common practice to assume Gaussian distributions or equal class priors due to their mathematical tractability, although they are not applicable to all contexts.

Biomedical datasets usually consist of features of *different modalities*. Data types can range from CT images, to EEGs, to blood tests, to microarray data or experts' opinions (Acharya, Wasserman et al. 1995). All have different dimensionalities and dynamic ranges and require different pre-processing, data mapping and feature extraction and representation methods.

The data is often collected through large scale multicenter studies. The participating centres are usually *distributed* geographically and have to submit new samples online. This creates the need for on-site data reduction in order to be able to transmit and store high volumes of patient data. A pattern classifier should be scalable or modular or be able to run in an automated way before the data has to be transmitted to a core facility.

In such a distributed setting a number of security topics become very important. In our case the term security is considered to be information systems security, which covers all aspects of data or information protection. The measures applied to protect information systems include:

- Authentication & Authorisation: Ensuring that the distributed data or other resources are accessible only to those people who are authorised to do so.

- Confidentiality: The data produced or handled is not exposed to unauthorised persons.
- Accountability: The ability to create an Audit Trail, to observe and chronologically log all actions undertaken by users (patient history data entry, treatment logging, etc).
- Non-repudiation: Strongly related to accountability is non-repudiation, which involves preventing an individual or entity from denying having performed a particular action related to data (records tampering, lowering security level, releasing).
- Privacy: Since much work to be carried out deals with clinical data concerning patients' private information, it should be assured that privacy is well protected. For example: In many cases researchers are able to complete their work without knowing the identity of the patients concerned in their study. This can be accomplished through de-identification and pseudonymisation (Taira, Bui et al. 2002).

The latter topic, patients' privacy, has a great impact on current research efforts due to the nature of post-genomic data and the existing legislation concerning this type of information. Data protection measures need to comply with laws and ethical guidelines, while reassuring patients about the proper protection of their sensitive information. Conversely, data protection measures must not be so restrictive that they inhibit the work of researchers.

Answering the above needs will evidently enhance the scope and results of medical informatics as a practical healthcare tool.

1.3 Methods for Processing Medical Data

Having established some of the main challenges to be addressed by the biomedical informatics research community, we will in the following section present the available methods for processing multilevel medical data.

1.3.1 Functional Taxonomies

Medical data dimensionalities can be high, 1-D time series, 2-D images, 3-D body scans, 4-D spatio-temporal sequences. How this data is processed depends on the eventual functionality required to be reached by the technological tools. The first high level division of techniques depends upon whether or not one is aware of the type or class of ailment, or whether one is more concerned with a generic investigation of health state.

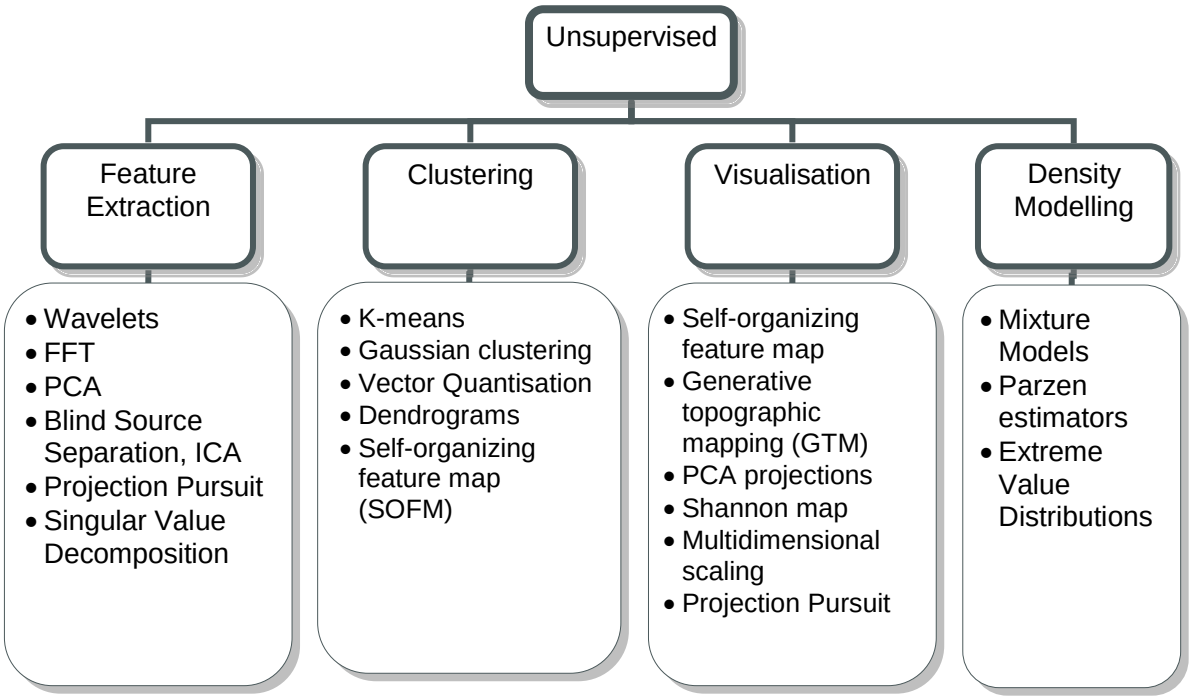


Figure 1.2 Functional separation of unsupervised tasks.

If the problem cannot be posed in a way that one can ascribe an explicit target value of a prognostic or medical indicator, then we are dealing with an *unsupervised* problem. For instance, in dealing with the visualization of a population of thousands of people based on features extracted from ECG waveforms, specific ailments might be irrelevant and we are more interested in how the population is dispersed as a distribution. This would be the visualization problem of unsupervised density estimation. Using such an unconditional distribution would be useful at the individualized level for detecting whether an individual should be considered somehow anomalous, or an outlier of the population. Such information could then be used as a warning signal for further investigation of that individual. Figure 1.2 depicts a functional hierarchy of unsupervised processing tasks, along with exemplar algorithms that implement these functional tasks.

In feature extraction, most techniques are focused either in decomposing the biomedical data into components (ICA, PCA) so that the noise and the signal may be more easily discriminated, or in transforming data into another more appropriate representation (Projection Pursuit, PCA), or simply in data reduction without reducing the relevant signal content. *Clustering* relies on the existence of similarity or dissimilarity measures for separating cases that exhibit a pathology from the general population, taking into account prior knowledge (Bathula, Papademetris et al. 2007) and cause-effect relationships. Some fuzzy classification techniques can additionally account for multiple concurrent pathological causes.

A common such application is the use of ICA for feature extraction from modalities such as electroencephalograms (EEGs) and functional magnetic resonance imaging (fMRI) signals. As an example, (Moosmann, Eichele et al. 2008) present such an application with extensions to synchronous visualization and data fusion. The two signal sources first undergo preprocessing to including correction of head motion related image offsets, temporal filtering and appropriate masking to deal with eye, pulse and magnetic susceptibility artifacts. Independent components are then estimated, back-

reconstructed, averaged across datasets and compared to the original sources with linear regression. This 2-stage process creates both robust feature representations and classification outcomes.

Visualization of high dimensional biomedical data is an important and under-researched area. It is a difficult task since the projection of high-dimensional data onto low dimensional spaces in which data can be visualized requires compromises that necessarily distort the data. Amongst the more common visualization methods are projections onto dominant principal component spaces, the self-organizing feature map (SOFM), the Generative Topographic Map, and a group of methods which explicitly exploit *relative dissimilarity* between feature vectors rather than their absolute distances (Serocka 2007). Another important application of unsupervised methods is related to *density modeling*. This is an explicit attempt to describe the probabilistic structure of medical data in situations that we do not have gold standard target data vectors. In this domain, one is usually interested in describing the unconditional probability distribution of patient's characteristics without knowledge of explicit disease characteristics or states.

If on the other hand one focuses on a specific class of ailment (e.g. benign or malignant cancer), then we are concerned with a *supervised* problem. Supervised problems are exemplified by classification and prediction problems where one is able to collect data involving *ground truth target vectors* for patients with known established outcomes. In supervised approaches labeled data is used as a source for constructing generic models which can then be applied to individuals from a population in which these target labels are not yet known. Many approaches in supervised biomedical tasks come under the categories of *prediction and classification*. Because we have access to a dataset of labeled patient data, parametric or nonparametric models can be constructed in attempting to reproduce the generator of the labeled data without reproducing the noise itself. The basic difference between classification and prediction tasks is in the nature of the target variable; in classification tasks one is concerned with the estimation of a binary vector, whereas in generic prediction the target is typically a continuously varying quantity, such as life expectancy. In this view, supervised classification extends to *function approximation*. Data modeling aims to provide some form of regularization or smoothing to allow noise and outliers to be eliminated. Figure 1.3 depicts this functional taxonomy of supervised tasks along with examples of common algorithms used in each area.

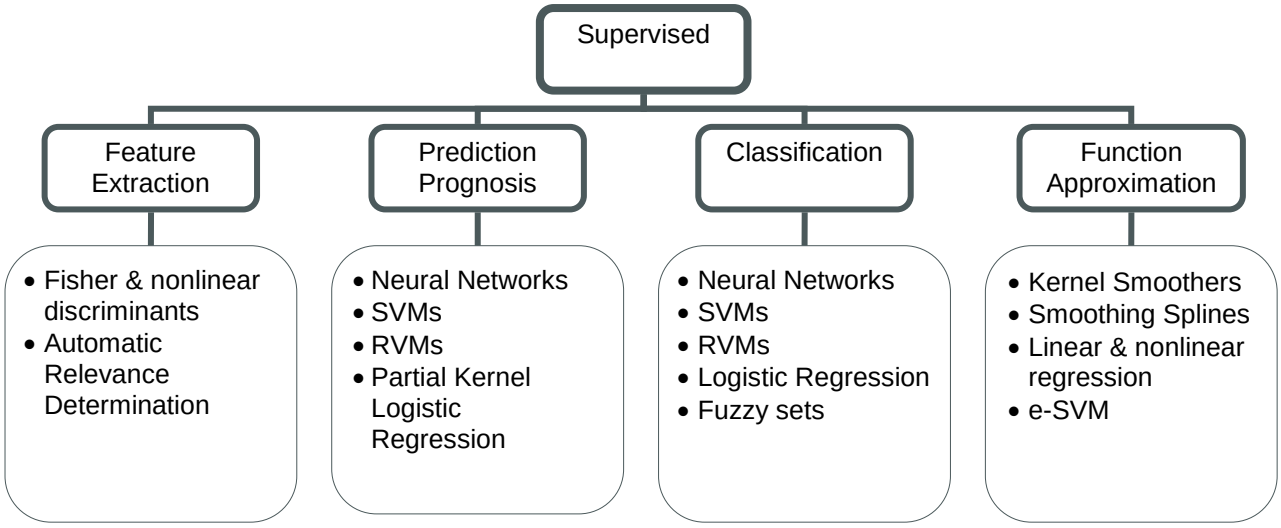


Figure 1.3 Functional separation of supervised tasks, with algorithm examples.

As an illustrative example we can briefly describe the case of classification of SPECT images of Alzheimer's disease patients (Fung and Stoeckel 2007). In this medical imaging classification problem certain preprocessing has to be applied in order to remove noise, amplify edges, centre the region of interest and perform spatial and space normalization before the actual classification mechanism is applied. A support vector machine was trained to identify regions using spatial information. The SVM was primarily used for feature selection by choosing subregions in which cerebral activity deviates highly from a set of baseline images (pre-labeled dataset). Figure 1.4 shows the extracted rectangular regions of voxels overlaid on the respective SPECT image.

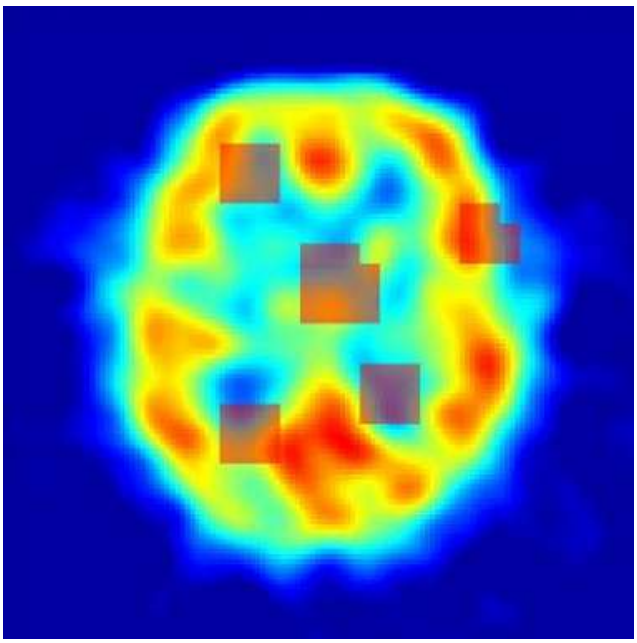


Figure 1.4 Rectangular regions picked by the algorithm overlaid on a SPECT image of an Alzheimer's disease patient.

1.3.2 Model Selection and Evaluation Techniques

Model selection and assessment are crucial common aspects that cut across the supervised/unsupervised boundary and relate to evaluation strategies of algorithmic approaches. In biomedical data processing it is essential to construct low bias models that are robust to fluctuations (in data and model parameters) for stability. Overtraining of adaptive models, or over-parameterisation of parameterized models are two examples of situations to be avoided, in particular in biomedical data processing. Methods for model assessment involve issues linked to the bias-variance dilemma and regularization, either explicitly through the cost functions being used or implicitly through restricting the model class. However, more and more in the intermediate levels of biomedical data processing, the use of single models is being replaced by methods of averaging. This is motivated from the Bayesian perspective of marginalization rather than selection. By averaging over model predictions or over models trained on different data samples drawn from the same distribution, it is possible to compensate for weak or overtrained models. Common methods of averaging include Bootstrap, Boosting, Bagging, Stacking, Bayesian averaging and approximate methods based on sampling, such as Markov chain Monte Carlo methods.

Many novel and promising models are presented and evaluated in the published literature in a way that leaves doubt regarding the variance, reproducibility and reliability of the results (Murphy 2004). Showing high accuracies on a specific dataset instance is not important unless supported by analysis that verifies statistical robustness. A large amount of research work and debate has been devoted to the choice of performance metrics (Eberhart and Dobbins 1990). Optimistically, diagnostic or prognostic models should aim to producing the class-conditional probabilities for combinations of disease occurrence and feature or test presence. Assessment and evaluation needs to also reflect the full tradeoffs between selecting prognostic classes depending upon a threshold, such as the Receiver Operating Characteristic (ROC) curve (Lasko, Bhagwat et al. 2005). Alternatives to the misclassification probabilities are the *predictive values*, which quantify the clinical relevance of the test and involve the reverse conditioning of probabilities. Another description of the prognostic value of a test (or model) is in terms of positive and negative *Likelihood ratios* (McGee 2002). These are specifically relevant since they quantify the increase in knowledge about a disease gained through the output of a diagnostic model.

Apart from the above performance metrics, the medical informatics community faces the need to standardize measures for clustering, visualization, model comparisons and optimization. Such common metrics that should be considered involve scoring metrics (Wilcoxon and Kruskal-Wallis statistics), prediction error variance, entropy measures, mutual information, Kullback-Leibler divergence and dissimilarity metrics, such as Standardised Residual Sum of Squares (STRESS).

1.3.3 Data and Decision Fusion

In pattern analysis it is known that there is no single best algorithm for a specific dataset. Classifier ensembles have in recent years produced promising results, improving accuracy, confidence and most importantly improved feature space coverage in many practical applications. In biomedical problems the combination of multiple classifiers has been effectively applied for diagnosis, gene selection and patient grouping (Dimou,

Manikis et al. 2006). The driving principle behind this type of approaches is that a combination of elementary classifiers can map a feature space effectively, provided that the outcomes are interpreted and combined in a statistically appropriate way. Classifier fusion methods are described and reviewed in (Ruta and Gabrys 2000). A common pitfall in using ensembles is that the base classifiers' outcomes are in many cases not interpretable as probabilities. This limits the choice of the available combiners if the objective is to provide statistical bounds on the fused output. Therefore, researchers should pay attention to the assumptions of each fusion method. The low ratio of positive to negative cases in clinical datasets poses an additional problem for both individual classifiers and combiners. Adjusting for prior class distributions is an efficient way to handle this asymmetry.

Data fusionwill also play an important part of clinical decision support, mainly in imaging related applications. Combining the various types of available information into a single decision boundary is important if all available information is to be utilized (Barillot, Lemoine et al. 1993). In achieving this goal the relative weight of each data source and contextual information need to be accounted for.

As the field becomes better explored and the techniques refined both in medical and analytical level, it is becoming apparent that future research will be focused on certain aspects. Firstly an open problem of crucial importance is the combination of expert knowledge (an experienced doctor's opinion) with statistical-analytical information (numbers in a dataset).

Secondly the feature extraction and selection process needs to be refined. Primarily in gene analysis but also in clinical datasets the number of available features has been increasing steadily. Many of these features are redundant and others are either prone to high noise or need to be combined in markers to provide good model performance. Too many features increase the cost of collection probably decreasing the number of potential cases to be studied.

1.3.4 Medical Data Integration

The nature and amount of information now available opens directions of research that were once in the realm of science fiction. Pharmacogenomics (Roses 2000), diagnostics (Sotiriou and Piccart 2007) and drug target identification are just a few of the many areas that have the potential to use this information to dramatically change the scientific landscape in the life sciences.

During this information revolution, the data gathering capabilities have greatly surpassed the data analysis techniques. If we were to imagine the Holy Grail of life sciences, we might envisage a technology that would allow us to fully understand the data at the speed at which it is collected. Sequencing, localization of new genes, functional assignment, pathway elucidation, and understanding the regulatory mechanisms of the cell and organism should be seamless. In a sense, knowledge manipulation is now reaching its pre-industrial age. The explosive growth in the number of new and powerful technologies within proteomics and functional genomics (Celis, Gromov et al. 2003) can now produce massive amounts of data. However, data interpretation and subsequent knowledge discovery require explicit and time consuming involvement of human experts. The ultimate goal is to automate this knowledge discovery process.

The process of heterogeneous database integration may be defined as "*the creation of a single, uniform query interface to data that are collected and stored in multiple, heterogeneous databases.*" Several varieties of heterogeneous database integration are useful in biomedicine. The most important ones are:

Vertical integration. The aggregation of semantically similar data from multiple heterogeneous sources. For example, a virtual repository that provides homogeneous access to clinical data that are stored and managed in databases across a regional health information network (Martin, Bonsma et al. 2007).

Horizontal integration. The composition of semantically complementary data from multiple heterogeneous sources. For example, a system that supports complex queries across genomic, proteomic, and clinical information sources for molecular biologists.

From the theoretical point of view, there exist three types of database integration methods (Sujansky 2001), namely 1) Information Linkage (IL), 2) Data Transformation (DT) and 3) Query Translation (QT). While IL uses cross references to establish proper links among the data sources, DT creates a centralized repository with a unified schema representing the integration (e.g. Data Warehouses). Conversely, QT focuses the transformation effort in the query and the retrieved results—i.e. a query formulated for the integration is divided into a set of sub-queries, appropriate for the underlying databases. After these queries are launched, retrieved results are integrated and presented to the user in a unified manner. IL is used by a variety of online sources, such as MEDLINE, GENBANK, OMIM, Prosite, etc. DT has been widely used in industrial solutions. However, given the disparate and evolving nature of data in the biomedical domain, the existing privacy issues and the significant size of the databases, QT approaches are more appropriate for mediation solutions in the field.

1.3.5 Approaches to solve syntactic and semantic heterogeneities among biomedical databases

Ironically, huge gains in efficiency in the "front end" of the discovery pipeline have created huge "down stream" inefficiencies because the data cannot be accessed, integrated, and analyzed quickly enough to meet the demands of drug R&D. The industry has outgrown traditional proprietary data capture and integration methods solve only part of the problem. First generation integration solutions that centred on the concept of local repositories have not scaled well, are costly to maintain, and ultimately are limited in long-term usefulness.

To achieve the aforementioned objectives and goals a new breed of techniques, systems and software tools are required for two main reasons:(a) to convert the enormous amount of data collected by geneticists and molecular biologists into information that physicians and other health-care providers can use for the delivery of care and the converse, and (b) to codify and anonymize clinical phenotypic data for analysis by researchers.

Towards the goal of seamless information and data integration (for sharing, exchanging and processing of the relevant information and data items) the need for uniform information and data representation models is raised. The Resource Description Framework (RDF) and XML technology offers the most suitable infrastructure framework towards seamless information/data integration. Based on an appropriate RDF Query Language the generated XML documents can be parsed in order to: (i) homogenize their

content (according to the adopted data-models and ontologies); and (ii) apply dynamic querying operations in order to generate sets of data on which intelligent data processing operation could be uniformly applied.

Syntactically homogeneous access to distributed data sources is typically provided by way of wrappers (Hernandez and Kambhampati 2004; Thiran, Hainaut et al. 2005). One of the main challenges in building wrappers is the variation in the query functionality of the underlying data sources. Data sources may not only use different data models and support syntactically different query mechanisms, but the query capabilities can differ as well. This makes it difficult to support a common query language, an essential step towards syntactic homogeneity. There are two extreme approaches. A highly expressive common query language can be chosen. This, however, makes it difficult to implement wrappers for sources with primitive query capabilities. On the other hand, if a very basic common query language is chosen, significant and unnecessary performance penalties are introduced as the capabilities of the underlying data sources are not effectively used.

As neither approach is ideal, an intermediate solution is proposed in (Martin, Bonsma et al. 2007). A powerful common query language is chosen, but wrappers may choose to only support a subset of the queries, based on the capabilities of the underlying data source. Each wrapper describes the queries it supports using the Relational Query Description Language (RQDL) developed for this purpose. An RQDL specification consists of a set of query templates that represent parameterized queries that are supported. Benefits of this approach are that wrappers can provide and expose exactly the query functionality that corresponds to that of the underlying data source. A drawback is the increased complexity associated with interpreting and reasoning about the query capabilities of each source. It is generally recognized that writing wrappers requires significant programming effort, and as a result research efforts have been devoted to automating parts of this (Thiran, Hainaut et al. 2005).

The largest barrier to heterogeneous database integration is the variety with which similar data are represented in different databases, i.e., semantic or representational heterogeneity. It is appropriate to consider several types of representational heterogeneity that schema integration techniques must resolve. The most general type of heterogeneity is related to the data models themselves. Aggregating data from relational, hierarchical, object-oriented, and flat file databases into a single representation is the first step in schema integration. However, even if all database systems were to use the relational model, significant semantic heterogeneity would remain. Semantic differences occur when the meanings of table names, field names and data values across local databases are similar but not precisely equivalent.

Semantic interoperability, as an important practical problem, has been tackled from many different angles (Martin, Bonsma et al. 2007). Methods to achieve semantic interoperability largely fall into the following three categories: model alignment, using semantic tags or metadata, and developing shared conceptual references or ontologies. The first approach, model alignment, creates mappings among models to support their semantic interoperability (Klein 2001). The second method is to use semantic tags or metadata, such as the Dublin Core Metadata Initiative. The third approach, which is also the ideal solution to semantic interoperability, is to develop core ontology or a shared conceptual reference model to serve as the common ground for all systems.

1.4 Conclusions

In this chapter we have identified common algorithmic aspects in biomedical data processing by reference to a more generic taxonomy of approaches to pattern recognition. In biomedical data analysis, whether the task is time series or image analysis, microarray processing or histology analysis, there are common themes that emerge, such as the desire to reduce noise, reduce dimension, transform to more suitable representations for subsequent interpretation, extract similarities, and exploit dissimilarities.

In many areas there have been significant advances in international research, notably in the areas of neural networks, optimization and image analysis, and Bayesian approaches to inference. However a several bottlenecks can be identified. Among them is the development of methods that have been genuinely devised to support medical decision making. To this end, tools and techniques that engage clinicians to all stages from data acquisition, to processing, validation and interpretation of results should be largely encouraged.

Biomedical data is notoriously unreliable, noisy, distributed and incomplete. Many tools and methods, however, assume data integrity. There are a proportionally insufficient number of methods which explicitly deal with uncertainty, both in the inputs and the outputs. Only a few methods exist that present predictions along with uncertainties in those predictions.

Developments in other areas, such as complexity, communications and information theory probably have a great deal to offer to biomedical data processing, as algorithmic requirements cross the discipline boundaries. What we have presented in this report is merely a summary of current common aspects reflecting future promise along with an indication that much more research is required in effectively dealing with the pattern processing waterfall. Advances are required in areas such as (a) biomedical ontologies, (b) ontology based integration of heterogeneous biomedical data, and (c) service oriented computational frameworks capitalizing on modern technologies (i.e. Grid) enabling the fast and efficient processing of biomedical data.

Equation Chapter (Next) Section 1

2 Data Imputation in medical datasets

2.1 Introduction

Neglecting missing values in diagnostic models can result in unreliable and suboptimal performance on new data. As shown in a highly cited review paper (Burton and Altman 2004), researchers tend to avoid addressing in depth the problem of missing data in biomedical applications. Frequently, ad hoc solutions are used such as ignoring incomplete variables, omitting cases with missing values, or using a naive imputation technique (e.g. unconditional mean imputation). Such techniques may create more problems than they solve, distorting estimates, standard errors and hypothesis tests (Rubin 1987). Applying a more elaborate imputation technique can provide statistically sound values for the missing cases, thus increasing data quality for developing diagnostic models.

2.2 Background

In females, ovarian cancer is one of the most lethal cancers (Jemal, Siegel et al. 2008). However, an accurate pre-surgical diagnosis of the tumor is essential for effective treatment and improved survival (Vergote, De Brabanter et al. 2001). Therefore, the International Ovarian Tumor Analysis (IOTA) group conducted a multi-center study to obtain extensive standardized diagnostic information for a large number of patients (Timmerman, Valentin et al. 2000),(Timmerman, Testa et al. 2005).

CA-125 is an important tumor marker that is widely used in clinical practice to diagnose ovarian tumors, with elevated levels raising concern for malignancy. However, there is discussion concerning the necessity of recording CA-125 levels (Timmerman, Van Calster et al. 2007). For a subset of patients in the IOTA data, the CA-125 level was not recorded. This is caused by differing practices of the participating centers: some consistently do or do not measure CA-125 levels, while others do not measure CA-125 when the tumor looks clearly benign on ultrasound examination.

In this study, we applied four different imputation techniques for the missing CA-125 values: regression imputation (Little and Rubin 2002), expectation-maximization (EM)(Dempster, Laird et al. 1977), data augmentation (DA)(Schafer 1997), and hotdeck imputation(Ford 1983). We developed models to predict tumor malignancy using imputed data sets, a data set without CA-125, and a complete case data set. This methodology allowed us to evaluate (a) the effect of imputation and (b) to investigate the necessity of CA-125 information for predicting malignancy.

A statistical limitation in real-life imputation studies, as compared to simulation studies, is that the missing values and the missing data mechanism are unknown such that the validity of the imputations cannot be checked. In the present study, the effectiveness (rather than the validity) of the imputations was primarily evaluated through the diagnostic performance of the developed models. Additionally, imputation and complete case analysis (CCA) were compared by investigating the estimated diagnostic model parameters.

In this chapter we explore practical solutions for handling this problem in a real medical diagnostic context. The primary objective involves the evaluation of missing data

imputation for the development of clinical decision support systems. More specifically, we impute missing values of the CA-125 tumor marker for ovarian tumors and evaluate the imputations by developing diagnostic models predicting malignancy of the tumor. Few studies have applied and compared imputation techniques in this context (Pérez, Dennis et al. 2002),(Van der Heijden, Donders et al. 2006).

2.3 Types of missingness

The approach to the above problem depends largely on the missingness mechanism. Three main missingness mechanisms have been identified: Missing Completely At Random (MCAR), Missing At Random (MAR) and Missing Not At Random (MNAR).

Missing Completely At Random (MCAR) refers to the probability of an observation being missing does not depend on observed or unobserved measurements. If data are MCAR, then consistent results with missing data can be obtained by performing the analyses we would have used had there been no missing data, although there will generally be some loss of information. In practice this means that, under MCAR, the analysis of only those units with complete data gives valid inferences.

Missing At Random (MAR) refers to the case that given the observed data, the missingness mechanism does not depend on the unobserved data. In practice MAR reflects the most general conditions under which a valid analysis can be done using only the observed data, and no information about the missing value mechanism.

Missing Not At Random (MNAR) is the more general case when neither MCAR nor MAR hold. This means that in this missingness mechanism the phenomenon is non-ignorable. Even accounting for all the available observed information, the reason for observations being missing still depends on the unseen observations themselves. To obtain valid inference, a joint model of the data and the missingness mechanism is required.

Unfortunately we cannot tell from the data at hand whether the missing observations are MCAR, NMAR or MAR. Additionally in the MNAR setting it is very rare to know the appropriate model for the missingness mechanism.

2.4 Imputation methods

The four imputation techniques that were applied assume MAR. Some other techniques were deemed unsuitable for this specific study due to experimental limitations or the underlying assumptions. For example, multiple imputation (MI) (Rubin 1987; Schafer and Olsen 1998; Bernaards, Farmer et al. 2003) is a statistical method that accounts for uncertainty in the imputations by generating multiple imputed data sets such that standard errors of estimated quantities are not underestimated. MI was not used in our study because it was impractical for this study's experimental setup, and because our primary focus was to investigate how well we predict malignancy in new data.

In order to describe the imputation methods, some mathematical notation is required. Suppose we have *N* data points $(x_n, y_n), n = 1,...,N$ where $x_n \in \mathbb{R}^q$ consists of all q measured variables and $y_n \in \{0,1\}$ indicates the outcome variable. Then, x_n can be split up in the observed values a_n and the missing values m_n.

2.4.1 Regression imputation

Regression imputation (RI) aims at imputing missing values using a regression model based on the complete cases. The regression model predicts CA-125 using the other variables (except the tumor outcome since that is the outcome of the diagnostic models developed later). A regression model was built using stepwise variable selection. CA-125 was bilog-transformed to approach a normal distribution. Regression imputation is a form of conditional mean imputation.

Intuitively, one may be tempted to exclude a significant missingness predictor from a regression imputation model, because the regression parameter for the available cases could be different from the parameter for the incomplete cases. However, the use of missingness predictors is encouraged because it renders the MAR assumption more plausible (versus MNAR) (Schafer and Olsen 1998).

2.4.2 EM imputation

Each x_n is generated by the underlying distribution $p(x|\theta) = p(a,m|\theta)$ where θ are the distribution's parameters. If there are no missing values, θ can be easily estimated. When missing values exist, the EM algorithm is a deterministic method to circumvent the problem of estimating θ without the missing values m by integrating $\log\left(p(a,m|\theta)\right)$ over m yielding the expected log-likelihood $E(l)$ (Dempster, Laird et al. 1977; Schneider 2001). In practice, one iteratively alternates between an E-step in which the missing values m are estimated using initial parameter values θ_0 (first iteration) or estimated parameter values θ_{t-1} (iteration *t*), and an M-step in which $E(l)$ is maximized with respect to θ. The M-step yields θ_t, the updated estimate of the model parameters to be used in the next E-step.

The algorithm's convergence relies on the data set and the stopping criteria. In practice, EM's convergence rate depends on the ratio of missing information and the size of the data set. A variation of the EM algorithm, called regularized EM, was used in this study (Schneider 2001). For the imputation using EM, multivariate normality was assumed. We relied on the robustness of the algorithm against violations of this assumption. A bilog-transformation was introduced for the highly skewed CA-125 levels in order to approach a normal distribution.

2.4.3 Data augmentation

DA (Schafer 1997) is a stochastic rather than deterministic method to circumvent the problem of estimating θ without the missing values m. In DA, missing values are randomly imputed given a missing data distribution based on assumed values for θ. Given a and the imputed values for m, a posterior distribution for ϑ is constructed from which updated estimates for θ are drawn. This procedure is implemented using the Gibbs sampling framework. In this two-step iterative process, the distributions for the missing values and the model parameters stabilize leading to convergence in distribution whereas EM converges in its estimates. For the convergence rate of DA, the arguments

given for EM apply here too. The difference is that, for DA, the values still change after convergence but their distribution remains fixed.

Both the EM and DA algorithms are based on a Bayesian estimation framework. As a consequence they rely on the choice of a prior distribution for the model parameters reflecting prior knowledge of the data. In the existing literature, often an uninformative prior is selected to reflect lack of a priori knowledge (Schafer 1997; Burton and Altman 2004). We experienced little variation when using a different prior. The DA implementation also assumed multivariate normality, and Ca-125 measurements were bilog-transformed to approach normality. As for EM, we relied on the algorithm's robustness against violations of this assumption.

As in the case of EM imputation described above, the CA-125 variable was bilog transformed to account for its highly skewed distribution, whereas other predictors used in the imputation model had to be log transformed to alleviate skewness. Also, to avoid numerical issues, the variables for both EM and DA methods were scaled to the $[-1,1]$ interval.

2.4.4 Hotdeck imputation

Hotdeck refers to the imputation of missing values with observed values from other cases (the donors). Many approaches for selecting a donor exist, and in this work we used a nearest neighbor method (Chen and Shao 2000). In this method, the donor for an incomplete case is the most similar complete case according to a similarity metric. A commonly used similarity metric is the sum of absolute differences for all variables (i.e. except the outcome). This method depends on the existence of similar cases in the data set, which is more likely when *N* is large. To improve the performance, we rescaled the numerical variables to the $[-1,1]$ interval.

Of all *N* cases, there are *C* complete cases $(x_c, y_c), c = 1,...,C$. The missing CA-125 level of patient i ($i = 1,...,N-C$) is imputed using sample x_c, where c is chosen using

$$\arg\min_c \left[d\left(\mathbf{a}_c, \mathbf{a}_i\right) \right] \qquad (2.1)$$

The distance *d*, as explained above, is

$$d(\mathbf{a}_n, \mathbf{a}_i) = \left| \mathbf{a}_n - \mathbf{a}_i \right| \qquad (2.2)$$

In this study, we imputed missing values for the CA-125 tumor marker in a large data set of ovarian tumors that was used to develop models for predicting malignancy. Four imputation techniques were applied: regression imputation, expectation-maximization, data augmentation, and hotdeck. Models using the imputed data sets were compared with models without CA-125 to investigate the important clinical issue concerning the necessity of CA-125 information for diagnostic models, and with models using only complete cases to investigate differences between imputation and complete case strategies for missing values. The models are based on Bayesian generalized linear models (GLM) and Bayesian least squares support vector machines. Results indicate that the use of CA-125 resulted in small, clinically nonsignificant increases in the AUC of diagnostic models. Minor differences between imputation methods were observed, and imputing CA-125 resulted in minor differences in the AUC compared with complete case

analysis (CCA). However, GLM parameter estimates of predictor variables often differed between CCA and models based on imputation. We conclude that CA-125 is not indispensable in diagnostic models for ovarian tumors, and that missing value imputation is preferred over CCA.

2.5 Case study: The IOTA dataset

In females, ovarian cancer is one of the most lethal cancers (Jemal, Siegel et al. 2008). An accurate pre-surgical diagnosis of the tumor is essential for effective treatment and improved survival (Timmerman, Valentin et al. 2000). CA-125 is a tumor marker used to diagnose ovarian tumors, with elevated levels in serum raising concern for malignancy. However, the value of CA-125 in diagnostic models is uncertain (Timmerman, Van Calster et al. 2007) and its use in practice varies widely. Researchers using practice-based data to develop or test diagnostic models often encounter subjects with missing CA-125 values. As shown in a review paper (Burton and Altman 2004), researchers tend to use ad hoc solutions for missing data, such as ignoring incomplete variables, omitting cases with missing values, or using a naive imputation technique (e.g. unconditional mean imputation). Such techniques may create more problems than they solve, distorting estimates, standard errors and hypothesis tests (Rubin 1987). Applying a more elaborate imputation technique can provide statistically sound values for the missing cases, thus increasing data quality for developing diagnostic models.

We conducted this study to compare different methods for missing data imputation for the development of clinical decision support systems and to compare models with and without CA-125. Specifically, we imputed missing values of the CA-125 tumor marker for ovarian tumors and evaluated the imputations by developing diagnostic models predicting malignancy of the tumor.

Few studies have applied and compared imputation techniques in this context (Perez, Dennis et al. 2002; Van der Heijden, Donders et al. 2006).

We used data from the International Ovarian Tumor Analysis (IOTA) group multi-center study, which obtained extensive standardized diagnostic information for a large number of patients (Timmerman, Valentin et al. 2000; Timmerman, Testa et al. 2005). Reflecting the differences concerning the use of CA-125 in clinical practice, some IOTA investigators by default did or did not measure CA-125 levels, whereas others omitted measuring CA-125 levels only when the tumor looked clearly benign on ultrasound examination.

We applied four different imputation techniques for the missing CA-125 values: regression imputation (Little and Rubin 2002), expectation-maximization (EM)(Dempster, Laird et al. 1977), data augmentation (DA)(Schafer 1997), and hotdeck imputation (Ford 1983). We developed models to predict tumor malignancy using imputed data sets, a data set without CA-125, and a complete case data set. This methodology allowed us to evaluate (a) the effect of imputation and (b) to investigate the necessity of CA-125 information for predicting malignancy.

A statistical limitation in real-life imputation studies, as compared to simulation studies, is that the missing values and the missing data mechanism are unknown such that the validity of the imputations cannot be checked. In the present study, the effectiveness (rather than the validity) of the imputations was primarily evaluated through the diagnostic performance of the developed models. Additionally, imputation and

complete case analysis (CCA) were compared by investigating estimated model parameters.

2.5.1 Patients and Sources

The IOTA data set consists of 1066 non-pregnant women with at least one persistent ovarian tumor, collected at nine clinical centers throughout Europe. When a tumor was present on both ovaries, data from the most complex mass were included in the study. The main outcome of the study was the diagnosis of the tumor as benign or malignant. The variables collected in the IOTA data set involved demographic and cancer history information, pain during examination, around 40 morphologic and blood flow measurements to describe the tumor, and the serum tumor marker CA-125. Apart from CA-125, the measurement of which was merely encouraged, investigators had to provide complete data. An extensive description of the data set and the data collection protocol can be found in (Timmerman, Testa et al. 2005). Except for CA-125, missing values in the IOTA data set were structural missings, in the sense that nothing could be measured. For example, if no papillary structures were present, none of the variables describing the papillary structures were applicable. After discussion with the clinical experts, these structural missings were imputed with zero.

Most imputation techniques assume that missing data are 'missing at random' (MAR) or 'missing completely at random' (MCAR). MAR implies that missingness depends on observed information but not on missing information (Little and Rubin 2002), while MCAR implies that the missing values are a random subsample of the data set. MCAR is usually unrealistic. Under MAR, a missing CA-125 value may depend on values of other measurements but not on the (unavailable) value of CA-125. Note that MAR allows that missingness of CA-125 is related to CA-125 levels, yet only indirectly through associations between both CA-125 levels and CA-125 missingness with other variables (Schafer and Olsen 1998). Data are 'missing not at random' (MNAR) when missingness depends on the missing values. Imputation methods assuming MNAR are considerably more complex, and no uniform methods are available. Based on the available information, CA-125 missingness in our data set is likely to largely follow the MAR assumption.

Variable	p-value
Negatively associated with missingness	
Colour Score of intratumoral flow (1-4)	0.039
Papillations (yes-no)	0.038
Time-averaged maximal velocity (cm/s)	0.016
Number of Papillations	0.014
Number of locules (0, 1, 2, 3, 4, 5-10, >10)	0.009
Maximal diameter of papillation (mm)	0.006
Fluid in pouch of Douglas (mm)	0.006
Hysterectomy (yes-no)	0.005
Septum thickness (mm)	0.003
Volume of largest solid component (mL)	0.002
Ratio solid component to lesion	0.001
Volume of the ovary (mL)	0.001
Volume of the lesion (mL)	0.001

Years postmenopause (years)	<0.001
Ascites (yes-no)	<0.001
Postmenopausal (yes-no)	<0.001
Age (years)	<0.001
Max. diam. of largest solid component (mm)	<0.001
Maximal diameter of the lesion (mm)	<0.001
Maximal diameter of the ovary (mm)	<0.001
Positively associated with missingness	
Resistance index	0.045
Pulsatility index	0.015
Acoustic shadows (yes-no)	0.001
Categorical variables	
Origin of tumor (ovarian or other)	0.005
Locularity (6 categories)	<0.001

Table 2-1 Significant CA-125 missingness predictors.

The missingness mechanism of CA-125 was investigated by checking the association of CA-125 missingness (a binary indicator with value 1 if CA-125 was missing and 0 otherwise) with other variables using likelihood ratio chi-squared tests. The statistically significant missingness predictors of CA-125 are shown in Table 2-1. These results indicate that there was a strong relationship between CA-125 missingness and other covariates, such that it is unlikely that the missing values are MCAR. When MCAR does not hold, CCA very often leads to biased results. Because CA-125 was less often measured for tumors that looked clearly benign on ultrasound, CCA is likely to be biased in our application.

2.5.2 Evaluation using diagnostic models

Two types of diagnostic models were implemented: least squares support vector machines (LS-SVM) (Suykens, Van Gestel et al. 2002) and generalized linear models (GLM) (Hosmer and Lemeshow 2000). Support vector machines are flexible models that can automatically account for non-linearity, therefore we used LS-SVMs to complement the GLM models.

2.5.2.1 Developing diagnostic models using generalized linear models

A linear model for binary classification typically links the logit of the probability of 'success' (malignancy) to a linear combination of input variables:

$$\text{logit}\left[P\left(Y=1|\,\mathbf{x}\right)\right]=\mathbf{w}^T\mathbf{x}+b\,. \qquad (2.3)$$

This model has the form of a logistic regression model, a member of the family of generalized linear models (GLM) (Hosmer and Lemeshow 2000). Note, however, that we used this algorithm in a simple way without consideration of interactions or transformations of variables (similar to (6)). A Bayesian approach based on the evidence procedure was used for the GLM models, which naturally protects for overfitting by keeping model parameters small (MacKay 1995).

2.5.2.2 Experimental setup

The IOTA data set is split up in a training set of 754 patients and a test set of 312 patients, using stratification for outcome and clinical center. All four imputation methods were applied to the training and test set data simultaneously.

Diagnostic models were developed for six different scenarios: once for each imputation method, once for a complete case analysis (CCA), and once for a complete variable analysis in which CA-125 is ignored (CVA). Binary variables were coded as -1 versus +1. In the training set, ordinal and continuous variables were rescaled into the $[-1,1]$ interval. In the test set, these variables were transformed accordingly. Variable selection proceeded in two steps. First, 20 variables were selected based on previous variable selection processes done on the same training set using logistic regression and LS-SVMs (Timmerman, Testa et al. 2005; Van Calster, Timmerman et al. 2007). These 20 variables (including CA-125) can be seen as the most important variables for predicting tumor malignancy. Second, separately for each scenario, the 20 variables (19 for CVA) were ranked according to their importance. This was done using automatic relevance determination (20), a Bayesian variable selection method. For the imputation and CCA scenarios, CA-125 was by definition ranked as the most important variable.

After the variable selection process, we developed diagnostic models using the 3, 4, …, 20 (or 19 for CVA) most important variables, resulting in 18 (17 for CVA) models. These models were applied to the test set. Test set performance was evaluated using the area under the receiver operating characteristic curve (AUC).

In a next step, we constructed 100 new train-test set splits (repeated data-splitting) with stratification for outcome. For each of the six scenarios, the 18 (17 for CVA) models were trained on each training set and tested on the accompanying test set. The 100 test set AUCs were summarized by their median.

Finally, to compare model parameters, we fitted a GLM model with all 20 variables on all data: 1066 cases for imputed data sets, and complete cases for CCA. For each imputation method, we computed the percent differences of the parameter estimates with the CCA parameter estimates.

2.5.3 Results

2.5.3.1 Preliminary results

Of the 1066 tumors, 800 were benign (75%) and 266 were malignant (25%). CA-125 was unavailable in 233 patients with a benign tumor (29%) and in 24 of the patients with malignant tumor (9%). This reflects the fact that some clinical investigators did not measure CA-125 when the tumor looked clearly benign. Figure 2.1 presents the estimated distributions of CA-125 for complete cases as well as for all cases after imputation of missing values.

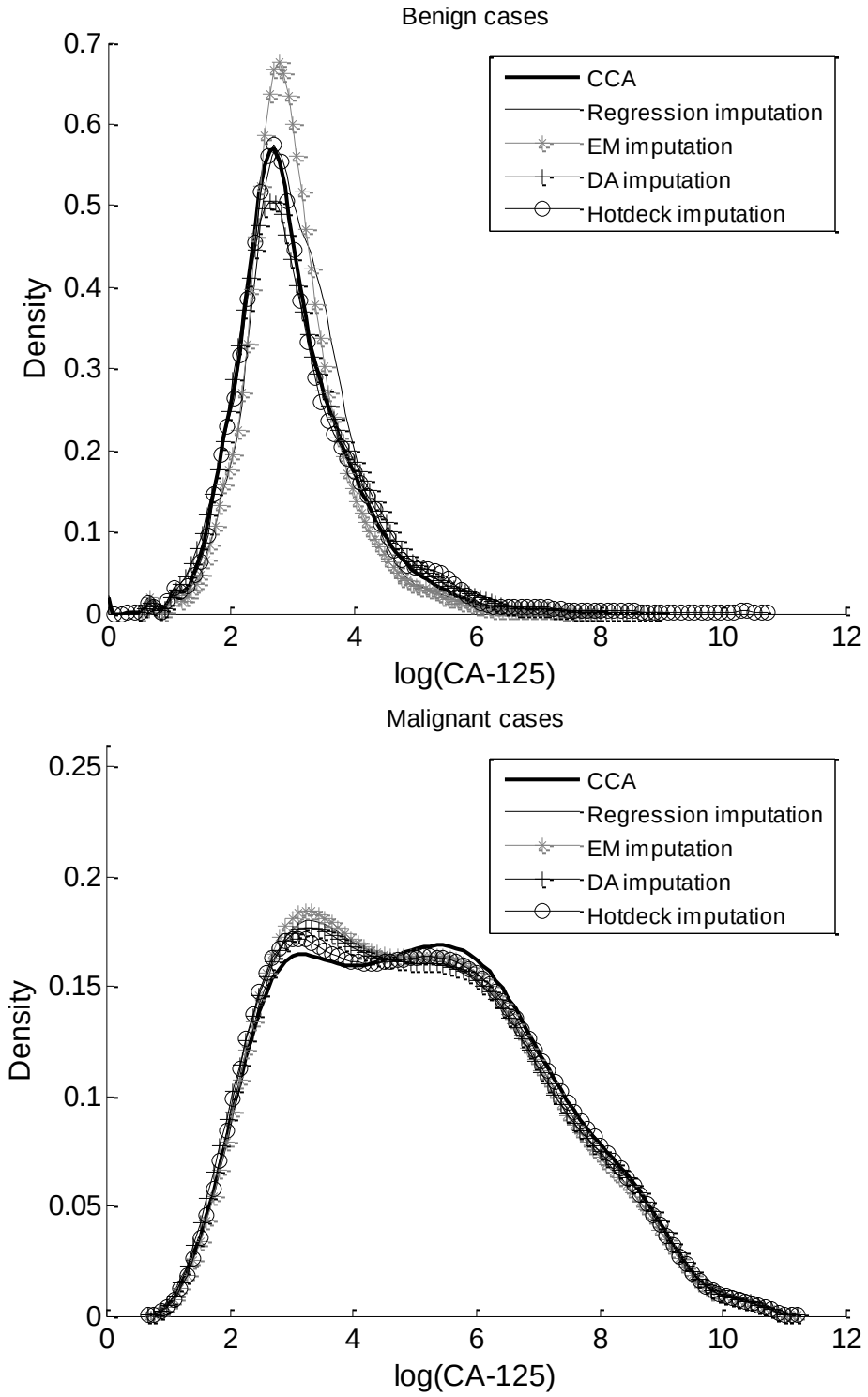


Figure 2.1 Distribution of log(CA-125) in the complete case and imputed data sets for benign (top) and malignant (bottom) tumors.

Using CCA (*n*=809), CA-125 alone has an AUC of 0.826. The AUC slightly dropped when missing values were imputed (*n*=1066): 0.807 for regression imputation, 0.821 for EM, 0.800 for DA, and 0.805 for hotdeck.

2.5.3.2 Diagnostic models

The AUC results for the diagnostic models based on GLMs are given in Figures 2.2-2.3. The results when using LS-SVMs were highly similar, as seen in Figures 2.4-2.5. Figures 2.2 and 2.4 show the performance for the imputation scenarios in the original test set and after repeated data-splitting. Figures 2.3 and 2.5 compare the results after regression imputation with the results based on CCA and CVA. As expected, the original test set results exhibit less performance stability than the results after repeated data-splitting.

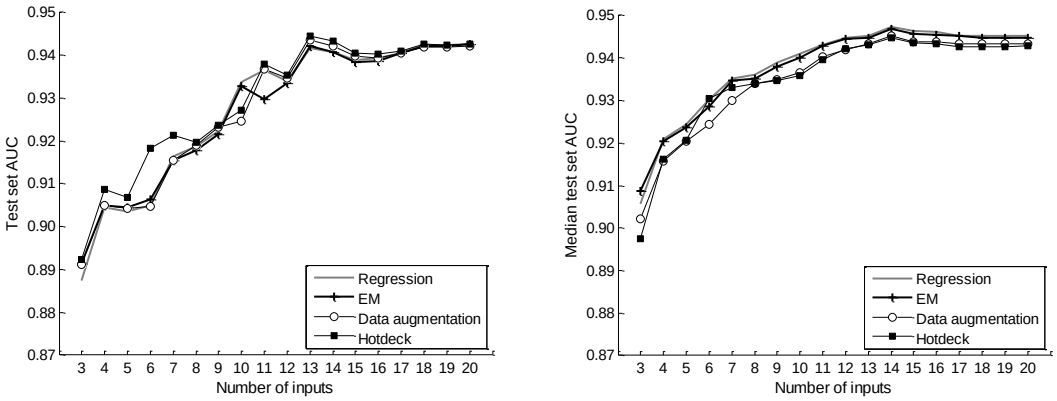


Figure 2.2 GLM diagnostic models' performance evaluated on single test set cases (left) and evaluated using resampling (right).

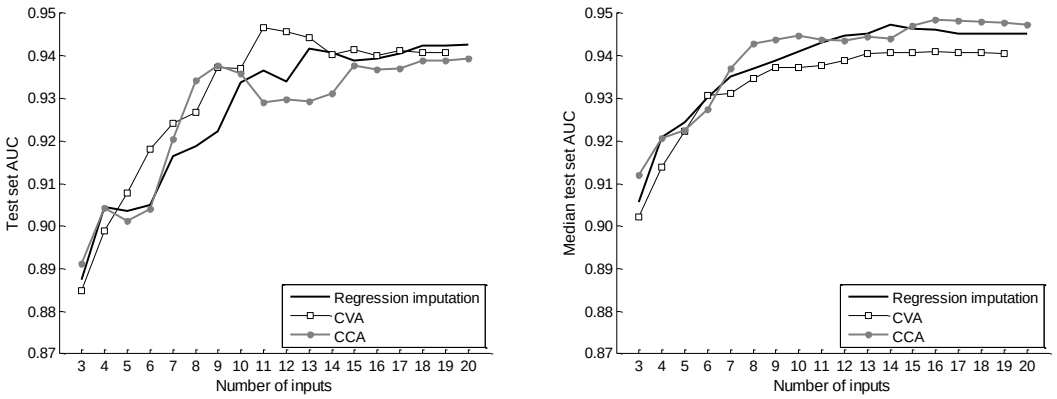


Figure 2.3 GLM performance of best imputation method versus case and variable discard options evaluated on single test set cases (left) and evaluated using resampling (right).

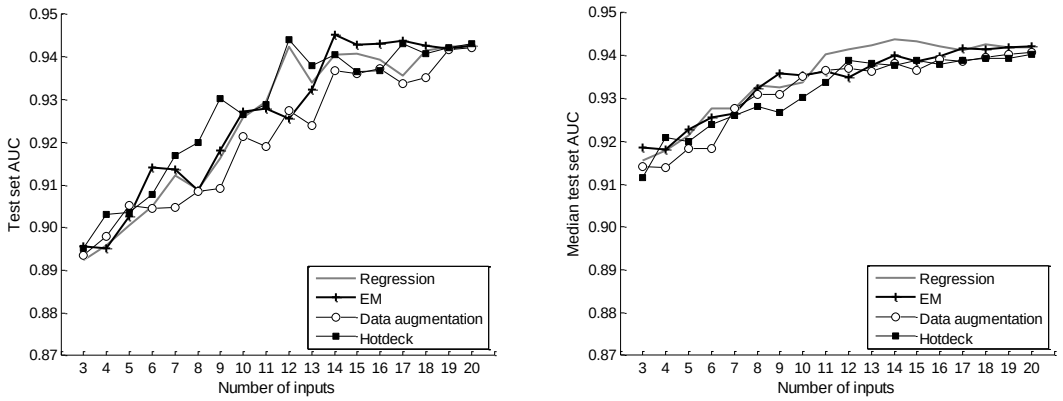


Figure 2.4 Bayesian LS-SVM diagnostic models' performance evaluated on single test set cases (left) and evaluated using resampling (right).

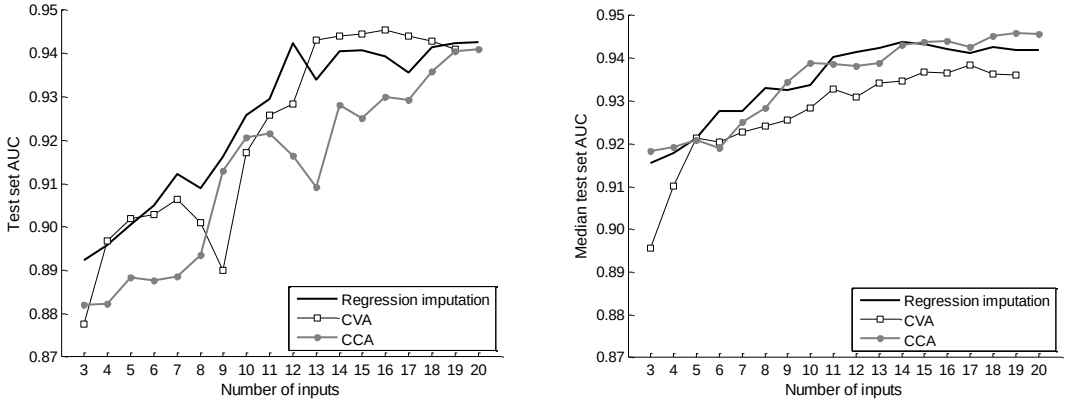


Figure 2.5 Bayesian LS-SVM performance of best imputation method versus case- and variable-discard options evaluated on single test set cases (left) and evaluated using resampling (right).

Differences between imputation methods – Focusing mainly on the repeated data-splitting results, there are minor, not meaningful differences in AUC between the various imputation scenarios. De facto, regression imputation very often resulted in the best median test set AUC. For this reason, the results for the regression-imputed data set were shown in Figures 2.3 and 2.5.

Differences between imputation and CCA – The CCA results were comparable with the results for the regression imputation scenario. After repeated data-splitting, no meaningful differences in median test set AUC were observed. Yet, imputing missing values has the added benefits of making available the whole dataset, and of alleviating bias when estimating particular parameters. Such bias is suggested by the parameter estimates obtained by fitting GLM models on all data. Table 2-2 presents the parameter estimates as well as the percent differences of the CCA estimates with the estimates obtained after missing value imputation. For five variables, all imputation-based estimates suggested that CCA had underestimated the value by at least 20%. For one variable, all imputation-based estimates suggested overestimation of at least 20% when CCA was used. Further, the variable rankings also differed between CCA and the imputation scenarios, suggesting that different variables would be selected after missing value imputation. Variable rankings for the four imputation scenarios correlated

between 0.97 and 0.998, whereas the variable ranking based on CCA correlated only between 0.73 and 0.79 with the rankings after imputation.

Variable[$]	CCA *β* (SE)	Regression *β* (SE)	EM *β* (SE)	DA *β* (SE)	Hotdeck *β* (SE)	% difference between imputation methods and CCA
Intercept	0.834	0.694	1.03	0.827	0.402	
Unilocular-solid tumor*	-0.049 (0.143)	-0.107 (0.138)	-0.081 (0.132)	-0.102 (0.132)	-0.110 (0.139)	+66% to +125%
Age (years)	0.527 (0.306)	0.809 (0.303)	0.795 (0.301)	0.823 (0.296)	0.812 (0.294)	+51% to +56%
Pain*	-0.266 (0.140)	-0.355 (0.132)	-0.355 (0.132)	-0.342 (0.130)	-0.334 (0.130)	+26% to +34%
Max. diam. of lesion (mm)	0.915 (0.427)	1.17 (0.403)	1.17 (0.402)	1.14 (0.388)	1.12 (0.382)	+22% to +28%
Multilocular-solid tumor*	-0.247 (0.136)	-0.328 (0.130)	-0.301 (0.124)	-0.299 (0.123)	-0.317 (0.130)	+21% to +33%
Max. diam. solid part (mm)	1.06 (0.436)	1.26 (0.474)	1.25 (0.468)	1.29 (0.460)	1.27 (0.446)	+18% to +22%
Acoustic shadows*	-0.668 (0.346)	-0.871 (0.308)	-0.768 (0.306)	-0.789 (0.301)	-0.808 (0.286)	+15% to +30%
Unilocular tumor*	-0.791 (0.287)	-0.960 (0.287)	-0.908 (0.283)	-0.900 (0.271)	-0.929 (0.263)	+14% to +21%
Irregular cyst walls*	0.519 (0.162)	0.538 (0.148)	0.545 (0.148)	0.561 (0.145)	0.572 (0.144)	+4% to +10%
Entirely solid tumor*	0.415 (0.163)	0.422 (0.155)	0.450 (0.147)	0.466 (0.145)	0.415 (0.154)	+0.2% to +12%
Hormonal therapy*	-0.303 (0.144)	-0.297 (0.133)	-0.293 (0.133)	-0.305 (0.132)	-0.319 (0.132)	-3% to +5%
Ascites*	0.683 (0.182)	0.647 (0.168)	0.639 (0.169)	0.686 (0.165)	0.734 (0.163)	-7% to +7%
Postmenopausal bleeding*	0.219 (0.228)	0.212 (0.195)	0.218 (0.195)	0.221 (0.192)	0.177 (0.192)	-19% to +1%
History of ovarian cancer*	0.628 (0.416)	0.694 (0.352)	0.670 (0.350)	0.631 (0.343)	0.520 (0.355)	-17% to +11%
Log(CA-125) (U/ml)	1.86 (0.431)	1.94 (0.414)	1.87 (0.397)	1.44 (0.331)	1.37 (0.329)	-26% to +5%
Papillary blood flow*	0.396 (0.187)	0.344 (0.159)	0.350 (0.159)	0.369 (0.157)	0.355 (0.157)	-7% to -13%
Color score	0.791 (0.165)	0.673 (0.149)	0.671 (0.148)	0.684 (0.146)	0.681 (0.145)	-14% to -15%
Nr of papillary projections	0.430 (0.231)	0.374 (0.207)	0.378 (0.206)	0.367 (0.204)	0.360 (0.201)	-12% to -16%
Lesion unequal to ovary*	-0.248 (0.194)	-0.201 (0.162)	-0.201 (0.163)	-0.231 (0.162)	-0.235 (0.159)	-5% to -19%
Multilocular tumor*	-0.245 (0.174)	-0.164 (0.161)	-0.128 (0.159)	-0.129 (0.155)	-0.150 (0.158)	-33% to -48%

* Binary variables: –1 = no, +1 = yes

$ Ordinal and continuous variables were scaled to the [–1, +1] interval

Table 2-2 GLM parameter estimates (β) for CCA analysis and the four missing value imputation methods. Standard errors (SE) were estimated using 1000 bootstrap samples.

The percentage differences between the imputation-based estimates and the CCA estimate are presented as a range.

Necessity of CA-125 – An important clinical question is whether CA-125 is indispensable when constructing diagnostic models for ovarian tumors. Focusing on the repeated data-splitting figures, missing value imputation resulted in slightly higher AUCs compared with CVA. Difference in median test set AUC between the regression imputation scenario and CVA is never larger than 0.01, which is clinically nonsignificant.

2.6 Discussion

Missing values are frequently encountered in medical research, and can result in biased inferences and in suboptimal performance of diagnostic models. In this study, we investigated the use of CA-125 for predicting ovarian tumor malignancy in a data set with 24% missing values for this tumor marker. Next to analyzing complete cases, we also imputed the missing CA-125 values using four single imputation techniques. Diagnostic models were developed using LS-SVMs and GLMs, and were evaluated using the test set AUC. We found that missing value imputation and CCA resulted in comparable AUCs with minor AUC differences between the four imputation scenarios. However, CCA clearly differed from imputation regarding model parameter estimates and variable selection. In addition, the use of imputation resulted in minor, clinically nonsignificant improvements over CVA. We will now elaborate on these findings and their implications.

Firstly, all four imputation methods led to diagnostic models with very similar AUCs. Yet, regression imputation often resulted in slightly higher AUCs de facto. This does not mean that this imputation method produced the most accurate imputations, however. Imputation accuracy could not be verified because the true CA-125 values for missing observations were not known, hence it is hard to conclude which imputation method was preferred for our application. Moreover, the AUC performance is not the optimal index to investigate imputation accuracy. Methods such as EM and DA are superior in theory (Schafer and Graham 2002). For example, EM and DA imputations take into account uncertainty in the imputations by averaging over the distribution for the missing values.

Secondly, it was also observed that CCA and missing value imputation resulted in similar AUCs. Notwithstanding this result, the obtained GLM model parameters suggested that imputation was preferred over CCA. For six out of 20 variables, all four imputation methods suggested that CCA overestimated (or underestimated) the model parameter by at least 20%. Also, CCA resulted in different variable selection results compared with the imputation scenarios. This suggests that missing value imputation has alleviated bias when estimating particular parameters. Note in this respect that this is conditional on the use of a reasonable imputation model, i.e. a rich model including variables related to the imputed variable, variables related to the missingness of the imputed variable, and variables of interest in the analysis following imputation (Rubin 1996; Schafer 1997).

Thirdly, since the advantage in AUC of models including CA-125 over models excluding CA-125 (CVA) was very small, we conclude that it is not necessary to use this marker in diagnostic models for ovarian tumor malignancy. This is an important clinical message that is consistent with earlier findings based on CCA using logistic regression modeling (Timmerman, Van Calster et al. 2007). Our results strengthen these findings because we imputed missing values rather than omitting them, and because we also used the flexible and regularized LS-SVM classifier to build diagnostic models.

Recently, it has been suggested to use the outcome when imputing missing values (Moons, Donders et al. 2006). Even though this may seem a ‘self-fulfilling prophecy’, it makes the imputation model more reasonable because otherwise the imputation model assumes that the association between the outcome and the variable that is being imputed is zero. Therefore, we added the outcome to the regression imputation model to obtain new imputations. Redoing the analyses using GLMs did not alter the

conclusions: the median test set AUCs hardly increased, and were at most 0.01 higher than those based on CVA. This is most likely due to the fact that the imputation model used several other variables that were predictive of malignancy.

Equation Chapter (Next) Section 1

3 Generalized-Space Support Vector Machines

3.1 Introduction

Assuming that the preprocessing and data conditioning stages yield a dataset of adequate quality, the key factor that affects the diagnostic capability of a medical decision support system is the algorithm's match to the data. This chapter introduces the novel concept of Generalized-Space Support Vector Machines within the context of composite kernels explores the properties and practical advantages of the method. The inherent limitation of the underlying algorithm of Hidden-Space SVMs for a linear second stage kernel is surpassed within the context of a more general formulation, which also allows indefinite kernel matrices to be used as feature kernels. This formulation broadens the choice of possible mapping functions and allows the incorporation of useful prior knowledge as invariance to the model, thus providing an improved chance of higher generalization capability of the trained classifiers. An extensive statistical evaluation of the benefits of the proposed extension is presented for multiple evaluation scenarios.

3.2 Background

Since its introduction more than a decade ago (Vapnik 1995), the concept of Support Vector Machines (SVMs) has gained ground as a mainstream pattern analysis tool. The unique mapping capabilities of SVMs balance generalization error with learning error using a two-stage approach that decouples the choice of a mapping function (kernel) from the solution algorithm (quadratic or linear optimization, sequential minimization). The kernel function is a similarity function satisfying the properties of being symmetric and positive-definite (PD) and its choice affects the overall performance of the pattern analysis system (Mika, Ratsch et al. 1999; Qingshan Liu, Hanqing Lu et al. 2004).

Major concerns regarding the efficient use of SVMs relate to the specification of parameters and the selection of the kernel. The first issue is mainly dealt with evolutionary computing approaches. Among the efforts to systematically address the problem of rationalized design of kernel/mapping there are models based on prior knowledge kernels (Fung, Mangasarian et al. 2003), soft subspace linear mappings (Leski 2003) and extensions to classification and clustering methods (Zhou, Zhang et al. 2006). The common denominator of such methods is the utilization of more complex data-dependent kernels. In practice, there are applications that require the use of non-PD kernels (either generic or custom designed). Such examples include kernels that quantify similarity between sets (Eichhorn and Chapelle 2004; Haasdonk 2005), sigmoid kernels (Lin and Lin 2003), sinusoidal kernels used in image classification (Eichhorn and Chapelle 2004), kernels between statistical distributions (Moreno, Ho et al. 2002) and data specific kernels based on Fisher discriminant analysis (Mika, Ratsch et al. 1999; Cristianini, Kandola et al. 2002; Qingshan Liu, Hanqing Lu et al. 2004). Nevertheless, kernels that appropriately characterize the data might not always form admissible kernels that satisfy the Mercer conditions (Vapnik 1995; Burges, Schölkopf et al. 1999).

Hidden-Space Support Vector Machines (HS-SVMs) provide a way to circumvent such design shortcomings and extend the choice of kernels to more general functions (Li,

Weida et al. 2004). The architecture of HS-SVM can be viewed as an effort to incorporate two-level processing on the data. The first layer through the primary kernel attempts to efficiently cluster the data, whereas the second one derives linear boundaries for class-separation of the clustered data. Furthermore, in (Liang 2010) the linear separability of samples in the SVM hidden space is examined and multiple parameterizations of polynomial second stage kernels are evaluated. The findings indicate that the linear decision plane in the hidden space might not be adequately optimal and thus more complex decision mapping may be needed. The authors conclude that, given a linear or low degree polynomial first stage (clustering) kernel, there is a small (~2%) margin of classification accuracy improvement by employing low degree polynomial second stage (classification) kernels. These findings are consistent with (Zhou, Zhang et al. 2006) and focus mainly on adjusting the decision hyperplane induced by using a more complex second-layer kernel. The concept of combining two stages of processing has also been examined in the context of Neural Networks with notable examples the cases of RBF and LVQ networks (Bishop 1995). Both the above types of neural networks employ a first layer nonlinear mapping to cluster the data and a second layer linear mapping to classify the clustered points. In the first layer, the RBF network employs a function-approximation strategy over the data space to derive soft labels describing the degree of belonging of each sample to the classes, whereas the LVQ network operates in a competitive fashion to assign the best-matching class to each sample. SVMs on the other hand resort to high dimensional projections of the input space at a first level aiming to achieve linear separability in the resulting hidden space at a second level. Despite their operational principles, SVMs and Neural Networks have been proven equivalent representations differing primarily in the solution method (Kecman 2001). Thus, the concept of two layer SVM design as in the case of HS-SVMs offers advantages worth of exploring. However, the selection and/or design of kernels are issues that need to be addressed in order to achieve considerable accuracy in HS-SVM classifiers.

In our study, we focus in kernel design rather than linearity properties in the hidden or kernel spaces. We introduce the design of general second-layer kernels and evaluate the use of RBF kernels, towards the incorporation of arbitrary classification kernels. In related literature (Cristianini and Shawe-Taylor 2000; Schölkopf and Smola 2002), this term refers to kernel functionals that are derived by subjecting known simpler kernels (referred to as minor or feature kernels) to specific transformations within the reproducing kernel Hilbert space (RKHS). As mentioned above, the need of the second-layer kernel arises as to satisfy Mercer's conditions by complex, non-positive-definite, or data-dependent kernels. Even though the approach of HS-SVM covers this requirement, the design may suffer from performance inefficiencies that are primarily attributed to the linear form of the secondary kernel. Two interesting aspects relate to the predictive power and the ability to incorporate prior knowledge in the design of kernels.

The combined two-level kernel mapping is actually a "distance-of-distances" (or meta-distance) mapping, which can be beneficial to specific real-world applications. Such mappings are used in DNA analysis for feature reduction (Sammon maps), and lately in social-network analysis to measure cluster compactness and affinity of single points to neighboring clusters. In view of its utility as a distance mapping, the two-layer kernel approach presents specific peculiarities that limit the predictive power of HS-SVMs compared to SVMs, as reported in (Liang 2010)(Liang 2010). More specifically, the first

layer aims to cluster the input data in the augmented space (often denoted by $\varphi(x)$), by means of the distances among the samples defined by corresponding kernel. The second-layer SVM operates on distance vectors and attempts to perform a new clustering based on its own kernel. In this form, the new kernel attempts to perform yet a different clustering on the already clustered vectors, which is not always efficient with the linear kernel. The more general structure of GS-SVM classifier alleviates this problem and can achieve higher performance in terms of cross-validated accuracy.

Another issue facilitated by the design of GS-SVM relates to the incorporation of prior knowledge into SVMs. In (Fung, Mangasarian et al. 2003) the attempt to incorporate prior knowledge by directly adjusting the kernel matrix results to an intractable optimization problem due to non-convex constraint surfaces. In essence, let prior knowledge be in the form of a polyhedral set defined by the set of constraint-boundary weights. The constraint of this form onto an SVM classifier implies a composite kernel combining the original kernel defined on the training data space with another kernel defining distances between the data and weight sets, by considering the projections (inner product) of pairs of vectors for the data and weight spaces, respectively. The composite kernel is not guaranteed to satisfy Mercer's conditions, so that the incorporation of prior knowledge into kernel approaches is not straightforward (Fung et al.). By reformulating the SVM into a two-layer GS-SVM, the knowledge-based kernel can be utilized as a first level kernel, while the actual classification is performed by a nonlinear second-layer kernel. The proposed solution preserves the prior-knowledge constraint isolated in the first level and tolerates possible nonlinearities of the decision boundary at the second stage.

The proposed GS-SVM approach addresses the above problems under the prism of composite kernels, which refers to kernel functionals that are derived by subjecting known simpler kernels to specific transformations within the reproducing kernel Hilbert space (RKHS). The concept of Generalized-Space Support Vector Machines (GS-SVMs) that is developed in this work maintains the primary kernel options provided by HS-SVMs, while providing enhanced second-level mapping capabilities, offering advantages in the freedom of design and achieving increased predictive power. In particular, we demonstrate the increased predictive power of GS-SVM, where the presentation proceeds as follows. Section 3.3 provides the necessary background of SVMs, which is required in order to relate with the proposed expansions, whereas Section 3.4 illustrates the advantages of (positive) indefinite kernels. Section 3.5 provides the definition of the GS-SVM concept and draws comparisons with the HS-SVM formulation. Experimental results on several datasets are provided in Section 3.6 and the chapter concludes with a discussion and summary in section 3.7.

3.3 SVMs background

SVMs have been proposed in part to overcome the problem of poor classifier generalization performance. They form a flexible tool that inherently incorporates nonlinearity in the classification model, which can be influenced by data. In order to formulate the key concepts of GS-SVMs a brief review of the basic SVMs formulation is given.

Let $X=\left\{\mathbf{x}_1,\mathbf{x}_2,..\mathbf{x}_{N_{trn}}\right\}$, $\mathbf{x}\in\mathbb{R}^d$ denote the set of independently and identically distributed training patterns each of dimension d and $y_i\in\mathbb{R}$, $i=1,..N_{trn}$ are the associated hard class labels. The d-dimensional variable space is mapped to a high-dimensional feature space using a mapping $\boldsymbol{\varphi}:\mathbb{R}^d\rightarrow\mathbb{R}^m$.

$$\mathbf{x}_i\xrightarrow{\phi}\mathbf{z}_i=\boldsymbol{\varphi}(\mathbf{x}_i)=\left[\varphi_1(\mathbf{x}_i),\varphi_2(\mathbf{x}_i),...,\varphi_m(\mathbf{x}_i)\right]^T \tag{3.1}$$

In this feature space, a linear (in the parameters) separation function $f(\mathbf{x})$ between the two classes is constructed. The classifier takes the following (primal) form:

$$f(\mathbf{x}_i)=\text{sign}\left[\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i)+b\right] \tag{3.2}$$

where, $\mathbf{x}_i\in X\subset\mathbb{R}^d$, $\mathbf{w}\in\mathbb{R}^m$ is the weight vector, b the bias term. Finally, $f(\mathbf{x}_i)$ is the hard prediction of the model after applying a $sign(\ )$ function to the soft output. Under the standard SVMs formulation (Cristianini and Shawe-Taylor 2000; Schölkopf and Smola 2002), the optimization problem associated with the problem's solution in the primal space is defined as

$$\min_{\mathbf{w},b,\xi}\frac{1}{2}\mathbf{w}^T\mathbf{w}+C\sum_{i=1}^{N_{trn}}\xi_i \tag{3.3}$$

subject to the constraints

$$\begin{aligned}&y_i\left(\mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}_i)+b\right)\geq 1-\xi_i\ ,\\&\xi_i\geq 0,\ i=1...N_{trn}\end{aligned} \tag{3.4}$$

for an input-output pair $\left\{\mathbf{x}_i,y_i\right\}$, where C is the regularization parameter and ξ_i are the slack variables which account for possible outlier samples. This model formulation attempts to express a balance between the margins separating the two classes, which is expressed as regularization (term 1) and minimization of misclassifications (term 2) in eq. (3.3). By solving the corresponding Lagrangian structure, the problem can also be formulated in the dual space as the equivalent classifier

$$f(\mathbf{x}_i)=\text{sign}\left[\sum_{j=1}^{N_{trn}}\alpha_j y_j\boldsymbol{\varphi}^t(\mathbf{x}_i)\boldsymbol{\varphi}(\mathbf{x}_j)+b_i\right] \tag{3.5}$$

where α_j are the weight parameters (support values) of the training cases and y_j denote the corresponding hard labels. Since the mappings $\boldsymbol{\varphi}(\mathbf{x})$ can be complex or unspecified by design, it is more convenient to implicitly work in the feature space by defining a positive definite scalar kernel function that measures the distance between two points $(\mathbf{x}_i,\mathbf{x}_j)$

$$f(\mathbf{x}_i)=\text{sign}\left[\sum_{j=1}^{N_{trn}}\alpha_j y_j k(\mathbf{x}_i,\mathbf{x}_j)+b_i\right] \tag{3.6}$$

which allows us to formulate appropriate hidden spaces that ensure better separability. The corresponding matrix $\mathbf{K}$ with elements $\mathbf{K}_{ij}=\left\{k(\mathbf{x}_i,\mathbf{x}_j)\right\}$, usually referred to as the kernel matrix, can be calculated and stored once for the training set and once for the test set and used repetitively in (3.6).

3.4 Indefinite kernels

The fact that the Mercer conditions significantly limit the pool of applicable kernel functions is increasingly recognized as diverse new application domains of the SVM algorithms are beginning to emerge (Haasdonk 2005; Jean-Philippe Vert, Jian Qiu et al. 2007). This has resulted in the widespread use of only a few admissible kernels, the most prominent of which is the RBF kernel. Despite its optimality and theoretically infinite mapping capability (VC dimension) as defined in (Vapnik 1995), this kernel can suffer from its own generic design. The bandwidth parameter σ is not easy to optimize and can result in considerable performance variations over diverse datasets. This effect, analogous to the "no free lunch" theorem (Duda, Hart et al. 2001), emphasizes the need to be able to adjust the kernel function to the specific problem at hand. The existing knowhow on statistical density estimation kernels has triggered the idea of reengineering and sharing kernel functions between statistics and kernel methods (Genton 2001), especially in time-domain analysis. In practice, this can often lead to analytical forms that are not guaranteed to comply with the SVM's PD conditions.

Additionally many authors have proposed kernels that attempt to incorporate domain-specific feature invariances in the classification model by utilizing approximate estimates of the Kernel matrices. Typical examples of such kernels found in literature include the following.

The sigmoid (hyperbolic tangent) kernel (Lin and Lin 2003) has been evaluated in the past, due to its correspondence to neural networks' sigmoid activation function.

$$k(\mathbf{x}_i, \mathbf{x}_j) = \tanh\left(a_i \mathbf{x}_i^T \mathbf{x}_j + b\right) \qquad (3.7)$$

The above kernel is in general non-PD and has been shown to produce asymptotically equivalent results to the RBF kernel.

The Epanechnikov kernel (Li and Racine 2007) is defined as

$$k\left(x_i, x_j\right) = \frac{3}{4}\left(1 - \left\|x_i - x_j\right\|^2\right) \qquad (3.8)$$

Its analytical form does not lend itself to a proof of positive definiteness. The Epanechnikov kernel is asymptotically optimal with respect to the distribution of features and for this reason it is used as a measure of comparison with other kernels.

The negative distance kernel (Haasdonk 2005) has been proposed in an effort to incorporate translation invariance of feature measurements to the SVM model.

$$k_{ND}\left(\mathbf{x}_i, \mathbf{x}_j\right) = -\left\|\mathbf{x}_i - \mathbf{x}_j\right\|^{\gamma}, 0 < \gamma < 1 \qquad (3.9)$$

Yet, its negative eigenvalues often lead to approximate implementation methods (kernel jittering, kernel smoothing) that introduce minor modifications to the kernel matrix.

Additionally the need for indefinite kernels occurs in many distance-based metrics and when the data structure corresponds to non-Euclidean spaces (i.e. kernels on sets, kernels on trees) (Eichhorn and Chapelle 2004). In fact most-dissimilarity-based kernels of the form

$$k\left(\mathbf{x}_i, \mathbf{x}_j\right) = f\left(\left\|\mathbf{x}_i - \mathbf{x}_j\right\|\right) \qquad (3.10)$$

are not positive definite.

In any case, non-PD kernels can be employed in SVM models with the acknowledged limitation that there is a high risk of numeric overflows, non-convergence to the solution

or converging to a local minimum of the error function surface. This leads to numerous practical difficulties and negates one of the core advantages of SVMs over neural networks namely global error minimization. According to (Haasdonk 2005), "such kernels result to SVMs which cannot be seen as margin maximizers".

3.5 Defining GS-SVMs

Having outlined the baseline of nonlinear soft-margin SVMs and the problems of indefinite kernel functions, we can establish the proposed GS-SVMs formulation. GS-SVMs extend the earlier concept of HS-SVMs (Li, Weida et al. 2004) and they will be presented in parallel.

Both the above methods map the inputs to an N_{trn}-dimensional hidden space $\boldsymbol{\varphi}' : \mathbb{R}^d \to \mathbb{R}^{N_{trn}}$, where the mapping is defined on the basis of all data vectors n the training set. Thus, for an input vector $\mathbf{x}_i$, the mapping $\boldsymbol{\varphi}'(\mathbf{x}_i)$ can be seen as reflecting the relation (proximity or distance) of $\mathbf{x}_i$ with all other vectors in the training set.

Exploiting the functionality of the first-level kernel as a means of projection onto the training space, we may consider this new feature space as a data-dependent mapping defined by the columns of the kernel matrix as:

$$Z' = \left\{ \mathbf{z}'_i \middle| \mathbf{z}'_i \triangleq \left[k\left(\mathbf{x}_i, \mathbf{x}_1\right), ..., k\left(\mathbf{x}_i, \mathbf{x}_{N_{trn}}\right) \right]^T, \mathbf{x}_i \in X \right\}, \quad i = 1, .., N_{trn} \tag{3.11}$$

From (3.11) it becomes clear that the kernelized φ' mappings ($\mathbf{z}'_i$) are special cases of a general form of SVM hidden space mappings φ introduced to allow arbitrary. Proceeding in a way analogous to the standard SVMs' formulation, we can define the HS-SVMs decision function in dual space as

$$f'\left(\mathbf{x}_i\right) = \operatorname{sign}\left[\sum_{j=1}^{N} \alpha_j y_j k'\left(\mathbf{x}_i, \mathbf{x}_j\right) + b_i \right] = \operatorname{sign}\left[\sum_{j=1}^{N} \alpha_j y_j \sum_{n=1}^{N_{trn}} k\left(\mathbf{x}_i, \mathbf{x}_n\right) k\left(\mathbf{x}_j, \mathbf{x}_n\right) + b_i \right] \tag{3.12}$$

where the last part in vector notation forms an inner product of the feature kernels and, hence, is positive semi-definite, therefore qualifying as a valid RKHS kernel. Hence HS-SVMs (and GS-SVMs) are special types of SVMs using a feature transformation (first kernel) that is based on the entire dataset.

Formulating the GS-SVMs, we propose to extend this second-level kernel to a more general functional f" which will itself be a kernel k_g:

$$f'\left(\mathbf{x}_i\right) = \operatorname{sign}\left[\sum_{j=1}^{N} \alpha_j y_j k''(\mathbf{x}_i, \mathbf{x}_j) + b_i \right] = \operatorname{sign}\left[\sum_{j=1}^{N} \alpha_j y_j \sum_{n=1}^{N_{trn}} k_g\left(\mathbf{k}\left(\mathbf{x}_i, \mathbf{x}_n\right), \mathbf{k}\left(\mathbf{x}_j, \mathbf{x}_n\right)\right) + b_i \right] \tag{3.13}$$

resulting in two consecutive kernels as shown in Figure 3.1. The first $\mathbf{k}\left(\mathbf{x}, \mathbf{x}_i\right) = \boldsymbol{\varphi}'\left(\mathbf{x}_i\right)$ maps the original input vector on a space defined by the training dataset. The second one in the GS-SVM approach, i.e. $k''(\mathbf{x}_i, \mathbf{x}_j) = k_g\left(\boldsymbol{\varphi}'\left(\mathbf{x}_i\right), \boldsymbol{\varphi}'\left(\mathbf{x}_j\right)\right)$, defines the similarity metric between two vectors in a way similar to the kernel function in the original SVM formulation. We refer to those two kernels as the feature and the similarity kernels, respectively. Recall that the feature kernel is often termed as minor or mapping kernel.

Notice that the similarity kernel in the HS-SVM formulation is linear, whereas it extends to more general forms in the proposed GS-SVM approach.

In principle, a SVM model utilizing composite kernels in the form of GS-SVMs is a classifier operating on a new multidimensional feature space spanned by the distances of the test sample from all other samples in each class, as defined by the mapping kernel. Therefore, at the first level, the GS-SVM forms compact classes in this new space, while at the second level it computes the similarity of the test sample with the classes in the new space.

Figure 3.1 shows the mappings that constitute the three examined model classes. SVMs map each sample $\mathbf{x}_i$ to the kernel space through an implicit hidden space. HS-SVMs introduce the additional step that is needed to calculate the second-stage linear kernel. GS-SVMs extend this additional step to calculate a more general kernel space. All three methods ultimately map the kernelized features to the output soft labels space through the decision function (19).

In order to be able to utilize such complex kernels we first have to prove their admissibility as Reproducing Kernel Hilbert Space (RKHS) mappings. In general, the sufficient conditions for a function $k''\left(x_i, x_j\right)$ to correspond to a dot product in the feature space F are defined by the Mercer theorem (Mercer 1909):

$$\int_{X\times X} k''\left(x_i, x_j\right) f\left(x_i\right) f\left(x_j\right) dx_i dx_j \geq 0 \quad (3.14)$$

where $f\left(x\right) \in L_2\left(X\right)$.

However direct evaluation of the above expression is often infeasible. In the context of this work related to the formulation of GS-SVMs, the admissibility of the derived above functionals as SVM kernels are can be derived by setting

$$\varphi\left(\mathbf{x}_i\right) = \left[k\left(\mathbf{x}_i, \mathbf{x}_1\right), ..., k\left(\mathbf{x}_i, \mathbf{x}_N\right)\right]^T \quad (3.15)$$

and using the kernel property defined in (Cristianini and Shawe-Taylor 2000)

		Mapping function/ feature kernel	Similarity kernel	Decision function
SVM	$\mathbf{x}_i$	$\boldsymbol{\varphi}(\mathbf{x}_i)=\begin{bmatrix}\varphi_1(\mathbf{x}_i)\\ \vdots \\ \varphi_m(\mathbf{x}_i)\end{bmatrix}$	$k(\mathbf{x}_i,\mathbf{x}_j)=$ $\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$	$f(\mathbf{x}_i)=$ $\text{sign}\left[\sum_{j=1}^{N}\alpha_j y_j k(\mathbf{x}_i,\mathbf{x}_j)+b_i\right]$
HS-SVM	$\mathbf{x}_i$	$\boldsymbol{\varphi}'(\mathbf{x}_i)\triangleq\begin{bmatrix}k(\mathbf{x}_i,\mathbf{x}_1)\\ \vdots \\ k(\mathbf{x}_i,\mathbf{x}_{Ntrn})\end{bmatrix}$	$k'(\mathbf{x}_i,\mathbf{x}_j)=$ $\boldsymbol{\varphi}'(\mathbf{x}_i)^T\boldsymbol{\varphi}'(\mathbf{x}_j)=$ $\sum_{n=1}^{N_{trn}}k(\mathbf{x}_i,\mathbf{x}_n)k(\mathbf{x}_j,\mathbf{x}_n)$	$f'(\mathbf{x}_i)=$ $\text{sign}\left[\sum_{j=1}^{N}\alpha_j y_j k'(\mathbf{x}_i,\mathbf{x}_j)+b_i\right]$
GS-SVM	$\mathbf{x}_i$	$\boldsymbol{\varphi}'(\mathbf{x}_i)\triangleq\begin{bmatrix}k(\mathbf{x}_i,\mathbf{x}_1)\\ \vdots \\ k(\mathbf{x}_i,\mathbf{x}_{Ntrn})\end{bmatrix}$	$k''(\mathbf{x}_i,\mathbf{x}_j)=$ $\sum_{n=1}^{N_{trn}}k_g\left(k(\mathbf{x}_i,\mathbf{x}_n),k(\mathbf{x}_j,\mathbf{x}_n)\right)$	$f''(\mathbf{x}_i)=$ $\text{sign}\left[\sum_{j=1}^{N}\alpha_j y_j k''(\mathbf{x}_i,\mathbf{x}_j)+b_i\right.$

Fig. 1 Hidden space mappings for SVMs (*top*), HS-SVMs (*middle*) and GS-SVMs (*bottom*), where $\mathbf{x}_i$ and $\mathbf{x}_j$ represent the test sample and a second-level scanning of all the dataset's samples respectively.

$$k''(\mathbf{x}_i,\mathbf{x}_j)=k'(\phi(\mathbf{x}_i),\phi(\mathbf{x}_j)) \qquad (3.16)$$

The GS-SVM formulation has a twofold advantage. It satisfies the two contradicting requirements, namely (a) the need for arbitrary custom kernels and (b) the need for positive definiteness of SVMs. The key capability that the GS-SVM introduces relates to the fact that it can construct customized mappings to handle specific classification problems, depending on the choice of k' and k''. For demonstration purposes, we will analyze some basic combinations of known SVM kernels and justify their properties along with the use of a similarity kernel $k_g(.,.)$ in the GS-SVM approach.

3.5.1 Nonlinear similarity kernels in GS-SVMs

Even with traditional SVMs, linear class separability is seldom achievable. An alternative that balances computational requirements with mapping power at the second-level kernel (k'') can be introduced by GS-SVMs in the form of polynomial k'' kernels:

$$k''(\mathbf{x}_i,\mathbf{x}_j)=k_g(\mathbf{k}(\mathbf{x}_i,\mathbf{x}),\mathbf{k}(\mathbf{x}_j,\mathbf{x}))=\left(\mathbf{k}(\mathbf{x}_i,\mathbf{x})^T\mathbf{k}(\mathbf{x}_j,\mathbf{x})+1\right)^p,\ p\in\mathbb{N} \qquad (3.17)$$

An additional option for maximum flexibility of the decision boundaries is the use of a family of parameterized RBF (Gaussian) similarity kernels. This scenario leads to the following GS-SVMs formulation:

$$k''\left(\mathbf{x}_i,\mathbf{x}_j\right)=k_g\left(\mathbf{k}\left(\mathbf{x}_i,\mathbf{x}\right),\mathbf{k}\left(\mathbf{x}_j,\mathbf{x}\right)\right)=\exp\left(-\frac{\left\|\mathbf{k}\left(\mathbf{x}_i,\mathbf{x}\right)-\mathbf{k}\left(\mathbf{x}_j,\mathbf{x}\right)\right\|^2}{2\sigma^2}\right) \tag{3.18}$$

where σ is the function's spread/bandwidth parameter.

Even though any nonlinear function can be employed as a similarity kernel, the choice of RBF as a second-layer kernel is intended as a concept demonstrator for the feasibility and performance of the GS-SVMs using a strong nonlinear mapping. An additional benefit of using GS-SVMs is due to the ability of wrapping and utilizing indefinite feature kernels. For demonstration purposes we use the RBF as a similarity kernel and provide derivations for several data-mapping kernels, including the sigmoid, Epanechnikov, negative distance and Mahalanobis kernel functions.

Sigmoid mapping -RBF similarity kernel:

$$k_{sig-RBF}\left(\mathbf{x}_i,\mathbf{x}_j\right)=k_g\left(\mathbf{k}\left(\mathbf{x}_i,\mathbf{x}\right),\mathbf{k}\left(\mathbf{x}_j,\mathbf{x}\right)\right)=\exp\left(-\frac{\left(\tanh\left(a_i\mathbf{x}_i^{\,T}\mathbf{x}+b\right)-\tanh\left(a_j\mathbf{x}_j^{\,T}\mathbf{x}+b\right)\right)^2}{2\sigma^2}\right)$$

Epanechnikov mapping -RBF similarity kernel:

$$k_{Ep-RBF}\left(\mathbf{x}_i,\mathbf{x}_j\right)=k_g\left(\mathbf{k}\left(\mathbf{x}_i,\mathbf{x}\right),\mathbf{k}\left(\mathbf{x}_j,\mathbf{x}\right)\right)=\exp\left(-\frac{\left(\frac{3}{4}\left(1-\left\|\mathbf{x}_i-\mathbf{x}\right\|^2\right)-\frac{3}{4}\left(1-\left\|\mathbf{x}_j-\mathbf{x}\right\|^2\right)\right)^2}{2\sigma^2}\right)$$

Negative Distance mapping -RBF similarity kernel:

$$k_{ND-RBF}\left(\mathbf{x}_i,\mathbf{x}_j\right)=k_g\left(\mathbf{k}\left(\mathbf{x}_i,\mathbf{x}\right),\mathbf{k}\left(\mathbf{x}_j,\mathbf{x}\right)\right)=\exp\left(-\frac{\left(\left\|\mathbf{x}_i-\mathbf{x}\right\|^\gamma-\left\|\mathbf{x}_j-\mathbf{x}\right\|^\gamma\right)^2}{2\sigma^2}\right)$$

Mahalanobis mapping -RBF similarity kernel:

$$k_{Mah-RBF}\left(\mathbf{x}_i,\mathbf{x}_j\right)=k_g\left(\mathbf{k}\left(\mathbf{x}_i,\mathbf{x}\right),\mathbf{k}\left(\mathbf{x}_j,\mathbf{x}\right)\right)=\exp\left(-\frac{\left(\left(\mathbf{x}_i-\bar{\mathbf{x}}_i\right)^T C_{ij}^{-1}\left(\mathbf{x}-\bar{\mathbf{x}}\right)-\left(\mathbf{x}_j-\bar{\mathbf{x}}_j\right)^T C_{ij}^{-1}\left(\mathbf{x}-\bar{\mathbf{x}}\right)\right)^2}{2\sigma^2}\right)$$

3.5.2 Data-Dependent kernels

The term "data-dependent kernels" (DDK) refers to a set of kernels that utilize information about the whole training set in order to adjust the kernel's form and parameters (Bengio, Delalleau et al. 2004; Huilin 2007).

The kernelized transformation φ' described in (3.11) can be considered either as function-driven $k'\left(\cdot,\mathbf{X}\right)$, as given in section 3.5.1 or data-driven $k'_{\mathbf{X}}\left(\cdot,\mathbf{X}\right)$, corresponding to the concept of data dependent kernels. Data dependent kernels are important in classification problems involving non-stationary patterns or high difficulty problems where tailored kernel designs are unfeasible due to lack of prior knowledge.

Data dependent mappings such as Principal Component Analysis (PCA), Independent Component Analysis (ICA), supervised PCA, Projection Pursuit and Factor Analysis

(Bengio, Delalleau et al. 2004; Koscor and Toth 2004), can be considered as members of this set. In this work, we implement a DDK consisting of PCA mapping in the context of both HS-SVM and GS-SVMs, as an indicative example. We construct a principal component projection matrix along the dimensions with the highest variance $\mathbf{A}_{PCA}(X)$ and use it to construct projections of each data point before calculating the kernel function as shown in eq. (3.19)

$$k_{DDK}(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{A}_{PCA}(X)\mathbf{x}_i, \mathbf{A}_{PCA}(X)\mathbf{x}_j) \tag{3.19}$$

This data dependent kernel is in general not positive definite, therefore its performance in conjunction with GS-SVMs is of interest.

Notice that both HS-SVM and GS-SVM formulations admit all non-PD kernels but their difference is in the combination of the second-layer (similarity) kernel in order to partition the output space. The GS-SVMs provide an additional flexibility for nonlinear coupling of the first layer (mapping) functions.

While it cannot be assumed in theory that a double nonlinear composite SVM kernel is in all cases preferable, compared to a single nonlinear mapping (for the same reason that highly complex mapping functions are not by definition better classifiers for all problems), we provide a context that can incorporate additional combined kernel options and evaluate a subset of these kernel options experimentally.

3.6 Experimental Results

In order to evaluate the performance of GS-SVMs, we use typical feature kernels (linear, polynomial, RBF) along with known non-PD functions as feature kernels, coupled with the proposed second-level functionality of nonlinear similarity kernels. Comparisons are conducted among the formulations of SVM, HS-SVM and GS-SVM. In this section we aim at comparing the formulations on the basis of the several PD and non-PD feature kernels rather than optimizing the overall performance approaches considered. Thus, for fair comparison on the first-layer, we keep the classification kernel fixed to linear (HS-SVM) and RBF (GS-SVM). The evaluation of various schemes is reported on six artificial and three real biomedical benchmark datasets. The kernel functions used are partitioned into three distinct kernel families namely SVMs, HSSVMs and GSSVMs. All models were implemented using the PRTools toolbox (Duin 2000; Heijden 2004) for Matlab, which allows flexible object-oriented modeling of datasets, classifiers and metrics.

3.6.1 Artificial datasets

In this part we consider the efficiency of GS-SVMs on three pairs of artificial datasets. We use 2-class, 2-dimensional datasets generated as described in (Raudys and Jain 1991) using PRtools functions for Gaussian, Lithuanian and correlated-Gaussian clusters, parameterized to reflect a high and a low separability scenario for each type of dataset. Figure 3.2 shows the 3 pairs' class distributions in the 2-d feature space.

The three approaches are tested with three PD and four non-PD feature kernels. The evaluation results in terms of 100-fold cross validated accuracy are presented in Figure 3.3.

The typical SVM models behave consistently with theory achieving higher accuracies using the typical linear, polynomial and RBF feature kernels, and degraded performance

when using non-PD feature kernels especially in the low separability and correlated features scenarios. As explained in section 3.5.1, the SVMs' performance when using non-PD feature kernels on the above datasets presented high variance.

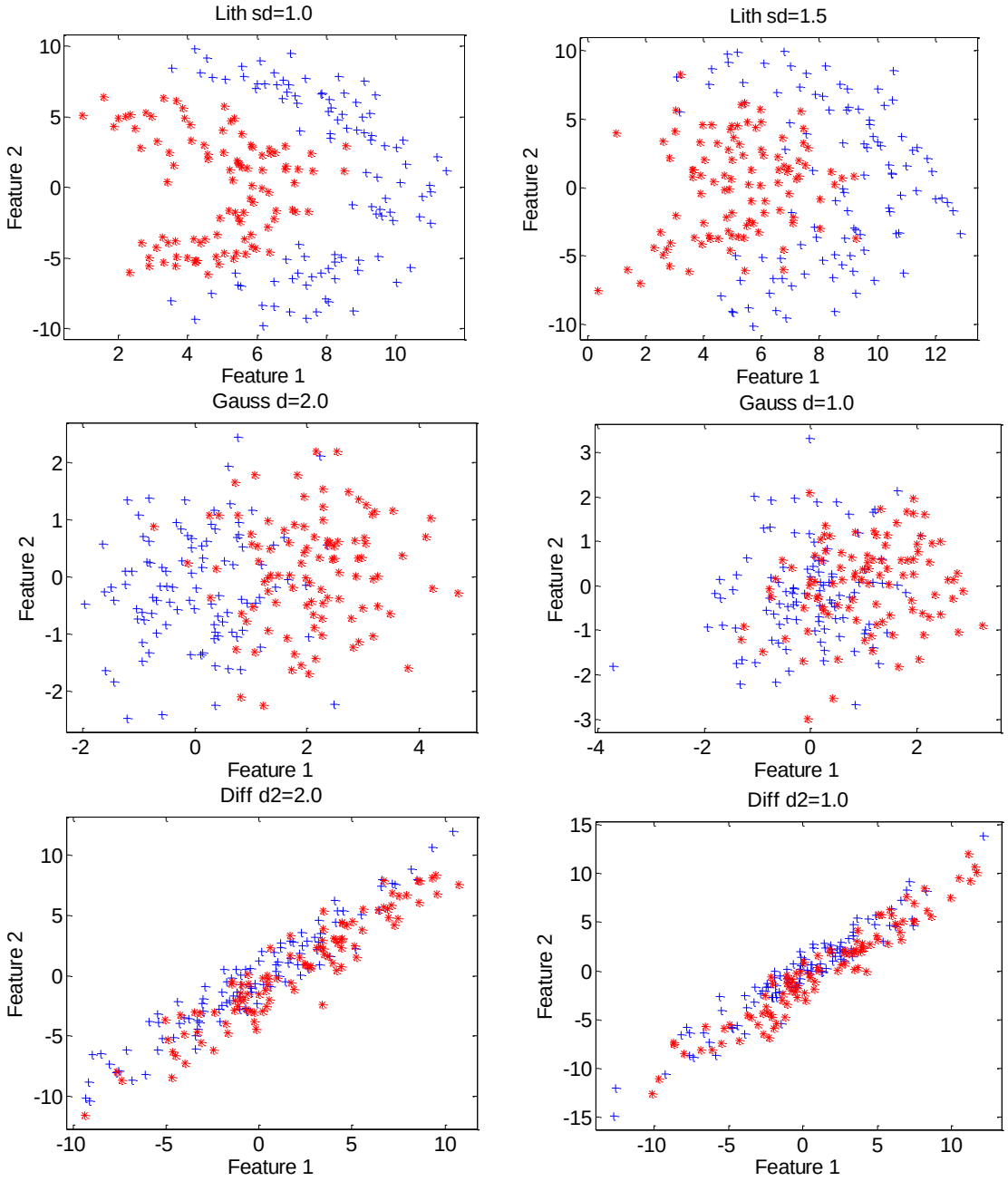


Figure 3.1 Artificial datasets with high (left) and low (right) separability.

Comparatively the HS-SVM appears to perform poorly in the majority of the artificial scenarios, despite the use of a two-layer approach that guarantees stability. This can be attributed to a potential inefficiency of the two layer strategy; the first layer by means of the primary kernels performs a distance-based clustering, whereas the linear second-level kernel attempts to enforce a different clustering in its own hidden space.

The added flexibility offered by the nonlinear RBF feature kernel of GS-SVMs allows defining appropriate class boundaries associated with the second-layer/similarity kernel and overcome such clustering inefficiencies. The experimental results indicate that in

most cases the GS-SVM models surpass its SVM counterpart in terms of the examined cross-validated accuracy metric.

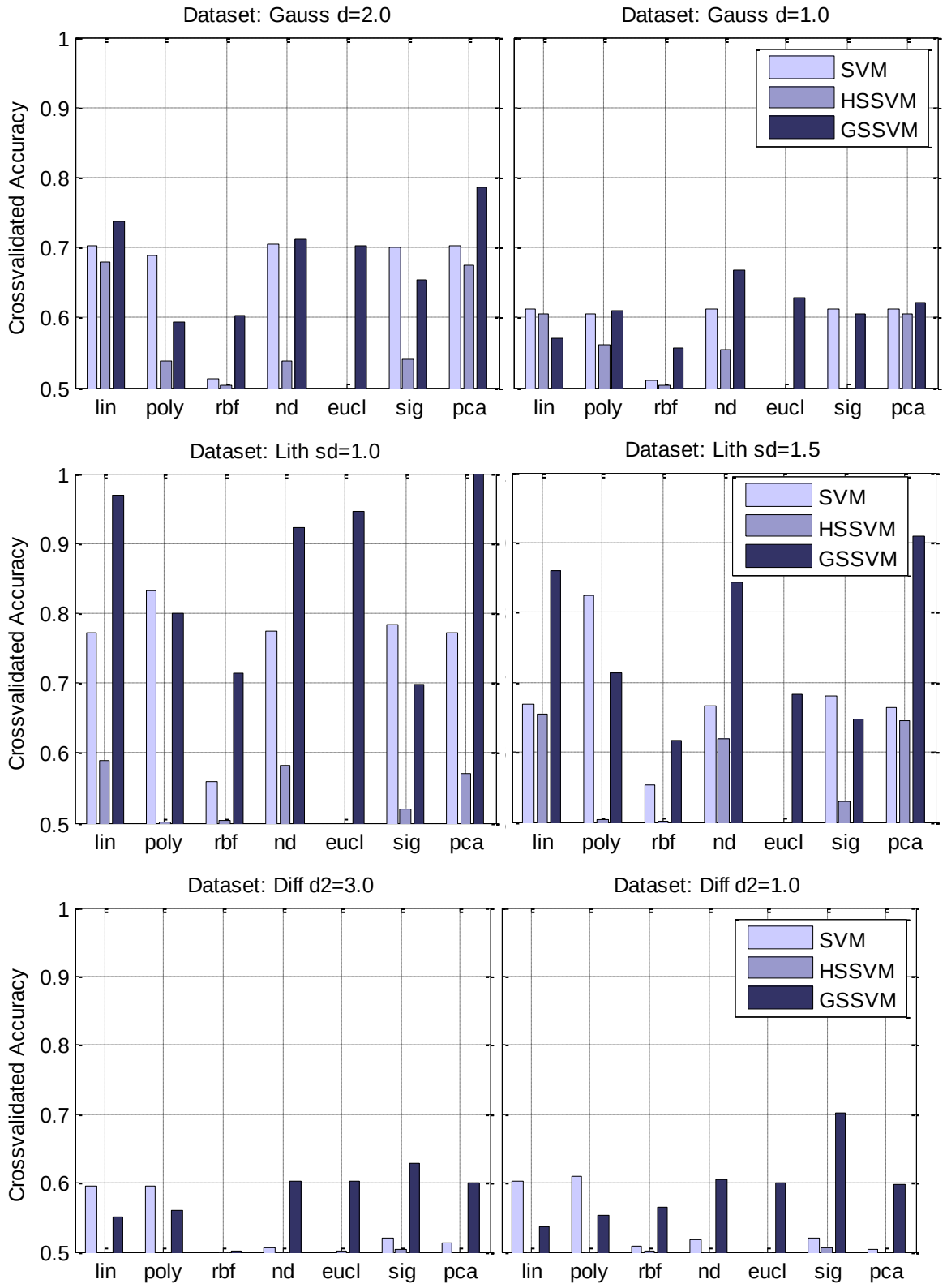


Figure 3.2 Cross-validated accuracy of 7 different feature kernels indicated on the horizontal axis. Left and right columns indicate high and low separability dataset respectively.

In order to obtain more detailed view of the statistical properties of the collected cross-validation results right-tail t-tests were performed on all kernel combinations. The visualizations of the right-tail t-tests' p-values at the standard 0.05 significance level are given in Figure 3.5 -3.7 (see Appendix). Darker tiles indicate significantly better performance of kernel in the corresponding row compared to the kernel in the corresponding column. The GS-SVMs kernels correspond to the bottom 7 rows of each plot (highlighted rectangle). Coherent low values in this region of the plot indicate a consistently superior performance of GS-SVMs over the other evaluated models.

In overall Figures 3.2 and 3.5-3.7 demonstrate that on artificial datasets GS-SVMs outperform SVMs and HS-SVMs in the majority of feature kernel scenarios; and show graceful degradation under worsening class separation conditions. The following section examines whether this trend carries along to a collection of real world biomedical datasets.

3.6.2 Real Datasets

The second test scenario included implementations of the composite kernels used on 3 benchmark datasets from the UCI repository (Newman, Hettich et al. 1998). The breast cancer diagnosis dataset from UCI repository contains 683 complete cytological tests described by 9 integer attributes with values between 1 and 10. The outcome is a binary variable indicating the benign or malignant nature of the tumor. It is considered a challenging medical dataset since reported optimized classifier accuracies generally do not exceed 85%. Another set we have used from UCI repository is the German Brain dataset consisting of 1000 cases by 20 features each distributed in two classes with a 70%-30% ratio. The third, diabetes-diagnosis dataset investigates whether patients show signs of diabetes according to the World Health Organization criteria. The population of 768 cases consists of female Pima Indians, aged 21 and older, living near Phoenix, Arizona, described by 8 continuous variables and a binary outcome variable. Detailed information on the datasets including classification performance is provided in (Lin and Lin 2003).

Each feature was normalized to zero mean and unit variance. Additionally random stratified validation sets were used wherever kernel parameter selection was performed. The smoothing parameter C of the SVM was optimized with grid search in the $[1,10]$ interval. The sigma parameter of the RBF kernel was optimized based on the class variances of the training set.

As in the previous section the evaluation process is performed for two groups of feature kernel functions. The first group includes the three established kernel forms (linear, polynomial, and radial basis function) which are by definition positive definite Gramm matrices, thus satisfying the Mercer conditions. The second evaluation group includes four kernels (negative distance, Epanechnikov, sigmoid, and PCA), which are not guaranteed to produce positive definite Gram matrices for an arbitrary dataset.

The above collection of models is intended to provide insight into the operation of composite kernels, especially in conjunction with non-PD feature kernels. Both HS-SVMs and GS-SVMs are capable of producing Mercer-compliant composite kernels, yet the latter retains the advantage of a nonlinear second-stage mapping. This assertion is supported by our simulation results. It must be explicitly stated that the experimental

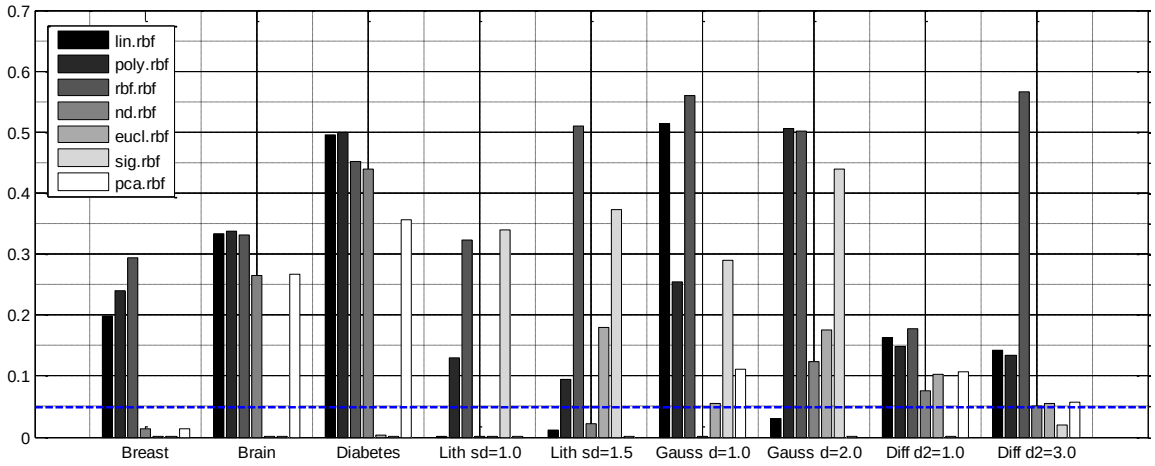


Figure 3.3 Averaged t-test p-values of GS-SVMs accuracy estimates (lower is better). P-values under the 0.05 threshold line indicates significantly higher performance over other (non-GS-SVM) models.

evaluations provide results on a single form (RBF) of the similarity kernel in the GS-SVM formulation. The extensive possibilities for other emerging second stage kernels will have to be evaluated accordingly. The kernel function at the decision level (similarity kernel) can highly affect the performance of classification. Nevertheless it is not the goal of this work to derive an optimal kernel for each dataset but rather to demonstrate the use and the benefits stemming from nonlinear kernels at the second layer of classification.

An intermediate preprocessing step involves the construction of the Gram matrices for each of the feature & composite kernels tested. The Kernel matrix of a given dataset is a similarity matrix as measured by a specific kernel and provides insight into the compactness of the classes and possibly the optimality of the chosen kernel classifier model. Kernel matrix normalization is also performed, according to (Graf and Borer 2001), resulting in unitary diagonal elements. In the classification phase we employed a 100x cross-validation scheme, with the accuracy being measured by means of error rate estimation.

In order to gain perspective about the magnitude of the problem imposed by non-PD kernel functions, the largest negative eigenvalues of the resulting Gram matrices are listed in Table 3-1 (only for kernels that produced non-PD eigenvalues.).

Kernels	Datasets		
	Breast	Brain	Diabetes
Negative	-8398,1	-36839,1	-2817,1
Epanechnikov	-5816,7	-26920,5	-1980,9
Sigmoid	-17,4	-22,2	-5,4

Table 3-1 Largest negative eigenvalue of kernels that produced non-PD eigenvalues. Only non-PD SVM kernels result in negative eigenvalues.

In the preprocessing stage we examine the non-PD kernels eligibility and performance on the problems considered. Apart from trivial numeric overflow artifacts, the 3 kernels that consistently produce large negative eigenvalues on the evaluated datasets are the negative distance, the sigmoid and the Epanechnikov kernels (as feature kernels). In

theory the existence of significant negative eigenvalues in the Kernel matrix violates the basic perquisites of kernel methods, hinders the convergence of an SVM model and reduces the overall generalization capability. From Table 3-1 it is also evident that the range of the eigenvalues is considerably higher for the negative definite and Epanechnikov kernels compared to the sigmoid kernel. The PCA kernel produced strictly positive eigenvalues at least in the specific datasets used, despite the fact that there is no relevant theoretical limitation.

Both HS-SVM and GS-SVM schemes suppress the appearance of negative eigenvalues in the composite formulation. GS-SVMs not only circumvented the kernel positive definiteness issue but also achieved consistent high classification accuracy results, as illustrated in Figure 3.7.

Figure 3.4 presents the evaluation results under 100-fold cross-validation. Depending on the dataset, increased complexity of the kernels in the SVM formulations can increase the mapping capability, but the appropriateness of kernels for each problem needs further verification.

In SVMs non-PD kernels achieve reasonably high accuracies (except from the last dataset), which further improve when the same non-PD kernels are utilized as part of the GS-SVMs formulation.

HS-SVMs achieve considerably reduced classification accuracies throughout all 3 real datasets.

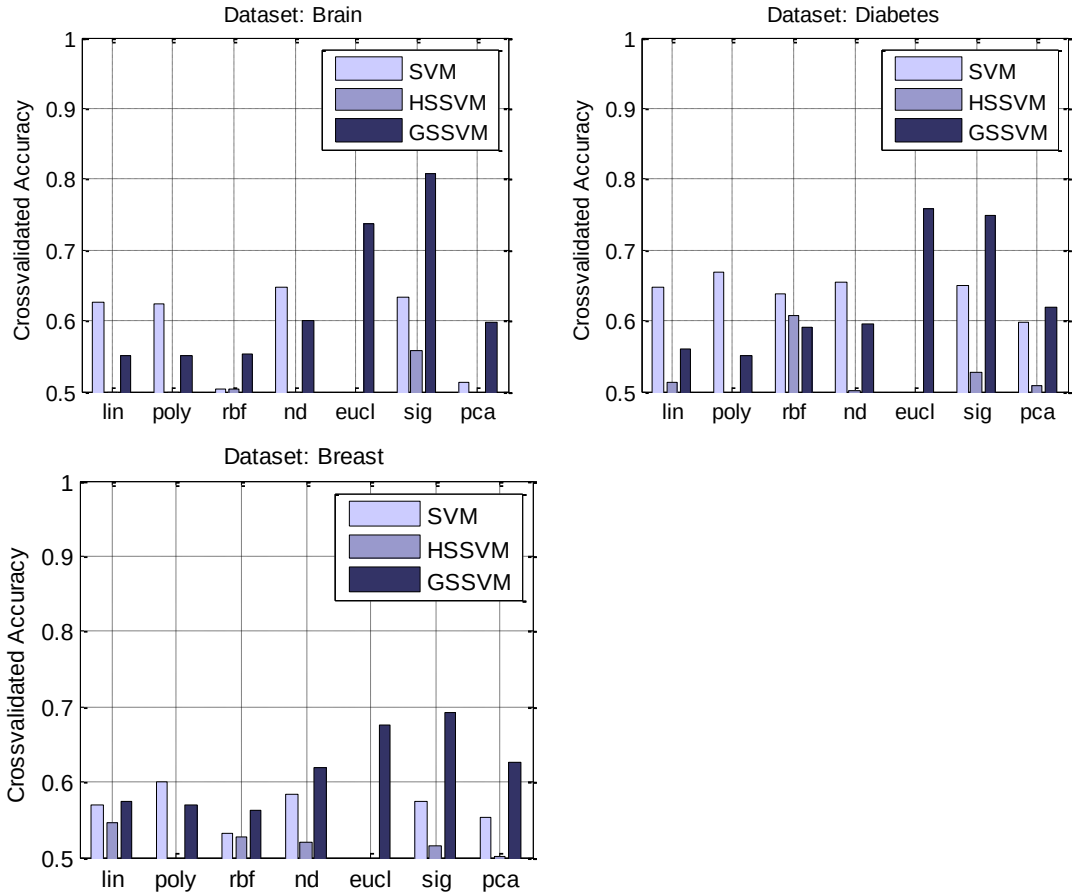
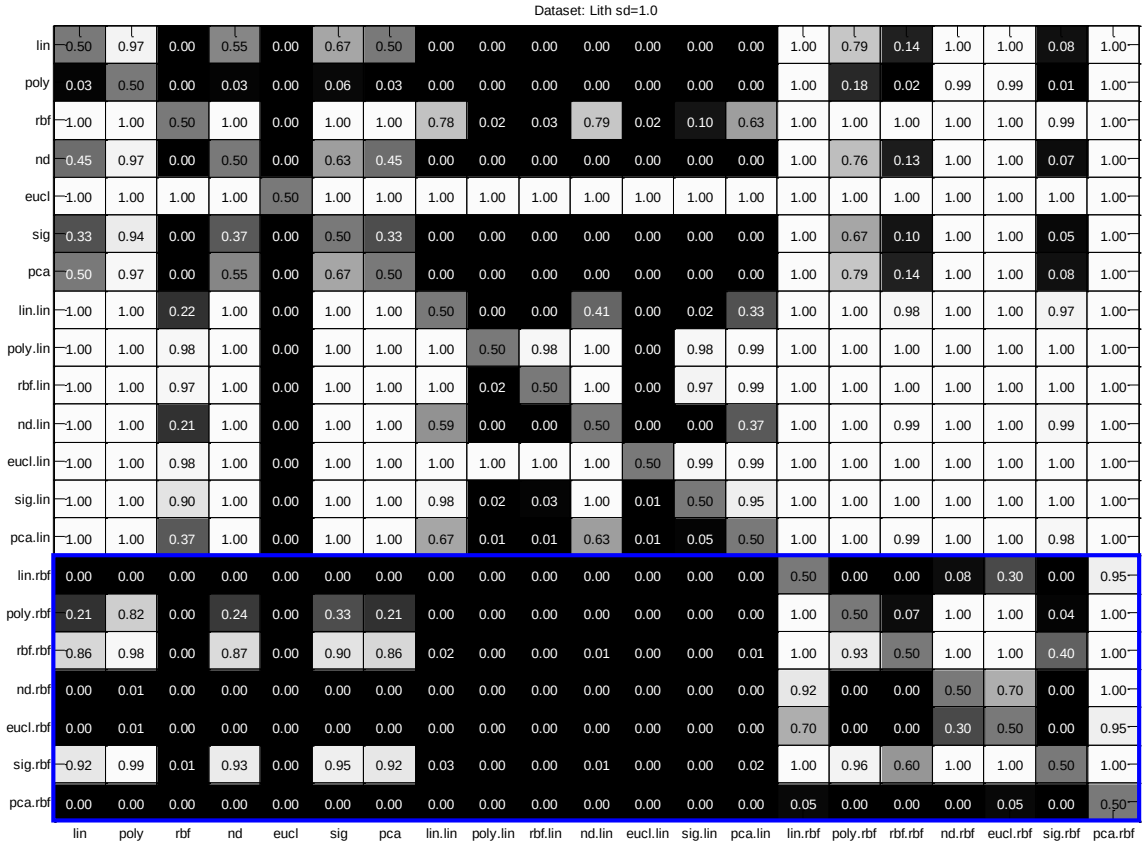


Figure 3.4 Cross-validated accuracy for all seven evaluated 1st level (feature) kernels. HS-SVMs employ linear and GS-SVMs nonlinear 2nd level (similarity) kernels.

While GS-SVMs are inferior to SVMs in typical linear/polynomial/RBF kernel models they appear to be able to leverage the mapping characteristics of non-PD feature kernels to achieve accuracies surpassing the standard SVM across all datasets.

In the same way as in the previous section the visualizations of the t-tests' p-values for the 3 bio-medical datasets at the standard 0.05 significance level are given in Figures 3.9-3.11 (included in the appendix). The corresponding right-tail t-test p-values indicate that GS-SVMs continue to exhibit statistical proven improvement in accuracy in all 3 biomedical datasets.

3.6.3 T-test results



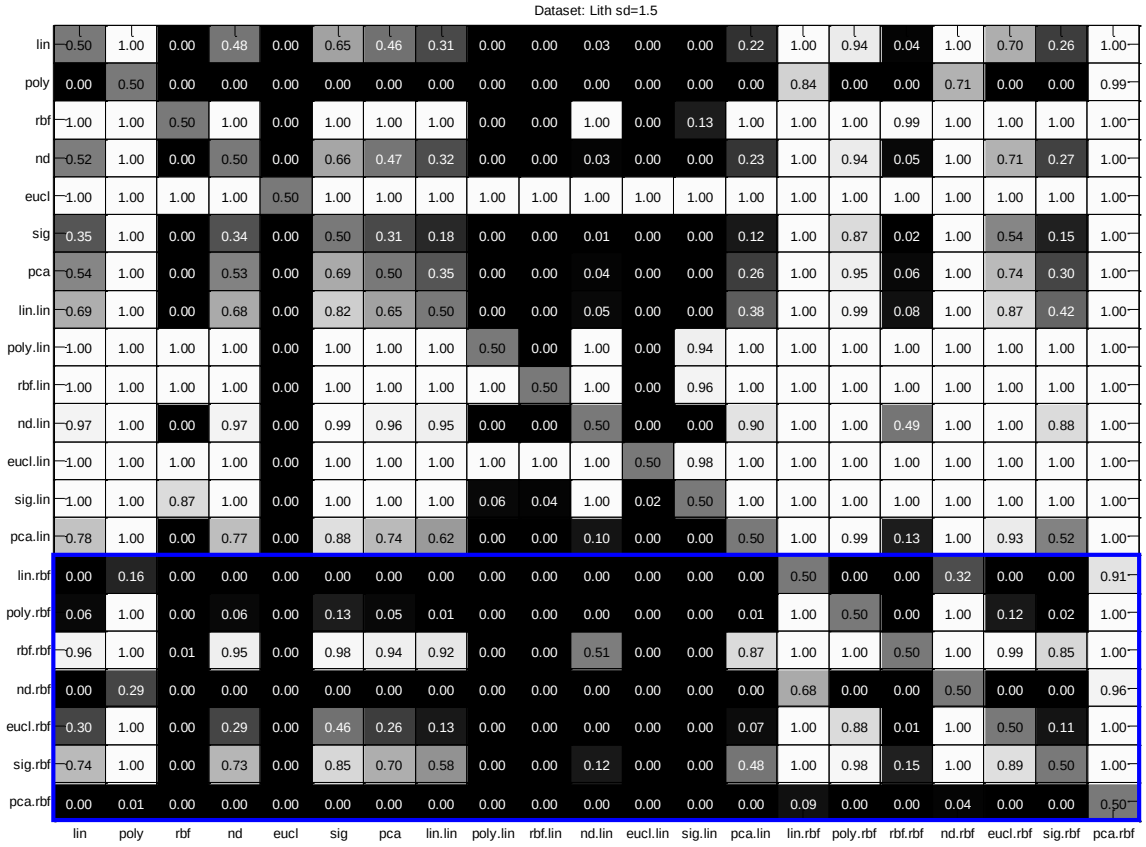


Figure 3.5 T-test p-values comparing accuracy estimates for Lithuanian clusters dataset.

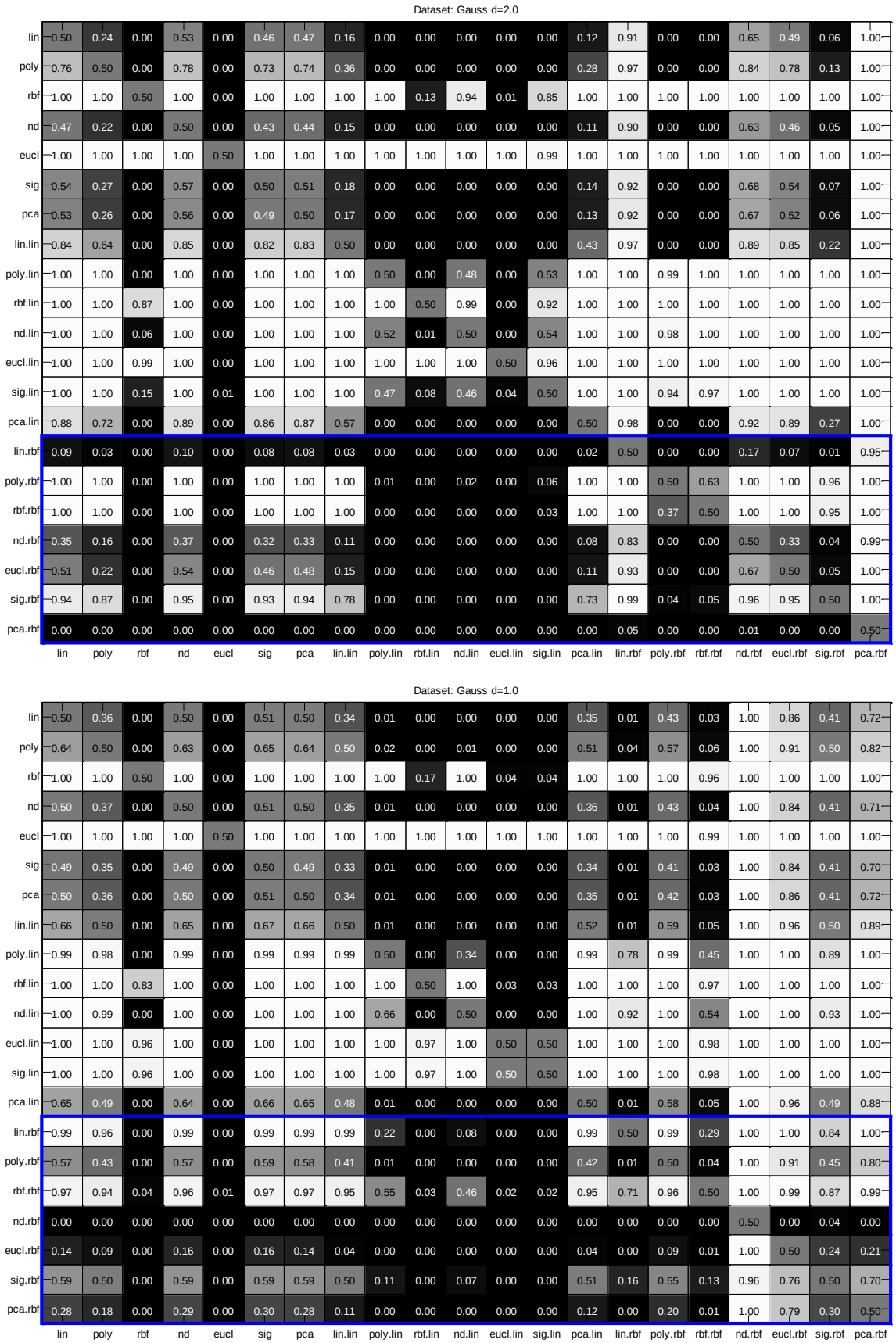


Figure 3.6 T-test p-values comparing accuracy estimates for Gaussian clusters dataset.

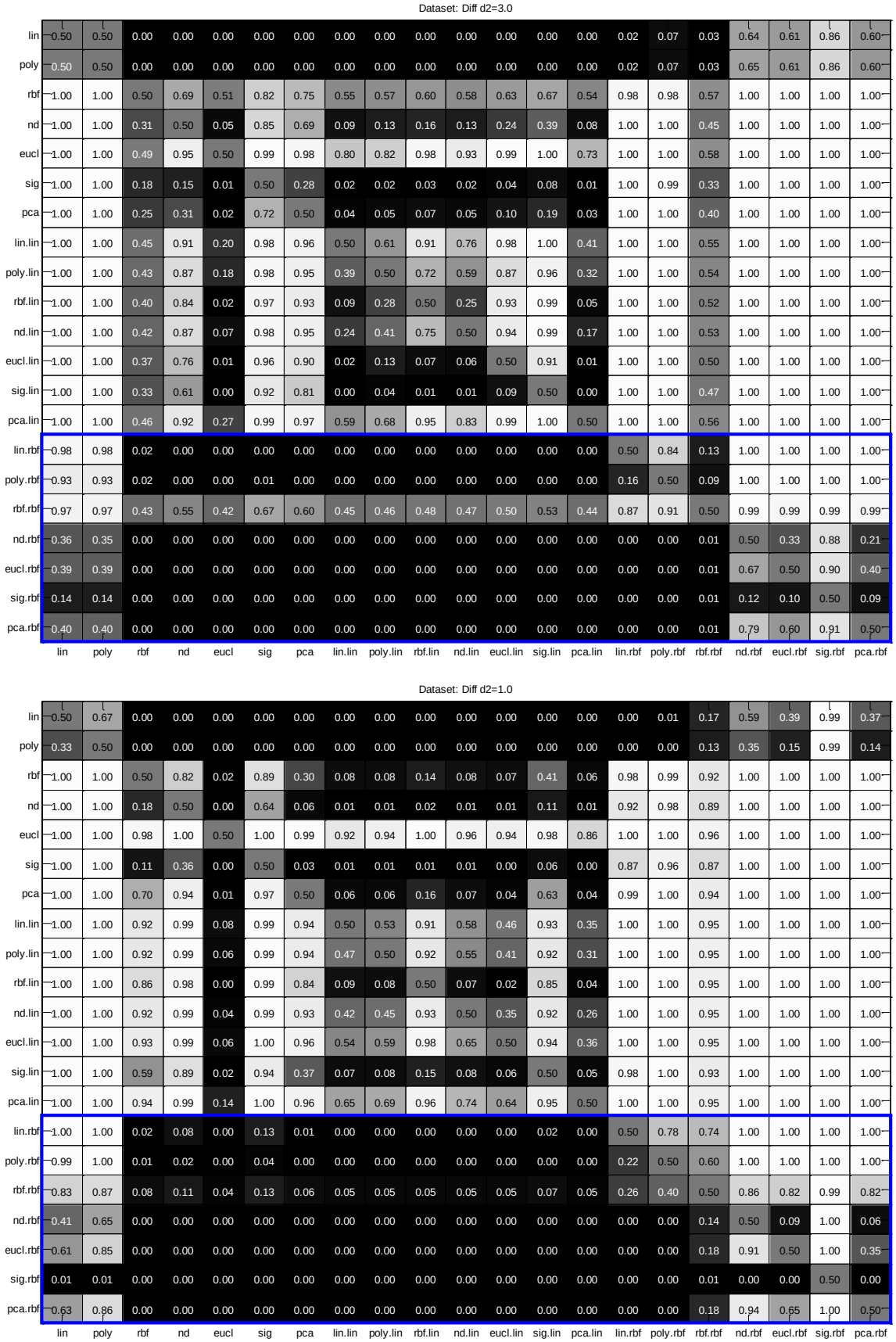


Figure 3.7 T-test p-values comparing accuracy estimates for Lithuanian clusters dataset.

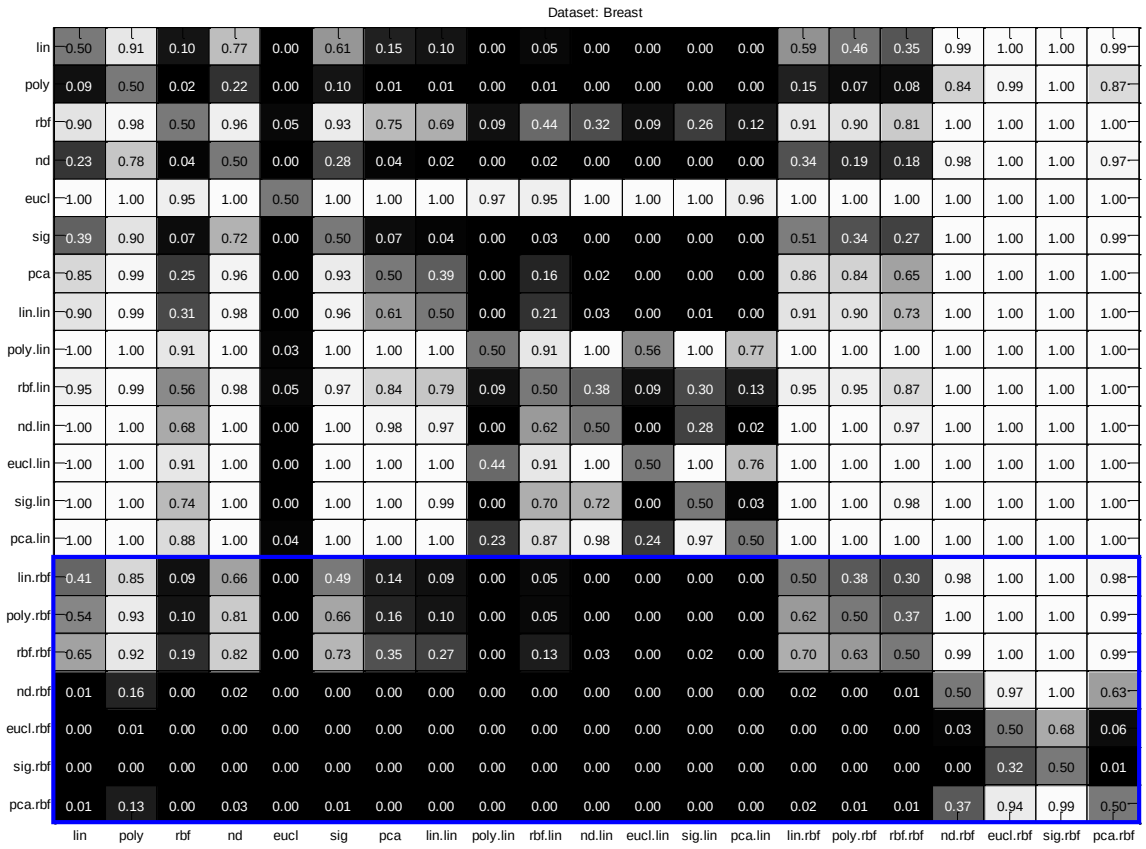


Figure 3.8 T-test p-values comparing accuracy estimates for the Breast cancer dataset.

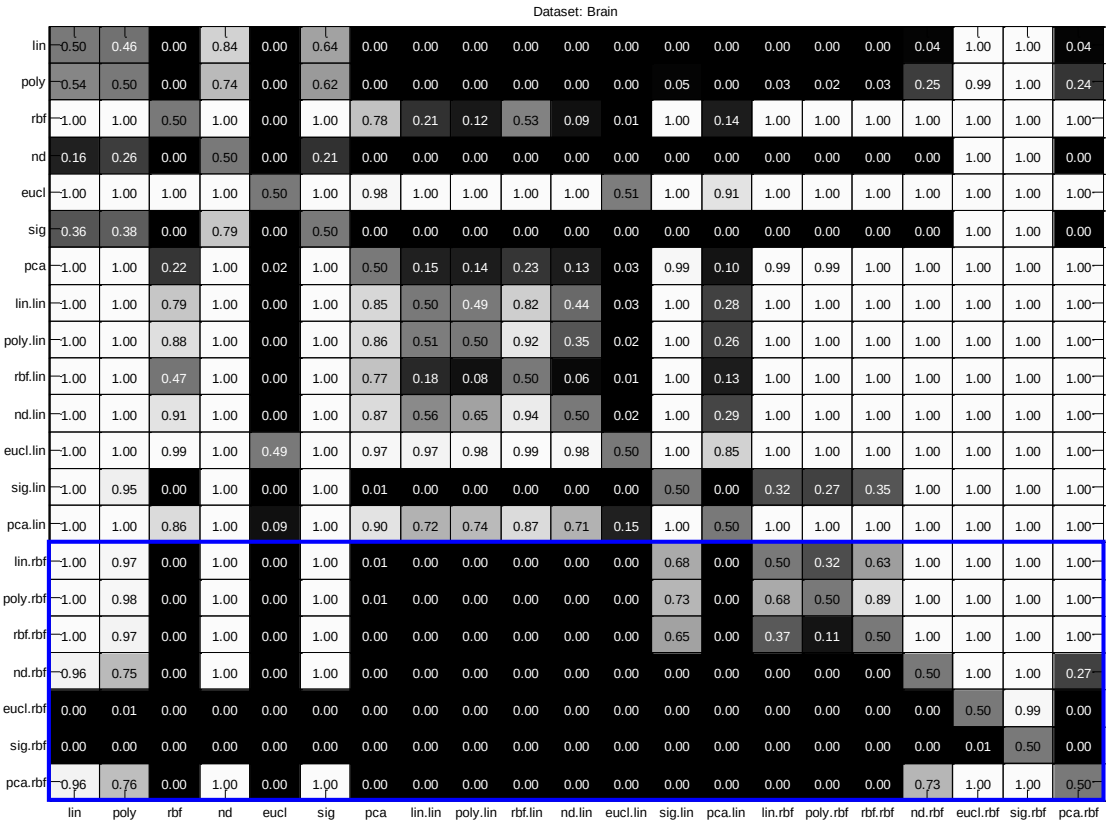


Figure 3.9 T-test p-values comparing accuracy estimates for the Brain disease dataset.

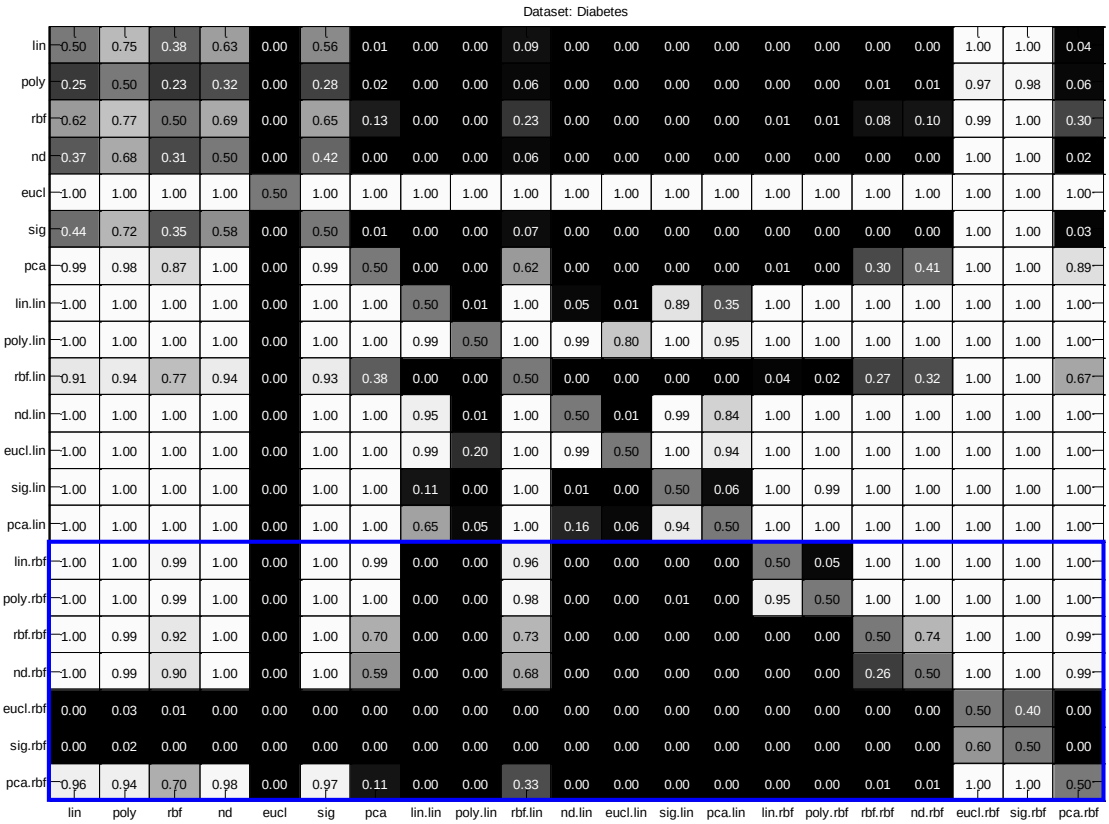


Figure 3.10 T-test p-values comparing accuracy estimates for the Diabetes dataset

Equation Chapter (Next) Section 1

3.6.4 Common trends

In order to summarize the t-test plots focusing on the GS-SVMs evaluation, the average p-value for each GS-SVM kernel is plotted in Figure 3.8-3.10. The plot shows more clearly that the use of GS-SVMs appears to be justified in the case of non-PD feature kernels and to a lesser extent in difficult datasets using linear/polynomial/RBF feature kernels.

Additionally Figures 3.5-3.7 and 3.8-3.10 show that the performance of GS-SVM depends on the choice of the feature kernels to a lesser extent (narrower ranges of p-values) than the performance of HS-SVMs or SVMs (larger ranges of p-values).

The treatment of non-PD feature kernels within the formulation of GS-SVMs, not only alleviates the problem induced by indefiniteness, but also additionally yields the highest classification performance. The improvement achieved by the GS-SVM scheme is attributed to its nonlinear ability in the second-level mapping through the RBF similarity kernel. While both HS-SVMs and GS-SVMs circumvent the numerical problems imposed by non-PD kernels, the use of GS-SVM is more flexible in achieving efficient data fitting and, thus, improved prediction accuracy.

In this study we have explored only one possible second-layer GS-SVM kernel, namely the RBF kernel. From a broader perspective, RBF kernels are metrics of the smooth weighted distances of a given sample $\mathbf{x}_i$ to all other samples in the same class. Accordingly, a GS-SVM formulation with RBF similarity kernel can be regarded as a second order statistic that measures the similarity of a given sample $\mathbf{x}_i$ with all other samples in the class. This similarity is measured in the more compact space defined by the feature kernel. Nevertheless, the identification of other, possibly more effective, similarity kernels remains an open research problem, based on our proposed formulation.

3.7 Conclusions

This chapter explores the design of multiple kernel classifiers and defines a general class of composite kernel functions in the formulation of GS-SVMs. The practical advantages of the proposed concept are highlighted through classifier performance in both artificial and real-world biomedical problems. The two key benefits of the proposed GS-SVM framework are the successful employment of non-positive definite Gram matrices, along with constraints, and the improved accuracies compared to SVMs and HS-SVMs.

The development of new, possibly non-PD, kernels incorporating explicit prior knowledge and being tailored to specific medical classification tasks can help reveal important diagnostic features and rank their effectiveness. GS-SVMs allow for the unconstrained mathematical formulation of expert knowledge rules into composite kernels.

The methodology presented here is not restricted to SVMs. Due to the nature of these algorithms the derived composite functionals can be used as modules in any other kernel method. Nevertheless, such utilization requires further investigation regarding its advantages and limitations in each kernel approach.

4 Designing Kernels for biomedical problems

4.1 Brain lesion classification using 3T MRS spectra and paired SVM kernels

Pattern analysis in the biomedical domain is far from a standard and trivial task. Multimodal datasets, complex data schemas, time domain interactions and complex underlying biological mechanisms are concurrently creating difficulties. In many cases, the flexibility of kernel methods can provide set of tools to bind the established knowledge and solution mechanisms to custom data forms. Building on the concepts of kernel design of the previous chapter, a case study of diagnostic application is described in this part with an emphasis in reusability and rationale.

The increased power and resolution capabilities of 3T Magnetic Resonance (MR) scanners have extended the reach of Magnetic Resonance Spectroscopy as a non-invasive diagnostic tool. Practical sensor calibration issues, magnetic field homogeneity effects and measurement noise introduce distortion into the obtained spectra. Therefore, a combination of robust preprocessing models and nonlinear pattern analysis algorithms is needed in order to evaluate and map the underlying relations of the measured metabolites. The aim of this work is threefold. Firstly we propose the use of a paired support vector machine kernel utilizing metabolic data from both affected and normal voxels in the patient's brain for lesion classification problem. Secondly we quantify the performance of an optimal reduced feature set based on targeted CSI-144 scans in order to further reduce the data volume required for a reliable computed aided diagnosis. Thirdly we expand our previous formulation to full multiclass classification. The long term aim remains to provide the human expert with an easily interpretable system to assist clinicians with the time, volume and accuracy demanding diagnostic process.

4.2 Background

Magnetic Resonance Spectroscopy (MRS), has been studied for more than a decade as a promising diagnostic tool for a variety of pathologies (Devos, Lukas et al. 2004; Lukas, Devos et al. 2004; Jan Luts, Poullet et al. 2008; Kounelakis, Zervakis et al. 2008). Coupled with the morphological features provided by Magnetic Resonance Imaging (MRI) techniques (Georgiadis, Cavouras et al. 2008), it can provide accurate identification and quantification of biologically important compounds in soft tissue.

The transition of MRS from experimental evaluation studies to clinical practice relies heavily on the implementation and standardization of robust methodologies that decouple the diagnostic problem from inter and intra-patient variations, sensor calibration and procedural issues and the varying expertise of clinical personnel.

The classification problem itself is recognized as a nonlinear multiclass problem with varying difficulty depending on the specific class labeling (Devos, Lukas et al. 2004; Lukas, Devos et al. 2004; Georgiadis, Cavouras et al. 2008). In particular the partitioning of gliomas versus metastatic tumor classes is notably more challenging relative to other class pairs.

Additionally, a new generation of 3Tesla MRS scanners calls for adaptation of existing classification models, optimized on 1.5T equipment, and evaluation of possible performance gains (Kousi, Tsougos et al. 2009). Regardless of the sensor technology used, the inter and intra patient variations of the collected spectra for each pathology class hinder the establishment of simple visual markers and outline the need for the development of adaptive nonlinear decision support tools.

Building upon our previous results (Dimou, Tsougos et al. 2009; Dimou, Tsougos et al. 2009; Kousi, Tsougos et al. 2009) we propose the use of a SVM kernel that leverages the information conveyed by intra patient metabolite measurements. We also extend our analysis to full multiclass classification, utilizing an updated more extensive dataset of high resolution 3T spectra obtained from patients at the Larisa University Hospital.

We additionally evaluate the use of a reduced metabolic feature set as an alternative to continuous spectrum classification in an effort to address practical compatibility, transferability and speed issues involved with the coupling of MRS scanners and clinical decision support systems. The utilized MRS scanner does not provide a well documented file access interface, therefore hindering automated high throughput access to the large volume of raw data obtained during each exam. Recording a minimal subset of exam metabolites circumvents this problem and evaluate their effectiveness as classification features.

The rest of this paper is organized as follows. The following section outlines the data mining and pattern analysis tools that we employ along with the proposed problem-specific SVM kernel. Section 4.4 provides an overview of the experimental results that we obtained on the Larisa MRS dataset and section 4.5 summarizes the key findings and provides general guidelines and pointers to future research.

4.3 Data mining and pattern analysis tools for MRS

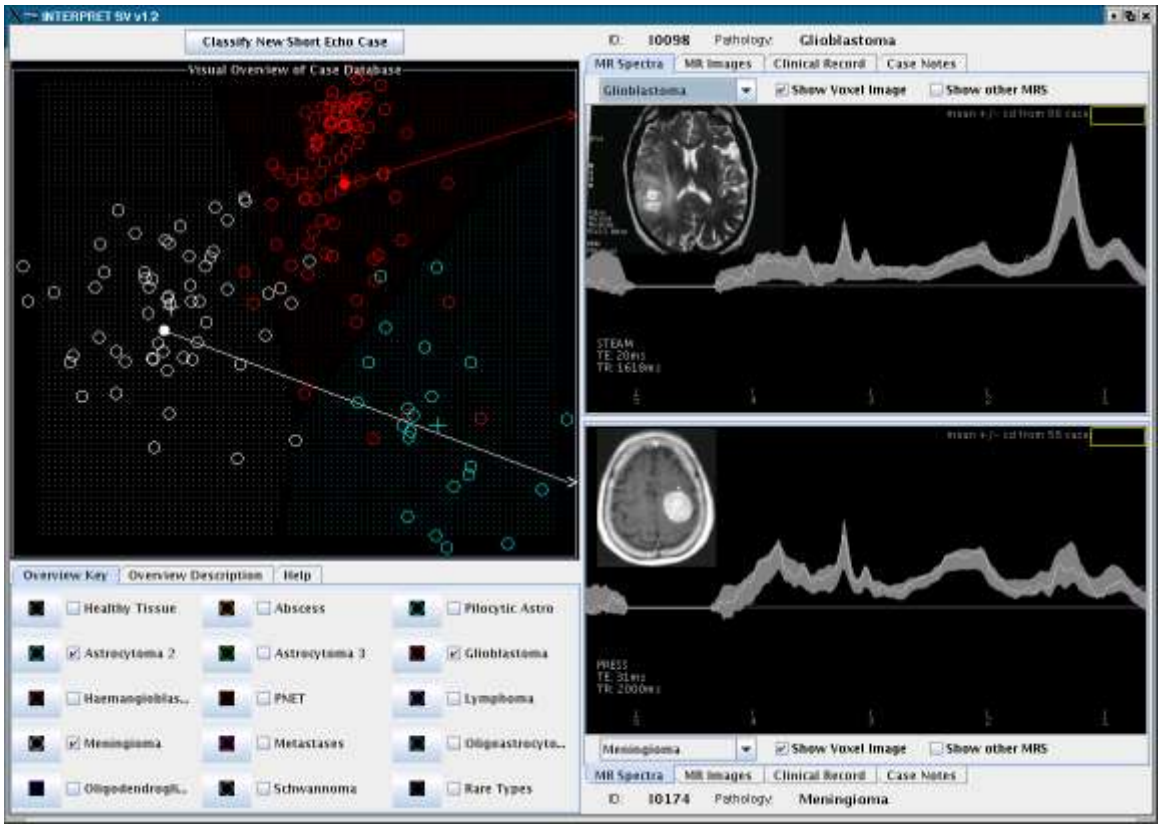


Figure 4.1 Interpret MRS classification tool

4.3.1 Dataset and preprocessing

In order to verify the applicability of the proposed scheme the classification model was evaluated on a dataset collected at the MR Dept., Larisa University Hospital, Greece using a GE Healthcare Signa® HDx MRS Scanner.

A total of 84 consecutive patients (age 8-77 years) under investigation of brain lesions (tumors, multiple sclerosis, gliosis, leukoencephalopathy, meningiomas etc) were enrolled in this study before any surgical biopsy and/or resection.

The typical proton MR spectroscopic features for the aforementioned lesions are NAA, Cho, Cre, mI, lipids and lactate. N-Acetyl-aspartic acid (NAA) is present in the health brain parenchyma and is the highest peak in the normal spectrum, resonating at 2.02ppm. The utility of NAA as an axonal marker is supported by the loss of NAA in many white matter diseases, including multiple sclerosis and leukoencephalopathy. Malignant tumours cause destruction of neurons and thus a loss of NAA and purely extra-axial tumours, such as typical meningiomas, demonstrate no NAA.

Choline (Cho) is a metabolic marker of membrane density and integrity, with its peak located at 3.22 ppm. Inta-axial and extra-axial tumours show an increase in the Cho peak because of increased cellularity. Increases in Cho relative to NAA are also noted in gliosis and multiple sclerosis. Therefore, difficulty may be encountered in interpreting results in some lesions, such as tumefactive multiple sclerosis.

In simplistic terms, Creatine (Cre) is a marker of ‘‘energy metabolism”. The central peak on the spectrum at 3.02 ppm represents the sum of creatine and phosphocreatine. In the clinical setting, Cre is assumed to be stable and is used for calculating metabolite ratios (Cho:Cre and NAA:Cre ratios). It may be useful to note that Cre itself does not originate in the brain, and hence systemic disease (such as renal disease) may impact on Cre levels in the brain.

Myo-inositol (mI) is a simple sugar, with a peak found at 3.56 ppm. It is considered a glial marker. An increase in mI content is believed to represent glial proliferation or an increase in glial cell size, both of which may occur in inflammation. It is elevated in the setting of gliosis, astrocytosis, and in disorders such as Alzheimer’s dementia. mI has also been labeled as a breakdown product of myelin present in tumourous lesions and multiple sclerosis.

Membrane lipids have very short relaxation times and are not usually visualized on intermediate or long TE, but are visualized on short TE. They produce peaks between 0.8 and 1.5 ppm and are usually large broad peaks. The presence of lipids may indicate voxel contamination by diploic space fat, scalp and subcutaneous tissues (when the voxel is placed near these structures). Lipid signals in pathology are generally associated with necrosis such as in high-grade brain tumors or metastases. In addition, lipid signals have been observed in brain MR spectra of patients with multiple sclerosis (Van der Graaf 2009) and lipomatous meningiomas (Qi and Li 2008).

Under normal circumstances, lactate is present only in minute amounts in the brain and is not resolved using the normal spectroscopic techniques. However, under conditions where the aerobic oxidation mechanism fails and anaerobic glycolysis takes over, such as brain ischaemia, hypoxia, seizures, metabolic disorders, and areas of acute inflammation, lactate levels increase significantly. Lactate also accumulates in tissues that have poor washout, like cysts and necrotic and cystic tumours. When present, it is recognized as a doublet (twin peak) at 1.33 ppm. Lactate is characterized by variable projection of the peak at different TEs. On acquisitions using intermediate TEs (135/144ms), the doublet peak is inverted below the baseline, but at very short or very long TE (30 or 288 ms), the doublet peak projects above the baseline (Soares and Law 2009).

All patients gave a written informed consent to participate in the study. 1H-MR spectroscopy studies were performed on a 3-Tesla MRI whole body unit (GE, Healthcare, Signa® HDx) using both automated PROBE single voxel and multivoxel (Chemical Shift Imaging) spectroscopy packages before contrast administration.

Single Voxel (SV) spectroscopy was performed using the point-resolved spectroscopy (PRESS) pulse sequence, provided by the manufacturer at an echo time of 35msec at axial, sagital and coronal planes. The repetition time was 1500msec. Chemical Shift Imaging (CSI) was performed using PRESS pulse sequence, in an axial plane at an echo time of 144msec and a repetition time of 1000mesc. The CS imaging slice was positioned in areas of maximum extension of the lesion.

In both cases of SV and CSI the regions of interest were defined as follows: 1) inside the lesion, 2) outer diameter of the lesion (if possible), 3) contralateral side, and 4) normal appearing white matter.

In cases of pathology we avoided inclusion of obvious necrosis, cyst, hemorrhage, edema, calcification and normal appearing brain tissue in the voxel, to avoid lesion's underestimation. Thus ROIs with potential contamination with cerebrospinal fluid, subcutaneous fat, or eye motion have been excluded from analysis.

For voxel positioning, fluid attenuated inversion recovery (FLAIR, TR=9502msec, TE=128msec) or a home-designed T2-weighted fast spin echo (TR=4520msec, TE=102msec) sequence in axial, coronal and sagittal planes were preceded using 26cm field of view, 5mm slice thickness and NEX equal to 1. The size and location of the voxels were carefully adjusted inside the lesion or in healthy brain parenchyma for the best possible shimming and spectra accuracy. The exact voxel positioning protocol is indicated in Figure 4.1. Due to data quality and availability limitations the resulting features used for classification included spectral measurements from areas 1 (inside the lesion) as pathological and 3 (contralateral) as normal. The features (metabolite measurements) obtained from each area were utilized either as a single feature vector or separately as part of the proposed composite kernel, described in section 4.3.2.

The MRS sensor data were preprocessed using standard statistical methods for outlier detection, normalization and peak integration. Peak integration was performed in the ranges 3.35-3.17ppm (Choline), 3.15-2.99 (Creatine), 2.23-1.97 (NAA), 3.02-3.31 (Creatine+Choline), 1.30-0.90 (Lipids and Lactate) and 3.69-3.54 Myolnositol.

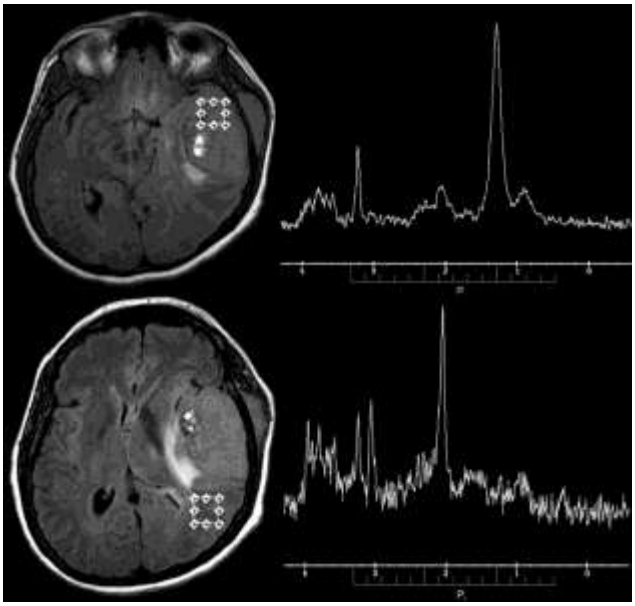


Figure 4.2 SV spectra from the inner (upper) and from the periphery of a neoplasm (lower), as indicated by the 2x2x2cm voxel.

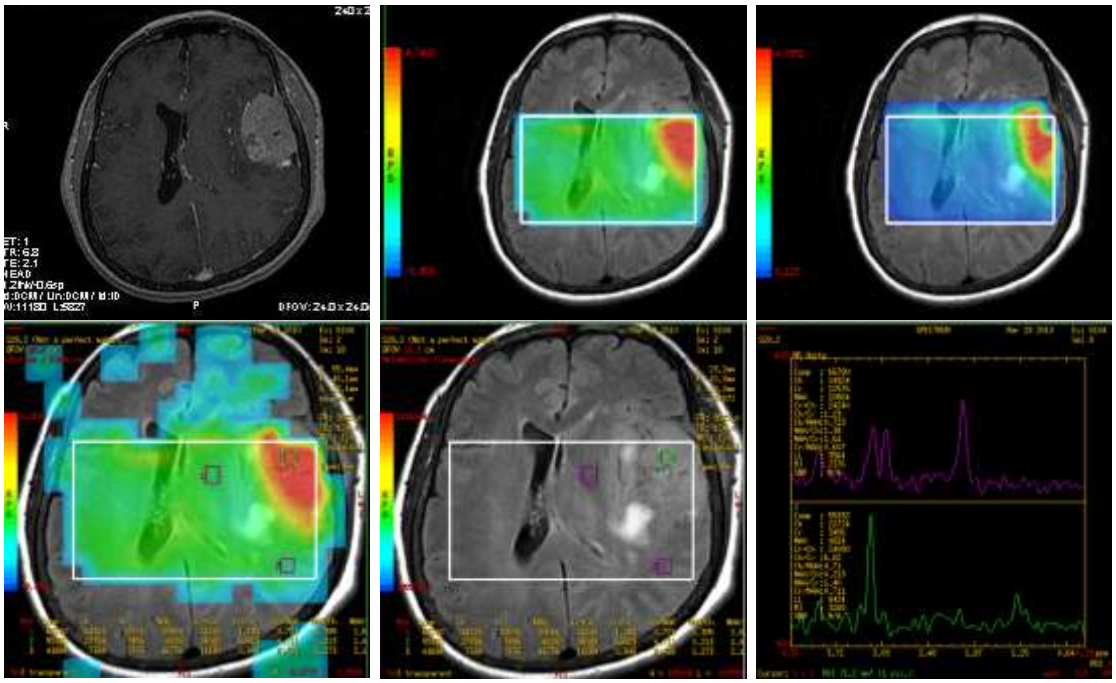


Figure 4.3 Multivoxel imaging of the same case.

Top left illustrates a T1 weighted contrast enhancement image. The top middle and right images represent Choline/Creatine and Choline/NAA maps respectively. The relative levels of specific signals are shown in colour. Red denotes the most elevated signal and blue the lowest signal. Lower left and middle images indicate the positioning of the voxels used for evaluation, with and without the map's guidance. Lower right image depicts the outcome of the CSI within the tumor and from the periphery of the lesion.

The diagnostic class labelings were obtained from three sources 1) histological examination (where available), 2) radiologist expert assessment, 3) physicist's expert assessment. Histological data were available for only ~42% of the patients. The resulting confusion matrices indicated a full agreement of the radiologist's and physicist's class labelings whereas there are notable differences between the above labelings and histological class labeling for the given patient subset. The correspondence between the radiologist diagnosis and histological results are shown in Figure 4.2. Despite the fact that histological evaluation results are considered the gold standard, the final classification system's training was performed using the radiologist's diagnosis due to the extensive data missingess of the histological labelings. Under this prism, the proposed model is at this phase evaluated from the aspect of optimally mapping a radiologist's diagnostic behavior. This classifier is intended as a primary expert mapper to be integrated in the future in a multi classifier system that will simulate a committee of clinical experts.

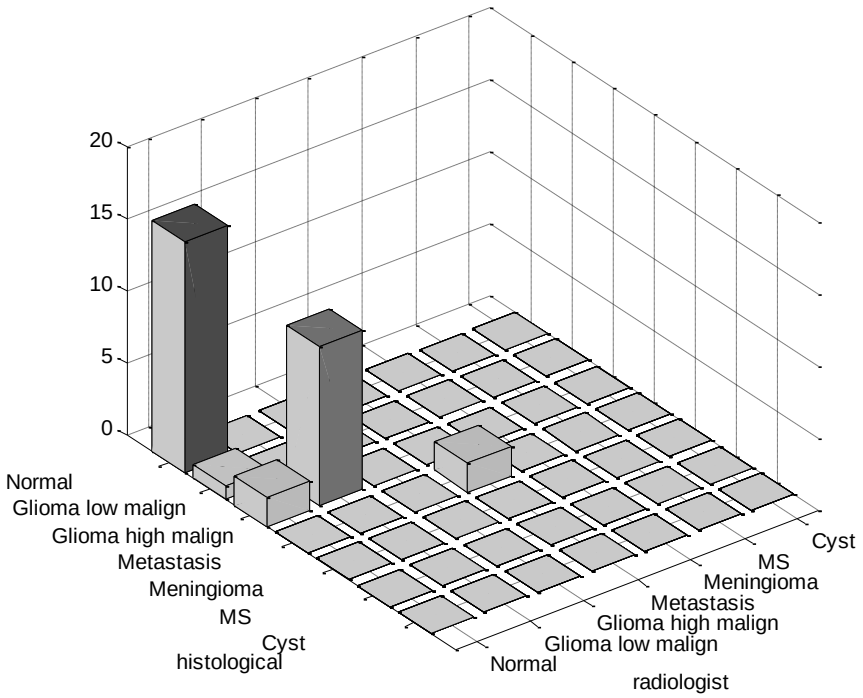


Figure 4.4 Distribution of class labeling provided by human experts.

Since brain tumor classification is inherently a multiclass problem, one has to resort to techniques that allow for handling such problems with a pool of classifiers that provide binary outcomes. There are numerous approaches to multiclass mapping in literature, the most prominent being one-against-all and pair training (Duda, Hart et al. 2001).

The diagnostic outcomes were grouped into 7 distinct classes corresponding to the pathological states given in Table 4-1. The "normal" state is derived from spectral samples from voxels placed on the unaffected areas (areas 2,3,4) of the dataset's patients. The largest affected pathological group, namely gliomas was further partitioned into two distinct classes, low and high malignancy gliomas in line with common clinical practice which calls for identification of disease staging. Classes 4 to 7 constitute of important yet undersampled pathologies that usually pose a difficult classification problem both for human experts and for automated decision support systems. In all cases, the clinical experts select one representative voxel from each area to be used for classification.

Class	**Pathology**	**Radiologist/Physicist**
Class1	Normal	42
Class2	Glioma (low mal)	8
Class3	Glioma (high mal)	20
Class4	Metastasis	4
Class5	Meningioma	3
Class6	MS	5
Class7	Cyst	2

Table 4-1 Pathological Classes

Using the labeling provided by radiology clinical experts we were able to train 7 SVM classifiers using the one-against-all scheme (Kuncheva 2004) and combine the resulting class membership estimates based on class conditional Bayesian probabilities. The SVM implementation was derived from the PRTools (Duin 2000) "svc" module that inherently supports multiclass problems of this type.

4.3.2 Support vector machine classifiers

The diagnostic models used to classify tumor types were developed using support vector machine classifiers (Cristianini and Shawe-Taylor 2000; Schölkopf and Smola 2002; Abe 2005). The SVMs model formulation in the primal space is

$$\min_{\mathbf{w},b,\mathbf{e}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \gamma\sum_{n=1}^{N} e_n^2 \qquad (4.1)$$

s.t.

$$y_n\left[\mathbf{w}^T\varphi(\mathbf{x}_n)+b\right] \geq 1-e_n,\ n = 1, ..., N \qquad (4.2)$$

for the classifier

$$y(\mathbf{x}) = \text{sign}\left[\mathbf{w}^T\varphi(\mathbf{x})+b\right] \qquad (4.3)$$

where y_n is the binary class indication encoded as -1 versus +1, γ is the regularization parameter, w is a weighting vector, b is a bias term, and e_n is the error variable. The mapping $\varphi: \mathbb{R}^q \rightarrow \mathbb{R}^r$ maps the q-dimensional input space into a high r-dimensional feature space. By solving the Lagrangian, the problem can be formulated in the dual space for the classifier

$$y(\mathbf{x}) = \text{sign}\left[\sum_{n=1}^{N}\alpha_n y_n K(\mathbf{x}, \mathbf{x}_n)+b\right] \qquad (4.4)$$

Where α_1, …, α_N are the Lagrange multipliers. We implicitly work in the feature space by applying a positive definite kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i)^T\,\varphi(\mathbf{x}_j) \qquad (4.5)$$

We explored two types of SVM kernels: the radial basis function (RBF) kernel and a modified paired RBF kernel given as:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\frac{\left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2^2}{\sigma^2}\right) \qquad (4.6)$$

and

$$k\left(\mathbf{x}_i,\mathbf{x}_j\right)=\exp\left(-\frac{\left\|\mathbf{x}_i-\mathbf{x}_j\right\|_2^2}{\sigma_{\mathbf{x}}{}^2}\right)+\exp\left(-\frac{\left\|\mathbf{x}_i'-\mathbf{x}_j'\right\|_2^2}{\sigma_{\mathbf{x}'}{}^2}\right) \quad (4.7)$$

In (7) each term's weighting factor $\sigma_{\mathbf{x}}$, $\sigma_{\mathbf{x}'}$ is calculated over a different area. This scheme can be expanded to arbitrary regions for each examination. The paired spectra SVM kernel given in Eq.7 takes advantage of an inherent invariance of the data model. In an MRS scan the clinician collects 1-3 "normal" spectra from distant areas in the brain along with the primary pathological area spectrum for reference. The paired SVM kernel utilizes the "normal" information $\mathbf{x}'$ in an effort to create an RBF map for the healthy reference baseline of the population in parallel to the pathological map reflected by the associated data x. This is expected to allow the resulting Gram matrix to obtain larger classification margins.[7]

Assuming the availability of scanning voxels placed on additional areas of interest relative to the lesion, the above framework can be extended to incorporate an arbitrary number of voxels.

This concept can be formulated as

$$k_m\left(\mathbf{x}_i,\mathbf{x}_j\right)=\sum_{a=1}^{N_{areas}}\exp\left(-\frac{\left\|\mathbf{x}^a{}_i-\mathbf{x}^a{}_j\right\|_2^2}{\sigma_{\mathbf{x},a}{}^2}\right) \quad (4.8)$$

In Eq.8 a is the area indicator variable and N_{area} is the total number of examined areas. In this study metabolic data were collected from $N_{area}=4$ brain areas, however due to missingness and homogeneity issues the classification layer utilizes $N_{area}=2$ thus reducing Eq.8 to Eq.7.

Based on the theory relating to "kernels on sets" (Cristianini and Shawe-Taylor 2000), it is also possible to extend the above formulation to include a varying set of scanned voxels so as to dynamically adjust the kernel to the available data.[7]

4.4 Experimental results

Our efforts focus on decoupling and fully automating the model and its parameter selection process, thus limiting the manual interaction to the initial selection of the affected area to be classified by the radiologist.

The overall evaluation of the diagnostic models was performed under a statistical-Bayesian perspective. The presented results are focused on the subset of CSI-144 mode features. These features are estimates of the metabolites' values at specific predefined frequency domains.

At an initial stage, feature selection was performed in order to evaluate the relative prognostic gain that is obtained by each feature separately. This was done using Automatic Relevance Determination (ARD) (Van Gestel, Suykens et al. 2001), a multi-level Bayesian variable selection method operating as a integral part of the SVMs' model. The

resulting features were combined sequentially based on the ARD criterion according to an "add-m remove-n" scheme to obtain an optimal feature subset which reduces the overall error estimates while minimizing feature redundancy. The feature selection algorithm was parameterized with m=1, n=0 and initialized with the full feature set. The criterion for each added feature on every iteration was the feature relevance (negative cost) function defined by ARD. The comparative relevance of the MRS covariates is given in Figure 4-3.

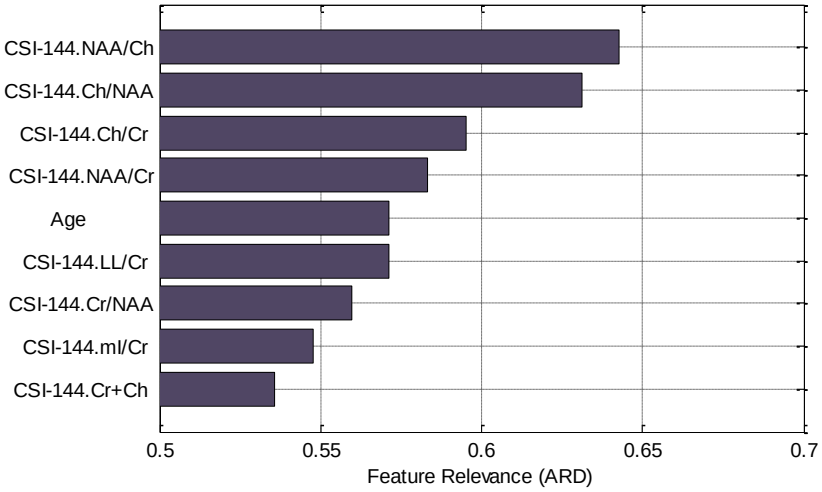


Figure 4.5 Feature ranking via Automatic Relevance Determination

Our analysis shows that the use of the 5 most informative features (NAA/Ch, Ch/Cr, NAA/Cr, Age, Ch/NAA) accounts for 82% of the dataset's variance and minimizes the mean overall misclassification rate for the given data. Inclusion of additional features increased the models' sensitivity to parameters and hindered further classification performance improvement. Both types of SVM classifiers used the 5 most informative features based on the feature selection analysis. This was done to provide a fair comparison between them.

The actual evaluation of the resulting reduced feature SVM classification system was performed by partitioning the dataset into disjoint subsets containing 70% (training set) and 30% (independent test set) of the available cases respectively. This process was carried out iteratively in order to create a 10x stratified sampling scenario that is expected to provide robust evaluation results. This is equivalent to creating 10 random partitionings of the data into 70%-30% groups with replacement in order to ensure adequate representation from each class. The reported accuracy estimates are the averaged values of the 10 test set calculated accuracies for each class combination.

Optimal SVM parameter estimation constitutes an optimization problem, which usually presents local minima in the parameter space. In most cases, a range of parameters where the performance is stable can be defined. In the context of the work presented here for each classifier we first selected the expected ranges of γ (C) and ε based on input range and estimated noise (Staelin 2003). The σ_x and $\sigma_{x'}$ parameters were coarsely estimated using Fisher Discriminants. Using the above limiting conditions all parameters were estimated using the procedure of iteratively refined grid search (Wang 2003). The small sample size did not allow for a separate parameter optimization set.

However, the same procedure was used on both types of kernels. While this does not guarantee optimal generalization, it suffices in this case as the main point of interest is the comparison of the paired and non-paired RBF kernels.

Figure 4.4 visualizes the SVMs' classification accuracy per pathological class by means of the corresponding confusion matrix. Better diagnostic capability for a class is indicated by a higher relative value in the corresponding diagonal element and proportionately increased color intensity

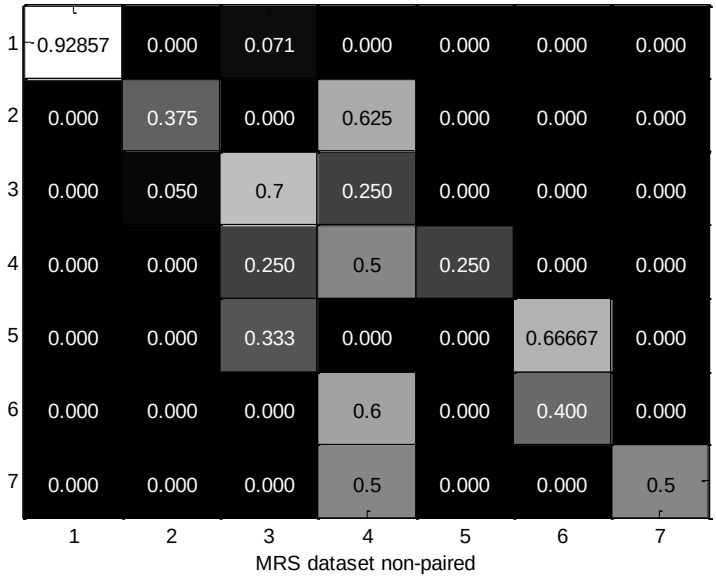


Figure 4.6 Multiclass classification performance for non-paired MRS features' SVM kernels.

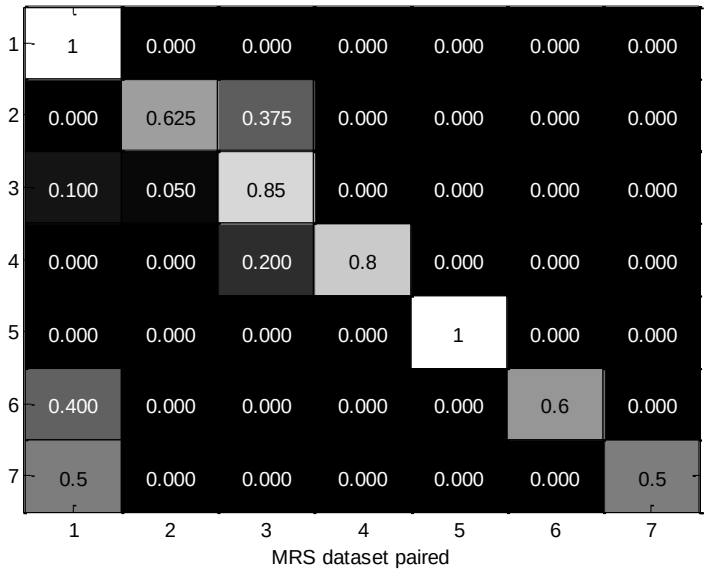


Figure 4.7 Multiclass classification performance for paired MRS features' SVM kernels. Paired kernel representation exhibits improved accuracy and reduced class overlap.

The classification performance achieved by standard (non-paired) feature vector RBF-SVMs, shown in Figure 4.6, indicate that the system can at an initial stage differentiate the healthy (class 1) from pathological (classes 2-7) cases with high confidence.

At a more granular level high accuracies are also observed in the selection of class 3 (Glioma high), while classes 2 (Glioma low), 4 (Metastasis), 5 (Meningioma), 6 (MS), and 7 (Cyst) suffer from poor performance. The partitioning of cases with Gliomas of different grades is also difficult under the given model. The reduced accuracy in class 5 (Meningiomas), is attributed to the missingness of the key indicator for this pathology, namely metabolite Alanine. The key concept of this work, which is the evaluation of paired SVM kernels for MRS diagnosis, is justified by the improved class specific accuracy of the paired SVM kernel. The overall accuracy reached 0.73 for the non-paired and 0.88 for the paired kernels thus indicating an overall improvement in classification performance.

Figure 4.7 visualizes the confusion matrix for the proposed paired-SVM kernel. The diagonal elements (true positive rates per class) are higher compared to the standard SVM model and class overlaps are reduced. Yet classes 6 and 7 appear to map a large portion of their cases to the "normal" class 1. This fact is attributed to the asymmetrically large size of class 1 and can be counter balanced by adjusting for unequal class prior distributions. This however increases the variance of the accuracy estimator, due to the small sample size and has been avoided until there are adequate samples for classes 2, 4, 5, 6 and 7.

Extensive previous research efforts in this domain as presented in (Lukas, Devos et al. 2004; Jan Luts, Poullet et al. 2008), reported accuracies in the range of 0.90-0.98 by utilizing continuous 1.5T spectra for a more limited set of outcomes (4 classes) and more

extended feature sets thus operating on a simplified version of this problem. The current approach is aimed as a preliminary demonstrator for the prospect of expanding the SVMs' classification to additional pathological classes with a more precisely chosen reduced feature set. The feature selection scheme has shown improved performance as the number of patients in the dataset increases.

Entry (5,5) in the confusion matrix in Fig 4.7 corresponds to the classification of meningiomas and shows a striking change in accuracy from 0 to 100% using the proposed paired kernel. Meningiomas are in the current version of our dataset a very small class consisting of only 3 samples and therefore the variance of the estimated accuracy is high. Despite the research team's debate over this result, it was deemed appropriate to publish it as is for consistency and re-evaluate the meningiomas class as soon as additional samples become available.

The second key benefit of the proposed model is that it provides a structured way to represent different scanned regions in terms of a paired SVM kernel, which opens the prospect of applying more elaborate MRS classification kernel designs.

Comparison with the corresponding clinical expert assessments is not yet feasible since the former are actually used as classification targets at this stage.

4.5 Conclusion

Conventional MRI is limited in its ability to highlight the most aggressive portions and the entire extent of a tumour, particularly in those which lack contrast enhancement. Although proton MRS is far away from replacing surgical biopsy for the diagnosis of brain tumour, it offers the advantage of selecting the appropriate target for biopsy and detects the extent of tumour, especially in cases of infiltrative gliomas before being detected from conventional MRI. Areas of high mI/Cr and Cho/Cr ratio in the contralateral side of multiple glioblastoma (GBM) have been shown to correlate with high tumour proliferative index before having any MR imaging suspicion. The assessment of residual disease after surgery is another important application of proton MRS particularly for low-grade gliomas or nonenhancing portion of high grade tumours. In addition, the more detailed identification of tumour type can provide valuable information prior to surgical or treatment planning. For example, metastatic brain tumours, high grade gliomas or some cases of lipomatous meningiomas are sometimes difficult to differentiate because of similar appearances on conventional MRI of their focal portion. Furthermore, cystical metastatic tumours can have also similar appearance with pure cystic lesions. Proton MRS aids in the differential diagnosis of the aforementioned lesions by detecting the metabolic features in the peritumoral region adding useful metabolic information to that obtained with conventional MRI. Peritumoral Cho/Cr and Lipids/Cr ratios provide a useful index in tumour type differentiation of high grade gliomas and metastasis, as well as of pure cystic lesions and cystical metastatic tumours. Thus, even if proton MRS does not change the leading diagnosis, it may rule out differential diagnosis and thereby reduce the need for biopsy when is not necessary.

Overall, the clinical information from MR imaging alone has an accuracy of lesion classification of 74%. Among several sophisticated imaging techniques proton MRS have been proved superior to distinguish benign from malignant brain tumours. Mishra et al

differentiated 52 histopathologically proved tumour cysts, abscess or benign cysts by using single voxel proton MRS and diffusion-weighted MR imaging. The authors reported the sensitivity and specificity of proton MRS to be 96% (95% CI) and 100% (95% CI) respectively. This compares favorably with diffusion-weighted imaging where specificity remained high (100%) but sensitivity was diminished (72%). [21]

In the near future, it is unlikely that radiologists will make a diagnosis based solely on conventional decision rule. The benefit of including proton MRS within a standard MRI examination have been shown a 15.4% increase in clinical diagnosis, 6.2% fewer incorrect diagnoses and 16% fewer equivocal diagnoses than for MRI alone. Especially with an automated decision support system that will analyse and classify proton MRS data improved differential diagnosis will definitely upgrade patient outcome.

The above analysis verifies that the performance of distance based nonlinear classifiers in the form of SVMs can amplify the diagnostic power gains achieved by current 3T MRS scanning technology. We demonstrate the applicability of modified SVM kernel in this task in conjunction with the above more powerful data acquisition layer.

The applicability of the proposed reduced feature set can be applicable in scenarios where the acquisition, transfer and utilization of a full resolution MR spectrum from the scanner to a clinical decision support system is impractical or time demanding. In this context the use of 4-5 features' estimates can provide an estimate of the patient's status.

While additional performance gains may be achievable by optimizing the above methodology at the data acquisition, preprocessing, classification and visualization levels, the practical use of such decision support tools in a clinical environment relies heavily on the ability to provide coherent performance across multiple datasets and pathological classes.

To this end future research directions might include the utilization of data and decision fusion methods, the use of full spectral information to create visual diagnostic aids and the evaluation of an interactive graphical user interface to communicate and adapt the decision support system's statistical estimates to the clinician's feedback in real time

Equation Chapter (Next) Section 1

5 Classifier fusion

5.1 Introduction

The question of how we can exploit the ability to combine different learning entities is fundamental to the core of automated pattern analysis and dictates contemporary research efforts in the field of decision fusion. While the broad class of information fusion methods is constantly enriched, their proper consideration on the basis of data or information distribution lacks a common framework and develops around ad-hoc methods that cannot justify the overall effectiveness of fusion methods. In this context, the present work aims at uncovering analogies between decision fusion methods and established primary classifiers. Such correspondence of specific fusion methods to base classifiers allows us to utilize knowledge from the field of data mining as to summarize and model the statistical performance of combiners and possibly provide best practices and optimality criteria for their use. As case studies, we focus on two main categories of classifiers, namely distance and discriminant-function based, when applied to the problem of classifier fusion. The Decision-Templates fusion method is examined as a representative distance based technique and compared with the Support-Vector-Machine scheme as representative of discriminant-function hyper classifiers. Based on theoretical performance measures, we advocate the use of SVMs as an efficient and extensible framework that can be adapted to specific application domains.

5.2 Background

In related literature, the terms “classifier fusion”, “multi classifier systems” (MCS), “classifier ensembles”, “classifier aggregation” and “classifier combiners” have been used to describe a common research domain. In the following analysis, we consider the above terms as equivalent and primarily use the terms “ensemble methods” or “combiners” to reflect the wide class of fusion algorithms according to (Valentini and Masulli 2002). On the other hand, adjacent concepts including feature and parameter selection, combination of binary classifier outputs for multiclass systems, and structured output designs is considered complementary to our analysis.

The notion of classifier-output diversity has been extensively explored from statistical (Goebel and Yan 2004), generative (Shipp and Kuncheva 2002) and information theoretic (Brown) viewpoints and is by now a known driving factor affecting the quality of the primary (or base) classifier pool. Based on the different states of a classifier ensemble, such dependence can be related with the quantification of the error at three levels, namely the data or feature space, the base-classifiers’ class labels and the combiner’s class labels. This multi-level consideration is visualized in Figure 1, where the above three levels of knowledge representation are presented along with the corresponding qualitative statistical distributions.

Decision fusion in the form of classifier outputs combination can be implemented via a variety of methods (Brown 2010). Currently available methods extend beyond simple non-parametric arithmetic operators (min, max, product, sum), to possibilistic (Dempster-Shäfer) and information theoretic combiners (BKS) ((Ruta and Gabrys 2000)). Moreover, the introduction of the “Decision-Profile” (DP) framework ((Kuncheva, Bezdek et al. 2001)) enabled the consideration of many of these approaches within a unified context. The resulting Decision-Templates (DT) method compares each sample’s decision profile with a pre-calculated decision template for each class, providing the option to utilize a wide pool of distance metrics. Kuncheva in (Kuncheva 2004) provides a more extensive taxonomy of

available methods, mainly grouping them into class-conscious versus class-indifferent and trainable versus non-trainable schemes. Other recent advances include cluster ensembles (Strehl and Ghosh 2003), dynamical systems learning ("concept drift") (Kuncheva 2004), and diverse and multimodal data classification.

Criticism on classifier fusion methods ((Kittler 2000)) has been primarily based on the increased complexity that a two-level classification model introduces and the limited space of improvement in view of the latest high-performance kernel machines.

More radical approaches ((Kuncheva 2008), (Raudys 2002), (Kittler 1998)) examine the duality between base classifiers and advanced feature extraction algorithms, a concept that reduces the overall combiner design to that of a generic classifier. Yet this idea has been leveraged only as a tool to exchange mutual information/diversity criteria and model selection schemes between the two domains and is not considered from a design or statistical perspective. In parallel to statistical relations, general pattern recognition algorithms have been tested as hyper-classifiers, using as input data the feature space created by the primary classifiers' hard labels or soft support values (Wang and Zhang 2001; Kittler and Messer 2002; Chang 2005). Following these developments, we proceed deeper with examining fundamental relations between base classifiers and combiners. Our analysis is based on the aspects of mathematical duality and statistical error propagation.

5.2.1 Motivation

The driving principle behind the development of classifier ensembles has been expressed in two directions. Firstly, despite the advances of data mining approaches, there are still open problems related to limited data availability (Kittler 2000), inadequate feature representation, dynamical nature and/or complex class separation hyperplanes, which prevent single classifiers from achieving practically usable performances. A classifier ensemble can be employed to alleviate these problems by dividing the task into simpler sub-problems requiring low performance and latter fusing the produced support values at a more accurate level. Secondly, despite inter and intra dataset variations, many sporadic results on empirical data show a general trend that classifier ensembles improve the overall classification performance and are more robust to feature outliers, missing data and concept drifts. Empirical studies pointing to the benefits of combiners include biomedical decision support systems (Kuncheva 1993; Dimou, Manikis et al. 2006; Altmann, Rosen-Zvi et al. 2008), network intrusion detection (Giacinto, Roli et al. 2003), handwriting recognition (Huang and Suen 1995), and remote sensing and person recognition (Oza and Tumer 2008). Other reasons in favor of ensemble methods include global optimality, minimal bias and/or variance and increased generalization (Valentini and Masulli 2002).

In most cases, the effects of fusion approaches have been documented with application-dependent tests based on experiments on specific datasets. Nevertheless, a more theoretic view of the overall error bounds is still needed as to justify the use of combiners and fusion schemes. In practice, each classifier fusion approach possesses different characteristics that affect the overall performance and stability depending on the dataset, the pool of base classifiers and the employed optimality criterion. In many cases, the statistical properties and assumptions of both base classifiers and combiners are not adequately analyzed (Chang 2005; Li, Yin et al. 2005) or have been based on overly strict assumptions (Kuncheva 2002) to be widely interpretable. This inefficiency results in large or even inestimable uncertainties in the combiner's overall accuracy and its generalization ability. Most research work in the field presents accuracy improvements in specific applications and algorithms while giving limited statistical justification. Where provided, the confidence intervals stem from the output variance of

within the test-set samples, which is subject to bias. Additionally, the skewed shape of the combiner feature-space invalidates many linear or Gaussian-based classifiers. More specifically, outputs of multiple classifiers employed as ensemble inputs possess a large concentration around the set of class labels $\{0,1\}$, which defies the assumptions of base classifiers and does not allow the creation of good final mappings. Furthermore, the occurrence of highly correlated soft labels in decision fusion applications may provide consensus but intensifies the problem of designing ensembles to improve the classification of the remaining difficult outlier cases.

Towards the effective design of ensembles, (Kittler 2000) proves that "fusion of raw expert outputs is a nonlinear function" motivating the use of nonlinear classifiers for this task. Initial theoretic approaches (Kittler, Hatef et al. 1998; Schapire 1999; Kleinberg 2000) account for the effectiveness of the overall combiner only from limited perspectives. In this work, we attempt to provide a theoretical basis for the design and analysis of classifier ensembles.

In practice based on the above work it can be argued that hyper-classifiers provide the following benefits:

- They can attain better coverage of the feature space.
- They provide more reliable estimates for error and confidence, whereas base classifiers usually fail to give more than a label per case.
- Every machine-learning algorithm has a learning capacity, beyond which it is incapable of learning and storing a given mapping. Ensembles can assign different weights to a number of classifiers and thus let each one specialize in a portion of the problem.
- Most importantly: It is proven that a chosen ensemble of even moderate performing base-classifiers can outperform a single fine-tuned one.

The combination of the above factors has led to the adoption of this comparatively new algorithmic direction that has emerged from the increasing need to provide robust results in a variety of difficult machine learning problems, where no individual algorithm suffices.

5.2.2 Contribution

As in (Kuncheva, Bezdek et al. 2001), we view the role of combiner as a traditional classifier operating on initial decisions (first-order classification labels) as in Figure 1. In this form, the difference of operation between classifiers and combiners stems from the input distribution; the input of the 2nd level (L2) classifier (combiner) assumes a skewed distribution concentrated around the 1st level (L1) class labels. The contribution is summarized in the following three aspects:

- Attempts to define a taxonomy of trainable combiners based on their class discrimination operator in relation with base classifiers. The two classes considered involve combiners based on discriminant functions and combiners based on distance functions.
- Associates these categories with relevant combiners that have been proposed in the relevant literature
- Derives error bounds for each category, where possible, by associating the concepts of combiners with those of classifiers.

(Valentini and Masulli 2002) have manifested the need to research hidden commonalities between existing classifier-fusion methods as a first step to designing more robust ensemble frameworks. By drawing analogies between trainable combiners and well known classifiers, we can model the probability of error for the decision-fusion layer based on established theoretical bounds from the

layer of base classifiers. In particular, two general classes of classifiers have dominated the area of primary classifiers. The use of minimum-distance classification in decision temples (DT) and in other specialized combiner methods (Dempster-Schafer, Naïve Bayes) provides specific benefits for mapping decision to the feature space, but can be suboptimal from a pattern analysis aspect. Alternatively, experimenting with the power of discriminant functions, (Chang 2005) evaluated Radial Basis Neural Networks as an alternative fusion scheme, resulting in increased fusion performance, comparable only to the Dempster-Schafer scheme. The potential of discriminant functions has been widely explored along with Support Vector Machines (SVMs) (Cristianini and Shawe-Taylor 2000). We build on these two classes by extending their notion from the classifier to the combiner space. Furthermore, we demonstrate this concept by comparing the DT and SVM fusion approaches on publically available datasets.

The error statistics of various classifiers have been thoroughly considered in the literature. Nevertheless, probabilistic models aiming to describe the combiners' error statistics have been proposed only for simple, non-trainable combiners ((Kuncheva 2002; Kuncheva, Whitaker et al. 2003)), which have been used extensively in the past to fuse decision outcomes in lack of more elaborate data driven algorithms. This class of methods includes minimum, maximum, sum, average, median and product combiners, with their key advantage being the implementation simplicity and speed. Majority voting belongs to this group with the difference that it utilizes hard classifier outcomes as inputs and has provided acceptable results in related tasks (Kuncheva, Whitaker et al. 2003). The average combination rule has shown to perform relatively well even compared to far more advanced methods (Kuncheva 2002). However, it degrades steeply when non-descriptive features or outliers from the base-classifier outputs are present. Under the assumption of either clipped uniform or normal probabilities, (Kuncheva 2002) derives theoretical error estimates of the above six methods. (Kittler, Hatef et al. 1998) provides a derivation of the average rule from the sum rule and suggests that the former is much more resilient to outliers compared to the product rule. He also suggests that there are high-noise scenarios where the information contributed by weak features is far less important than the prior class probabilities and therefore the sum/average rule is superior despite its simplicity. We exploit the above formulations, as well as the proposed generalization of classifier classes into combiner categories in order to draw error bounds for the widely used combiners. More specifically, considering the output distribution and error pdfs of primary classifiers, we derive the overall error bounds of combiners; results are summarized in Appendix II.

In the following sections we deploy the above ideas, building on the necessary background on classifier ensemble methods. In Section 5.3 we explore the key analogies between trainable combiners and established classifier models that are based on distance metrics. In Section 5.4 we extend the analogy to nonlinear combiners based on the concept of discriminant functions and focus on SVMs as representative examples. We explore the statistical and mapping properties of these models and outline the benefits of using SVMs as combiners. Section 5.5 describes the comparative experimental framework and discusses the obtained results from the above classifier fusion methods. In section 5.6 we summarize and review our findings in conclusion of this work. Detailed notation used throughout this work is discussed in the next sub-section and summarized in Appendix I. Appendices II and III outline our conclusions on the statistical properties and the suggested use-scenarios of ensemble methods, respectively, with pointers to related literature.

5.2.3 Notation

Vectors are represented in lowercase bold notation, e.g. $\mathbf{x} \in \mathbb{R}^N$, and their scalar components in italic script, e.g. $x_1, x_2, ..., x_N$. Matrices are represented in uppercase bold script, e.g. $\mathbf{DP} \in \mathbb{R}^{L \times C}$. The sizes of matrices, spaces and sets are represented in uppercase italic script, e.g. N.

Let $\{\mathbf{X}, \mathbf{Y}\}$ be a dataset consisting of N feature vectors $\mathbf{x}_i \in \mathbb{R}^D$ and the corresponding class labels $\mathbf{y}_i \in \mathbb{R}$, $i = 1, .., N$ and let $f = \{f_1, ..., f_L\}$ $l = 1, .., L$ be a set of L classifiers each producing a set of C soft class labels $d_{lc}(\mathbf{x}_i)$ corresponding to classes labels $\omega_c, c = 1, .., C$. The index i in every feature vector x_i will be omitted, when it is clear from the context.

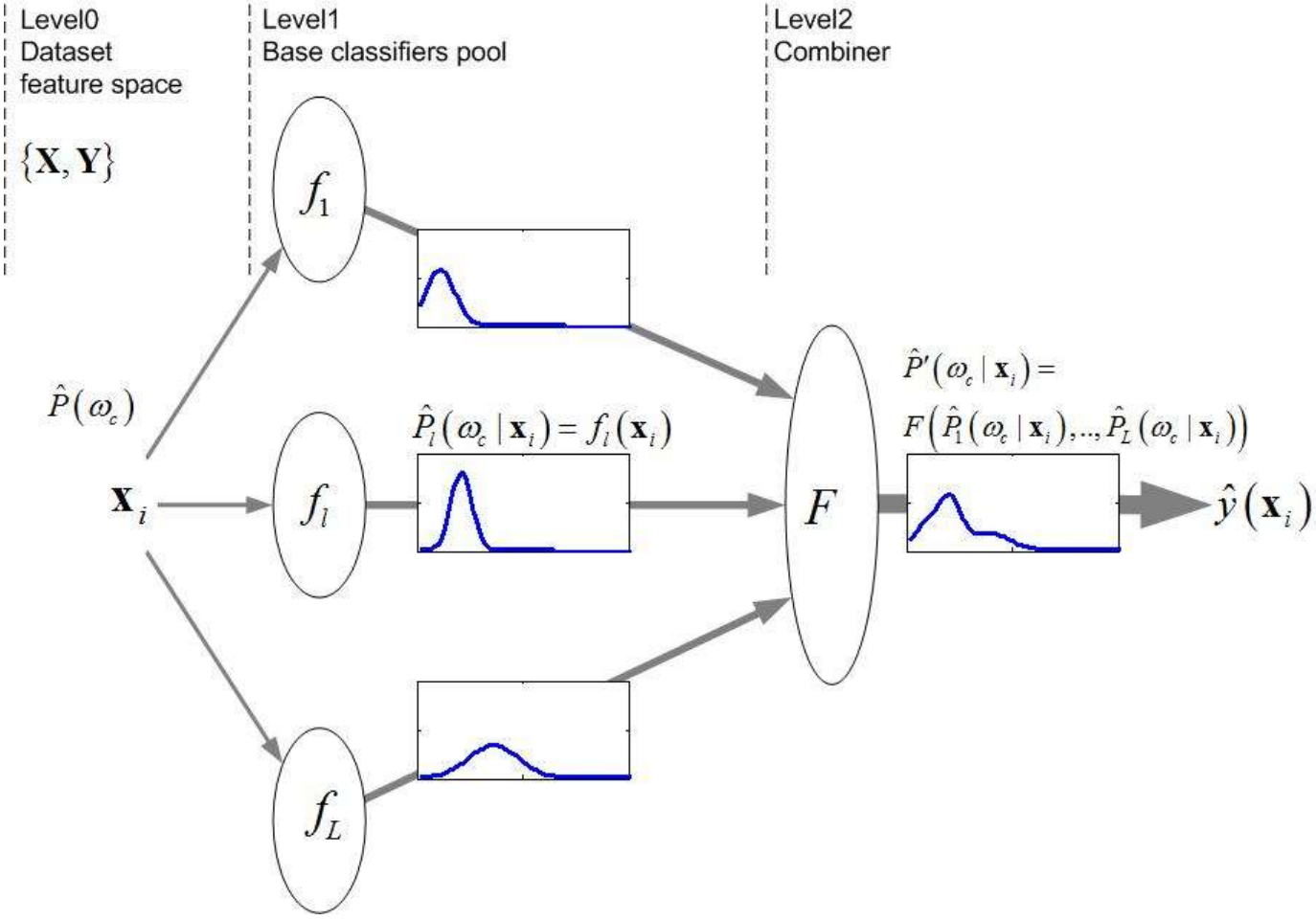


Figure 5.1 Probabilistic information flow in a decision fusion system

From a Bayesian perspective, for any base classifier the optimal decision regarding sample $\mathbf{x}$ amounts to selecting the class with the maximum a posteriori probability, that is:

$$assign\ \hat{\omega}_c \rightarrow \omega_c\ \ if\ \ P(\omega_c \mid \mathbf{x}) = \max_j P(\omega_j \mid \mathbf{x}) \qquad (5.1)$$

The decision error arises from the fact that the estimate on the class posterior probability is suboptimal due to a variety of reasons including the features' poor descriptive ability; the classifiers' limited mapping capability etc.

For a classifier ensemble the above equation becomes

$$assign\ \hat{\omega}_c \rightarrow \omega_c \quad if\ P'(\omega_c \mid \mathbf{x}) = \max_j P'(\omega_j \mid \mathbf{x}) \tag{5.2}$$

where P' denotes the corresponding class posterior probability estimates at the decision fusion level.

Similarly the overall system's error arises from the fusion scheme's suboptimal class posterior probability estimate $\hat{P}'(\omega_c \mid \mathbf{x})$ due to poor base classifier performance $\hat{P}_l(\omega_c \mid \mathbf{x})$ (which includes the above error sources), and the combiner's limited mapping capability.

Unless otherwise stated for the derivation of the statistical formulations, we use probabilistic soft outputs from the base classifiers. For classifiers that do not directly support this we employ either the logistic link function (for the soft support values) or counting estimators (for hard outcomes) as described in (Kuncheva 2004).

We also assume $C = 2$ class problems $P_l(\omega_1 \mid \mathbf{x}) + P_l(\omega_2 \mid \mathbf{x}) = 1$ for which the error estimates are denoted by

$P_l(e \mid \mathbf{x}) = P\left(\hat{P}_l(\omega_c \mid \mathbf{x}) < 0.5\right)$ for the base classifiers level and

$P'(e \mid \mathbf{x}) = P'\left(\hat{P}'(\omega_c \mid \mathbf{x}) < 0.5\right)$ for the combiner's level

We assume that the base classifiers commit independent and identically distributed (i.i.d.) errors in estimating $\hat{P}_l(\omega_c \mid \mathbf{x})$. Overall, we can estimate

$$\hat{P}_l(\omega_c \mid \mathbf{x}) = P_l(\omega_c \mid \mathbf{x}) + P_l(e \mid \mathbf{x}) \tag{5.3}$$

Similarly, assuming i.i.d. error pdfs we can estimate the combiners' posteriors:

$$\hat{P}'(\omega_c \mid \mathbf{x}) = P'(\omega_c \mid \mathbf{x}) + P'(e \mid \mathbf{x}) \tag{5.4}$$

Using the Bayes rule, we can rewrite the fusion system's class posterior as

$$P'(\omega_c \mid \mathbf{x}) = \frac{\hat{P}'(\mathbf{x} \mid \omega_c)\hat{P}(\omega_c)}{P(\mathbf{x})} \tag{5.5}$$

Furthermore, it can be shown that combining the base classifiers' estimates and omitting the class-independent terms, leads us to the formula

$$P'(\omega_c \mid \mathbf{x}) \propto \hat{P}(\omega_c)^{L-1} \prod_{l=1}^{L} \hat{P}_l(\omega_c \mid \mathbf{x}) \tag{5.6}$$

which is equivalent to the optimal Bayesian decision rule.

Assuming that per-class conditional probabilities can be safely estimated from the available data, each base classifier's error is given by:

$$P_l(e \mid \mathbf{x}) = \int_{\mathbf{x} \in \omega_2} P_l(\omega_1 \mid \mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} + \int_{\mathbf{x} \in \omega_1} P_l(\omega_2 \mid \mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} \tag{5.7}$$

Similarly the overall error can theoretically be calculated as

$$P'(e \mid \mathbf{x}) = \int_{\mathbf{x} \in \omega_2} P'(\omega_1 \mid \mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} + \int_{\mathbf{x} \in \omega_1} P'(\omega_2 \mid \mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} \tag{5.8}$$

The above formulations are generic and do not rely neither on classifier's (or combiner's) properties nor on specific distributions. In the following sections, we will employ more specific formulations for the above statistical quantities. We will consider base classifiers using identical feature sets (as opposed to discrete feature representations). Therefore, all the required diversity will be derived either through selection of a variety of classifiers, or through their parameterization.

5.2.4 Classifier fusion principles

Before applying a specific family of aggregation strategies, it is important to put the problem into perspective and identify some key concepts.

Firstly, there are many different approaches to classifier fusion based on:

1. the assumptions about the classifier dependencies
2. the type of base classifiers used
3. the aggregation strategy (local/global)
4. the aggregation procedure

We can also identify two different approaches in using the dataset, namely selection and fusion. In the first one, we regard each particular classifier as a local expert for an area in the feature space and train and use them in a complementary way to each other. In the second, we train and use all base classifiers in the same full data and then merge all their outputs for a specific data point.

One very important in the process is the type of L1 classifier outputs. Although we can scale any classifier's outputs to [0,1] for consistency, it is also necessary to take into account the way each algorithm's outcome is supposed to be interpreted. Due to the different assumptions, other learners' outcomes can be seen as labels, posterior probabilities, conditional probabilities, or just support values for a specific class. For example, linear regression models require Gaussian covariates whereas Neural Networks do not. In context with our previous notation, we denote the L1 outputs for the c classes as

$$\mu_D(x) = \left[\mu_D^1(x),...,\mu_D^c(x)\right]^T, \mu_D^1(x) \in [0,1]$$

Three basic ways to interpret μ^i_D are given in literature (Kuncheva, Bezdek et al. 2001) hard, fuzzy and possibilistic.

In a more general approach we can handle the fusion problem using any 2nd level classifier, using the L1 outputs as the intermediate features (Kittler 1998). This solution however has an inherent drawback. The distributions of the L1 outputs are strongly concentrated around 0 and 1, thus invalidating many linear algorithms.

5.3 Distance-Based combiners

Distance-based combiners are methods that in essence utilize distance metrics of each point in the L2 feature space to each class's focal points to determine the final (fused) class membership. A variety of combiners belong to this algorithmic scheme, where the information utilized in the combiner is encoded in the soft class labels or posterior probabilities of the primary classifiers. In particular, the decision templates combiner derives a class-related template for all classifiers derived from the training set, the Bayes classifier estimates conditional class likelihood, whereas the Behavior-Knowledge-Space combiner computes the occurrence counts per class. Then, for the input vector, each type of combiner derives an appropriate matching prototype and compares this sample-specific with the corresponding class-specific metric, so as to control the distance-based classification outcome. In the following subsections we outline the design and details of this extended class of combiners.

5.3.1 Decision profiles as a framework for fusion

The Decision Profile (Kuncheva, Bezdek et al. 2001) is in essence a two dimensional structuring of the soft class labels indexed by class and classifier number in a way appropriate to facilitate subsequent stages of processing.

Let $\{f_1,..., f_L\}$ be the set of L classifiers used in the ensemble trained on a problem with C output classes. We denote the output of the l^{th} classifier as $\mathbf{D}_l(\mathbf{x}) = [d_{l,1}(\mathbf{x}),..., d_{l,C}(\mathbf{x})]$, where $d_{l,c}(\mathbf{x})$ is the degree of support given by classifier f_l to the hypothesis that a sample $\mathbf{x}$ comes from class c. Where it is clear from the context we use $\mathbf{x}$ instead of $\mathbf{x}_i$ to denote the i^{th} sample of the dataset. The fused output of the L first-level classifiers is constructed using an aggregation function *F* as

$$F(\mathbf{x}) = F\left(f_1(\mathbf{x}),..., f_L(\mathbf{x})\right) \quad (5.9)$$

Assuming a two-class problem ($C = 2$), the decision profile for each sample $\mathbf{x}$ in the dataset is a $L \times 2$ array of the soft labels given by each of the L classifiers.

$$\mathbf{DP}(\mathbf{x}) = \begin{bmatrix} d_{11}(\mathbf{x}) & d_{12}(\mathbf{x}) \\ \cdots & \cdots \\ d_{l1}(\mathbf{x}) & d_{l2}(\mathbf{x}) \\ \cdots & \cdots \\ d_{L1}(\mathbf{x}) & d_{L2}(\mathbf{x}) \end{bmatrix}$$

There are methods that calculate the support for class c using only the c^{th} column of $\mathbf{DP}(\mathbf{x})$. Such fusion methods that use the DP class-by-class are called *class-conscious*. If on the other hand the whole DP is used to calculate the support for each class the fusion method is called *class-indifferent.*

The distribution of the soft labels of the primary classifiers $d_{l,c}(\mathbf{x})$ creates difficulties for hyper classifiers. In essence, the rows of the decision template become highly dependent so that the covariance matrix of the DP becomes singular and non-invertible as each primary classifier's accuracy increases. This is in part the reason why a diverse ensemble of low accuracy algorithms is in general preferable to a highly correlated high-accuracy ensemble.

5.3.2 Decision templates combiner

The decision templates combiner relates to the notion of a focal point for the cluster of decision profiles of each class. This quantity is defined as:

$$dt_{c'}(l,c)(\mathbf{X}) = \frac{\sum_{i=1}^{N} ind(\mathbf{x}_i, c') d_{l,c}(\mathbf{x}_i)}{\sum_{i=1}^{N} ind(\mathbf{x}_i, c')} \quad (5.10)$$

$c, c' = 1, \ldots C$

$l = 1, \ldots L$

$$ind(\mathbf{x}_i, c') = \begin{cases} 1, \text{ if } \mathbf{x}_i \text{ has crisp label } \omega_{c'} \\ 0, \text{ otherwise} \end{cases} \qquad (5.11)$$

The decision template for each class c' *represents a matrix* $\mathbf{DT}_c$ with dimensions LxC, which is built from the individual degrees of support $dt_{c'}(l,c)(\mathbf{X})$ the same way as the decision profile is built from $d_{l,c}(\mathbf{x})$. When a test sample $\mathbf{x}_i$ is submitted for classification, the DT algorithm matches $\mathbf{DP}(\mathbf{x})$ to $\mathbf{DT}_c$, $c = 1, \ldots C$, and produces the soft class labels that correspond to posterior probability estimators for each class

$$\hat{P}'(c \mid \mathbf{x}) = S\left(\mathbf{DT}_c, \mathbf{DP}(\mathbf{x})\right), \quad c = 1, \ldots, C \qquad (5.12)$$

where S is a similarity or inverse-distance measure. Most often S corresponds to a Euclidean, normalized Euclidean or Mahalanobis distance, although a number of other distance metrics have been proposed in (Kuncheva, Bezdek et al. 2001).

Many other combiners utilize this concept. The Dempster-Schafer combiner uses the decision templates to define a possibilistic measure of membership. Naïve Bayes and Probabilistic Product combiners use probabilities as distance metrics. The common ground is that they rely on the notion of distance measures minimization. For the primary-classifier outputs $d_{l,c}(\mathbf{x}_i)$ of each sample, a weighted distance is calculated between its DP and each class's DT. Based on the sum of this distance and a mapping function, a class membership value $\hat{P}'_{D'_l}(c \mid \mathbf{x}_i)$ is calculated. The sample is then assigned to the class with the highest class membership value (smaller distance). This way all other points of each class take part indirectly to the classification of the sample.

Depending on the distance measure used, we can establish analogies between the DT and known classifiers as analyzed below. Most published studies utilize DTs with Euclidean distance metric, while the Mahalanobis distance can also be used to account for non-equal covariance matrices of the DPs. In addition, (Kuncheva, Bezdek et al. 2001) proposes fuzzy similarity measures and indices of inclusion and consistency for flexible implementations of DTs.

Statistical estimates on the performance of DTs using the above metrics can be derived from their equivalent classifiers, namely the minimum distance classifier and the quadratic discriminant classifier (QDC). More specifically the expressions for $\hat{P}_{D'}(error)$ and $\hat{P}_{D'}(c \mid \mathbf{x})$ are formulated as follows:

The Euclidean metric to assign DPs to a classes' DT is defined as

$$S_{Eucl}\left(\mathbf{DP}(\mathbf{x}_i), \mathbf{DT}_c\right) = 1 - \frac{1}{LC}\sqrt{\sum_{l=1}^{L}\sum_{c'=1}^{C}\left(d_{l,c'}(\mathbf{x}_i) - dt_c(l,c')\right)^2} \qquad (5.13)$$

and the equivalent classifier is the minimum Euclidean distance, also known as nearest mean classifier NMC. In (van Otterloo and Young 1978) a nonparametric upper bound for the error of the NMC classifier utilizing the Euclidean distance is developed.

$$P_{NMC}\left(error \mid \omega_c\right) \leq \frac{D}{R_c^2} \quad (5.14)$$

where R_c is the radius of the minimum enclosing hypersphere that contains all samples from class c and D is the features' dimensionality.

The same bound can also be used for the combiner, at the L2 classification level. Additionally, introducing the Gaussian assumption for the L1 soft labels results in a tighter bound

$$P_{NMC}\left(error \mid \omega_c\right) \leq \frac{\Gamma\left(N/2, R_c^2/2\right)}{\Gamma\left(N/2\right)} \leq \frac{N}{R_c^2} \quad (5.15)$$

where Γ denotes the gamma function.

If we formulate the above bounds in terms of the DTs notation, where the new features' dimensionality is $D' = L \cdot C$ we arrive at

$$P_{DT}\left(error \mid \omega_c\right) \leq \frac{LC}{R_c'^2} \quad (5.16)$$

where R_c' is the radius of the minimum enclosing hypersphere that contains all the DPs derived from samples of class c.

The obvious limitation of such minimum distance methods lies in the resulting decision surface. The optimal estimation of this surface results in a hyperplane perpendicular to the line connecting the centers of the two classes (DTs) in the L-dimensional space. Information that would be useful in creating more complex decision surfaces is discarded. This observation is irrespective of the metric used to define the centers. This inherent limitation of distance based combiners can be overcome by employing flexible mapping algorithms based on discriminant functions of high polynomial degree. As it will be described in Section 5.4, such alternative combiners in practice could improve class separation of both core and outlier cases mapped to soft labels.

5.3.3 Naïve Bayes Fusion

The Naïve Bayes Fusion (NBF) is a reliable decision fusion technique (Xu, Krzyzak et al. 1992), despite the fact that it assumes independent base classifiers (L1 outcomes). It exploits the L1 hard labels to estimate the conditional class likelihood as:

$$P\left(\mathbf{y} \mid \omega_c\right) = P\left(y_1, ..., y_L \mid \omega_c\right) = \prod_{l=1}^{L} P\left(y_l \mid \omega_c\right) \quad (5.17)$$

By applying the base classifiers to the training set, we obtain a confusion matrix CM_l for each classifier, whose elements correspond to the counting estimates of the probabilities that the input sample from true class ω_c is mapped to class $y_{c'}$:

$$\mathbf{CM}_l = \begin{matrix} \\ \omega_1 \\ \vdots \\ \omega_C \end{matrix} \begin{matrix} \begin{matrix} y_1 & \cdots & y_C \end{matrix} \\ \begin{bmatrix} \hat{P}\left(y_1 \mid \omega_1\right) & \cdots & \hat{P}\left(y_1 \mid \omega_C\right) \\ \vdots & \hat{P}\left(y_c \mid \omega_{c'}\right) & \vdots \\ \hat{P}\left(y_C \mid \omega_1\right) & \cdots & \hat{P}\left(y_C \mid \omega_C\right) \end{bmatrix} \end{matrix} = \begin{matrix} \\ \omega_1 \\ \vdots \\ \omega_C \end{matrix} \begin{matrix} \begin{matrix} y_1 & \cdots & y_C \end{matrix} \\ \begin{bmatrix} cm_{1,1}^l & \cdots & cm_{1,C}^l \\ \vdots & cm_{c,c'}^l & \vdots \\ cm_{C,1}^l & \cdots & cm_{C,C}^l \end{bmatrix} \end{matrix} \quad (5.18)$$

Thus, by summing the c'^{th} column of $\mathbf{CM}$ we can estimate the probability of classifier *l* outputting class c' regardless of the real class c. Using this formulation, we can construct a label matrix $\mathbf{LM}$, which takes the form of an output-class normalized $\mathbf{CM}$ and contains the estimated probabilities of getting label c' when the actual label is c, i.e.

$$lm_{c,c'}^{l} = \hat{P}\left(\omega_c \mid y_i = \omega_{c'}\right) = \frac{cm_{c,c'}^{l}}{cm_{:,c'}^{l}} \tag{5.19}$$

Now using the classifier independence assumption, we can combine (multiply) the entries from all L classifiers and calculate the posterior probability of class c as:

$$\hat{P}_{NBF}\,'\left(\omega_c \mid \mathbf{x}\right) = \hat{P}_{NBF}\,'\left(\omega_c \mid y_l = \omega_{c'}\right) = \prod_{l=1}^{L} \hat{P}\left(\omega_c \mid y_l = \omega_{c'}\right) = \prod_{l=1}^{L} lm_{c,c'}^{l} \tag{5.20}$$

Subsequently, the combiner selects the class c with the maximum probability. One of its problems relates to the fact that a zero element in $\mathbf{CM}$ leads to $\hat{P}_{NBF}\,'\left(\omega_c \mid \mathbf{x}\right) = 0$, which can be avoided using slightly biased estimators of CM. Considering the above scheme in terms of a distance based classifier we can regard the label matrix $\mathbf{LM}$ as the class template and the posterior probability estimates $\hat{P}_{NBF}\,'\left(\omega_c \mid \mathbf{x}\right)$ as the distance to be used for classification.

The Naïve Bayes Fusion algorithm resembles the standard Bayes Classifier (BC), in the sense that it aggregates the (assumed i.i.d.) support values of the $\mathbf{CM}$ into a product, in the same way as the Bayes classifier aggregates the values of each feature $\mathbf{x}$ into posterior class probabilities. The corresponding posterior probabilities are given in the form:

Naïve Bayes Fusion

$$\hat{P}'_{NBF}\left(\omega_c \mid y_{c'}\right) = \frac{\hat{P}\left(\omega_c\right)\hat{P}'\left(y_{c'} \mid \omega_c\right)}{\hat{P}'\left(y_{c'}\right)} \tag{5.21}$$

Bayes Classifier

$$P_{BC}\left(\omega_c \mid \mathbf{x}\right) = \frac{P\left(\omega_c\right)P\left(\mathbf{x} \mid \omega_c\right)}{P\left(\mathbf{x}\right)} \tag{5.22}$$

where $y_{c'} = f\left(\mathbf{x}\right)$. This implies that the difference of the two schemes lays in the fact that the BC expresses the model's likelihood and evidence in terms of feature vectors while NBF expresses them in terms of true vs estimated class frequencies. In this sense, the BC model conveys more information regarding each base classifier's class-specific performance.

The Bayes Classifier (Duda, Hart et al. 2001; Heijden 2004) is also the minimum error rate classifier, provided that the class conditional probability estimates are accurate. The NBF algorithm is Bayes-optimal when the input labels are hard. Based on the error estimates for the Bayes Classifier

$$P_{BC}\left(e\right) = \int_x p\left(e \mid \mathbf{x}\right) p\left(\mathbf{x}\right) d\mathbf{x} = \int_x \left(1 - \max_{c \in C}\left(P\left(\omega_c \mid \mathbf{x}\right)\right)\right) d\mathbf{x} = \sum_{i=1}^{N} \left(1 - \max_{c \in C}\left(P\left(\omega_c \mid \mathbf{x}_i\right)\right)\right) \tag{5.23}$$

we can derive a corresponding upper bound for the Naïve Bayes combiner :

$$P_{NBF}\left(e\right) \le \sum_{c'=1}^{C} p\left(e \mid y_{c'}\right) p\left(y_{c'}\right) = \sum_{c'=1}^{C} \left(1 - \max_{c \in C}\left(\hat{P}\left(\omega_c \mid y_{c'}\right)\right)\right) = \sum_{c'=1}^{C} \left(1 - \max_{c \in C}\left(\prod_{l=1}^{L} lm_{c,c'}^{l}\right)\right) \tag{5.24}$$

utilizing the NBF's features, i.e. the labels of base classifiers. The above error rate estimator can be calculated from the trained classifiers, without running the combiner. Behavior-Knowledge-Space Combiner

The concept of utilizing the classifiers' soft labels as features is quite often used in many combiner schemes albeit not always directly acknowledged (Jacobs 1995). Considered as a Bayesian learning problem in a $L \times C$ feature space, the aim of classifier fusion is to estimate the soft label of a sample $\mathbf{x}$ using the joint class-conditional likelihood of the L classifiers.

$$P(\omega_c \mid \mathbf{x}) \propto p(DP(\mathbf{x}) \mid \omega_c) P(\omega_c), c = 1, ..., C$$

When the features used are the hard class-labels, then the above scheme is equivalent to a multinomial combination method called Behavior Knowledge Space (BKS). BKS is a trainable ensemble method based on counting label occurrence frequencies. During the training phase, the algorithm builds a $3 \times L^C$ lookup table in which there is a column for each class-label combination, a count of occurrences of the combination and the resulting (majority class) label respectively. While this algorithm assigns fused labels for each possible set of L1 labels based on the majority rule, it is considered able provide a more detailed mapping than standard majority voting since it can account for multiple different label combinations corresponding to a single majority result. A known drawback of this scheme is that it is susceptible to ties, in which case part of the outcome patterns is assigned randomly.

For each combination of indicated labels $\{y_l\}, l = 1, ..., L$ with the combiner (column), the corresponding elements of rows 1, 2 and 3 of the BKS lookup table are defined as follows:

$R_1 = \{y_l\}, l = 1, ..., L$ (indicating combination with the full set of labels *l*)

$R_2(c) = \sum_{l=1}^{L} ind(y_l = \omega_c), c = 1, ..., C$ (computing the relative frequencies of assigned labels for each nomination)

$R_3 = \arg\max_c (R_2(c))$ (revealing the estimated label using majority voting).

As an example, assume that there are $L = 2$ classifiers, $C = 2$ classes and N samples. This setting results to $N' = L^C = 2^2 = 4$ cases (of label combinations) in a $L = 2$-dimensional label space (BKS space) as shown in Table 5-1.

indicated class label combinations	$\{1,1\}$	$\{1,2\}$	$\{2,1\}$	$\{2,2\}$
Class frequencies for each label combination	$N_{\omega_1}(\{1,1\})$, $N_{\omega_2}(\{1,1\})$	$N_{\omega_1}(\{1,2\})$, $N_{\omega_2}(\{1,2\})$	$N_{\omega_1}(\{2,1\})$, $N_{\omega_2}(\{2,1\})$	$N_{\omega_1}(\{2,2\})$, $N_{\omega_2}(\{2,2\})$
L2 majority label (example data)	1	1	2	1

Table 5-1 Example of BKS lookup table for 2 classes and 2 base classifiers

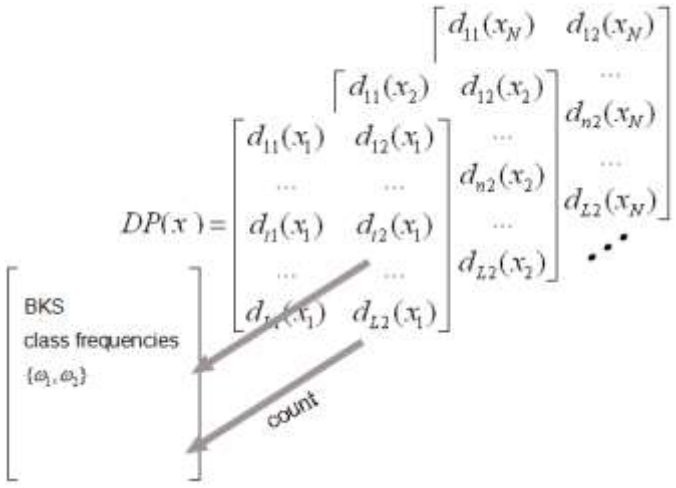


Figure 5.2 Calculation of class frequencies from DP

BKS is analogous to a modified "*voting nearest neighbor* classifier" (Fukunaga 1990; Huang and Suen 1995) which is a variation of the k nearest neighbor (k-NN) classifier with class frequencies used instead of sample distances to neighboring points. As shown in Figure 5.2, the class support values provided by the DP of each sample are transformed to hard labels and counting estimates of the class posterior for each bin are stored in the second row of the BKS lookup table. The analogy of BKS to a voting k-NN classifier is evident if we parameterize the k-NN classifier with a variable number of neighbors consisting of the samples producing the same combination of soft labels in the L-dimensional feature space of the combiner, and we use voting to select the local majority class of each label combination set. The above mapping holds true irrespective of the number of input cases. According to (Dasarathy 1991; Bishop 1995) the voting 1-NN classifier produces identical results with the standard (volumetric) k-NN classifier for $C = 2$ class problems.

Based on the above analogy, we can derive statistical error for the BKS combiner based on the voting 1-NN classifier and assume $\hat{P}'_{BKS}(e \mid \mathbf{x}) = \hat{P}_{1-NN}(e \mid \mathbf{x})$. For simplicity, we continue using $\mathbf{x}$ as the conditioned random variable of the new feature space, acknowledging that the analysis actually deals with the hard label space of the BKS combiner. More specifically as shown in (Fukunaga 1990) for a sample $\mathbf{x}$ and its closest neighbor $\mathbf{x}_{NN}$ the voting 1-NN classifier's error is

$$\begin{aligned}
&\hat{P}_{1-NN}(e \mid \mathbf{x}, \mathbf{x}_{NN}) \\
&= P\left[\{\mathbf{x} \in \omega_1 \text{ and } \mathbf{x}_{NN} \in \omega_2\} \text{ or } \{\mathbf{x} \in \omega_2 \text{ and } \mathbf{x}_{NN} \in \omega_1\}\right] \\
&= P\left[\{\mathbf{x} \in \omega_1 \text{ and } \mathbf{x}_{NN} \in \omega_2\}\right] + P\left[\{\mathbf{x} \in \omega_2 \text{ and } \mathbf{x}_{NN} \in \omega_1\}\right] \\
&= P(\omega_1 \mid \mathbf{x}) P(\omega_2 \mid \mathbf{x}_{NN}) + P(\omega_2 \mid \mathbf{x}) P(\omega_1 \mid \mathbf{x}_{NN})
\end{aligned} \tag{5.25}$$

Assuming that $\mathbf{x}$ and $\mathbf{x}_{NN}$ are close, or in the case of BKS identical (all in the same label combination column), the corresponding posterior class membership probabilities can be considered equal $P(\omega_c \mid \mathbf{x}) = P(\omega_c \mid \mathbf{x}_{NN}) = P_c(\mathbf{x}),\ \textit{for}\ c = \{1,2\},$ which results to

$$\hat{P}_{1-NN}(e \mid \mathbf{x}, \mathbf{x}_{NN}) = P_1(\mathbf{x}) P_2(\mathbf{x}) + P_2(\mathbf{x}) P_1(\mathbf{x}) = 2P_1(\mathbf{x}) P_2(\mathbf{x}) \tag{5.26}$$

Additionally, in the case of BKS we can directly estimate $P(\omega_c \mid \mathbf{x})$ as the counting frequencies of each class within each bin $P(\omega_c \mid b)$.

$$P(\omega_1, b) = P_1(b) = \frac{N_{\omega_1,b}}{N_{\omega_1,b} + N_{\omega_2,b}} \tag{5.27}$$

Therefore, it is suitable to reformulate the above error estimate in terms or bins rather than samples

$$\hat{P}_{BKS}(e \mid b) = P_1(b)P_2(b) + P_2(b)P_1(b) = 2P_1(b)P_2(b) \tag{5.28}$$

Averaging over all BKS bins we get the total average BKS error

$$\hat{P}_{BKS}(e) = \frac{1}{Nbins} \sum_{b=1}^{Nbins} \hat{P}_{BKS}(e \mid b) \tag{5.29}$$

The two algorithms should not be considered fully identical since BKS requires a variable number of nearest neighbors (samples producing the same label set) whereas 1-NN assumes a fixed $k = 1$ neighbor. Their partial analogy is adequate for using equation (25) as a basis to derive the above error bound for BKS.

Additionally, considering the above scheme in terms of a distance based classifier; we can regard the label set of each column of the BKS lookup table as the class template (in the L-dimensional label space) and the counting estimates for each class as the distance to be used for classification. Using the max operator we compare each input $\mathbf{x}$ with all class templates that form the columns of the BKS matrix. This reflects the fact that the BKS can be seen as a minimum distance classifier operating in a new space defined by the labels set of the combiner, using the majority voting frequencies as class support values and the maximum operator as the distance metric.

5.4 Discriminant-function based combiners

In a previous study, (Dimou, Manikis et al. 2006) demonstrated the effectiveness of certain generic classifiers in learning decision profiles and then classifying unknown DPs as hyper-classifiers. This approach defines a mapping function for the decision profiles that, similar to the discriminant function, enables their direct classification in the prescribed classes. Having available the decision profiles for all training samples we are able to train generic classifier algorithms to map each sample to a class and overcome the limitations of distance based combiners. The difference of the distance-based vs the discriminant-function approach at the DP space (Level2 combiner) is graphically demonstrated in Figure 5.3. For an unknown DP, the decision in the former scheme is based on the distance metric from the class DTs, whereas in the second scheme the decision is obtained by means of the outcome of the discriminant function. Based on this consideration and the known analogy between distance and discriminant-based classifiers, we can build combiners for classifier ensembles operating on the principle of discriminant functions, so as to overcome certain limitations of distance-based combiners.

Some of the classifiers that can be used from this class of algorithms include linear and quadratic discriminant analysis, Bayesian classifiers, artificial neural networks and support vector machines (Tumer and Ghosh 1996; Wang and Zhang 2001; Kuncheva 2008). The corresponding combiners utilize the same discriminant function but in the DP space.

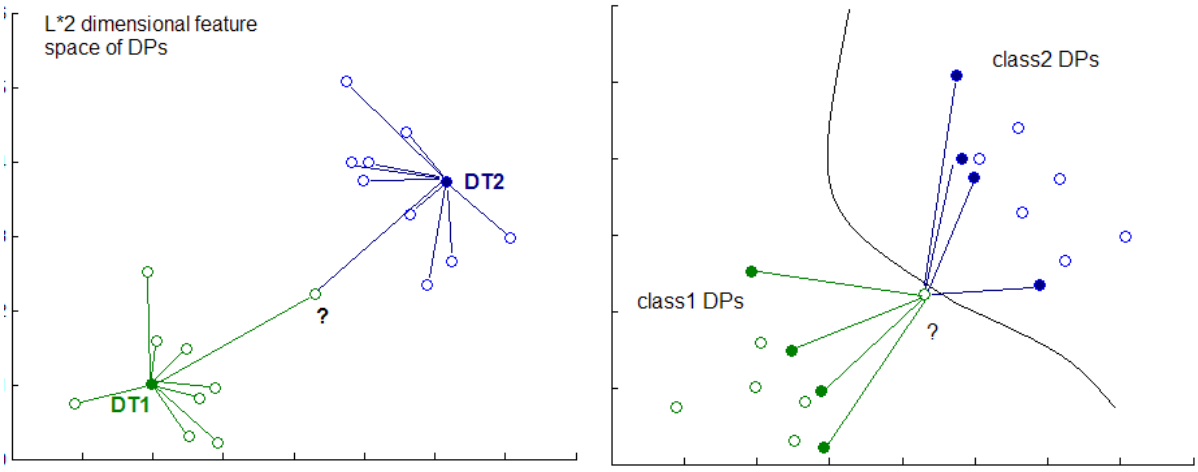


Figure 5.3 Principle of operation of the distance based combiners where DTs are the class centers (left) vs the support vector machine combiner where each sample's DP

As a work case in this study, we analyze from this class of combiners a support vector machine, which combines the arbitrary mapping capacity of artificial neural networks with a firm statistical background stemming from statistical learning theory (SLT) (Vapnik 1995). A neural network classifier could alternatively be employed as a nonlinear discriminant combiner of comparable mapping capability to SVMs.

The description of the SVM's theory and principle of operation is beyond the aim of this work. For a primer and overview material for the SVMs field refer to (Cristianini and Shawe-Taylor 2000). For the purposes of our analysis it will suffice to outline that the SVMs are based on the thresholded decision function that incorporates nonlinearity in a mapping function $\varphi(\mathbf{x})$ that is defined indirectly via the concept of kernels. Thus

$$y(\mathbf{x}) = \text{sign}\left[\mathbf{w}^T \varphi(\mathbf{x}) + b\right] \quad (5.30)$$

where y is the binary class indication encoded as -1 versus +1, w is a weighting vector, b is the bias term, The mapping $\varphi: \mathbb{R}^q \to \mathbb{R}^r$ maps the q-dimensional input space into a high r-dimensional feature space. By solving the Lagrangian, the problem can be formulated in the dual space for the classifier

$$y(\mathbf{x}) = \text{sign}\left[\sum_{n=1}^{N} \alpha_n y_n K(\mathbf{x}, \mathbf{x}_n) + b\right] \quad (5.31)$$

where α_1, …, α_N are the Lagrange multipliers. We implicitly work in the feature space by applying a positive definite kernel

$$K(\mathbf{x}_i, \mathbf{x}_j) = \varphi(\mathbf{x}_i)^T \varphi(\mathbf{x}_j) \quad (5.32)$$

The concept of kernels provides a very powerful way to design and parameterize an SVMs classifier for specific problem domains. A typical option if to use a generic RBF kernel

$$k_{RBF}(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma^2}\right), \sigma \in \Re \quad (5.33)$$

We can reformulate the above kernel in order to operate on a DP feature space as a combiner as follows

$$k_{RBF}(d_i, d_j) = \exp\left(\frac{\left\|d_i - d_j\right\|^2}{\sigma^2}\right),\ \sigma \in \Re \qquad (5.34)$$

where d_i is the decision profile for sample *i* (rearranged to form a 1-dimensional feature vector) and σ is the radial basis function's regularization parameter. This formulation is visualized in Figure 5.3, which indicates a possible distribution of DP samples, the corresponding nonlinear discriminant function and an unknown sample to be classified. Recall that this concept is broader than SVMs and applies to all discriminant based classification algorithms.

One of the first bounds on SVM error was proposed in (Vapnik 1995) and is known as the "radius-margin bound". It limits the test error of an SVM below a threshold expressed in terms of the radius of minimum enclosing hyper sphere B and the geometrical margin of separation of the samples in the feature space $1/R$.

$$P\theta(f) \le D^2 R^2 \qquad (5.35)$$

In (Bartlett and Mendelson 2002) a margin-based estimate of the misclassification probability of SVMs is calculated.

For a margin cost function $c: \mathbb{R} \to [0,1]$

$$c(a) = \begin{cases} 1 & \text{if } \alpha \le 0 \\ 1-\alpha/\gamma & \text{if } 0<\alpha \le \gamma \\ 0 & \text{if } \alpha>\gamma \end{cases}$$

and assuming that the dataset's samples are chosen independently according to some probability distribution P on $X \times \{\pm 1\}$, then with probability >1-δ every function f of the form

$$f(\mathbf{x}) = \sum_{i=1}^{n} a_i k(\mathbf{X}_i, \mathbf{x}) \quad (5.36)$$

with

$$\sum_{i,j} \alpha_i \alpha_j k(\mathbf{X}_i, \mathbf{X}_j) \le B^2 \qquad (5.37)$$

satisfies

$$P\left(Y \cdot f(\mathbf{x}) \le 0\right) \le \hat{E}_n \phi\left(Y \cdot f(\mathbf{x})\right) + \frac{4B}{\gamma n}\sqrt{\sum_{i=1}^{n} k(\mathbf{X}_i, \mathbf{X}_j)} + \left(\frac{8}{\gamma}+1\right)\sqrt{\frac{\ln(4/\delta)}{2n}} \qquad (5.38)$$

The above result holds for any distribution $P(\mathbf{X}, \mathbf{Y})$ of $(\mathbf{x}_1, y_1), .., (\mathbf{x}_N, y_N)$.

Also a slightly different bound has been proposed by (Blanchard, Bousquet et al.) and is given by the equation

$$P\theta(f) \le P_N\left[l(f) \ge 1\right] + \frac{4R}{\sqrt{N}}\sqrt{\frac{1}{N}\sum_{i=1}^{N} k(\mathbf{x}_i, \mathbf{x}_i) + 9\sqrt{\frac{\mathbf{x}}{2N}}} \qquad (5.39)$$

This bound holds for any discriminant classification function B(R), not only for SVMs, hence the lack of dependence on the parameter γ.

One key benefit of applying an SVM to classifier fusion is that it can take advantage of the feature set's characteristics and invariances through the introduction of specifically designed kernels (Dimou and Zervakis 2008). Such kernels can be designed to exploit information on the shape of the output distributions from the primary classifiers. Composite kernels (kernels from simpler kernels) can be adapted to learn separately subset of each class DPs in an effort to find more separable representations. Additionally, Zhou (Zhou and Chellappa 2006) has proposed kernels that work on diversity metrics for data fusion. This idea can be extended to design diversity kernels for classifier fusion.

An additional benefit is that SVMs do not rely on statistical assumptions. While probabilistic product, naïve Bayes, behavior knowledge space and decision templates combiners require the primary classifiers' outcomes (or for discrete outcomes, their underlying distributions) to be valid probabilities, SVM based combiners can operate on arbitrary (yet normalized) support values. This enlarges the pool of candidate primary classifiers thus allowing better ensemble designs with respect to diversity. To that respect, the DP is known to have non-uniformly distributed features, which restricts the selection of primary classifiers as combiners at the L2 classification level. Indeed, due to the nature of the classification task, most L1 outputs are concentrated near the edge values $\{0,1\}$ of the allowable output label range and thus invalidate Gaussian assumption based hyper-classifiers. The DP itself is composed of such values as depicted in Figure 5.4, for an exemplary 2-classifier ensemble. In fact, the skewed distribution of the DP values can improve class separability in the DP feature space as in the top part of Figure 5.4. However, for low performing base classifiers shown at the bottom part of Figure 5.4, there is increased class overlap in classifier labels that requires increased flexibility in the operation of the combiner. This is the case in "difficult" classification problems where there is a larger margin of improvement using classifier fusion. As demonstrated in the following section, SVMs provide the capability to create flexible discriminant-based separating hyperplanes that handle this problem effectively.

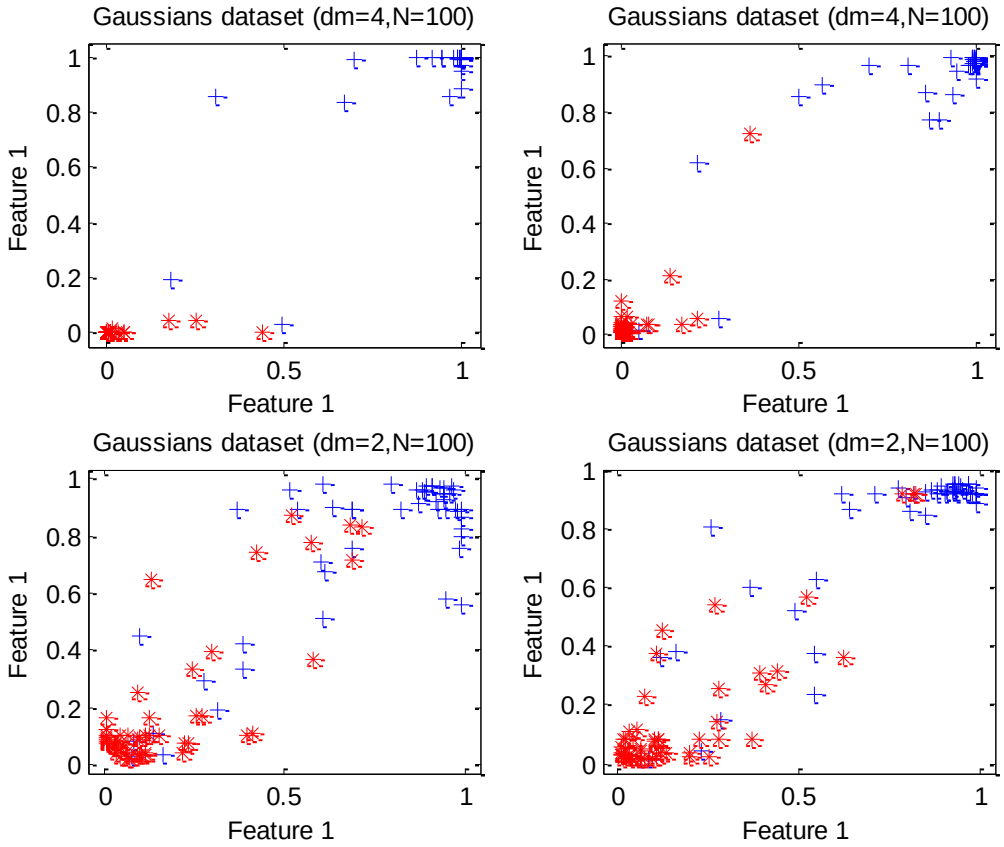


Figure 5.4 Indicative pairs of feature spaces for combiners using $L = 2$ base classifiers. In high separability (easy) datasets, or highly performing base classifiers (top), the soft labels form compact clusters with occasional outliers. In low separability (difficult) datasets, or poorly performing base classifiers (bottom), the soft labels create complex clusters and multiple outlier cases requiring more flexible combiners.

5.5 Experimental Validation

5.5.1 Testing Framework

For the experimental evaluation of the above theoretical concepts under varying test scenarios, we resorted to artificial datasets with configurable class-separability, distribution and number of samples. The characteristics of the datasets employed in this study are shown in table 2.

	Artificial					
	Lith1	**Lith2**	**Lith3**	**Ban1**	**Ban1**	**Ban3**
# features	2	2	2	2	2	2
# samples	200	200	200	200	200	200
variance	1	2	4	1	2	4
Dmean ($\mu_1 - \mu_2$)	1	1	1	1	1	1

% of positives	50%	50%	50%	50%	50%	50%

Table 5-2 Dimensionality and characteristics of the synthetic benchmark datasets

Datasets L1 to L3 are generated as “Lithuanian” clusters defined by Raudys (Raudys and Jain 1991) with increasing difficulty (variance). Datasets B1 to B3 consist of two typical ”banana shaped” classes with increasing class overlap (variance). The benchmark dataset “Phoneme” (Newman, Hettich et al. 1998) is also used for more realistic comparison of results with existing literature.

The datasets used for the performance evaluation of algorithms are partitioned according to a two-level hierarchical scheme. This allows the statistical evaluation of both the classification and fusion modules using cross validation, as to reduce variance on the performance estimates. More specifically, the initial dataset is partitioned into a 90% training set and a 10% test set as illustrated in Figure 5.5. This is done iteratively with replacement 10 times. Each resulting training set is further partitioned into two equal training sets to be utilized for the training of the primary classifiers (L1) and trainable combiners (L2). For the base classifiers’ evaluation, the first train subset (45%) is used for training and the common test subset (10%) for testing.

For the combiners’ metrics, the second training subset (45%) is used on trainable combiners and the common testing subset (10%) is used for testing. This scheme has the drawback of utilizing a smaller part of the original dataset for training (hence accuracies may be underestimated) however it ensures objective evaluation of both classifiers and combiners.

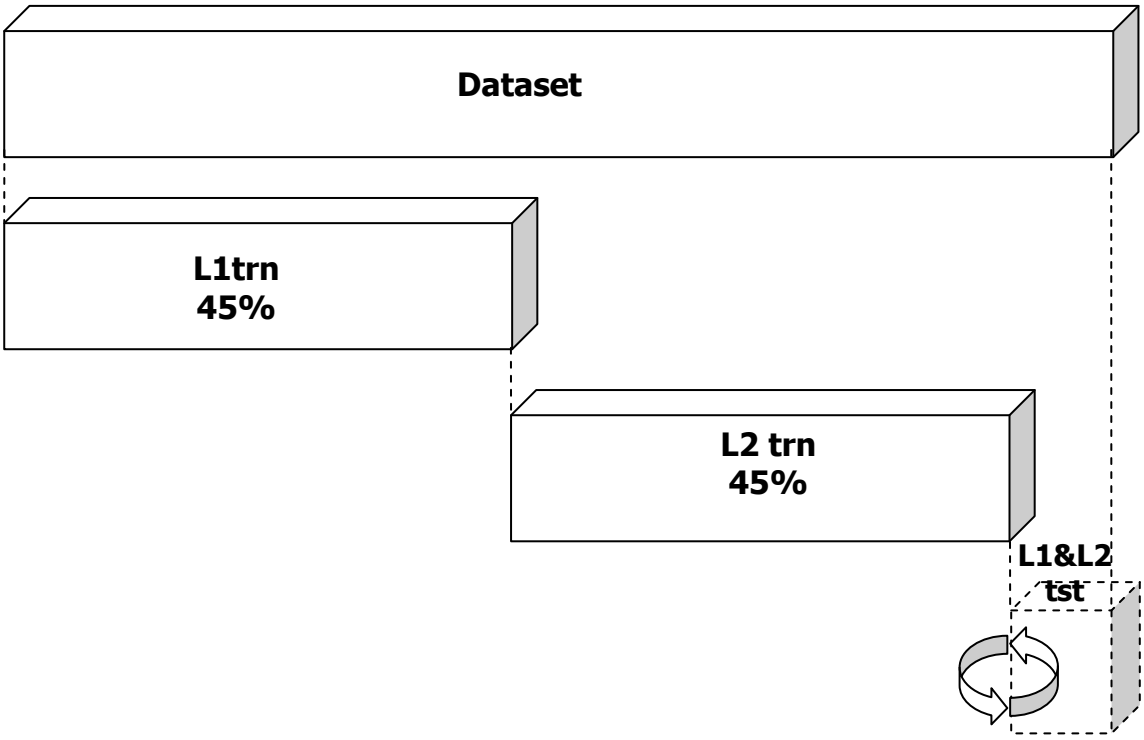


Figure 5.5 Two-level hierarchical partitioning of the dataset for training and testing both base classifiers (L1) and combiners (L2)

In order to compare the properties and performance of the two classes of combiners, we implemented the 2-stage classification scheme shown in Figure 5.6. The system is structured as follows. At the first level (L1) a pool of base classifiers is trained using the benchmark datasets. The pool of base classifiers includes QDC and SVM classifiers with varying parameterizations in order to create a set of both low and high performance algorithms and examine how the fusion model might affect performance, especially in small and diverse sample-size scenarios.

For each type of L1 classifier, three sets of parameters are chosen to provide a balanced representation of a standard parameter range. For the QDC base classifiers, three regularization parameters $R \geq 0, S \leq 1$ were used to compute the covariance matrix for the results presented in Figures 5.7-5.13. Three SVM base classifiers are also presented in the same figures with their regularization parameters $C \geq 0, s \leq 1$.

The number of samples in each dataset is at the low end compared to typical classification problems. For such dataset sizes an additional benefit of classifier fusion is that it allows leveraging multiple class mappings based on limited information. The ensemble of support vector machines with radial basis function (RBF) kernel of different parameterization provides different error levels, which derive from the varying degree of fitness to training data that the regularization parameters allow for. The base classifiers are not explicitly optimized, since our aim is to provide a wide range of L1 decisions, rather than highly customized and skewed outcomes. In a similar experimental evaluation, (Kuncheva, Bezdek et al. 2001) used quadratic discriminant (QDC) classifiers that favor compact feature clouds, in order to create the same effect of varying performance.

At the fusion level (L2), the combiner methods are applied on the L1 soft labels. As shown in Figure 5.6, the two stage classification process consists of first obtaining the DPs and performance statistics of the base classifiers for all dataset partitionings. At the second level, the DPs are applied to the pool of fixed, distance, and discriminant-based combiners and additional performance statistics are gathered.

The key aspects that are examined using the experimental setup include the accuracy of combiners compared to classifiers, the performance differences between distance and discriminant combiners, the verification of the DT-NMC equivalence and BC-NBC and BKS-1NN statistical analogies and the validity and usability of the derived error bounds. Collateral aspects considered are the extraction of usage prerequisites and guidelines for each algorithm. Theoretically it is expected that ensemble methods can improve cross-validated accuracy especially in the difficult datasets and that DTs produce exactly the same results as the Euclidean NMC used as a generic combiner. In general ensemble methods are expected to show increased robustness to degrading input features compared to base classifiers.

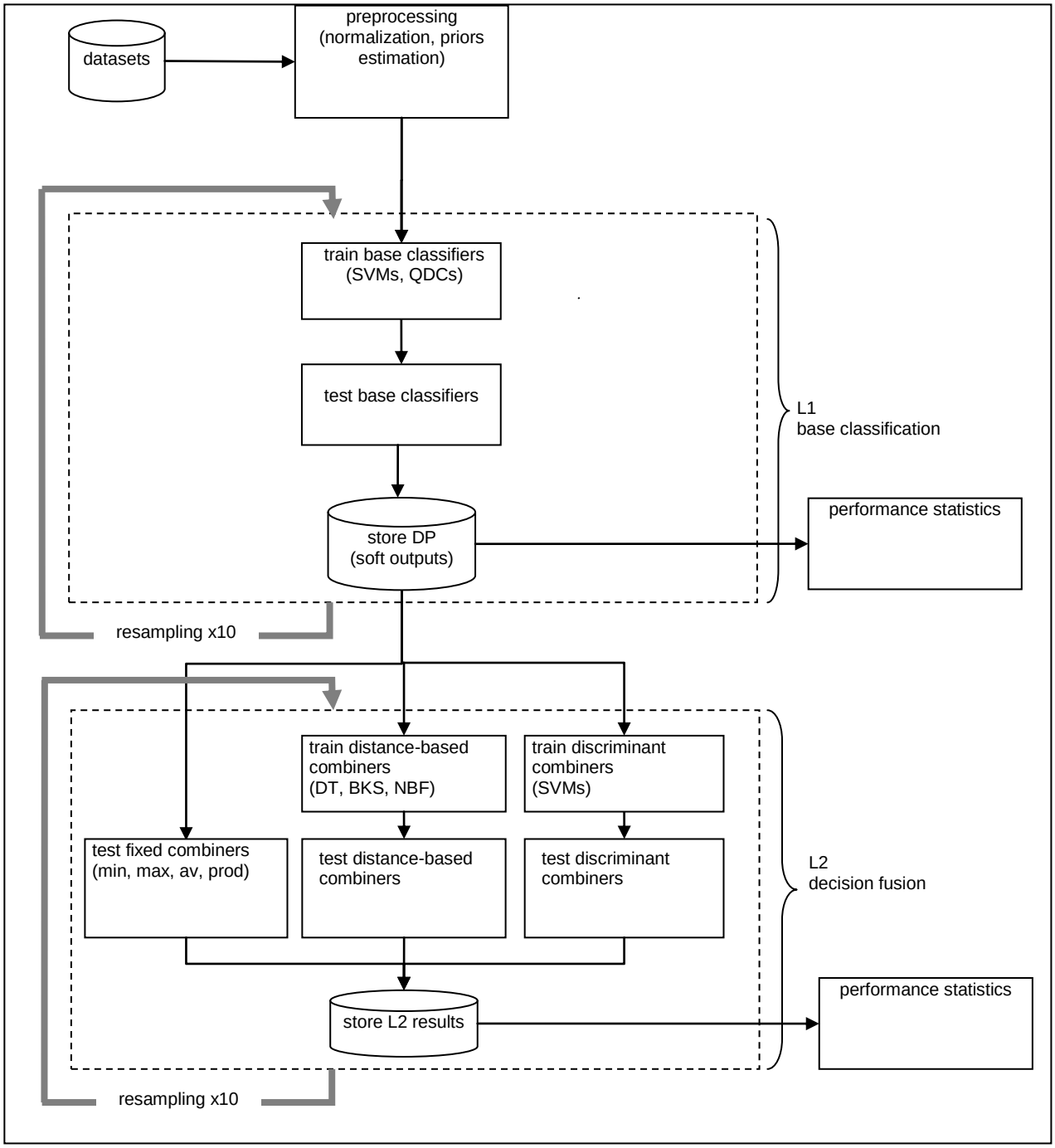


Figure 5.6 Block diagram of the 2-stage classification scheme

5.5.2 Results

In this section we present results from various tests performed starting from general error bound estimates to more detailed accuracy boxplots and then to statistical discrimination tests (t-tests). With respect to the experimental objectives, a number of general observations can be summarized from the detailed presentation in the following subsections.

First, the derived upper error-bounds have proven to be comparative indicators of the worse-case performance. With the exception of the NBF/BC, the calculated error bounds are also tightly coupled with the actual accuracies achieved for each dataset.

From the accuracy viewpoint, the classification of artificial datasets appeared to benefit from the use of ensemble methods, while the effects vary significantly with the method itself. In particular, the use of combiners appears to improve cross-validated accuracy, especially for the difficult datasets. In general, combiners show increased robustness depending on the quality of input features, as compared to base classifiers.

Within the group of distance-based combiners, the obtained accuracy values of Decision Templates exactly match the ones of the Euclidean Nearest Mean, providing additional evidence for their functional equivalence. The statistical analogies of BC with NBC algorithms cannot be verified at the experimental level, since they operate on different feature spaces. The same holds for BKS and KNN algorithms; although using similar feature sets, BKS uses the set of samples producing identical L1 labels to infer class membership while KNN always employs a wider set of nearest neighbors for the same task.

In terms of statistical discrimination, the single sided t-tests on the accuracies prove the hypothesis that discriminant combiners outperform distance-based combiners for the experimental scenarios tested. Distance-based combiners produce linear boundaries, so that problems with skewed outcome distributions are not handled efficiently, since they require nonlinear boundaries implemented through the use of nonlinear Discriminant functions.

Based on these observations and associated literature, we can extract some general usage guidelines of these fusion algorithms. There exist algorithms, such as Majority voting and Decision Templates, which benefit from consensus. Most other combiners exploit the inherent diversity of more challenging datasets in order to improve performance on outlier cases. Especially combiners counting frequency estimates (Naïve Bayesian Fusion, BKS) require a large number of samples, preferably evenly distributed across classes, to ensure non-zero frequencies and adequately small variance within each bin. Despite this shortcoming, the BKS combiner along with majority voting are the best as non-parametric fusion solutions for problems in which the pdfs of the outcomes of base classifiers cannot either be estimated or fall under the implicit Gaussian assumption.

On the contrary, in scenarios where few samples are available the use of these combiners may be problematic. SVMs can be employed in such problems, in conjunction with a larger pool of base classifiers, since then can effectively handle asymmetric datasets of high dimensionality.

The product and probabilistic-product combiners should be avoided in cases with diminishing soft support values $\hat{P}_l\left(\omega_c \mid \mathbf{x}\right) \to 0$ since a single classifier's zero soft-label would diminish the class posteriors over all base classifiers. A similar argument holds for the Naïve Bayes fusion algorithm, where low counts in certain true-assigned class combinations in the confusion matrix are likely to invalidate the corresponding outcomes. Additionally, in high noise datasets where classification has to be based primarily on class priors, the average and sum combiners are both simple and effective in producing robust fused labels.

More detailed use case scenarios and references can be extracted from the comparative table in Appendix III, which summarizes the properties of combiners. The following sections present the experimental results and conclusions in detail.

5.5.2.1 Comparisons of Error bounds

In order to evaluate the performance range of each trainable combiner the error bounds are calculated using the actual data from each dataset.

$P(e) <$	L1	L2	L3	S1	S2	S3
DT (Eucl.NMC)	0,180	0,390	0,430	0,080	0,240	0,290
Naïve Bayes Fusion (BC)	0,274	0,341	0,453	0,270	0,330	0,353
BKS (1NN)	0,382	0,436	0,463	0,387	0,348	0,410
SVM combiner	0,386	0,390	0,475	0,388	0,390	0,328

Table 5-3 Bounds on the mean error for each dataset

The resulting upper bounds on the mean error are shown in Table 5-3 and in Figures 5.7-5.13 as black boxes. For the DT (via NMC), NBF, BKS (via 1NN) and SVM combiners these accuracy estimates are consistent with simulation results and provide a usable indication for each combiner's performance range. It should be noted that since the above bounds are calculated based on specific data instances, they are actually mean estimates of the error bounds over all dataset partitionings.

Based on the results of Table 5-3, the DT and SVM combiners appear to have the narrower error margins followed by BKS. Moreover SVM combiner results in narrower mean error bounds than the DT indicating better performance.

To the authors' best knowledge, there is limited work associated with the error bounds of trainable combiners. While acknowledging the indicative nature of the presented bounds, we aim at using them to provide an additional insight for the expected performance of each fusion algorithm and the variation of results depending on the increasing complexity of problems.

5.5.2.2 Comparisons based on true Accuracies and CIs

Running the above fusion schemes using 10-fold cross validation, we obtain the accuracies (red line) and 95% confidence intervals (blue boxes) and outliers ("+") shown in Figures 5.7-5.13. The QDC and SVM with various parameters are used as base classifiers. The classifier fusion is performed through several schemes including fixed (product, mean, median, max, min and voting) as well as trainable combiners (DT, Naïve Bayes Combiner, BKS, NMC, KNN and SVM). It should be clear that the Naïve Bayes Combiner (NBC) is a generic Naïve Bayes classifier trained on the base classifiers' outputs (DP). The Naïve Bayesian Fusion (NBF) is a fusion algorithm defined in (Xu, Krzyzak et al. 1992) which uses confusion matrix probabilities for the classifier pool to estimate the fused class membership probabilities. These combiners cover a wide range of fusion philosophies and are used as means of identifying structural differences and exploring their potential. In the resulting figures a number of trends can be identified. First, the use of classifier fusion algorithms achieves equal or better performance than individual classifiers. Our study verifies earlier results claiming that even simple non-trainable combiners can be effective (Kuncheva 2002).

Furthermore, trainable combiners and especially discriminant classifiers perform superior to simple combiners, albeit with increasing variance with the difficulty of the dataset. More specifically, in the first scenario (Figures 5.7-5.9) the trainable classifier accuracies reach a range of 1.0 in Figure 5.7, 0.7 in Figure 5.8 and 0.6 in Figure 5.9. Similarly, in the second scenario (Figures 5.10-5.13) the trainable classifiers achieve accuracies in the range of 1.0 in Figure 5.10, 0.9 in Figure 5.11 and 0.8 in Figure 5.12.

The high-variance parameterization used to produce the most difficult datasets on both scenarios (Figures 5.9 and 5.12) possibly creates unrealistic clusters with random overlap, in which case the combination of nearly random outlier soft-labels at the fusion level is inherently problematic. Therefore, trends in combiners' performance in Figures 5.9 and 5.12 are difficult to distinguish.

A notable exception is evident in the case of the BKS combiner, which degrades steeply under worsening classification conditions. This is evident both in Figures 5.8-5.10 and in Figures 5.10-5.12, where BKS consistently shows the lowest mean accuracy within the trainable combiners group.

Additionally, the equivalence of DTs with NMC classifier combiner is verified experimentally as shown by their identical accuracy boxplots both in the first (Figures 5.7-5.9) and the second scenario (Figures 5.10-5.12). Examining the complete set of soft labels provided the DT and NMC, the authors were able to verify that the two methods are not only equivalent based on performance but also produce identical outcomes for all samples.

With respect to the "Phoneme" dataset in Figure 5.13, While a direct comparison with (Kuncheva, Bezdek et al. 2001) is not meaningful due to different feature sets, we were able to verify that the trend of increased trainable classifier performance identified in the aforementioned work carries over to our results. Additionally SVM and KNN combiners achieve top performance.

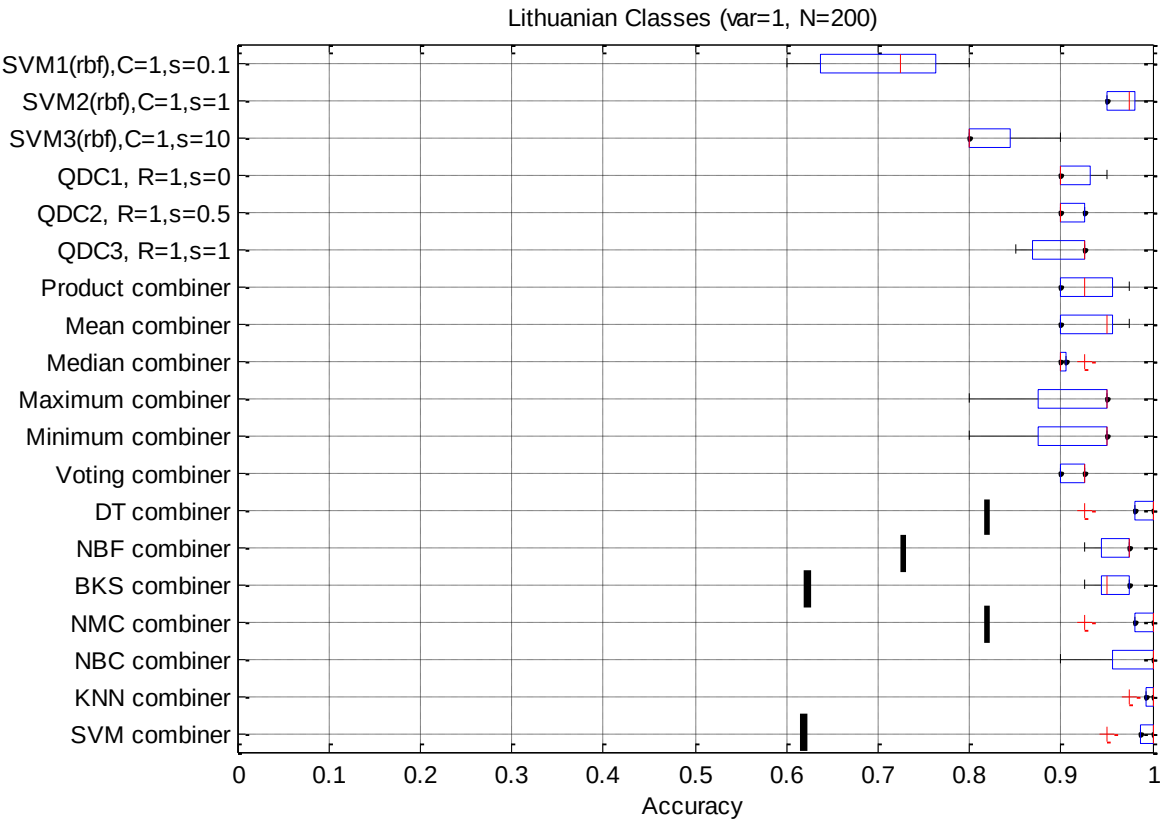


Figure 5.7 Accuracy estimates for the "Lith1" artificial dataset. High separability is achieved by all fusion methods, even under varying performance of the base classifiers.

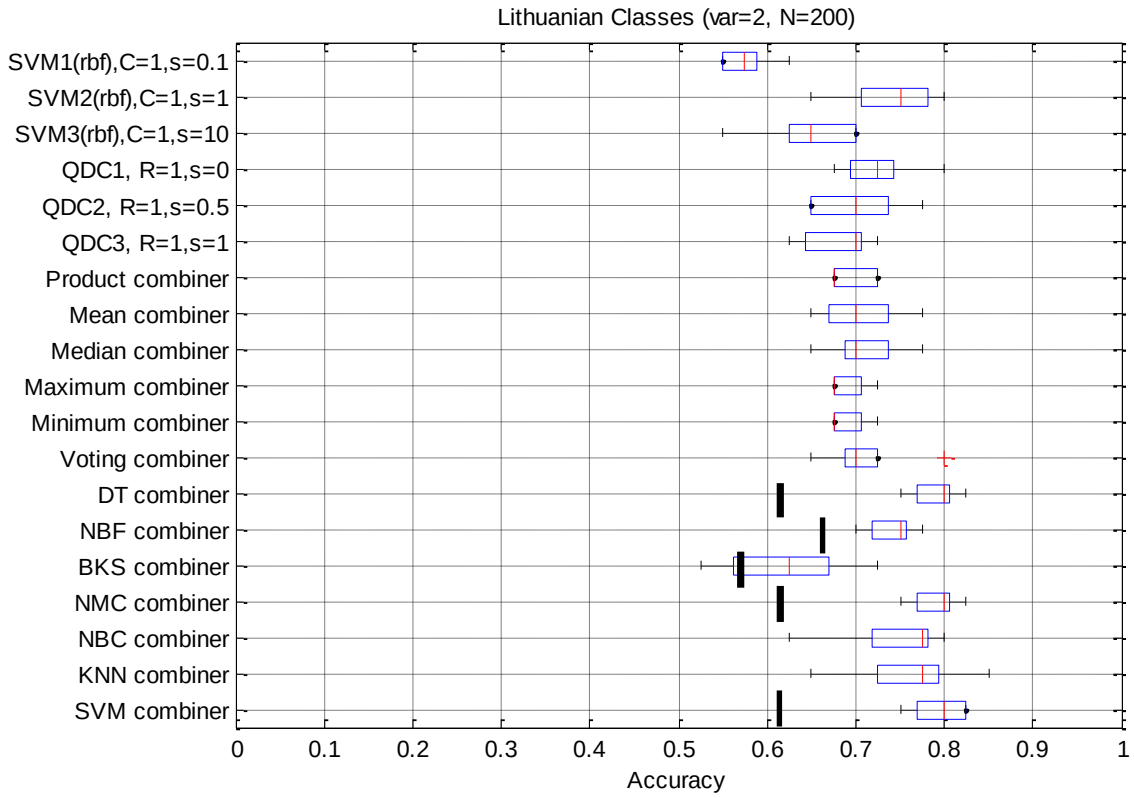


Figure 5.8 Accuracy estimates for the "Lith2" artificial dataset. As the variance increases, fusion methods degrade, while the performance of base classifiers varies depending on their parameterization.

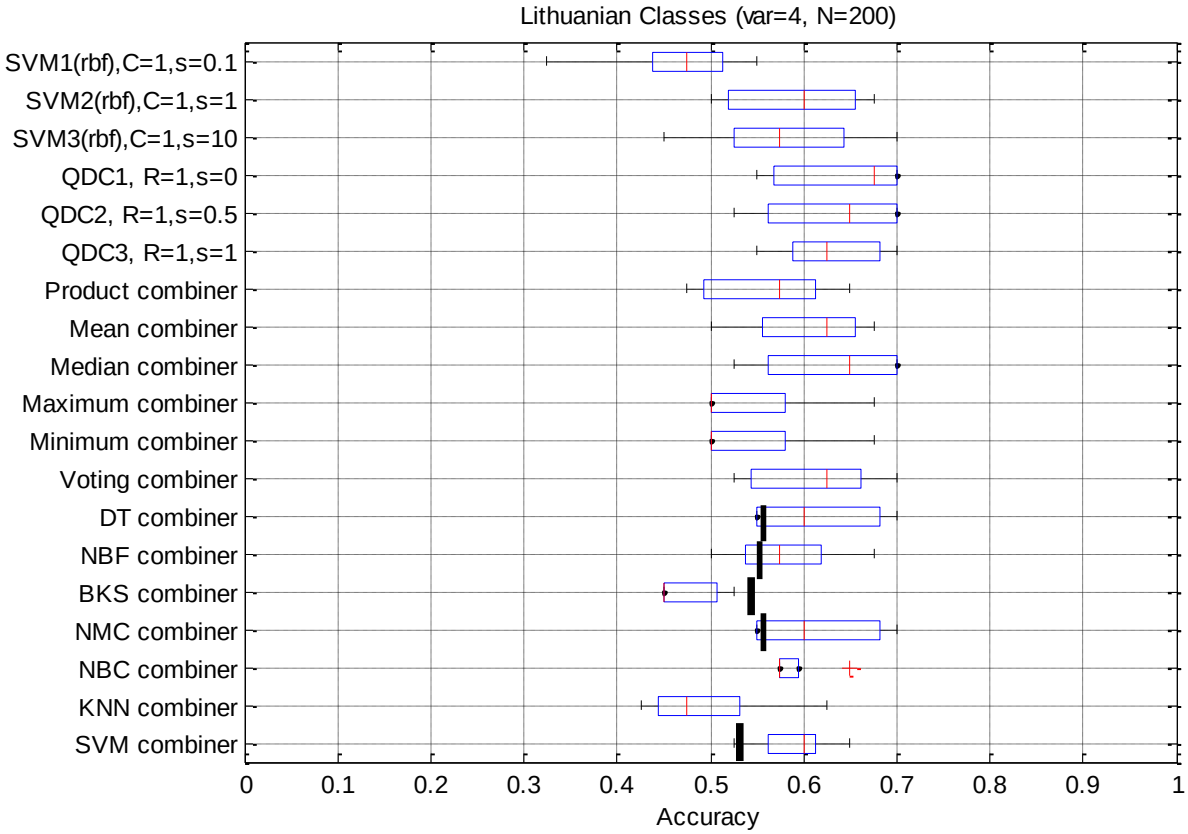


Figure 5.9 Accuracy estimates for the "Lith3" artificial dataset. For even higher feature variance values compared to Figures 5.7 and 5.8, fusion methods appear less effective in improving the classification accuracy.

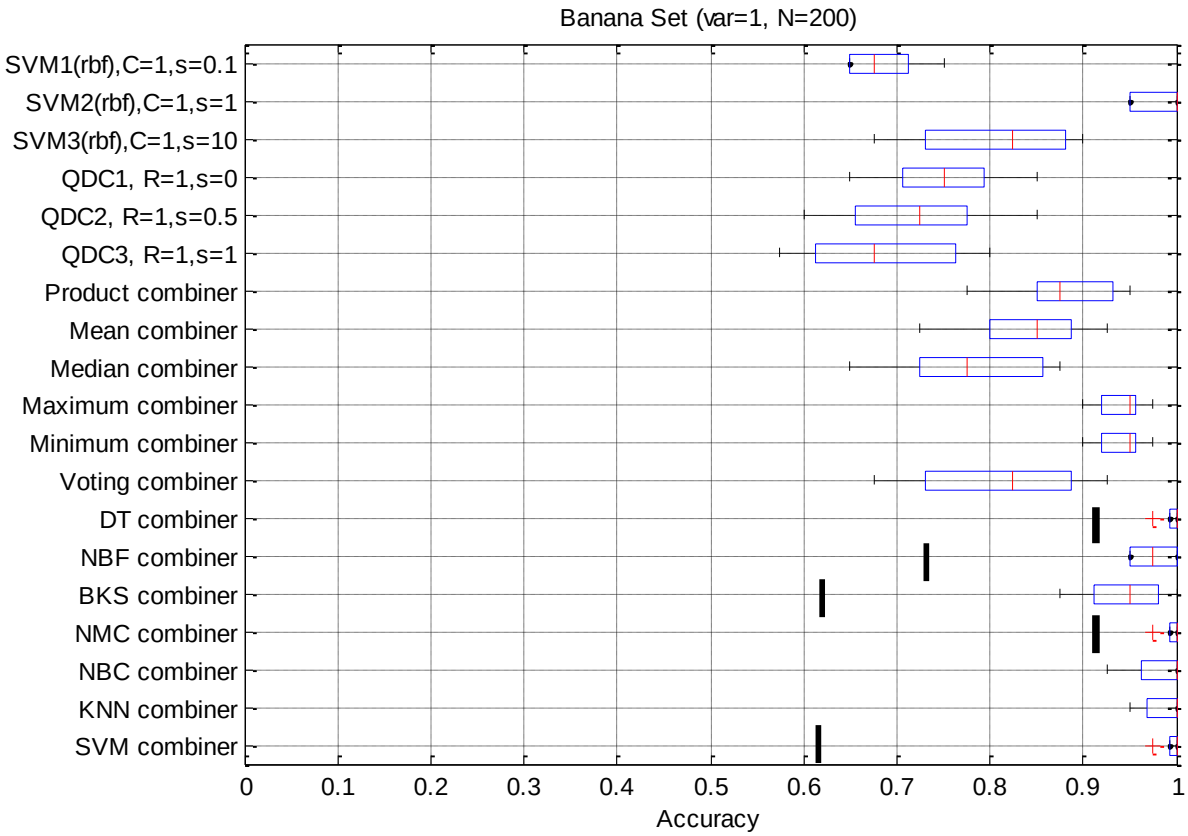


Figure 5.10 Accuracy estimates for the "Ban1" artificial dataset. This skewed, yet highly separable dataset allows better performance for the SVM base classifiers with poor performance of the QDC base classifiers. Some fusion methods are affected by this discrepancy (product, mean, median, majority voting, BKS) while most trainable combiners achieve high accuracies.

.

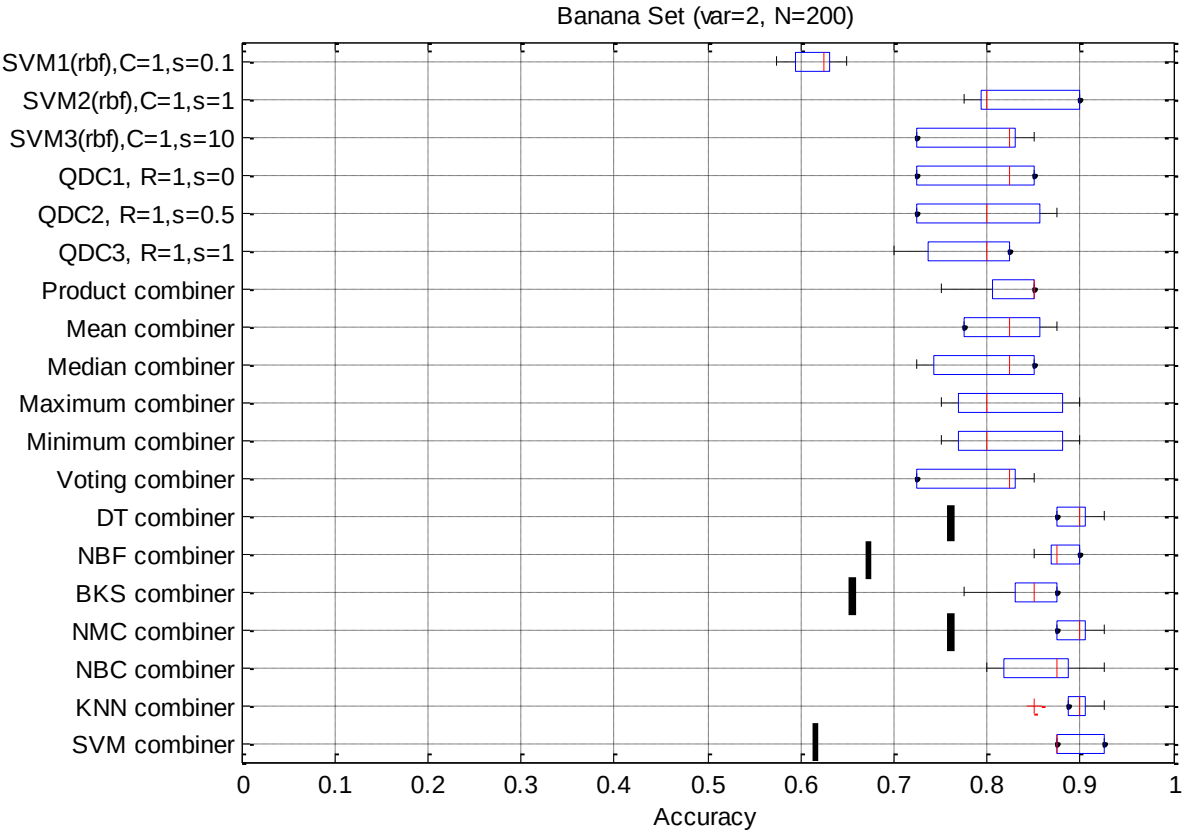


Figure 5.11 Accuracy estimates for the "Ban2" artificial dataset. As the dataset's variance increases, base classifiers achieve lower performances in the range of 0.8. Similarly, fixed combiners achieve low accuracies, whereas trainable combiners reach accuracies in the range of 0.9 with considerably narrower variances.

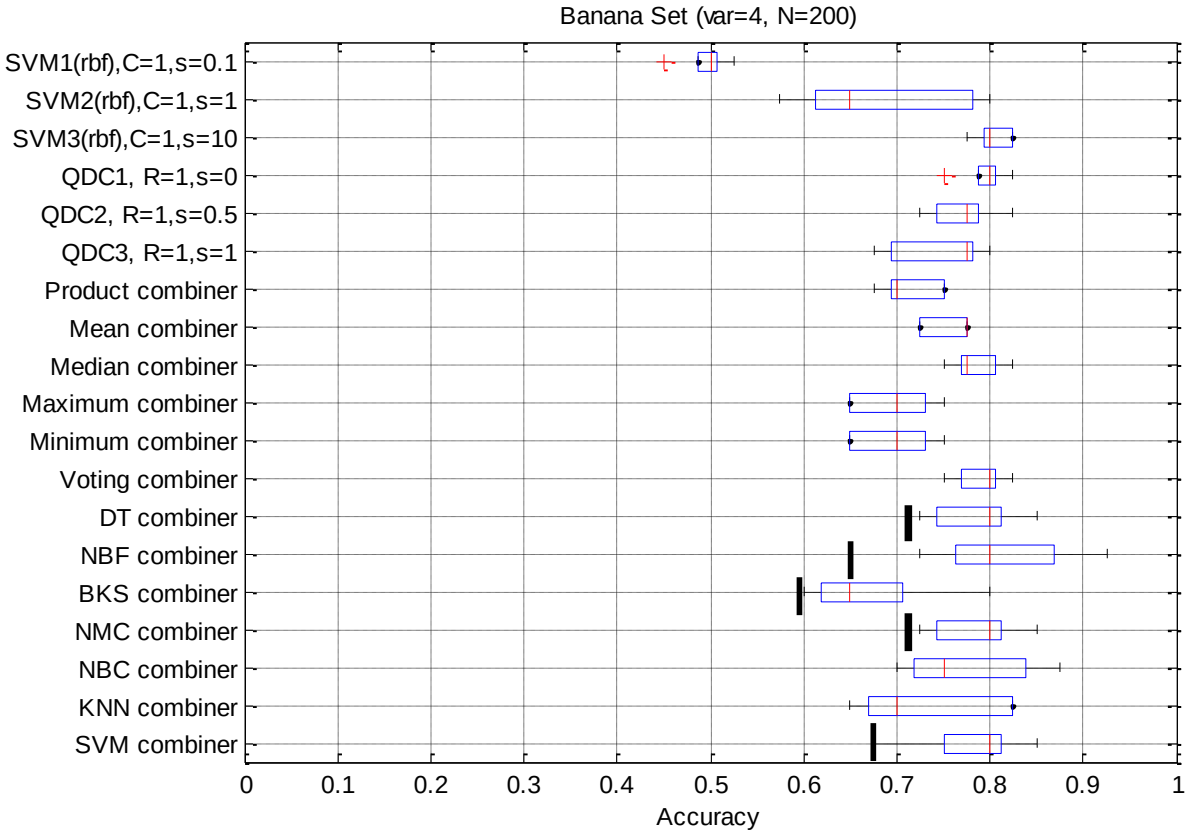


Figure 5.12 Accuracy estimates for the "Ban3" artificial dataset. This skewed dataset with high class overlap diminish the performance of fusion algorithms by 0.1 on average, compared to Figure 5.11. The combiner's accuracy variance is also increased, while BKS and KNN combiners appear to degrade more drastically.

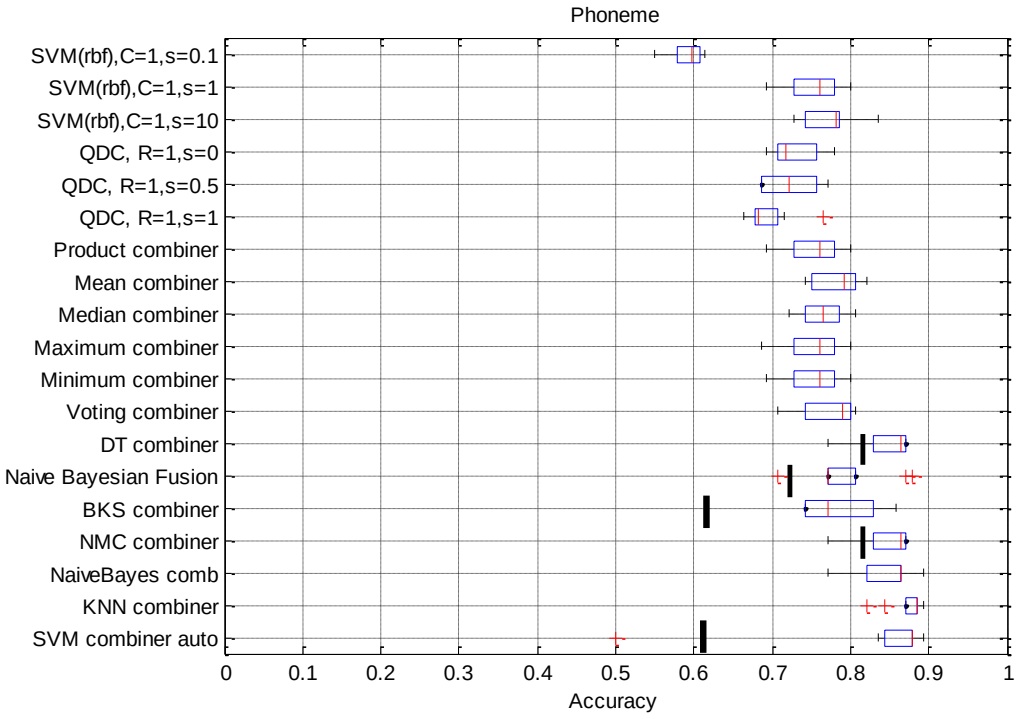


Figure 5.13 Accuracy estimates for the "Phoneme" dataset. In this dataset trainable fusion algorithms exhibit a clear advantage compared to base classifiers and fixed combiners. SVM and KNN combiners achieve top performance.

5.5.2.3 Comparisons of classification schemes based on T-test

In order to further exemplify the performance trends of the various models, the performance results per pair of combiners were statistically compared for significant differences through the t-test, to highlight improvements attributed to specific models and problem context. More specifically, Figures 5.14-5.20 present the p-values of such comparisons obtained by performing a right-sided t-test on the hypothesis that the algorithm indicated on the vertical axis produced higher accuracy than the corresponding algorithm on the horizontal axis. Thus, p-values lower than the standard 0.05 threshold indicate a significant improvement of the corresponding method on the vertical axis. For visualization purposes, p-values are color-coded in gray tiles with intensity ranging from white (corresponding to p=1.00) to gray (p=0.5) and black (corresponding to p=0.00). Base classifiers are listed in rows 1-6, fixed combiners in rows 7-12 and trainable combiners in rows 13-19. To highlight their performance, the area of trainable combiners is outlined by a thick rectangle.

Figures 5.14-5.16 reflect the first scenario of problem context, which uses a Lithuanian-type dataset parameterized with increasing variance. Figures 5.17-5.19 present the corresponding t-tests for the second scenario, which employs a benchmark skewed (banana shaped) dataset also parameterized with increasing variance. With the notable exceptions of BKS (row 15) and –to a less extent- KNN algorithms (row 18), all trainable ensemble methods consistently show low p-values compared to base classifiers, in both the easy (Figures 5.14,5.17) and medium-difficulty (Figures 5.15,5.18) datasets of both scenarios. This exemplifies with strong statistical significance their improved performance compared to the base classifiers (columns 1-6) and fixed combiners (columns7-12). Possible deviations

from this trend may be attributed to occasionally high-performing base classifiers (i.e. the SVM in column 2). Within the group of fixed combiners, the product and mean models appear to be the most effective, as they produce consistently low p-values (visualized by darker rows in tables) when compared to other fixed combiners.

In the group of trainable, distance-based combiners, DTs/NMCs are the leading performers both in the first (Figures 5.14-5.16) and the second scenario (Figures 5.17-5.19) and are the last to degrade their performance for the difficult sets in figures 5.16 and 5.19. Therefore, they can be used in benchmark tests between distance-based with discriminant-based SVM combiners. Overall DTs/NMCs (rows 13/16) and SVMs (row 19) achieve p-values lower than the 0.05 in Figures 5.14, 5.15, 5.17 and 5.18, in their comparison against base classifiers and fixed combiners.

BKS is the first model to degrade its performance under the decreased-separability context of the difficult datasets in Figures 5.16 and 5.19, resulting in p-values around 0.9 for any comparison. KNN is slightly better with p-values in the range of 0.3 under the same context. Notice in these comparisons that BKS/1NN and KNN operate on different feature sets. BKS/1NN use the set of samples producing identical L1 labels to infer class membership, while KNN employs a wider set of nearest neighbours for the same task. By examining the SVM (row) vs DT (column) tiles in each graph, we observe lower p-values compared to the corresponding DT (row) vs SVM (column) tiles. While the p-values in the first case are not below 0.05 to justify significant improvement of SVMs vs DTs, the comparison always favour the SVMs. Considering that the SVM combiner is not explicitly optimized in parameters, the t-test results indicate that the ensemble combination can highly benefit from the generalization capability of SVMs.

In easy to medium-difficulty problems (Figures 5.14-5.15 and 5.17-5.18), a single base classifier (SVM in column 2) significantly outperforms the fixed fusion methods (rows 7-12) possibly due to its successful parameterization. However, this does not hold when comparing with trainable combiners (rows 13-19), which show equal or better performance without the need to rely on parameter optimization.

With respect to the “Phoneme” dataset in Figure 5.20, we observe improved performance of all trainable combiners, especially the KNN and DT combiners.

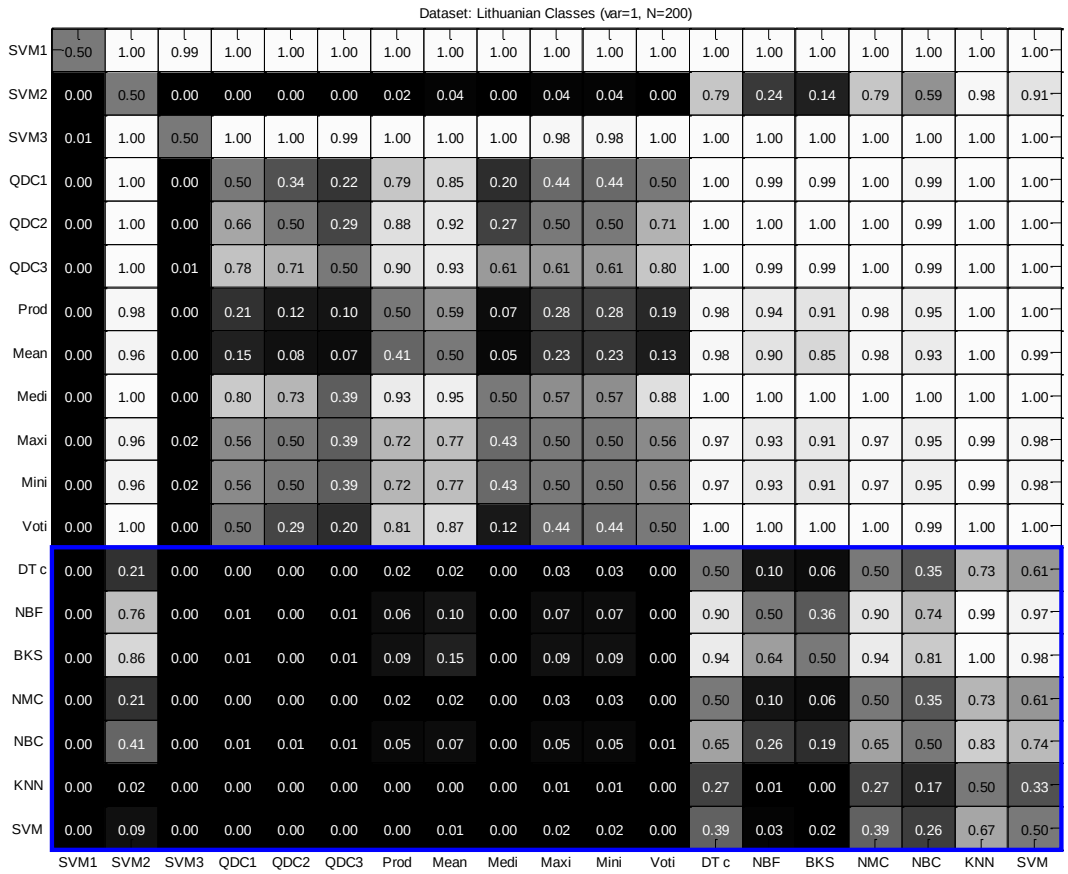


Figure 5.14 P-values measuring differences between classification schemes in estimated accuracy achieved on the "Lith1" artificial dataset. High separability is achieved by all fusion methods, even under varying performance of the base classifiers.

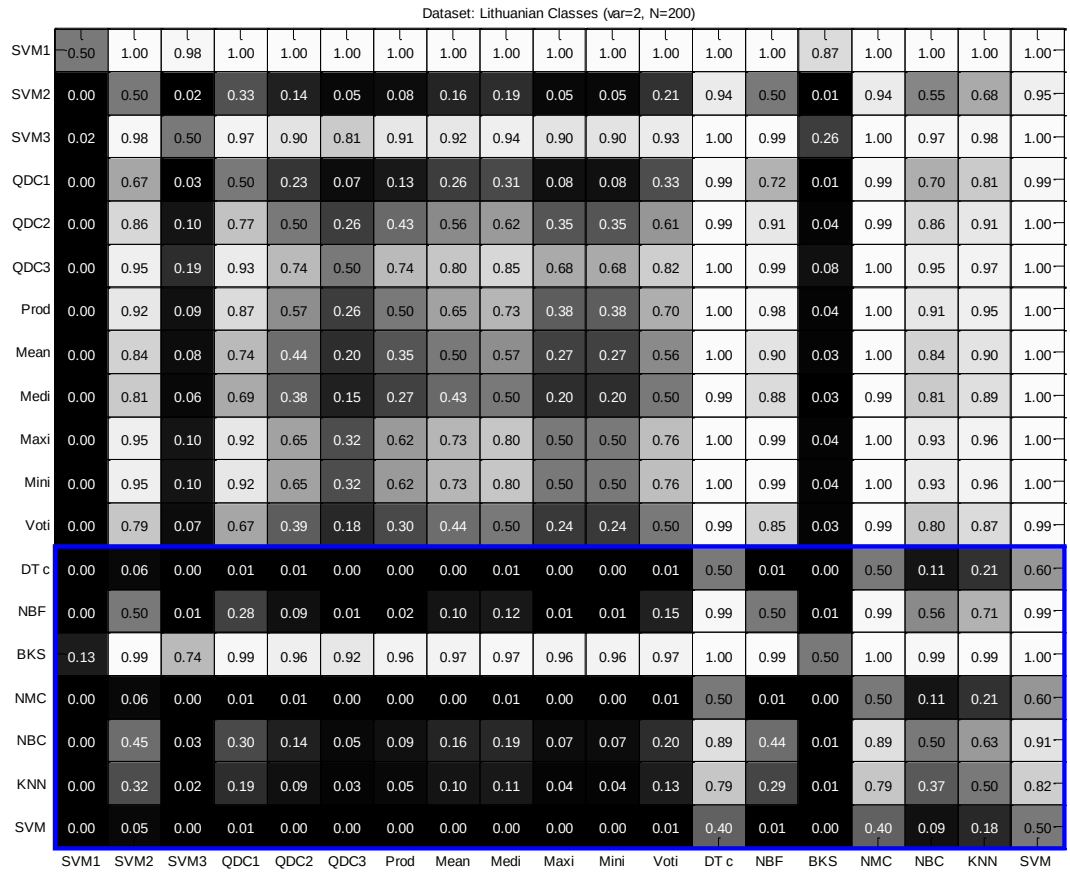


Figure 5.15 P-values measuring differences in estimated accuracy for the "Lith2" artificial dataset. As the variance of the dataset increases compared to Figure 5.14, the performance of base classifiers varies depending on their parameterization, while fusion methods degrade gracefully but still retaining improved performance...

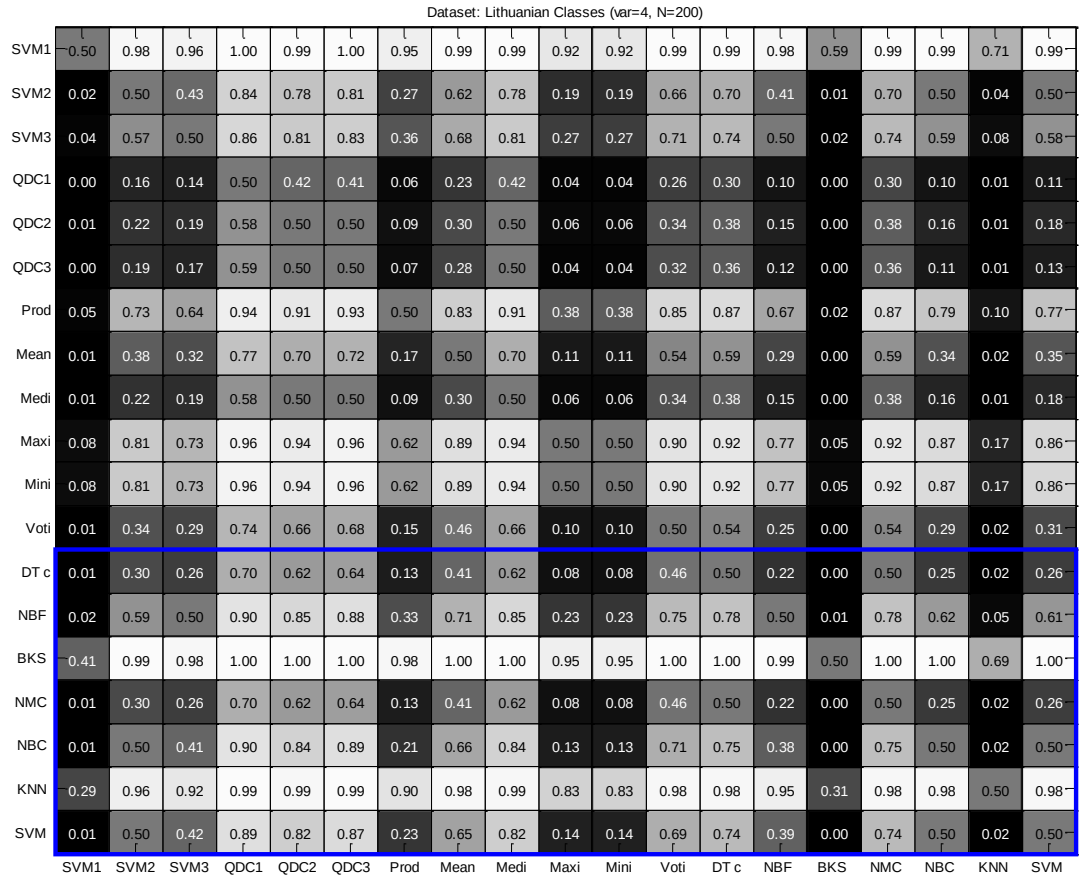


Figure 5.16 P-values measuring differences in estimated accuracy for the "Lith3" artificial dataset. For even higher variance compared to Figures 5.14 and 5.15, only DT/NMC and SVM combiners retain significantly better performance.

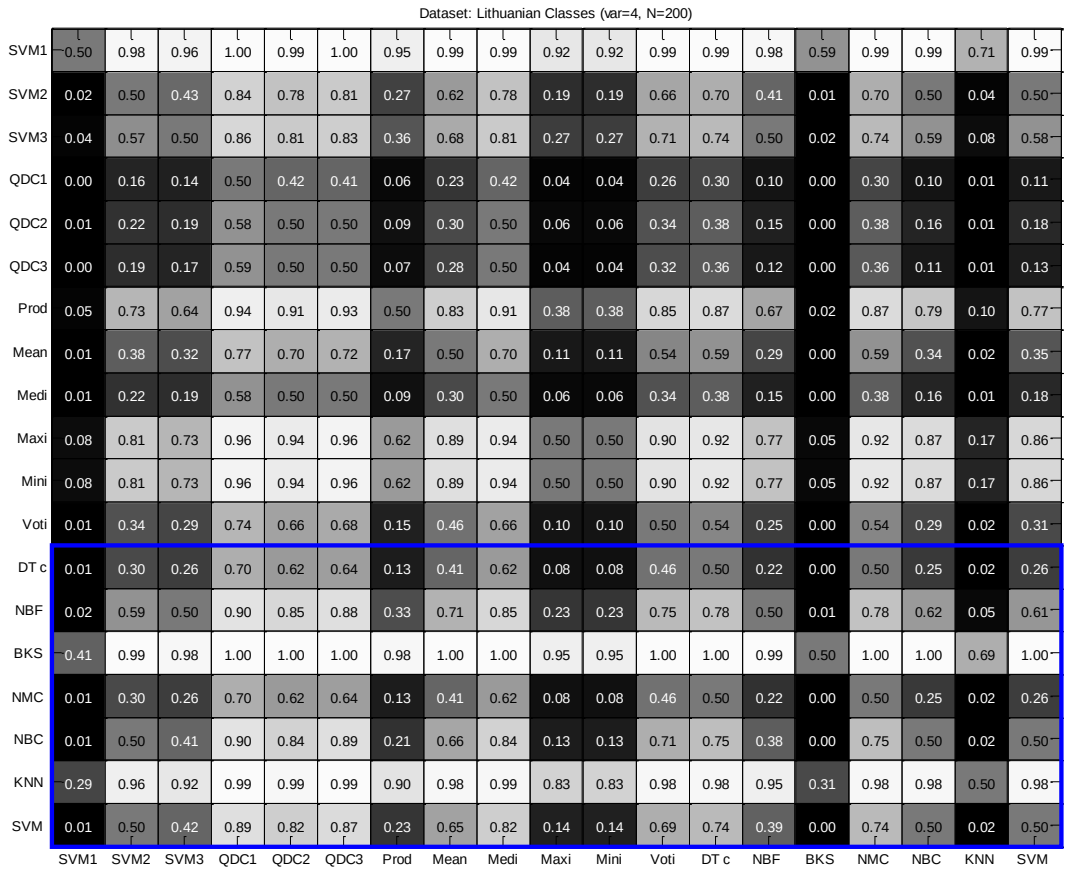


Figure 5.17 P-values measuring differences in estimated accuracy for the "Ban1" artificial dataset. This skewed yet highly separable dataset allows better performance for the SVM base classifiers and poor for the QDC base classifier. This is demonstrated by the large p-value (dark cells) in the 2nd and 3rd lines, indicating superior performance of the SVM over other base and even some non-trainable combiners. Some fusion methods are affected by this discrepancy (product, mean, median, majority voting, BKS), while most trainable combiners (last few rows) achieve high accuracies consistently better than other base and fusion schemes.

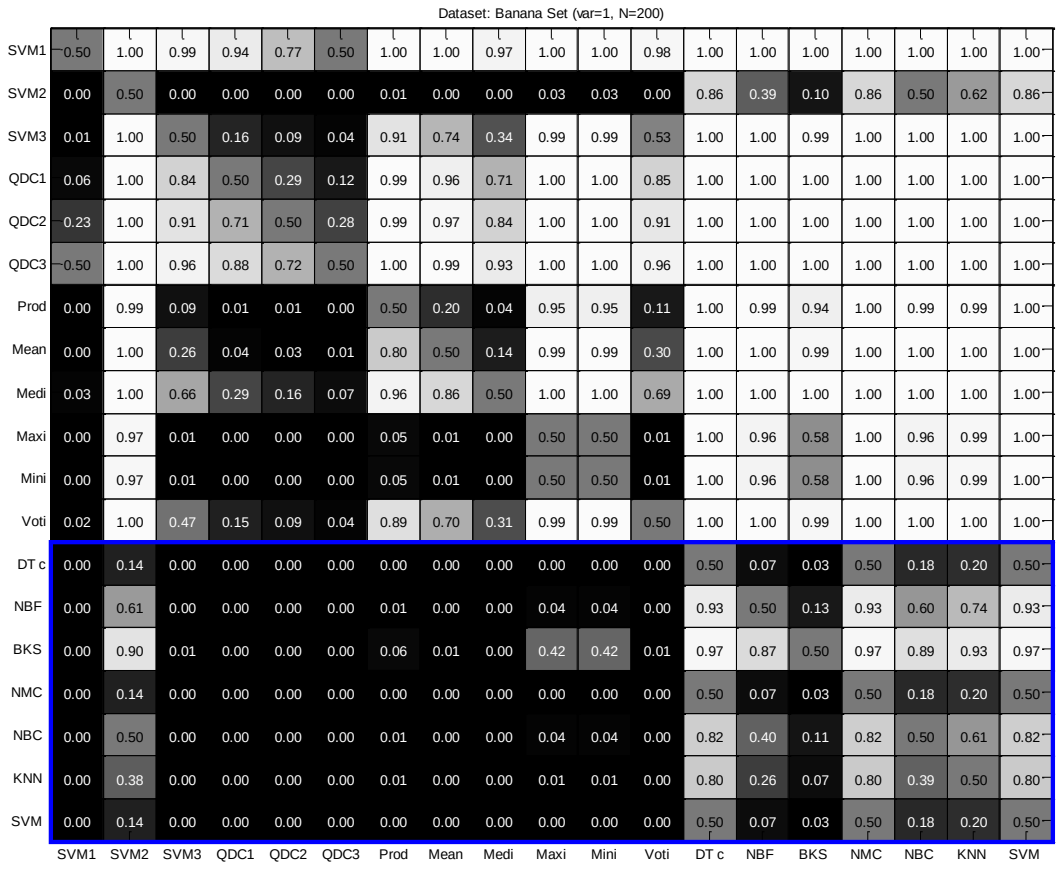


Figure 5.18 P-values measuring differences in estimated accuracy for the "Ban2" artificial dataset. As the variance of the dataset increases, the performance of base classifiers and non-trainable combiners decreases. Only trainable fusion algorithms reach p-values lower than 0.05 in comparison to others, showing significantly better accuracy over base classifiers.

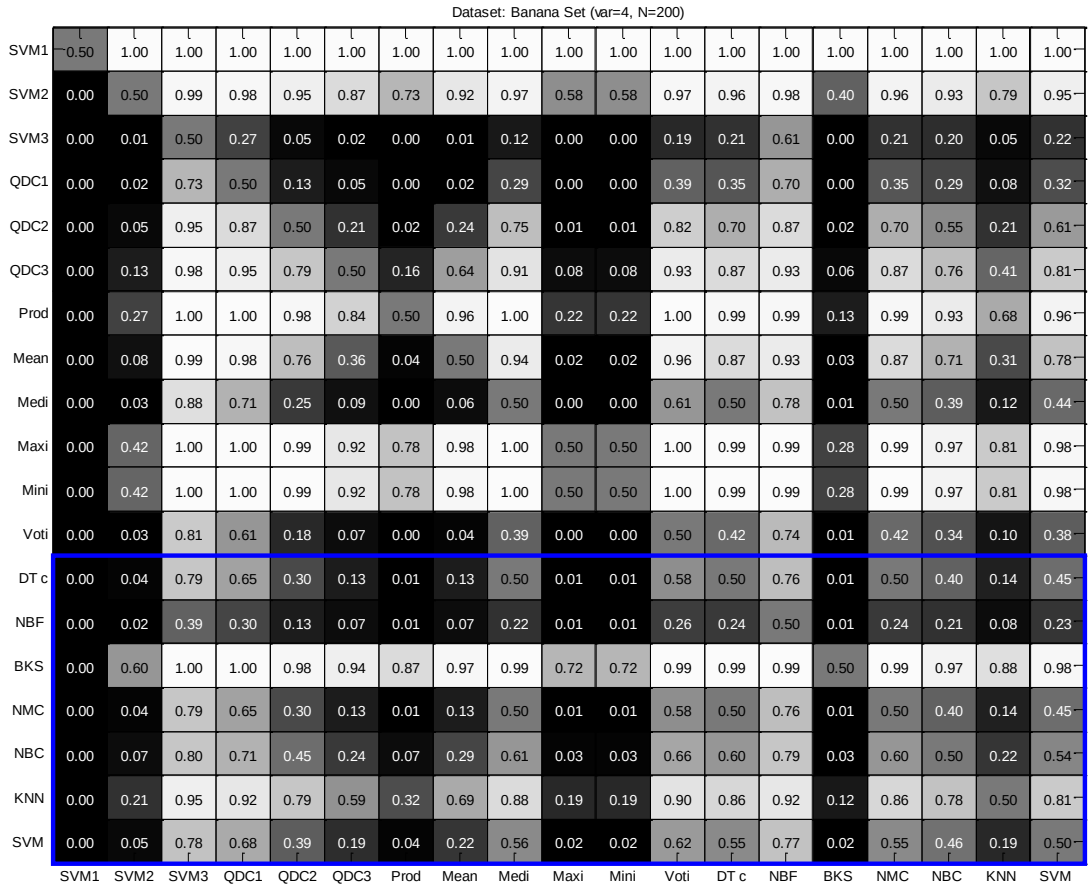


Figure 5.19 P-values measuring differences in estimated accuracy for the "Ban3" artificial dataset. This skewed dataset with high class overlap forces most fusion algorithms to achieve only slightly better performance than base classifiers (p-values>0.05). Only NBF and SVM combiners appear to be the least affected by this trend.

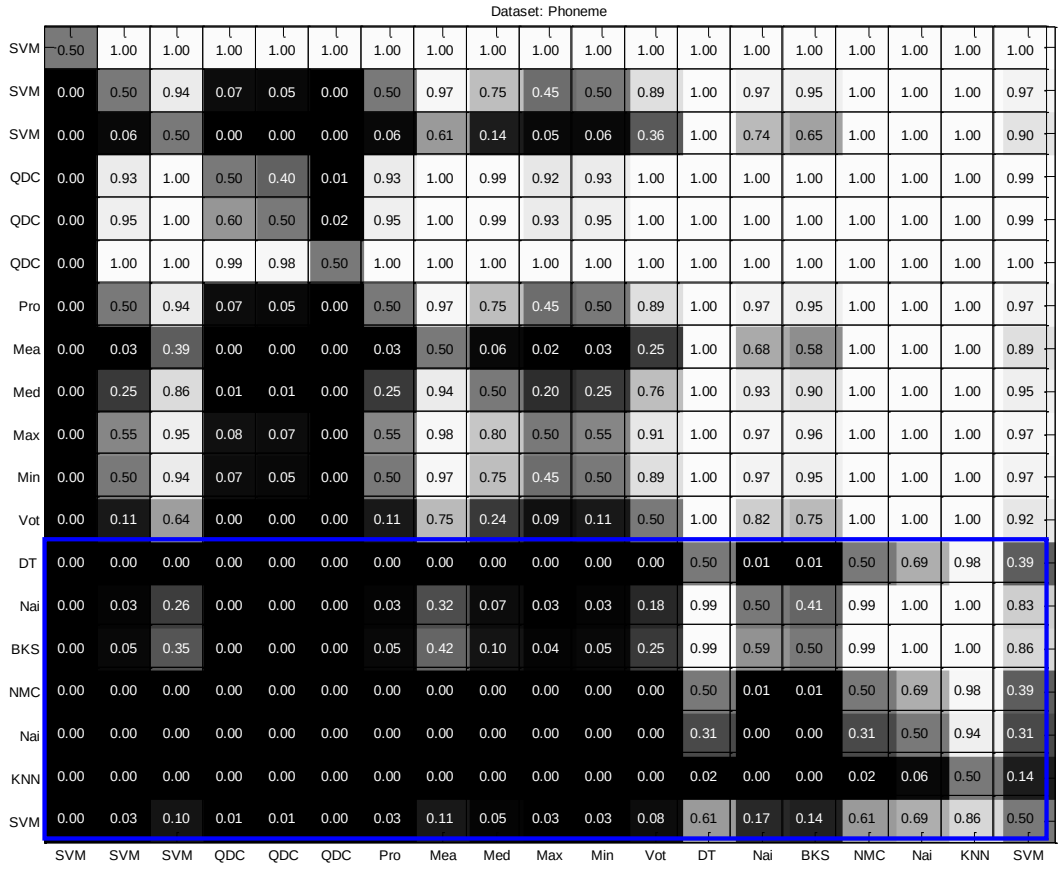


Figure 5.20 P-values measuring differences in estimated accuracy for the "Phoneme" dataset. Trainable fusion algorithms and especially KNN reach p-values lower than 0.05 and outperform fixed combiners and base classifiers.

5.6 Discussion and Conclusions

In this work we explore the use of two general categories of classifiers, namely distance and discriminant-based, as decision fusion (combiner) models. Many widely-used combiners are covered by these categories that originate from complementary philosophies, the former based on area partitioning and the latter on boundary discrimination. Comparing the above two main categories of decision fusion schemes, the error at the combiner lever appears to originate from either insufficient estimation of $\hat{P}(\omega_c \mid \mathbf{x})$ (mainly in the case of distance-based combiners) or from the discriminant function drifting away from the Bayesian decision boundary (for discriminant combiners).

Despite the source of error, the ultimate use of combiners is to leverage the insufficiency of generic classifiers through the processing of the L1 outcomes at a second level. This concept can be further stretched, advocating the use of arbitrary level multi classifier systems. The experimental results can help clarifying the effects of this process, at least up to the second level examined. Moreover, the use of a kernel based discriminant classifier in the form of SVM at the fusion level can extend the application domain of decision fusion methods to non-vectorial L1 feature representations. The kernel method's capability to operate on non-vectorial inputs implies that we can possibly use L1 outputs with representations other than those used in DP. For example, we can use Bayesian Trees ((Pearl 1988)) as primary models and use their structure (a directed graph) as input to an SVM hyper-classifier.

Alternatively, qualitative, fuzzy, or comparative decisions from human experts could be incorporated in an SVM combiner by employing appropriately designed kernels. Thus, SVMs not only provide equivalent or better performance, but are also open to implementation in new dataset domains.

In conclusion, this work addresses the relation of state-of-the-art classifier fusion techniques and generic classifier models. It further shows how such correspondences can be leveraged to estimate error bounds for ensemble fusion methods and select the most appropriate algorithm based on the dataset and combiner properties. In particular, useful error bounds can be derived by extending the model operation from the feature space to the output space of primary classifiers. Through the artificial problem sets considered and the results obtained, we present comparisons of performance from different viewpoints of interest. The advantage of discriminant-based models and, in particular, the superior performance of SVMs as combiners is supported by most of these experiments. The results presented here can be generalized to other classifier families such as neural networks and generalized logistic models.

6 Conclusions

6.1 Discussion

Building on the overall research effort presented in the previous chapters the more general conclusions that can be drawn are presented here.

Firstly, while the core research in the field of computer-aided diagnosis is focused on analytic tasks, our experience has repeatedly revealed that the experimental results and consequently the performance of a prediction system depend largely on the underlying data-conditioning mechanisms. The primary problem in this stage is data scarcity either in the feature or sample space of the dataset. Having investigated the use of a key predictor (CA-125 tumor marker) of ovarian tumor malignancy in a data set with high missingness, it is evident that imputation up to the 25% missingness threshold is a viable alternative to case omission (CCA). The LS-SVMs and GLMs classifier models yield minor AUC differences between the four imputation scenarios. However, CCA clearly differed from imputation regarding model parameter-estimates and variable selection. While it was observed that CCA and missing value imputation resulted in similar AUCs, the obtained GLM model parameters suggested that imputation was preferred over CCA since missing value imputation alleviates bias when estimating particular parameters.. Therefore, we can safely conclude that justified imputation of missing data can produce different classifier models with identical high accuracies. As a side issue, using the outcome when imputing missing values is more essential for imputation because otherwise the association between the outcome and the imputed variable is assumed to be zero.

Regarding the design of multiple kernel classifiers, it has been shown that a general class of composite kernel functions (HS-SVMs. GS-SVMs) exhibits practical advantages including the successful employment of non-positive definite Gram matrices and the improved accuracies compared to standard SVM implementations. The development of new, possibly non-PD, kernels incorporating explicit prior knowledge on the data and being tailored to specific medical classification tasks can help reveal important diagnostic features and rank their effectiveness. GS-SVMs allow for the unconstrained mathematical formulation of expert knowledge rules into composite kernels. In the author's perspective, the key benefit of GS-SVMs formulation extends beyond initial experimental results and is not restricted to SVMs. The primary gain is that the derived composite functionals can be considered as reusable components extending any other kernel method to unconstrained feature kernels.

A case study of SVMs is considered in MRS classification. Besides the assessment of the classification accuracy, the evaluation of features in the SVM scheme offers the advantage of selecting the appropriate target for biopsy and detect the extent of a tumour, especially in cases of infiltrative gliomas before being detected from conventional MRI. While far away from replacing surgical biopsy for the diagnosis of brain tumour, this technique is useful for the assessment of residual disease after surgery, particularly for low-grade gliomas or nonenhancing portion of high-grade tumours. In addition, the more detailed identification of tumour type can provide valuable information prior to surgical or treatment planning. Example applications are in

metastatic brain tumours, high-grade gliomas or some cases of lipomatous meningiomas, which are sometimes difficult to differentiate because of similar appearances on conventional MRI of their focal portion.

The performed analysis verifies that the performance of distance-based nonlinear classifiers in the form of SVMs can further amplify the diagnostic power achieved by current 3T MRS scanning technology. The author's research demonstrates the applicability of modified SVM kernel in this task in conjunction with more powerful data acquisition schemes, such as the 3T devices. The proposed SVM kernel and reduced feature set can be applicable in scenarios where the acquisition, transfer and utilization of a full-resolution MR spectrum from the scanner to a clinical decision-support-system is impractical or time demanding. In this context, the use of 4-5 features can provide a first estimate of the patient's status.

While the clinical information from MR imaging alone has a mean accuracy of lesion classification in the range of 75%, proton MRS in our case study has been proved to distinguish benign from malignant brain tumours with a mean accuracy of 95%. This compares favorably with diffusion-weighted imaging while there is still margin for improvement with the utilization of data fusion techniques. In the near future, it is unlikely that radiologists will make a diagnosis based solely on a conventional decision rule. The complexity of variable interactions calls for the combined utilization of proton MRS within a standard MRI examination and additional data sources. An automated decision support system that will analyze and classify proton MRS data will offer improved differential diagnosis and upgrade patient outcome.

The capitalization of decisions from multiple modalities will take systems such as the above to the next fusion level for advanced decision support. Merging the background of established classifiers with the concepts of combiners allows extracting useful error bounds and selection criteria. Combining the outcomes of a wide pool of base classifiers using trainable combiners can minimize the error at the combiner lever originating from estimation of $\hat{P}(\omega_c \mid \mathbf{x})$.The ultimate use of combiners is to leverage the insufficiency of generic classifiers through the processing of the L1 outcomes at a second level. Our two-level approach to fusion illustrates how correspondences of predictors in these layers can be leveraged to estimate error bounds for ensemble fusion methods and select the most appropriate algorithm based on the dataset and combiner properties. In particular, useful error bounds can be derived by extending the model operation from the feature space to the output space of primary classifiers. Through the artificial problem-sets considered and the results obtained, we present comparisons of performance from different viewpoints of interest. The advantage of discriminant-based models and, in particular, the superior performance of SVMs as combiners is supported by most of these experiments. The results presented here can be generalized to other classifier families such as neural networks and generalized logistic models.

Moreover, the use of a kernel-based discriminant classifier in the form of SVM at the fusion level can extend the application domain of decision-fusion methods to non-vectorial L1 feature representations. The kernel method's capability to operate on non-vectorial inputs implies that we can possibly use L1 outputs with representations other than those used in DP. For example, we can use Bayesian Trees as primary models and

use their structure (a directed graph) as input to an SVM hyper-classifier. With recent developments in the biomedical field, we also envision that new relevant data in the form of interaction graphs will soon become available for decision making. Such graph structures encode interaction at the genomic/proteomic level, but also capture dynamic information regarding the activation of different brain regions as depicted in EEG/fMRI. Interaction networks are increasingly studied in genomics providing more consistent information regarding the contribution of groups of genes and/or proteins in pathology. Thus, genome wide association studies move from markers including isolated genes to markers composed of groups of genes and their interaction pathways and networks, which are represented in graph structures. Furthermore, extensive neurophysiological studies demonstrate that the progression from one state of a brain pathology to the other (e.g. Alzheimer's disease) shows intense loss of corticocortical connectivity. Thus, markers of functional interactions, which can be effectively described by neuronal graph structures, may characterize better these changes than quantitative EEG markers, which show considerable cross-population overlap. In addition, string data in genomics are becoming more and more exploited for gene and protein localization in the form of binary or symbol strings from sequenced genomes. Such data in need of appropriate processing tools can be effectively handled by means of kernels. Alternatively, qualitative, fuzzy, or comparative decisions from human experts could be incorporated in an SVM combiner by employing appropriately designed kernels. Thus, SVMs not only provide equivalent or better performance, but also are open to implementation in new dataset domains.

6.2 Future Research and Implementation Directions

While additional performance gains can be achievable by optimizing the above methodology at the data acquisition, preprocessing, classification and visualization levels, the practical use of such decision support tools in a clinical environment relies heavily on the ability to provide coherent performance across multiple datasets and pathological classes.

To this end, perceived future research directions might include the utilization of data and decision fusion methods, the use of full feature information to create visual diagnostic aids and the evaluation of an interactive graphical user interface to communicate and adapt the decision support system's statistical estimates to the clinician's feedback in real time.

From the collected experience with the data collection process and interfacing with the clinicians, it is evident that a large portion of human expert knowledge is not leveraged within computer aided diagnosis systems. This fact is attributed to the difficulties in transforming such knowledge to a structured format and to the boundaries to adoption of such systems by caregivers. The production systems used in hospitals (Laboratory-, Radiology- and Patient- Information Systems) need to be augmented with business intelligence capabilities embedded in the clinical workflow.

The coherence of collected data needs to be improved especially to counterbalance variations among different clinicians. The variations themselves can be seen as a by-product in a clinical evaluation process.

One additional aspect in which there is a clear gap is the confidence levels provided by automated diagnostic tools. The outputs of decision support tools should be accompanied by quality indicators in the form of error bounds, secondary outputs or performance statistics. This practice will aid the adoption of such systems in a per-case context based on the difficulty level of each specific patient and pathology.

Finally, the author emphasizes the highly promising research opportunities from “kernels on sets”. Such kernels can be defined on loosely structured sets with minimal closure properties instead of standard kernels being defined on vector spaces. Their use will enable the inclusion of additional data types as features to diagnostic models.

In overall the common algorithmic aspects in biomedical data processing include the requirement to reduce noise, reduce dimension, transform to more suitable representations for subsequent interpretation, extract similarities, and exploit dissimilarities.

In areas such as neural networks, optimization and image analysis, and Bayesian approaches to inference, there have been significant advances. However, several bottlenecks can be identified. Among them is the development of methods that have been genuinely devised to support medical decision making. To this end, tools and techniques that engage clinicians to all stages from data acquisition, to processing, validation and interpretation of results should be largely encouraged.

While biomedical data is notoriously unreliable, noisy, distributed and incomplete, most analytic tools and methods, however, assume data integrity. There is a proportionally insufficient number of methods which explicitly deal with uncertainty, both in the inputs and the outputs. Developments in other areas, such as complexity, communications and information technologies probably have a great deal to offer to biomedical data analysis, as algorithmic requirements cross the discipline boundaries. Advances are required in biomedical ontologies, ontology based integration of heterogeneous biomedical data, and service oriented computational frameworks capitalizing on modern technologies enabling the fast and efficient access to and processing of biomedical data.

6.3 Lessons learnt

A long tradition at K.U. Leuven, where part of this work has been carried out, requires that a PhD thesis is accompanied by a list of practical lessons learnt aside with the main research effort.

Following this tradition the lessons acquired as part of this work include the following:

1. No preprocessing, normalizing and choosing right variable mappings can cripple some pattern recognition algorithms while leave others unaffected. Since these steps do not cost anything, do use them anyway.
2. It is essential that the training set has never met the validation set and they both have never met the test set.

3. Using the right tools, source control, descriptive variable names, modular development and unit testing are things that keep you sane.
4. Have a plan and check back to it regularly.
5. Talk to the application domain experts. Knowing your dataset and choosing features are essential to getting good results and interpreting them correctly.
6. Running a simulation once and getting good results proves nothing. You have to prove that this was not mere coincidence. AUCs, confidence intervals and cross-validation provide sufficient proof.
7. Working in a group is always more effective. Communicate. Genius is collaborative.
8. Better complete a task than do it perfect.

Equation Chapter (Next) Section 1

Appendices

Appendix I Notation

$\{\mathbf{X},\mathbf{Y}\}$	labeled dataset
$\mathbf{X}_a$	available cases in a dataset
$\mathbf{X}_m$	cases with missing values in a dataset
$\mathbf{X}_c$	complete (available+imputed) cases in a dataset
$\mathbf{x}_i$ (or $\mathbf{x}$)	sample feature vector
$\mathcal{F}$	feature space
D	dimensionality of feature vector
N	number of cases in a dataset
i	index of a sample
ω_c	class c
C	number of classes
c	index of a class
f_l	classifier l
L	number of classifiers
l	index of a classifier
Z	classifier soft output
Y	classifier hard output
d_{lc}	DP element (L1 support value) of base classifier l for class c
dt_{lc}	DT element of base classifier l for class c
$\hat{P}_l(c \mid \mathbf{x}_i)$	(estimator of) class c posterior probability of sample $\mathbf{x}_i$ based on classifier f_l
F	combiner function
$\hat{P}'(c \mid \mathbf{x}_i)$	(estimator of) class c posterior probability of sample $\mathbf{x}_i$ based on the combiner
Φ	standard normal distribution.
$\mathcal{N}(\mu,\sigma^2)$	unıvariate Gaussian distribution with mean μ and variance σ^2
$\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$	multivariate Gaussian distribution with mean **μ** and variance **Σ**
σ^2	variance
$k(\cdot,\cdot)$	kernel function
$\mathbf{K}$	SVM kernel matrix
$\mathbf{K}_{mn}$	element of the nth row and mth column of matrix $\mathbf{K}$
H	Hessian matrix
Σ	covariance matrix
a	Lagrange multiplier

Appendix II Statistical properties of decision fusion methods

The statistical properties of classifier fusion methods, can provide insight to the models' output and error under given input conditions. The following table in spite of being sparse summarizes the available analytic expressions on the statistical properties of the examined fusion algorithms.

		Fusion algorithm	**Fused class posterior probability** $\hat{P}'(\omega_c \mid \mathbf{x}) =$	**Fusion error (or bound)** $\hat{P}'(e) = \int_{x \in \omega_2} P'(\omega_1 \mid \mathbf{x}) d\mathbf{x}$	**Assumptions** (for given stat formulations)
		Single classifier		$= \Phi\left(\frac{0.5-\mu}{\sigma}\right)$	Gaussian inputs
		Oracle (perfect combiner)		$= \Phi\left(\frac{0.5-p}{\sigma}\right)^L$	Gaussian inputs
L2 fixed		Min/Max (identical for C=2 classes)		$= \Phi\left(\frac{\sqrt{L}0.5-p}{\sigma}\right)$	Gaussian inputs independ. features
		Average	$\Phi\left(\frac{x-p}{\sigma^2/L}\right)$		
		Product			independ. features
		Median	$\sum_{j=\frac{L+1}{2}}^{L}\binom{L}{j}\Phi\left(\frac{0.5-p}{\sigma}\right)^j\left[1-\Phi\left(\frac{0.5-p}{\sigma}\right)\right]^{L-j}$		
		Majority voting	$\sum_{m=\lfloor L/2\rfloor+1}^{L}\binom{L}{m}p^m(1-p)^{L-m}$	$= \frac{2L(1-p)}{L+1}$	i.i.d. L1 outcomes equal L1 accur. p
L2 trainable	**distance based**	DT (Euclidean distance)	-	$\frac{\Gamma(LC/2, R_c^2/2)}{\Gamma(LC/2)} \le \frac{LC}{R_c^2}$	via correspondence to NMC
		"Naive" Bayes	$\prod_{l=1}^{L} lm_{c,c'}^l$	$= \sum_{c'=1}^{C}\left(1-\max_{c\in C}\left(\prod_{l=1}^{L} lm_{c,c'}^l\right)\right)$	minimize $E[\text{cost}]$ mutually independent L1
		Behavior Knowledge Space	$P_1 = P(X \in \omega_1) = \frac{k_1}{k}$	$2\sum_{i=1}^{N'}\sum_{j=1}^{C}\frac{k_j}{k_i}\left(1-\frac{k_j}{k_i}\right)$	via the correspondence to voting 1-NN
	discriminant	Support Vector Machine combiner	-	$\hat{E}_n\phi(Y\cdot f(\mathbf{x})) + \frac{4B}{\gamma n}\sqrt{\sum_{i=1}^{n} k(\mathbf{X}_i, \mathbf{X}_j)} + \left(\frac{8}{\gamma}+1\right)\sqrt{\frac{\ln(4/\delta)}{2n}}$	-

Table 6-1 Statistical properties of decision fusion methods

Appendix III Classifier fusion systems usage scenarios

The following table summarizes the cumulative theoretical and experimental findings on the characteristics of the most prominent fusion methods. It is aimed at being used as an indicative reference for choosing appropriate decision fusion methods depending on the problem context.

		Fusion algorithm	Equivalent classifier	input type	Output type [1]	L1 diversity	N [2]	L	Remarks / Assumptions	References
L2 fixed		Min/Max	-	soft	Hard	high	-	-	• derived from sum rule + assumption for equal priors • independent features (L1 outcomes)	(Kittler, Hatef et al. 1998)
		Average/sum	-	soft	Hard	high	-	high	• superior in high noise scenarios where decision is based primarily on priors • independent features (L1 outcomes)	(Kittler and Alkoot 2003)
		Product	-	soft	Soft	-	-	-	• best if $\hat{P}_l(\omega_c \mid \mathbf{x}) \to P_l(\omega_c \mid \mathbf{x})$ • sensitive to diminishing outliers $\hat{P}_l(\omega_c \mid \mathbf{x}) \to 0$	(Kittler, Hatef et al. 1998)
		Probabilistic product	-	soft	Soft	-	-	-	• similar to product with adjustment for class priors • sensitive to diminishing outliers $\hat{P}_l(\omega_c \mid \mathbf{x}) \to 0$	(Bordley 1982)
		Majority voting	-	hard	Hard	low	-	Low	• no L1 outcome pdf assumptions • resilient to heavy tailed pdfs	(Kuncheva 2004)
L2 trainable	**distance based**	DT Euclidean	Nearest Mean Classifier(NMC)	soft	Soft	low	-	High	• NMC is ML optimal for Gaussian features • Diagonal covariance matrix of DPs	(van Otterloo and Young 1978)
		"Naive" Bayes	Bayes classifier on L1 labels	hard	Soft	high	High	high	• Bayes optimal for hard i.i.d. L1 labels • assumes conditional independence of L1 labels	(Xu, Krzyzak et al. 1992; Duda, Hart et al. 2001)
		Behavior Knowledge Space	Voting 1-NN on L1 label space	hard	Hard	high	high	low	• no L1 outcome pdf assumptions • aka "supra Bayesian" approach (Kuncheva 2004) • Bayes optimal for adequate sample sizes	(Fukunaga 1990; Bovino, Dimauro et al.

[1] "soft" output means capable of producing soft and hard output (through thresholding)

[2] combiners are generally more efficient than base classifiers in smaller sample sizes N utilizing a high number of base classifiers L

										2003; Kuncheva 2004)
		Dempster-Schafer	-	soft	Soft possibilistic	-	-	high	• extends DTs and combines them with Bayesian combining of distances	(Rogova 1994)
	discriminant	QDC	QDC	soft	Soft	-	-	-	• Bayes optimal combiner if L1 outputs are normal • sensitive to input covariance matrix singularity	(Duda, Hart et al. 2001)
		Fisher		Soft	Soft	-	-	-	• Nonparametric. For cases where a strong departure from Bayesian assumptions is suspected	(Duda, Hart et al. 2001)
		SVM	SVM	Soft	Soft	-	low	high	• for low accuracy base classifiers that lead to complex combiner feature spaces • For small datasets with large ensembles	(Dimou and Zervakis 2008)

Table 6-2 MCS selection table based on problem attributes

Appendix IV The TSI Classifier Fusion Toolkit

In order to provide a common framework for the research group's pattern analysis tasks in classifier fusion, a toolkit of related functions was developed by the author in Matlab. This tool was made available in January 2007 to the Biopattern and research community through the DADS portal.

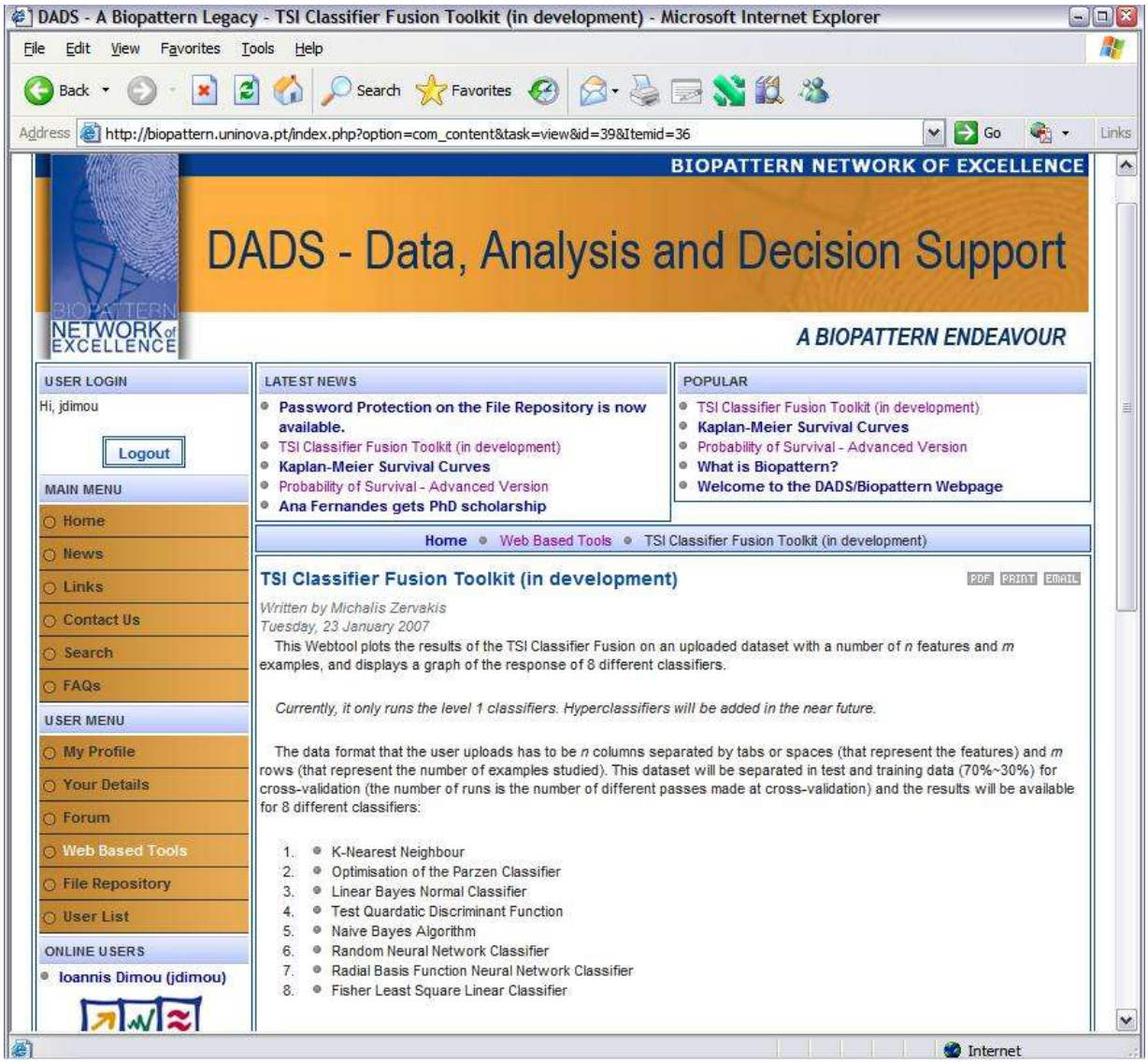


Figure 6.1 The TSI classifier fusion toolkit available through the Biopattern project website

The toolkit consists of a collection of Matlab functions, which implement a number of established classifier fusion algorithms and supporting utility functions. The functionality of the tools is designed to leverage an object-oriented framework, work modularly and be be expandable.

Perquisites:

The toolkit requires the (freeware) pattern recognition toolbox PRTools (http://prtools.org/) in order to accept the dataset in a structured format and to utilize some of the classifiers.

The toolkit has been successfully tested for compatibility with Matlab versions up to R14 and 2010b and with PRTools Version 4.0.14.

Included modules and descriptions

The functionality of the toolkit can be conceptually partitioned into the following modules:

Dataset insertion

Scenario definition

Iterative training and testing

Structured results persistence

Results Analytics

Implemented hyper classifiers

(jxxxxxclf.m)

jbksclf.m Function that implements Behavior Knowledge Space method.
The Behaviour Knowledge Space method is based on a matrix that contains counts the occurrence of all possible label combinations from each class. The fused class label is set to the class with the largest number of occurrences for the specific observed label combinations.

jnbclf.m Function that implements Naive Bayes method
The Naïve Bayes combination method assumes that all L1 classifiers are mutually independent. It works by forming a confusion matrix that maps the correspondence probabilities of observed to true labels. Using these probabilities the algorithm calculates for each new sample the posterior and selects the label with the larged posterior as the fused class label.

jdsclf.m Function that implements Dempster-Schafer method
The Dempster Schafer fusion method is a class indifferent (CI) method that makes direct use of the Decision Template, but not in the same way as the DT method. More specifically the DS algorithm assumes possibilistic L1 outputs and calculates a quadratic proximity measure between the Decision Template and Decision Profiles. In a second step the proximity measure is used to estimate belief degrees for each class. The

product of these belief degrees over all L1 classifiers gives the membership degrees vector from which the actual fused label is found.

jmajvoteclf.m Function that implements Majority Voting method

jmmapclf.m Function that implements Mean, Max, Average & Product methods

jprobpodclf.m Function that implements Probabilistic Product method.

The Probabilistic Product method assumes mutual classifier independence and used the soft L1 outputs as posterior probabilities. the prior-compensated product of the soft outputs over all classifiers gives the final membership function.

jdtclf.m Function that implements Decision Templates method

The Decision templates method calculates the Decision Template for each class by summing the soft labels of that class and normalizing. The fused class membership is calculated by matching a sample's Decision Profile to each classes Decision Template through various metrics.

The above functions (jxxxxxprtf.m) have been designed to implement the corresponding classifiers in a way compatible to the prtools format. The user is encouraged to formulate any new classifiers in the same way, although this is not mandatory.

Usage

To start testing various classifier fusion scenarios the user should do the following:

Step 1: Simulate

Run clfiterator.m to load dataset and run L1 & L2 classifiers.

This script provides the feature space (DPtst, Ytst) on which the L2 (hyper) classifiers will work on and executes a selected subset of combiners.

Step 2: Analyze.

Executing clfResultsAnalytics.m loads the raw results file, creates OLAP cude representation of the results and applies comparative statistics (t-tests, ANOVA) on various dimensions of the scenario. In this way the analytics module yields meaningful comparisons in the (L2) model, sample size, complexity etc dimensions of the experiment.

As a warning, most easy datasets give high accuracies and highly correlated classifiers. To evaluate classifier fusion algorithms one has to utilize difficult datasets or reduce the features/predictors.

The user should also pay attention to the difference between class-conscious (CC) and class indifferent (CI) soft-input hyper-classifiers. Class-conscious take into account L1 labels for the positive class only. Clas indifferent fusion algorithms take into account the

whole Decision Template and therefore in principle they use more available information and should be more effective.

For a detailed explanation on the operation, assumptions and limitations of the classifier fusion algorithms please refer to (Kuncheva, Bezdek et al. 2001).

Key advantages

Arbitrary kernel injection

Due to the requirement for non-PD kernel evaluation (GS-SVMs) the toolkit supports the injection of any type of kernel functional or precalculated matric. This extends beyond the functionality provided inherently by Matlab or via other toolboxes.

Results analytics

The multi-dimensional OLAP based analytics module promotes reusability in future experiments and consistent reporting of results irrespective of the underlying methods being evaluated.

Results visualization

The summarization of multiple result sets and the emphasis on outliers and important trends has let to the design of a new plotting function which provides visualization of the differences in the performance of models by creating a grayscale plot of t-test results. This functionality allows for easy identification of benefits of certain methods over others within the experimental scenario.

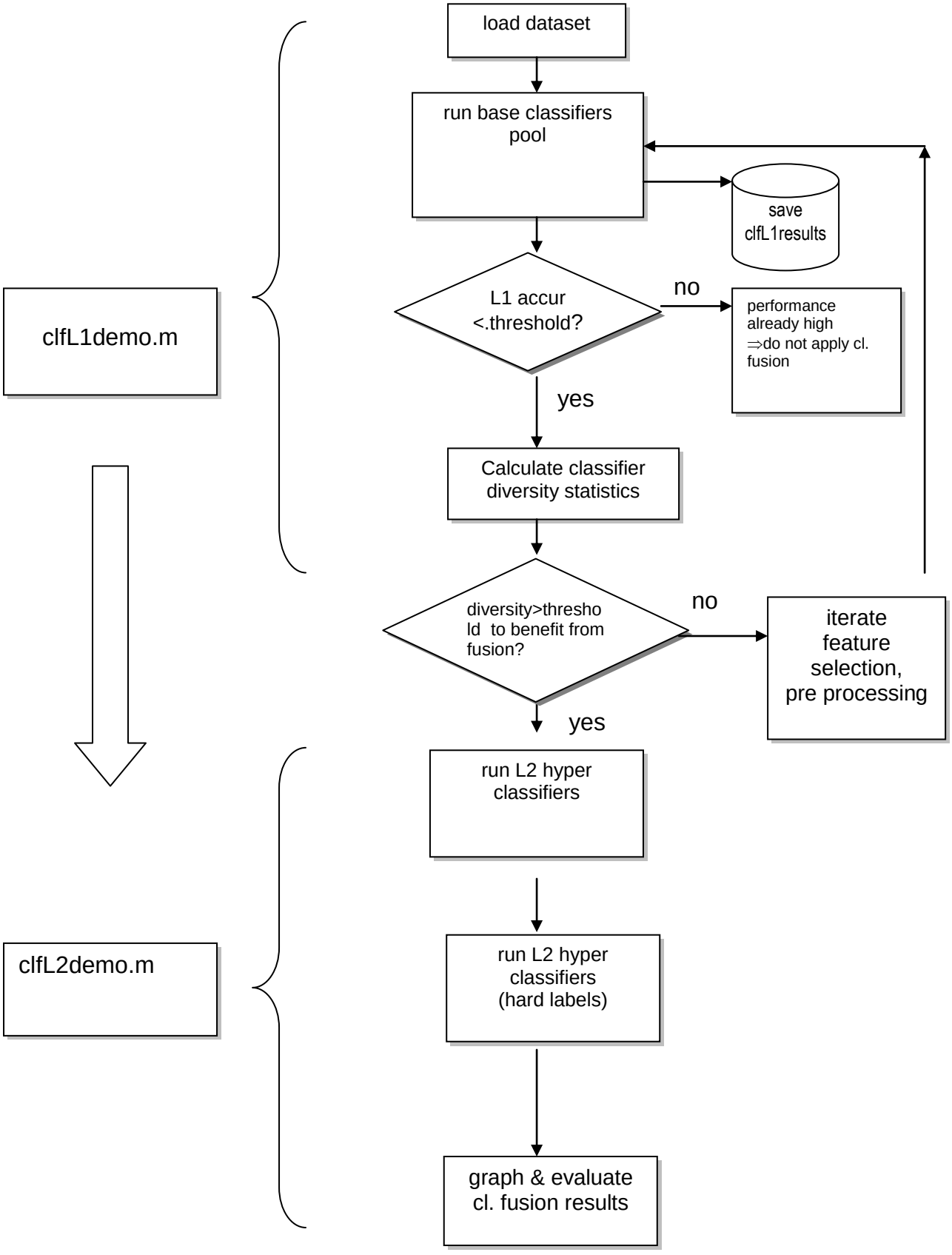


Figure 6.2 TUC Classifier Fusion toolkit block diagram

Equation Chapter (Next) Section 1

Appendix V Author's research work

The work presented in this thesis constitutes of a subpart of the author's continuing research effort that so far has yielded the following publications.

Data imputation

Dimou, I., B. Van Calster, et al. (2009). "Evaluation of Imputation Methods in Ovarian Tumor Diagnostic Models Using Generalized Linear Models and Support Vector Machines." Med Decis Making: 0272989X09340579.

Dimou, I., B. Van Calster, et al. (2007). Comparison Of Imputation Approaches In Ovarian Tumour Diagnostic Models Based On Ls-Svms

CIMED2007. Plymouth.

Van Holsbeke, C., B. Van Calster, et al. (2007). "External validation of mathematical models to distinguish between benign and malignant adnexal tumors: a multicenter study by the International Ovarian Tumor Analysis Group." Clin Cancer Res 13(15 Pt 1): 4440-4447.

Support Vector Machines - MRS

Dimou, I., I. Tsougos, et al. (2011). "Brain lesion classification using 3T MRS spectra and paired SVM kernels." Biomedical Signal Processing and Control In Press, Corrected Proof.

Dimou, I. N., I. Tsougos, et al. (2009). Classification of pathological human brain lesions using Magnetic Resonance Spectroscopy at 3T. World Congress on Medical Physics and Biomedical Engineering, September 7 - 12, 2009, Munich, Germany: 1368-1370.

Dimou, I., I. Tsougos, et al. (2009). Classification of 3T MRS spectra using Support Vector Machines. 9th International Conference on Information Technology and Applications in Biomedicine, Larnaca, Cyprus.

Kounelakis, M., I. Dimou, et al. (2011). "Strengths and Weaknesses of 1.5T and 3T MRS Data in Brain Glioma Classification." IEEE Trans Inf Technol Biomed.

Classifier Fusion

Dimou, I. and M. Zervakis (2008). Support Vector Machines versus Decision Templates in Biomedical Decision Fusion. Machine Learning and Applications, Fourth International Conference on.

Dimou, I. N., G. C. Manikis, et al. (2006). Classifier Fusion Approaches for Diagnostic Cancer Models. 28th Annual International Conference on Engineering in Medicine and Biology Society, New York, IEEE.

Dimou, I. N. and M. E. Zervakis (2009). "Error bounds of decision templates and support vector machines in decision fusion." Int. J. Knowl. Eng. Soft Data Paradigm. 1(3): 227-238.

Genomics

Pardalos, P. M. (2012). Data mining for biomarker discovery. New York, Springer.

Papanikolaou, V., E. Athanassiou, et al. (2011). "hTERT regulation by NF-kappaB and c-myc in irradiated HER2-positive breast cancer cells." Int J Radiat Biol 87(6): 609-621.

Papanikolaou, V., D. Iliopoulos, et al. (2010). "Survivin regulation by HER2 through NF-kappaB and c-myc in irradiated breast cancer cells." J Cell Mol Med.

Papanikolaou, V., D. Iliopoulos, et al. (2009). "The involvement of HER2 and p53 status in the regulation of telomerase in irradiated breast cancer cells." Int J Oncol 35(5): 1141-1149.

References

Abe, S. (2005). Support vector machines for pattern classification. London, Springer.

Acharya, R., R. Wasserman, et al. (1995). "Biomedical imaging modalities: a tutorial." Computerized Medical Imaging and Graphics **19**(1): 3-25.

Alonso-Calvo, R., V. Maojo, et al. (2007). "An agent- and ontology-based system for integrating public gene, protein, and disease databases." J Biomed Inform **40**(1): 17-29.

Altmann, A., M. Rosen-Zvi, et al. (2008). "Comparison of Classifier Fusion Methods for Predicting Response to Anti HIV-1 Therapy." PLoS ONE **3**(10): e3470.

Barillot, C., D. Lemoine, et al. (1993). "Data Fusion in Medical Imaging - Merging Multimodal and Multipatient Images, Identification of Structures and 3d Display Aspects." European Journal of Radiology **17**(1): 22-27.

Bartlett, P. L. and S. Mendelson (2002). "Rademacher and Gaussian Complexities: Risk Bounds and Structural Results." Journal of Machine Learning Research **3**: 463-482.

Bates, D. W. (2002). "The quality case for information technology in healthcare." BMC Med Inform Decis Mak **2**: 7.

Bathula, D. R., X. Papademetris, et al. (2007). Level set based clustering for analysis of functional MRI data. 2007 4th IEEE International Symposium on Biomedical Imaging: From Nano to Macro - Proceedings.

Bemmel, J. H. v. and M. A. Musen (1997). Handbook of medical informatics. AW Houten, Netherlands, Bohn Stafleu Van Loghum.

Bengio, Y., O. Delalleau, et al. (2004). "Learning eigenfunctions links spectral embedding and kernel pca." Neural Computation **16**(10): 2197-2219.

Bernaards, C. A., M. M. Farmer, et al. (2003). "Comparison of Two Multiple Imputation Procedures in a Cancer Screening Survey." Journal of Data Science **1**.

Bishop, C. (1995). Neural Networks for Pattern Recognition, Oxford University Press.

Blanchard, G., O. Bousquet, et al. "Statistical Performance of Support Vector Machines."

Bordley, R. F. (1982). "A Multiplicative Formula for Aggregating Probability Assessments." MANAGEMENT SCIENCE **28**(10): 1137-1148.

Bovino, L., G. Dimauro, et al. (2003). "On the combination of abstract level classifiers." International Journal on Document Analysis and Recognition **6**(1): 42-54.

Brown, G. "An information theoretic perspective on multiple classifier systems.".

Brown, G. (2010). Ensemble Learning. Encyclopedia of Machine Learning. C. Sammut and G. I. Webb, Springer**:** 1-24.

Bruce, E. N. (2001). Biomedical signal processing and signal modeling. New York, John Wiley.

Burges, C. J. C., B. Schölkopf, et al. (1999). Advances in kernel methods support vector learning. Cambridge, Mass., MIT Press**:** vii, 376 p.

Burton, A. and D. G. Altman (2004). "Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines." British Journal of Cancer **91**(1): 4-8.

Celis, J. E., P. Gromov, et al. (2003). "Integrating proteomic and functional genomic technologies in discovery-driven translational breast cancer research." Mol Cell Proteomics **2**(6): 369-377.

Chang, A. P. F. (2005). Comparison of different fusion approaches for network intrusion detection using SVMs. Fourth International Conference on Machine Learning and Cybernetics, Guangzhou.

Chen, H., S. S. Fuller, et al., Eds. (2005). Medical Informatics. Knowledge management and data mining in biomedicine., Springer.

Chen, J. and J. Shao (2000). "Nearest neighbor imputation for survey data." Journal of Official Statistics **16**: 113-131.

Cristianini, N., J. Kandola, et al. (2002). "On kernel target alignment." Journal of Machine Learning Research **1**.

Cristianini, N. and J. Shawe-Taylor (2000). An introduction to Support Vector Machines and other kernel-based learning methods. Cambridge ; New York, Cambridge University Press.

Dasarathy, B. (1991). "Nearest neighbour (NN) norms: NN pattern classification techniques." IEEE Computer Society Press: 1-30.

Dempster, A., N. Laird, et al. (1977). "Maximum likelihood from incomplete data via the EM algorithm." Journal of the Royal Statistical Society **39**(1): 1–38.

Devos, A., L. Lukas, et al. (2004). "Classification of brain tumours using short echo time 1H MR spectra." Journal of Magnetic Resonance **170**(1): 164-175.

Dimou, I. and M. Zervakis (2008). Support Vector Machines versus Decision Templates in Biomedical Decision Fusion. Machine Learning and Applications, Fourth International Conference on.

Dimou, I. N., G. C. Manikis, et al. (2006). Classifier Fusion Approaches for Diagnostic Cancer Models. 28th Annual International Conference on Engineering in Medicine and Biology Society, New York, IEEE.

Dimou, I. N., I. Tsougos, et al. (2009). Classification of pathological human brain lesions using Magnetic Resonance Spectroscopy at 3T. World Congress on Medical Physics and Biomedical Engineering, September 7 - 12, 2009, Munich, Germany**:** 1368-1370.

Dimou, I., I. Tsougos, et al. (2009). Classification of 3T MRS spectra using Support Vector Machines. 9th International Conference on Information Technology and Applications in Biomedicine, Larnaca, Cyprus.

Duda, R. O., P. E. Hart, et al. (2001). Pattern classification. New York, Wiley.

Duin, R. P. W. (2000). PRTools - A Matlab Toolbox for Pattern Recognition**:** www.prtools.org.

Eberhart, R. C. and R. W. Dobbins (1990). Neural network performance metrics for biomedical applications. Computer-Based Medical Systems, 1990., Proceedings of Third Annual IEEE Symposium on.

Eichhorn, J. and O. Chapelle (2004). Object categorization with SVM: Kernels for Local Features, Max Planck Institute for Biological Cybernetics. **137**.

Ferranti, J. M., R. C. Musser, et al. (2006). "The clinical document architecture and the continuity of care record: a critical analysis." J Am Med Inform Assoc **13**(3): 245-252.

Ford, B. L. (1983). An overview of hot-deck procedures. Incomplete data in sample surveys. W. G. Madow, I. Olkin and D. B. Rubin. New York, Academic Press**:** 185-207.

Freitas, A., A. Costa-Pereira, et al. (2007). Cost-sensitive decision trees applied to medical data. Lecture Notes in Computer Science (including subseries Lecture

Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). **4654 LNCS:** 303-312.

Fukunaga, K. (1990). Introduction to Statistical Pattern Recognition. New York, Academic Press.

Fung, G. and J. Stoeckel (2007). "SVM feature selection for classification of SPECT images of Alzheimer's disease using spatial information." Knowledge and Information Systems **11**(2): 243-258.

Fung, G. M., O. L. Mangasarian, et al. (2003). Knowledge-based nonlinear kernel classifiers.

Genton, M. G. (2001). "Classes of Kernels for Machine Learning: A Statistics Perspective." Journal of Machine Learning Research **2**: 299-312.

Georgiadis, P., D. Cavouras, et al. (2008). "Improving brain tumor characterization on MRI by probabilistic neural networks and non-linear transformation of textural features." Comput. Methods Prog. Biomed. **89**(1): 24-32.

Ghosh, A. K. (2004). "On the challenges of using evidence-based information: The role of clinical uncertainty." Journal of Laboratory and Clinical Medicine **144**(2): 60-64.

Giacinto, G., F. Roli, et al. (2003). "Fusion of multiple classifiers for intrusion detection in computer networks." Pattern Recogn. Lett. **24**(12): 1795-1803.

Goebel, K. and W. Yan (2004). "Choosing Classifiers for Decision Fusion."

Graf, A. B. A. and S. Borer (2001). Normalization in Support Vector Machines. Pattern Recognition: 23rd DAGM Symposium, Munich, Germany, September 12-14, 2001. Proceedings**:** 277.

Haasdonk, B. (2005). "Feature space interpretation of SVMs with indefinite kernels." Pattern Analysis and Machine Intelligence, IEEE Transactions on **27**(4): 482-492.

Heijden, F. v. d. (2004). Classification, parameter estimation, and state estimation : an engineering approach using MATLAB. New York, Wiley.

Hernandez, T. and S. Kambhampati (2004). "Integration of biological sources: current systems and challenges ahead." SIGMOD Rec. **33**(3): 51-60.

Hosmer, D. W. and S. Lemeshow (2000). Applied logistic regression. New York, Wiley.

Huang, C. H., V. Lanza, et al. (2005). "HealthGrid - Bridging life science and information technology." Journal of Clinical Monitoring and Computing **19**(4-5): 259-262.

Huang, Y. S. and C. Y. Suen (1995). "Combination of multiple experts for the recognition of unconstrained handwritten numerals. , 17:90–94, 1995. 7." IEEE Trans. on Pattern Analysis and Machine Intelligence **17**: 90-94.

Huilin, X. (2007). "Data-Dependent Kernel Machines for Microarray Data Classification." IEEE/ACM Transactions on Computational Biology and Bioinformatics **4**: 583-595.

Jacobs, R. A. (1995). "Methods For Combining Experts' Probability Assessments." Neural Computation **7**(5): 867-888.

Jan Luts, J.-B. Poullet, et al. (2008). "Effect of feature extraction for brain tumor classification based on short echo time H MR spectra." Magnetic Resonance in Medicine **60**(2): 288-298.

Jean-Philippe Vert, Jian Qiu, et al. (2007). "A new pairwise kernel for biological network inference with support vector machines." BMC Bioinformatics **8**(8).

Jemal, A., R. Siegel, et al. (2008). "Cancer Statistics, 2008." CA Cancer J Clin **58**(2): 71-96.

Katehakis, D. G., M. Tsiknakis, et al. (2002). "Towards an Integrated Electronic Health Record - Current Status and Challenges, Business Briefing: Global Healthcare

2002. Business Briefing: Global Healthcare 2002." The Official Publication of the World Medical Association.
Kecman, V. (2001). Learning and soft computing : support vector machines, neural networks, and fuzzy logic models. Cambridge, Mass., MIT Press.
Kittler, J. (1998). "Combining classifiers: A theoretical framework." Pattern Analysis and Applications **1**(1): 18-27.
Kittler, J. (2000). A Framework for Classifier Fusion: Is It Still Needed? Advances in Pattern Recognition. F. Ferri, J. Iñesta, A. Amin and P. Pudil, Springer Berlin / Heidelberg. **1876:** 45-56.
Kittler, J. and F. M. Alkoot (2003). "Sum versus vote fusion in multiple classifier systems." IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(1): 110-115.
Kittler, J., M. Hatef, et al. (1998). "On Combining Classifiers." IEEE Transactions on pattern analysis and machine intelligence
Computer Society Press **20**(3).
Kittler, J. and K. Messer (2002). Fusion of multiple experts in multimodal biometric personal identity verification systems. Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on.
Klein, M. (2001). Combining and relating ontologies: an analysis of problems and solutions. Workshop on Ontologies and Information Sharing, IJCAI'01.
Kleinberg, E. M. (2000). A Mathematically Rigorous Foundation for Supervised Learning. Multiple Classifier Systems. First International Workshop, MCS 2000, Cagliari, Italy, Springer-Verlag.
Koscor, A. and L. Toth (2004). "Kernel-Based Feature Extraction with a Speech Technology Application." IEEE Transactions on Signal Processing **52**(8).
Kounelakis, M. G., M. E. Zervakis, et al. (2008). Identification of significant metabolic markers from MRSI data for brain cancer classification. BioInformatics and BioEngineering, 2008. BIBE 2008. 8th IEEE International Conference on.
Kousi, E., I. Tsougos, et al. (2009). Proton Magnetic Resonance Spectroscopy at 3T - Evaluation of Metabolic Profile of Human Brain Lesions. World Congress on Medical Physics and Biomedical Engineering, September 7 - 12, 2009, Munich, Germany**:** 335-337.
Kuncheva, L. (1993). "Two-level classification schemes in medical diagnostics." Int J Biomed Comput **32**(3-4): 197-210.
Kuncheva, L. I. (2002). "Switching between selection and fusion in combining classifiers: an experiment." IEEE Trans Syst Man Cybern B Cybern **32**(2): 146-156.
Kuncheva, L. I. (2002). "A theoretical study on six classifier fusion strategies." IEEE Transactions on Pattern Analysis and Machine Intelligence **24**(2).
Kuncheva, L. I. (2004). Classifier Ensembles for Chaning Environments. International Workshop on Multiple Classifier Systems.
Kuncheva, L. I. (2004). Combining pattern classifiers : methods and algorithms. New York, Wiley-Interscience.
Kuncheva, L. I. (2008). Classifier Ensembles: Facts, Fiction, Faults and Future (Plenary talk). 19th International Conference on Pattern Recognition (ICPR). Tampa, Florida.
Kuncheva, L. I., J. C. Bezdek, et al. (2001). "Decision templates for multiple classifier fusion: An experimental comparison." Pattern Recognition **34**(2): 299-314.

Kuncheva, L. I., C. J. Whitaker, et al. (2003). "Limits on the majority vote accuracy in classifier fusion." Pattern Analysis and Applications **6**(1): 22-31.
Lasko, T. A., J. G. Bhagwat, et al. (2005). "The use of receiver operating characteristic curves in biomedical informatics." Journal of Biomedical Informatics **38**(5): 404-415.
Leski, J. (2003). "A fuzzy if-then rule-based nonlinear classifier." International Journal of Applied Mathematics and Computer Science **13**(2): 215-224.
Li, Q. and J. S. Racine (2007). Nonparametric econometrics : theory and practice. Princeton, N.J., Princeton University Press.
Li, Y., R.-P. Yin, et al. (2005). A New Decision Fusion Method in Support Vector Machine Ensemble. ICML.
Li, Z., Z. Weida, et al. (2004). "Hidden space support vector machines." Neural Networks, IEEE Transactions on **15**(6): 1424-1434.
Liang, X. (2010). "Feature space versus empirical kernel map and row kernel space in SVMs." Neural Computing and Applications **19**(3): 487-498.
Lin, H.-T. and C.-J. Lin (2003). A Study on Sigmoid Kernels for SVM and the Training of non-PSD Kernels by SMO-type Methods, National Taiwan University.
Little, R. J. A. and D. B. Rubin (2002). Statistical analysis with missing data. New York, Wiley.
Lukas, L., A. Devos, et al. (2004). "Brain tumor classification based on long echo proton MRS signals." Artificial intelligence in medicine **31**(1): 73-89.
MacKay, D. J. C. (1995). "Probable networks and plausible predictions – a review of practical Bayesian methods for supervised neural networks." Network: Computation in Neural Systems **6**: 469-505.
Martin-Sanchez, F., I. Iakovidis, et al. (2004). "Synergy between medical informatics and bioinformatics: facilitating genomic medicine for future health care." J. of Biomedical Informatics **37**(1): 30-42.
Martin, L., E. Bonsma, et al. (2007). Data access and management in ACGT: tools to solve syntactic and semantic heterogeneities between clinical and image databases. Proceedings of the 2007 conference on Advances in conceptual modeling: foundations and applications. Auckland, New Zealand, Springer-Verlag**:** 24-33.
McGee, S. (2002). "Simplifying likelihood ratios." Journal of General Internal Medicine **17**(8): 646-649.
Mercer, J. (1909). "Functions of Positive and Negative Type, and their Connection with the Theory of Integral Equations." Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character **209**(441-458): 415-446.
Mika, S., G. Ratsch, et al. (1999). Fisher Discriminant Analysis With Kernels. IEEE Signal Processing Society Workshop.
Moons, K. G. M., R. A. R. T. Donders, et al. (2006). "Using the outcome for imputation of missing predictor values was preferred." Journal of Clinical Epidemiology **59**: 1092-1101.
Moosmann, M., T. Eichele, et al. (2008). "Joint independent component analysis for simultaneous EEG-fMRI: Principle and simulation." International Journal of Psychophysiology **67**(3): 212-221.
Moreno, P. J., P. P. Ho, et al. (2002). "A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications."

Murphy, J. R. (2004). "Statistical errors in immunologic research." Journal of Allergy and Clinical Immunology **114**(6): 1259-1263.
Newman, D. J., S. Hettich, et al. (1998). "UCI Repository of machine learning databases." from www.ics.uci.edu/~mlearn/),Irvine.
Oza, N. C. and K. Tumer (2008). "Classifier ensembles: Select real-world applications." Inf. Fusion **9**(1): 4-20.
Pearl, J. (1988). Probabilistic reasoning in intelligent systems : networks of plausible inference. San Mateo, Calif., Morgan Kaufmann Publishers.
Pérez, A., R. Dennis, et al. (2002). "Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia." Statistics in Medicine **21**: 3885-3896.
Perez, A., R. J. Dennis, et al. (2002). "Use of the mean, hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in Colombia." Statistics in Medicine **21**: 3885-3896.
Potamias, G. and V. Moustakis (2001). Knowledge Discovery from Distributed Clinical Data Sources: The Era for Internet-Based Epidemiology. 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Istanbul.
Qi, Z. G. and Y. X. Li (2008). "Lipid signal in evaluation of intracranial meningiomas." Chin Med J (Engl) **121**(23): 2415-2419.
Qingshan Liu, Hanqing Lu, et al. (2004). "Improving Kernel Fisher Discriminant Analysis for Face Recognition." *IEEE Transactions on Circuits and Systems for Video Technology* **14**(1): 42-49.
Raudys, S. (2002). invited talk. Multiple Classifier Systems conference.
Raudys, S. J. and A. K. Jain (1991). "Small sample size effects in statistical pattern recognition: Recommendations for practitioners." IEEE Transactions on Pattern Analysis and Machine Intelligence **13**(3): 252-264.
Rogova, G. (1994). "Combining the results of several neural network classifiers." Neural Networks **7**: 777-781.
Roses, A. D. (2000). "Pharmacogenetics and pharmacogenomics in the discovery and development of medicines." Novartis Found Symp **229**: 63-66; discussion 66-70.
Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York, J. Wiley & Sons.
Rubin, D. B. (1996). "Multiple imputation after 18+ years." Journal of the American Statistical Association **91**(434): 473-489.
Ruta, D. and B. Gabrys (2000). "An Overview of Classifier Fusion Methods." Computing and Information Systems **7**: 1-10.
Schafer, J. and J. Graham (2002). "Missing data: our view of the state of the art." Psychological Methods **7**: 147-177.
Schafer, J. L. (1997). Analysis of incomplete multivariate data, Chapman & Hall.
Schafer, J. L. and M. K. Olsen (1998). "Multiple imputation for multivariate missing-data problems: a data analyst's perspective." Multivariate Behavioral Research **33**: 545-571.
Schapire, R. E. (1999). "A Brief Introduction to Boosting." Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence.

Schneider, T. (2001). "Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values." Journal of Climate **14**(5): 853-871.

Schölkopf, B. and A. J. Smola (2002). Learning with kernels : support vector machines, regularization, optimization, and beyond. Cambridge, Mass., MIT Press.

Serocka, P. (2007). Visualization of High-Dimensional Biomedical Image Data Advances in Multimedia Information Processing, Springer.

Shipp, C. A. and L. I. Kuncheva (2002). "Relationships Between Combination Methods and Measures of Diversity in Combining Classifiers."

Shortliffe, E. H. and J. J. Cimino (2006). Biomedical informatics : computer applications in health care and biomedicine. New York, NY, Springer.

Soares, D. P. and M. Law (2009). "Magnetic resonance spectroscopy of the brain: review of metabolites and clinical applications." Clin Radiol **64**(1): 12-21.

Sotiriou, C. and M. J. Piccart (2007). "Taking gene-expression profiling to the clinic: when will molecular signatures become relevant to patient care?" Nat Rev Cancer **7**(7): 545-553.

Staelin, C. (2003). Parameter selection for support vector machines. Technical Report HPL-2002-354 (R. 1), HP Laboratories Israel.

Steele, A. (2002). Medical Informatics Around the World: An International Perspective Focusing on Training Issues, Universal Publishers.

Strehl, A. and J. Ghosh (2003). "Cluster Ensembles - a knowledge reuse framework for combining multiple partitions." Journal of Machine Learning Research **3**: 583-587.

Sujansky, W. (2001). "Heterogeneous database integration in biomedicine." J Biomed Inform **34**(4): 285-298.

Suykens, J. A. K., T. Van Gestel, et al. (2002). Least squares support vector machines. Singapore, World Scientific.

Taira, R. K., A. A. Bui, et al. (2002). "Identification of patient name references within medical documents using semantic selectional restrictions." Proc AMIA Symp: 757-761.

Taktak, A. F., A. C. Fisher, et al. (2004). "Modelling survival after treatment of intraocular melanoma using artificial neural networks and Bayes theorem." Phys Med Biol **49**(1): 87-98.

Thiran, P., J.-L. Hainaut, et al. (2005). Database Wrappers Development: Towards Automatic Generation. Proceedings of the Ninth European Conference on Software Maintenance and Reengineering, IEEE Computer Society**:** 207-216.

Timmerman, D., A. C. Testa, et al. (2005). "A new logistic regression model to distinguish between the benign and malignant adnexal mass before surgery: a multicenter study by the International Ovarian Tumour Analysis (IOTA) Group." Journal of Clinical Oncology **23**: 8794-8801.

Timmerman, D., L. Valentin, et al. (2000). "Terms, definitions and measurements to describe the ultrasonographic features of adnexal tumors: a consensus opinion from the international ovarian tumor analysis (IOTA) group." Ultrasound in Obstetrics and Gynecology **16**: 500-505.

Timmerman, D., B. Van Calster, et al. (2007). "Inclusion of CA-125 does not improve mathematical models developed to distinguish between benign and malignant adnexal tumors." Journal of Clinical Oncology **25**: 4194-4200.

Timmerman, D., B. Van Calster, et al. (2007). "Inclusion of CA-125 does not improve mathematical models developed to distinguish between benign and malignant adnexal tumors." J Clin Oncol **25**(27): 4194-4200.

Tumer, K. and J. Ghosh (1996). "Analysis of decision boundaries in linearly combined neural classifiers." Pattern Recognition **29**: 341-348

Valentini, G. and F. Masulli (2002). Ensembles of Learning Machines. Neural Nets. M. Marinaro and R. Tagliaferri, Springer Berlin / Heidelberg. **2486:** 3-20.

Van Calster, B., D. Timmerman, et al. (2007). "Preoperative diagnosis of ovarian tumors using Bayesian kernel-based methods." Ultrasound in Obstetrics and Gynecology **29**: 496-504.

Van der Graaf, M. (2009). "In vivo magnetic resonance spectroscopy: basic methodology and clinical applications." Eur Biophys J.

Van der Heijden, G. J. M. G., A. R. T. Donders, et al. (2006). "Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example." Journal of Clinical Epidemiology **59**: 1102-1109.

Van Gestel, T., J. A. K. Suykens, et al. (2001). Automatic relevance determination for least squares support vector machine regression. Proceedings of the International Joint Conference on Neural Networks.

van Otterloo, P. J. and I. T. Young (1978). "A distribution-free geometric upper bound for the probability of error of a minimum distance classifier." Pattern Recognition **10**(4): 281-286.

Vapnik, V. (1995). The Nature of Statistical Learning Theory. N.Y., Springer.

Vergote, I., J. De Brabanter, et al. (2001). "Prognostic importance of degree of differentiation and cyst rupture in stage I invasive epithelial ovarian carcinoma." Lancet: 176-182.

W. Dean Bidgood, S. C. Horii, et al. (1997). "Understanding and Using DICOM, the Data Interchange Standard for Biomedical Imaging." Journal of the American Medical Informatics Association **4**: 199-212.

Wang, W. (2003). "Determination of the spread parameter in the Gaussian kernel for classification and regression." Neurocomputing **55**(3-4): 643-663.

Wang, Y. and H. Zhang (2001). Content-based image orientation detection with support vector machines. Content-Based Access of Image and Video Libraries, 2001. (CBAIVL 2001). IEEE Workshop on.

Xu, L., A. Krzyzak, et al. (1992). "Methods of combining multiple classifiers and their application to handwritting recognition." IEEE Trans Syst Man Cybern B Cybern **22**: 418-435.

Xu, L., A. Krzyzak, et al. (1992). "Methods of combining multiple classifiers and their applications to handwriting recognition." Systems, Man and Cybernetics, IEEE Transactions on **22**(3): 418-435.

Zhou, S. K. and R. Chellappa (2006). "From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing Hilbert space." Pattern Analysis and Machine Intelligence, IEEE Transactions on **28**(6): 917- 929.

Zhou, W., L. Zhang, et al. (2006). Hidden space principal component analysis. Singapore. **3918:** 801-805.

Index