Advanced Ensemble Methods for Automatic Classification of ^{1}H-NMR Spectra

Dissertation

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

der Technischen Universität Dortmund
an der Fakultät für Informatik
von

Kai Lienemann

Dortmund

2010

Tag der mündlichen Prüfung. 20.05.2010

Dekan: Prof. Dr. Peter Buchholz

Gutachter:
Prof. Dr.-Ing. Gernot A. Fink
Prof. Dr. Claus Weihs

ERKLÄRUNG

Gemäß § 8 Abs. 3 der Promotionsordnung der TU Dortmund für den Fachbereich Informatik (Stand: 26.11.2003) wird im Folgenden die Beteiligung von Kai Lienemann an in Kooperation erzielten Ergebnissen beschrieben.

Die vorliegende Dissertation ist während einer Kooperation der TU Dortmund mit Boehringer Ingelheim Pharma GmbH & Co. KG (BI) enstanden. Die Planung und Durchführung der benötigten Tierversuche, die Entnahme von Urinproben und die toxikologische Beurteilung aller Proben wurden in der Gruppe *General Pharmacology* (Abteilung *Drug Discovery Support*) der Firma BI durchgeführt. Die von diesen Proben gemessenen NMR Spektren und die toxikologische Beurteilungen wurden Kai Lienemann zur Durchführung seiner Dissertation zur Verfügung gestellt.

Diese Daten dienen als Grundlage für die Untersuchung und Entwicklung von Verfahren zur Verbesserung der Qualität von NMR Spektren und der automatisierten Vorhersage von Organschädigungen. Die in der vorliegenden Dissertation beschriebenen Untersuchungen und entwickelten Verfahren sind Eigenanteil von Kai Lienemann.

ACKNOWLEDGMENTS

Such as many things in life, also this work would not have been possible without the support of many people. First of all I would like to thank my supervisor Gernot A. Fink for offering me the possibility to write my thesis in his group. Although the type of data was quite new to him, he was always able to give new ideas for further developments. I'm very grateful to Claus Weihs who agreed to review my thesis and made valuable comments about my work.

Two people had a great influence on this thesis and also on my further life. I'm very grateful to Thomas Plötz, who supported me all the time, even when paper discussion took quite a long time and his schedule was tight. Sabine Pestel from Boehringer Ingelheim Pharma GmbH & Co. KG was always a great help when I had various questions on the data. Furthermore, although pattern classification approaches where quite new to her, she always showed a lot of interest in my work and was willing to understand even the details. Thomas and Sabine, I have learned a lot from both of you regarding scientific topics and beyond them. You two made the last years an unforgettable time and I'm happy to call you friends.

I would also like to thank Andreas Weith, Karsten Quast and Brian Guth from Boehringer Ingelheim Pharma GmbH & Co. KG for supporting me and the project all the time. I was able to present my work at the company and at many conferences, which were one-time experiences.

This work also needed a lot patience from my colleagues in the Intelligent Systems group at the Dortmund University of Technology when I had several experiments running on their machines. Thanks for supporting me all the time even in stressful times. Also thanks for proof reading my thesis and significantly improving the quality of this work. A special thanks goes to Jan Richarz, it were great three years with you in one bureau and beyond that.

A further thanks goes to the staff of the laboratory at Boehringer Ingelheim Pharma GmbH & Co. KG responsible for the collection of urine samples. Thanks a lot for the cordially receptions which made me always happy to visit all of you.

On the personal side, I would like to thank my parents, my brother and my friends for supporting me all the time. Also the work on this thesis was quite time consuming and sometimes exhausting, my wife Yvonne Lienemann emotionally supported me in every situation.

CONTENTS

ACRONYMS

ARSS	adapted random subspace sampling
ANN	artifical neural network
CART	classification and regression tree
CLOUDS	classification of unknowns by density superimposition
COMET	Consortium for Metabonomic Toxicity
COW	correlation optimized warping
DP	decision profile
DRC	drug related compound
DSS	2,2-dimethylsilapentane-5-sulfonic acid
DT	decision template
FID	free induction decay
GA	genetic algorithm
GFHT	generalized fuzzy Hough transform
HGA	hybrid genetic algorithm
HWHM	half-width at half-maximum
*k*NN	*k* nearest neighbor
KWV	Kohavi-Wolpert variance
LOO	leave-one-out
MC	Matthews Correlation Coefficient
MLP	multilayer perceptron
NCE	new chemical entity
NIR	near infrared
NMR	Nuclear Magnetic Resonance
NN	nearest neighbor
PARSE	peak alignment using reduced set mapping
PC	Principal Component

PCA	Principal Component Analysis
PCR	Principal Component Regression
PRESS	predicted residual sum of squares
PLS	Partial Least Squares
PLSR	Partial Least Squares Regression
PNN	probabilistic neural network
RBF	radial basis function
RSS	Random Subspace Sampling
SBS	sequential backward selection
SBFS	sequential backward floating selection
SFFS	sequential forward floating selection
SFS	sequential forward selection
SIMCA	soft independent modelling of class analogy
SMART	scaled to maximum aligned reduced trajectories
SNV	standard normal variate
SROI	spectral region of interest
SV	support vector
SVD	singular value decomposition
SVM	support vector machine
VAST	variable stability scaling

1 INTRODUCTION

Since the famous clinical trial done by James Lind in 1753 on the treatment of scurvy (reported in [Troh 03]), the assessment of pharmaceutical adverse effects has become a major issue in industrial drug development. Avicenna (980-1073) first introduced rules for the experimental testing of new pharmaceuticals in 1025 AD in his book *The Canon of Medicine*. This early work is still the basis of modern clinical trials in safety pharmacology for the analysis of new drugs [Brat 00].

adverse effects

safety pharmacology

The testing of a new chemical entity (NCE) starts with preclinical *in vitro* and *in vivo* tests. In vitro tests aim at the investigation of intended or unexpected interactions of the NCE in a controlled environment. After the efficacy is proven by in vitro tests and no undesired interactions could be observed, pharmacological and toxicological effects are tested by in vivo tests in animal studies. The organism used for testing is chosen in order to achieve results that can be transferred to the human organism. Rats are a typical example of an organism chosen for early studies in drug development. These studies are strictly regularized for ethical reasons and an important goal in industrial drug design is to reduce the number of animal experiments for ethical and cost reasons.

preclinical in vivo and in vitro studies

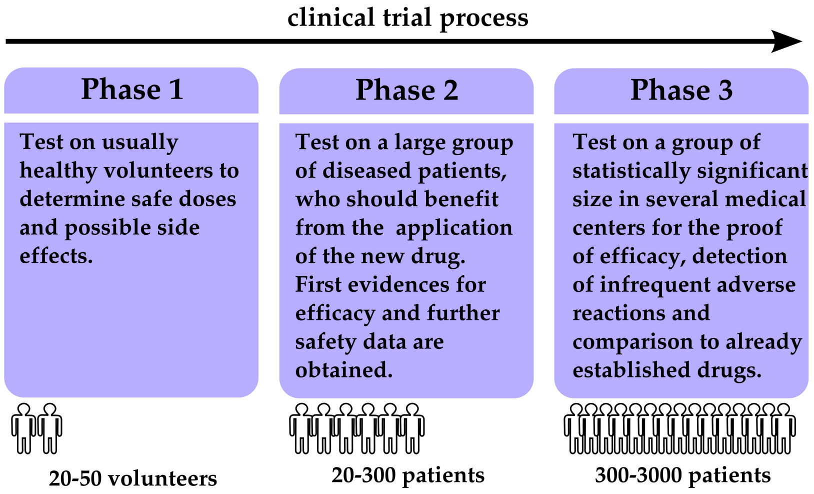
Reliable methods for the detection of adverse effects in preclinical trials are required for the reduction of animal testings and to allow for safe clinical testings of NCEs in humans. These clinical studies are usually performed in a stepwise procedure as shown in figure 1.1. Trials pursued in the first phase aim at the detection of adverse effects and the definition of the appropriate dose by the application of the NCE to usually healthy volunteers (20-50). Compounds showing no toxic adverse effects in the first phase are tested on a larger group of diseased patients (20-300) who should benefit from the pharmaceutical compound. Thereby, the efficacy is determined and adverse effects can be studied on a larger group of patients. Drug candidate failure usually occurs in this stage due to a low efficacy or toxic adverse effects. The drug candidate is applied to a group of significant size (300-3 000 volunteers) in different medical centers for the final determination of effectiveness and adverse effects. The third phase is the longest and most expensive one, and an approval for the marketing of the NCEs passing this phase is given by regulatory agencies.

clinical trials in three phases

determination of efficacy and adverse effects

Passing preclinical and clinical trials can last several years and costs between 500 and 2 000 million US dollars (cf. [Adam 06,

drug development costs up to 2 000 million US dollars



Test on usually healthy volunteers to determine safe doses and possible side effects.

Test on a large group of diseased patients, who should benefit from the application of the new drug. First evidences for efficacy and further safety data are obtained.

Test on a group of statistically significant size in several medical centers for the proof of efficacy, detection of infrequent adverse reactions and comparison to already established drugs.

Figure 1.1: Phases in clinical trials (cf. [DiMa 91]).

early detection of adverse effects

DiMa 03]). Candidate drug failure can occur even after the product release stage and endanger the safety of numerous patients [Kola 04]. Thus, an early detection of vital adverse effects, which is the main reason for drug withdrawal from the worldwide pharmaceutical market [Fung 01], can improve industrial drug development. Thereby, a faster drug design leads to safer pharmaceuticals and reduces the costs of their development.

current approaches: histopathology and interpretation of results from clinical chemistry

expert knowledge required

In order to prevent drug withdrawal in late phases of the drug development process, toxic adverse effects have to be reliably detected in preclinical trials. Current approaches are based on the visual examination of tissue samples by histopathology, or the analysis of the composition of biofluids such as blood or urine. While organ toxicities are identified by an expert in histopathology by visual examination based on abnormal shapes and structures visible under the microscope, analysis results concerning the composition of biofluids have to be interpreted by an expert with specific background knowledge. A small number of biofluid ingredients are known, which show a change in concentration in relation to the health status of a particular organ due to leakage of cell content, or changes in the metabolic activity. However, this procedure is dependent on the knowledge of threshold values for the *biomarker* substances. In order to give support in safety pharmacology, new analysis techniques independent of expert knowledge and a high sensitivity after single administration are required.

Metabonomics

New approaches for the detection of organ toxicities based on spectroscopic measurements of biofluids have been presented in recent years in the field of *Metabonomics*. Developments in Nuclear Magnetic Resonance (NMR) instrument and measurement techniques have led to highly reproducible spectroscopic data in a high resolution. Therefore, several institutions used NMR spectra of biofluids for the identification of particular effects in the metabolism induced by application of pharmaceutical compounds. NMR spectra of urine samples represent the concentrations of several molecules contained in the samples as a spectral profile composed of a multitude of spectral signals denoted as peaks.

NMR spectroscopy

The analysis of the urine composition allows for the collection of multiple samples from the same animal at different points in time, thereby reducing the amount of animals used in the studies. Changes in concentration are indicated by different intensities of the same peak in several spectra, thereby allowing for the detection of changes of the urine composition. However, the relevance of each peak for the detection of organ toxicities is not known and has to be determined with respect to a set of spectra with known classification as being non-toxic and toxic.

spectral profile represents sample composition

importance of spectral signals not known

Generally, methods from the field of pattern recognition can be applied in order to achieve a classification of spectra as being non-toxic or toxic. Data sets used in metabonomic applications are rather small due to the expensive measurement procedure. Furthermore, each spectrum contains approximately 1 000 peak signals. Thus, robust classification approaches achieving a reliable classification even in case of sparse and complex data sets have to be applied.

application of pattern recognition approaches

Ensemble methods are an emerging technique from the field of pattern recognition showing competitive classification results for several applications. For example, the Knowledge Discover and Data Mining (KDD) Cup is an annual competition for the development of the best classification approach to different applications. Algorithms based on ensemble methods won the first price in all categories in the year 2005, and the best approaches in at least one category were based on ensemble methods in the following years. Also the winner of the Netflix Prize is based on ensemble methods, improving by 10 % the prediction accuracy of how much a person likes a film based on their known movie preferences. The general approach of ensemble systems is to define a more accurate and robust classification system by combining several (simple) classifiers instead of relying on a single one. Due to the effectiveness of ensemble methods and the possibility of defining a classification system design for a particular problem, the application of ensemble methods is promising for assistance in drug development.

ensemble methods

competitive classification results for many applications

1.1 FOCUS

Interpretation of NMR spectra from urine samples as high-dimensional feature vectors allows for the application of pattern recognition approaches. Thereby, a classification of samples be-

development of a robust classification system

ing non-toxic or toxic for a particular organ can be achieved. Changes in peak intensity are expected to be present in the same dimension of all samples, thereby an inference on changes in concentration of specific molecules is achieved. This theoretical assumption does not completely hold in practice due to different sources modifying the exact position and intensity of peaks. These perturbations are reduced in this thesis by methods established in metabonomic applications. Thus, this thesis aims at the development of a classification system, which achieves a high classification accuracy even in case of sparse, complex and noisy data sets.

focus on ensemble methods for classification system design

Ensemble methods have shown in several applications competitive classification results in comparison to alternative approaches. Furthermore, ensemble combination and aggregation techniques can be designed for the particular application, thereby respecting particular characteristics of a given data set. Due to these advantages, the focus of this thesis is on the development and application of ensemble methods for the automatic classification of NMR spectra as being non-toxic or toxic for a particular organ. Thereby, an improved classification performance is the main goal of this thesis. However, also the interpretation of ensemble classifications is investigated in order to gain additional information on the grade of a detected toxic effect or the detection of peak signals mainly relevant to classification.

1.2 ORGANIZATION OF THE THESIS

related work from the field of metabonomics and ensemble systems

This thesis is an interdisciplinary work, applying methods from the field of pattern recognition for metabonomic applications in order to support drug development in industrial studies. Thus, chapters 2 and 3 give an overview of related work from metabonomic applications and ensemble theory in the first part of this thesis, respectively. The first two sections of chapter 2 deal with established approaches for the detection of adverse effects in safety pharmacology, and the fundamentals of NMR spectroscopy. In the following two sections, common approaches for the interpretation and classification of metabonomic data sets are discussed. In chapter 3 the focus is on the general concept of multiple classifier systems and common ensemble approaches.

materials and methods

development of advanced ensemble methods

The data sets and evaluation approaches presented in chapter 4 form the basis for the development and evaluation of advanced ensemble systems in the second part of this thesis. The general concept of multiple classifier systems is applied in an ensemble approach presented in the first part of chapter 5. Common ensemble methods are used in this approach, but the main potential of ensemble systems for metabonomic applications is the design of a multiple classifier system for the particular characteristics of the data. Therefore, two further approaches are presented, mainly aiming at an improvement of classification performance by focusing on spectral regions, which are most relevant to classification. These spectral regions are determined in an automated way, requiring no background knowledge on the relevance of signals in the spectrum. The classification performance and interpretability of the newly developed ensemble approaches is investigated in an experimental evaluation presented in section 6.

experimental evaluation

The thesis concludes with a summary in chapter 7, where the main achievements of this thesis are discussed. Finally, possible starting points for a further improvement of the presented ensemble approaches are discussed.

2 ANALYSIS OF METABOLIC RESPONSES

drug design

The design of new chemical entities (NCEs) having an effect on a biological target relevant in an investigated disease is the primary goal in industrial drug design. While chemical reactions between the NCE and the biological target can be studied in vitro, the overall effect on the target organism and possible adverse effects have to be studied in vivo.

adverse effects

safety pharmacology

Any pharmaceutical that has been applied orally, intravenously or in any other administration form is subject to further transportation and metabolic decomposition in the organism. Dependent on the applied product, transportation is achieved in the blood circulation or digestion system until the product is finally excreted from the body. While the product remains in the body it is further decomposed by enzymatic reactions. The product by itself or metabolic byproducts interact in various ways with organs and other molecules. These interactions can be intended in drug design in order to heal an illness or reduce its symptoms, or could have negative affects on the organism – commonly referred to as adverse affects. Since generally neither adverse affects of new substances nor enzymatic reactions and transportation systems of an organism are completely known, adverse affects of a new NCE have to be identified during experiments with model organisms. Organ toxicities, as one of the most frequent adverse effects of nowadays' pharmaceuticals, are detected during industrial drug design by safety pharmacology methods. These methods either detect organ toxicities by visual judgment in histopathology or analysis of metabolic changes reflected in the composition of biofluids such as blood or urine.

Metabonomics

biomarker identification

An emergent technology in the field of safety pharmacology is *Metabonomics*, defined as "the quantitative measurement of the time-related multiparametric metabolic response of living systems to pathophysical stimuli or genetic modification" [Nich 99]. Metabonomics is usually applied for the analysis of spectroscopic data from biofluids in order to give support to the interpretation of drug-induced metabolic changes. This analysis can be performed on different levels: a) detection of an induced organ toxicity, b) identification of the affected organ(s), or c) determination of metabolites relevant to the detection of a potentially observed organ toxicity. While the first two aspects give information on the applied NCE, the detection of metabolites indicating an induced organ toxicity (*biomarker*) provides a deeper insight into the underlying biological mechanisms. This knowledge can be used in further studies and organ toxicities can be detected by measurement of these identified metabolites.

In this chapter an overview of approaches for the detection of adverse effects in safety pharmacology and fundamental methods used in Metabonomics will be given. The main focus of this thesis is on the development of a classification system for the automatic classification of spectroscopic data using pattern recognition methods. Therefore, pharmacological and spectroscopic methods will be explained according to their basic principles. The following explanations will be sufficient to understand the concepts and relevant aspects of the methods, but for a detailed description key references will be given in the respective sections.

In section 2.1 two basic approaches for the identification of adverse effects in safety pharmacology will be outlined. First, the measurement and interpretation of metabolite concentrations in biofluids will be shown, followed by histological analysis of biosamples. The basic principles of NMR spectroscopy will be described in section 2.2, giving a deeper insight in the way of representing information on molecules in spectroscopic data and steps necessary for data preparation. Different sources can modify the data samples, and approaches to a reduction of these perturbations will be explained in section 2.3. Metabonomics has already been used for the interpretation of several data sets by multivariate data analysis methods. An overview of these methods will be given in section 2.4, followed by a summarizing section.

2.1 DETECTION OF ADVERSE EFFECTS IN SAFETY PHARMACOLOGY

The design of NCEs to influence biological targets in order to cure a disease or alleviate its symptoms is already a challenging task. But NCEs can also negatively effect other organs or metabolic reactions and these *adverse effects* have to be identified by methods from the field of safety pharmacology. Thus, NCEs are tested in preclinical and clinical trials to determine the efficacy and adverse effects in the target organisms.

preclinical and clinical trials

A NCE is administered within preclinical trials to experimental animals in a certain dose over a defined time period. Analysis of biofluids, which can be collected during the whole experiment, allows for the detection of metabolic changes by measurement and interpretation of certain molecules' concentrations. Furthermore, all organs can be investigated afterwards by histopathology and abnormal modifications of distinct organic regions can be detected.

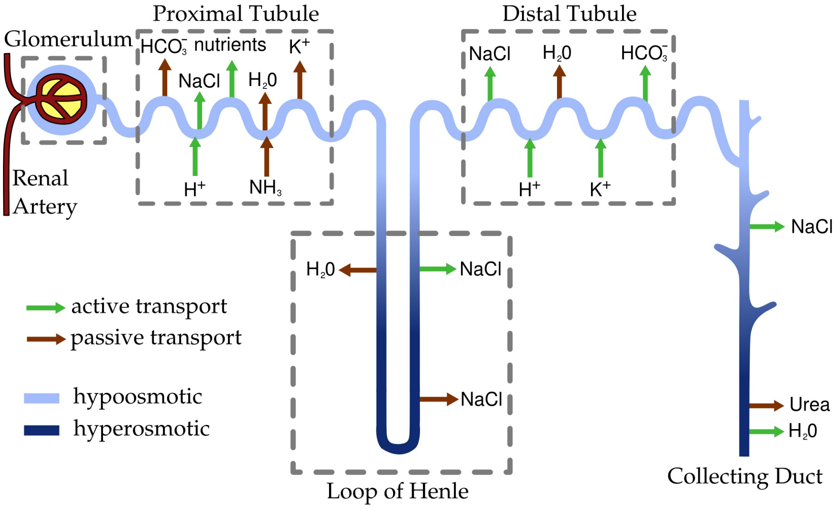


Figure 2.1: Structure of the nephron and functions of different segments of the transport epithelium (adopted from [Camp 03], p. 1136).

2.1.1 *Analysis of Metabolites by Clinical Chemistry*

An induced organ toxicity will generally affect the organ's functionality and the metabolic reactions it is involved in. Therefore, specific changes in the metabolite concentrations can be used as a marker for the functionality of distinct organs or even organ regions. These changes are expected to be recognizable by analysis of metabolite, enzyme or electrolyte concentrations in biofluids like urine or blood, simultaneously leading to an overview of the viability of several organs.

determination of organ viability by biofluid analysis

The functionality of the kidney, as an important organ for the regularization of urine composition, will be briefly summarized in the following part based on the explanations in [Camp 03]. Subsequently, the effect of an induced organ toxicity and its use for the detection of organ toxicities based on the analysis of metabolite concentrations will be presented.

Kidney Functionality

The kidneys of vertebrates play a key role in the osmoregulation and excretion. Essential ingredients like salt and water are extracted from the blood, a vital prerequisite for organisms living outside of the water to prevent dehydration. Each of the two human kidneys consists of tenuous blood vessels, collection ducts and around one million *nephrons* (cf. figure 2.1) – the functional unit of the kidney.

nephron

Water and solute up to a molecular weight of 5 000 Dalton permeate through particular cells in the glomerulum from the blood to the proximal tubule, forming the so-called ultrafiltrate.

glomerulum

proximal tubule

Blood cells and proteins remain in the blood because of their high molecular weight. The proximal and distal tubule are important segments for the regulation of the urine composition. Water and salt are filtered by reabsorption and the acid-base balance is adjusted by the secretion of protons (H^+) and the reabsorption of bicarbonates (HCO_3^-). Water is further extracted within the water-permeable descending limb of the loop of Henle due to the increased osmolarity of the hyperosmotic interstitial fluid. Salt is extracted by active and passive transport in the ascending limb, leading to a further dilution of the ultrafiltrate. Further reabsorption and secretion is carried out in the distal tubule and the ultrafiltrate is led into the collecting duct. The ultrafiltrate from several nephrons is further modified in the collecting duct by active transport of salt, and passive transport of water and urea into the core of the nephron. Finally, the urine is collected in the urinary bladder and excreted.

distal tubule

loop of Henle

collecting duct

Regional Markers for Kidney Damage

The kidney is divided into functional segments with a certain role in the regularization of urine composition as shown in the previous section. The cells of each region have a certain enzymatic composition according to their regulative function. A regional cell damage, one of the different types of organ toxicities, causes a decreased efficiency of metabolic reactions or a release of the cell content, leading to a change of the urine composition.

cell damage

An illustration of an induced cell damage in the region of the proximal tubule by an applied pharmaceutical is shown in figure 2.2. Apoptosis of epithelial cells leads to cellular desquamation and accumulation of cellular debris in the ultrafiltrate, increasing the concentration of enzymes specific for the proximal tubule cells (e.g. NAG, LDH, . . .) in the urine. Gaps between epithelial cells are covered by neighboring epithelial cells during recovery and the amount of cellular debris in the ultrafiltrate is decreased.

increasing enzyme concentration

Changes of urine ingredients' concentrations can be quantified in clinical chemistry by using photometric measurement methods in an automated assay. Concentrations of several sample ingredients can be quantified by the addition of chemicals, inducing a colorization of the sample, which indicates the concentration of a specific molecule and can be measured by a photometer. Certain changes in urine composition serve as markers for the damage of distinct organs or even organ regions and can be identified by an expert. Thus, a non-invasive monitoring of different organs' functionality is possible throughout an experiment, allowing for the detection of a drug-induced organ toxicity at different points of time.

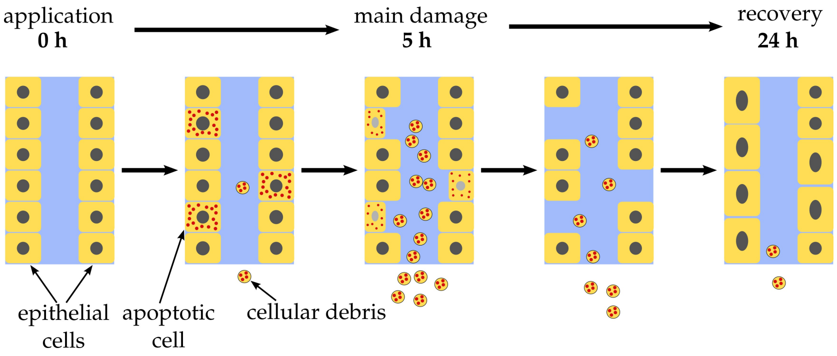


Figure 2.2: Illustration of a drug-induced organ damage. Apoptotic cells desquamate in the period of main damage and the gaps are filled with epithelial cells during recovery.

2.1.2 *Histopathological Analysis of Biosamples*

In addition, histopathology is a valuable technique for the detection of abnormal modifications of organs or distinct organic regions. Therefore, tissue samples are collected and prepared by histology procedures to be examined under a microscope by a qualified specialist. Different cellular components can be distinguished by staining procedures, leading to a certain color for cellular components or substances within the tissue.

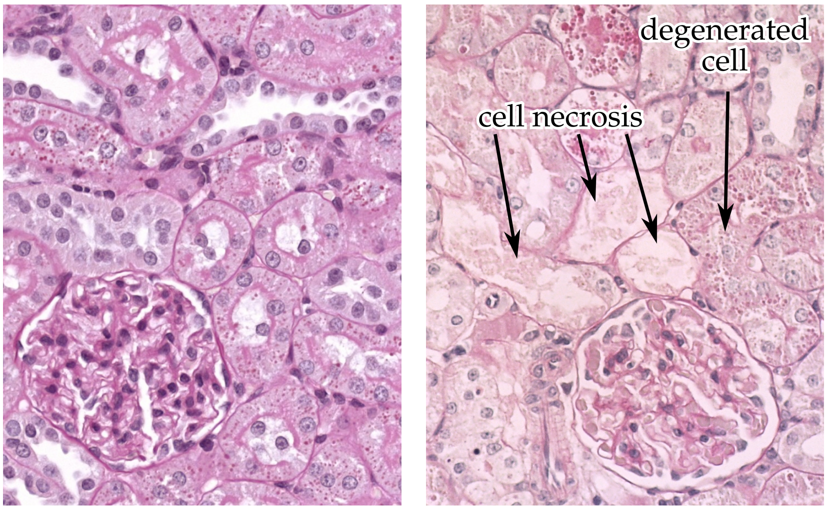
visual investigation of tissue samples

Histopathological images from a normal and damaged rat kidney are shown in figure 2.3. The damaged kidney originates from a rat treated over 7 days with a known nephrotoxicant. The circular glomerulum is visible in the lower part and the remaining shapes are cuts through the tubule. The bright center of the tubule are the interior for the transportation of the urine and the darker outer parts are the epithelial cells.

Three bright tubule cuts with nearly invisible epithelial cells are located in the center of figure 2.3b, indicating a cell necrosis. A degenerated tubule with an abnormal shape, size and staining is visible in the right part of figure 2.3b. These cell damages seriously affect the kidney functionality and the urine composition.

Histopathology is an important tool for the diagnosis of several diseases (e.g. cancer), but the application in preclinical or clinical trials has its disadvantages. Tissue samples or cells have to be removed by an invasive procedure, thereby producing stress for the subject and influencing the further progress of the test. Animal studies require euthanasia prior to biopsy, impeding the creation of time-series data without the use of a significant amount of experimental animals. The amount of experimental animals used in preclinical studies is also an important aspect and strictly controlled by ethical regulations.

invasive sample collection



(a) Healthy kidney

(b) Damaged kidney

Figure 2.3: Histopathology of a (a) healthy kidney and (b) kidney with cell necrosis and degeneration after a daily treatment with a nephrotoxicant over a period of 7 days (courtesy of Dr. Thomas Nolte, Boehringer Ingelheim Pharma GmbH & Co. KG).

2.2 FUNDAMENTALS OF ^{1}H-NMR SPECTROSCOPY

nuclear magnetic resonance spectroscopy

An emerging technology for the analysis of biofluids for a better understanding of metabolic processes in living systems is the NMR spectroscopy. Samples are exposed to a strong magnetic field and an additional external magnetic field with different orientation is applied. Several molecules and their concentrations in the sample can be identified with respect to their specific reaction of certain molecule's atoms.

non-selective measurement

highly reproducible

This non-invasive method allows for the simultaneous detection and quantification of various molecules without the prior definition of molecules to be measured. A further advantage of the NMR spectroscopy is its high reproducibility in case of controlled experimental conditions [Duma 06]. Keun et al. have shown the similarity of spectroscopic data, even when the same sample is measured in different institutions [Keun 02]. This is a major prerequisite, since even small spectral differences are used in metabonomic studies in order to analyze certain properties of the samples. Thus, the main variance in the data has to originate from biology and not from the measurement process. Otherwise, relevant information could be falsified and degrade a subsequent classification procedure.

Generally, all chemical elements with an odd number of protons or neutrons could be measured within NMR experiments. But specific elements are favorable due to their high concentration in the sample or an increased sensitivity in the measurement

ISOTOPE	ABUNDANCE	QUANTUM NUMBER	GYROMAGNETIC RATIO
^{1}H	99.985	$\frac{1}{2}$	26.75
^{12}C	98.9 %	0	-
^{13}C	1.108 %	$\frac{1}{2}$	6.73
^{16}O	99.96 %	0	-
^{17}O	0.037	$\frac{5}{2}$	−3.63

Table 2.1: Properties of isotopes with putative relevance for the measurement of NMR spectra from biological samples (adopted from [Frie 99], page 3).

process. The most frequent elements in samples from biology are carbon (C), oxygen (O) and hydrogen (H). However, among these elements hydrogen is the only one measurable in NMR spectroscopy due to its odd number of protons. Although isotopes of C and O exist with an odd number of protons or neutrons, they are quite rare in nature as shown in table 2.1.

Hydrogen atoms are widely distributed in chemical compounds to be analyzed in medical investigations and have a very strong nuclear magnetism. Thus, ^{1}H-NMR spectroscopy (henceforth denoted as NMR spectroscopy) will be the sole NMR spectroscopic method discussed in this thesis. The following explanations on the basic principles of NMR spectroscopy are based on [Free 03, Frie 99, Schw 96].

hydrogen NMR spectroscopy

2.2.1 *Nuclear Spin in a Magnetic Field*

Hydrogen atoms are made up of a nucleus and a surrounding electron, whereby the nucleus accounts for the main mass. One very important property of the nucleus is its spinning movement around a fixed axis. Thereby, an electric current and a magnetic field is induced as shown in figure 2.4a. The magnetic moment μ is defined according to the isotope-specific spin I, the gyromagnetic ratio γ and Planck's reduced constant $\hbar$ by

nuclear spin

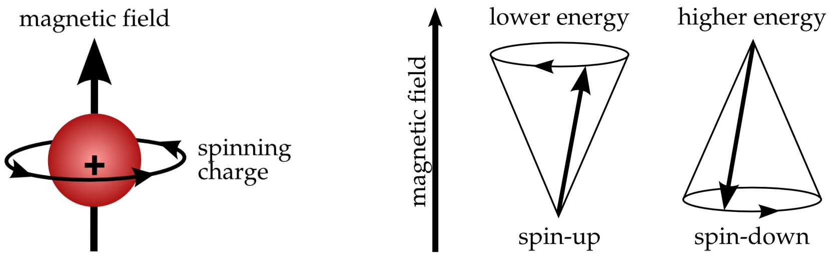
induced magnetic field

$$\mu = \gamma\sqrt{I(I+1)}\hbar .$$

Isotopes with a high γ have a high detection sensitivity within the NMR measurement process. The orientation and magnitude of the angular momentum is quantized according to the isotope-specific quantum number m (cf. table 2.1). Hydrogen has a spin of $\frac{1}{2}$ and m can take only the values $m = -\frac{1}{2}$ or $m = +\frac{1}{2}$.

Placing the sample into a strong magnetic field leads to a quantized orientation of the angular momentum in parallel or antiparallel to the applied magnetic field direction (cf. figure

angular momentum orientation



(a) Spinning nucleus.

(b) Orientations of the hydrogen nucleus in a magnetic field (adopted from [Free 03]).

Figure 2.4: Magnetism of nuclei and their reaction to an external magnetic field.

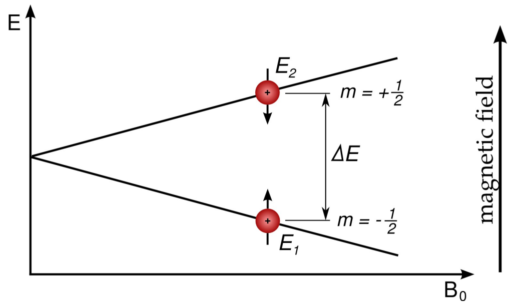


Figure 2.5: Correspondence between magnetic field strength B_0 and resonance energy ΔE for a state transition between the low-energy state E_1 and the high-energy state E_2.

2.4b). The parallel spin-up orientation has a lower energy level than the antiparallel spin-down orientation. A transition between these two states requires an energy of ΔE, which is dependent on the strength of the applied magnetic field B_0 (cf. figure 2.5). The lower energy level is slightly favored, leading to a macroscopic magnetism into the direction of the applied magnetic field. Actually, the spins do not align exactly to the magnetic field, but precess in the spin-up and spin-down direction on the surface of a cone. The precession frequency ν_L, also known as *Lamor frequency*, is proportional to B_0 by

Lamor frequency

$$\nu_L = \left|\frac{\gamma}{2\Pi}\right| B_0.$$

2.2.2 *Nuclear Magnetic Resonance and Relaxation*

The basic concept of NMR measurements is the nuclear magnetic resonance, which is based on the interaction of precessing nuclei in a strong magnetic field with an additional magnetic field. This phenomenon was first described and measured in the year 1938 by Isidor Rabi [Rabi 38] and was further refined by several

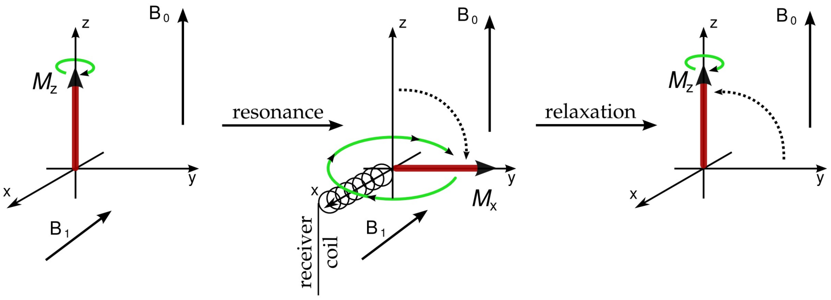


Figure 2.6: Spin resonance and relaxation during a NMR measurement.

researchers. Their work lead to a Nobel price in physics in the year 1952 for Felix Bloch and Edward Mills Purcell, a Nobel Price in chemistry for Richard R. Ernst in the year 1991, and John B. Fenn, Koichi Tanaka and Kurt Wüthrich in the year 2002.

nuclear magnetic resonance

In order to measure a nuclear magnetic resonance, the distribution of nuclei on the different energy levels is altered by an additional magnetic field B_1. The energy ν_1 of this field has to fulfill the following condition

$$\Delta E = \hbar\nu_1 = \hbar\left|\frac{\gamma}{2\Pi}\right| B_0 \,,$$

leading to changes of the energy levels from spin-up to spin-down by absorption and vice versa by emission. Both transitions have the same probability, but more changes from the lower energy level to the higher energy level occur due to the favored occupation of the lower energy level. These transitions induce a change of the macroscopic magnetism orthogonal to the direction of the external magnetic field B_0 as shown in figure 2.6.

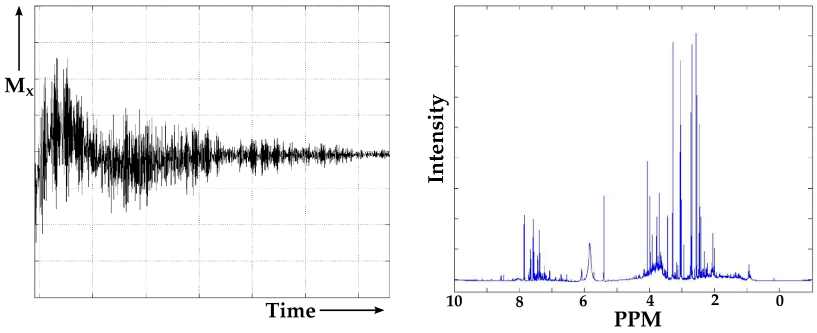
An equal occupation of the two energy levels would generally lead to a saturation and no macroscopic magnetism would be measurable. However, the nuclei do not precess statistically equally distributed on the cone surface and induce a magnetic field orthogonal to the z-direction by means of their grouped precession. Turning off the magnetic field B_1 leads to a gradual movement of the macroscopic magnetism in direction of the magnetic field B_0 and a decay of the measured signal intensity, known as *relaxation*. The precessing macroscopic magnetism is recorded by the receiver coil and converted into a spectrum representation, allowing for the identification and quantization of distinct molecules.

relaxation

2.2.3 *From Time to Frequency Domain*

The signal recorded in the receiver coil (cf. figure 2.6) is a mixture of several decaying complex-valued sinusoid signals, denoted as *free induction decay (FID)*, as shown in figure 2.7a. While the

FID



(a) Free induction decay.

(b) NMR spectrum.

Figure 2.7: Free induction decay and NMR spectrum of a representative rat urine sample measured at 600 MHz.

frequency of each signal is equal to the difference between the precession frequency of the particular molecule's atom Lamor frequency ν_L and the impulse frequency ν_1, the amplitude is dependent on the concentration of the molecule in the sample. Although the absolute concentration of each molecule cannot be determined in relation to the amplitude of a single signal, it can be quantified by relation to the signal intensity of a molecule with known concentration.

fourier transformation: FID → spectrum

Information contained in the FID is not easily accessible due to the superimposition of multiple sinusoid signals. Frequency and amplitude of these signals are the important signal characteristics and are extracted by fourier transformation. Thereby, the FID is expressed as a superimposition of fourier coefficient weighted sine and cosine signals and transferred from the time-domain to the frequency domain. The resulting NMR spectrum[1] (cf. figure 2.7b) represents single signals by peaks and the signal amplitude corresponds to the peak height.

2.2.4 *The Chemical Shift*

magnetic shielding

According to the precession frequency of hydrogen, there would be only a single measurable signal by nuclear magnetic resonance to an applied magnetic impulse and the method would not be of interest for the chemist. However, the precession frequency ν_1 is modified by the shielding properties of the extranuclear electron and interaction between neighboring electrons. Thereby, the effective magnetic field at the nucleus B_{eff} is slightly reduced

[1] PPM values on the horizontal axis are denoted in decreasing order. Signals are sorted according to their shielding and molecules with low shielding induce signals with a high chemical shift. Thus, these are located in the left part of the spectrum.

in comparison to the externally applied magnetic field B_0 with respect to the shielding parameter σ by

$$B_{\text{eff}} = (1 - \sigma)B_0.$$

Thus, the resonance frequency ν_r is slightly changed according to σ by

$$\nu_r = \left|\frac{\gamma}{2\Pi}\right| (1 - \sigma)B_0.$$

The shielding property σ is independent from B_0, and ν_r is proportional to the shielding term $(1 - \sigma)$ and B_0. Therefore, chemically non-equivalent nuclei produce peaks at different spectral positions, allowing for the discrimination of signals from different molecules.

The resonance frequency is not an absolute measure for the position of certain peaks due to the dependency of the external magnetic field B_0. Improvements in the development of new NMR spectrometers lead to higher magnetic field strengths, which generally allow for a better resolution of the spectra, but also cause different resonance frequencies of identical samples. Since the acquisition of samples and spectroscopic data is a laborious and sumptuous process, previous experiments performed with a spectrometer of different magnetic field strength should be conferrable for comparison with new samples. Thus, a normalization of the resonance frequency is required for compensation of changes in the magnetic field strength.

normalization of the resonance frequency

A first normalization regarding the zero point of the new scale is achieved by reference to a compound previously added to the sample. Nowadays, 2,2-dimethylsilapentane-5-sulfonic acid (DSS) is the agreed standard due to sharp signal[2] and the increased stability to sample condition changes in comparison to the former standard tetramethylsilane. The distance $\Delta\nu$ between the resonance frequency of the reference standard ν_{DSS} and the actual substance ν_r is still dependent on B_0. Thus, this difference is divided by ν_0 and multiplied by a constant scaling factor[3] of 10^6 and the final scale is denoted as *ppm* (parts per million). The scaled difference between ν_{DSS} and ν_r is called the chemical shift δ, the main information for discrimination of signals from chemically different hydrogen nuclei.

position normalization according to a reference compound

scaling of the chemical shift according to ν_0

$$\delta = \frac{\Delta\nu}{\nu_0} \cdot 10^6$$

Although it is hard to predict the peak pattern of a molecule with known structure, certain chemical groups produce signals

[2] DSS emits a strong signal with a low chemical shift due to the presence of nine identical methyl protons shielded by the low electronegativity of the silicon atom in the molecular structure.

[3] The scaling constant has originally been defined for a simpler notation of ppm values and is used henceforth as scaling constant for conversion of frequencies to ppm values.

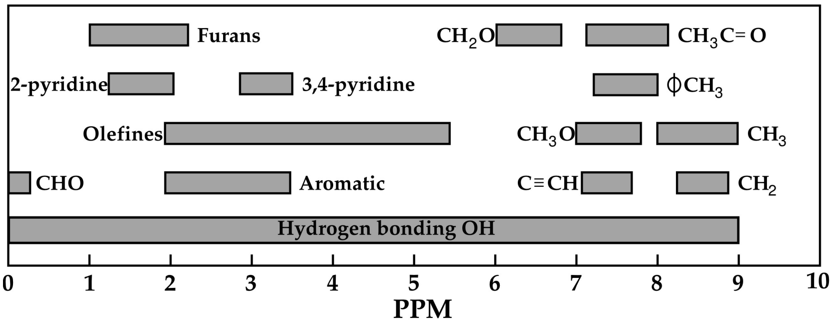


Figure 2.8: Chemical shift ranges for some organic groups in NMR spectroscopy (adopted from [Free 03], page 97).

specific regions of peak for chemical groups

in specific spectral regions as shown in figure 2.8. Thereby, a coarse categorization of corresponding molecules to putatively interesting peaks detected in a NMR spectrum can be achieved according to the organic group the signal emitting hydrogen nucleus is part of. Determination of the corresponding molecule to spectral peaks is mainly achieved by comparison to signals stored in databases containing NMR signals of various molecules. However, a single peak is practically not sufficient for a reasonable unique identification of a molecule. Rather all signals emitted by the molecule's hydrogen nuclei should be used for a database search. But all signals corresponding to the same molecule are not easy to determine in biofluids, since these are usually a mixture of several molecules. Thus, background knowledge on possible molecules contained in the sample has to be applied in order to curtail the list of possible molecules.

reliable identification requires all peaks of the same molecule

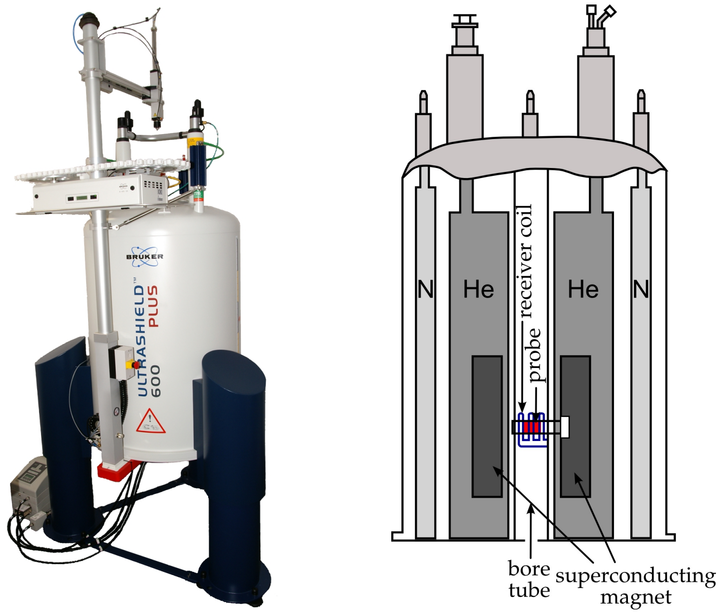
peak intensity depends on molecule concentration

Peak intensity is dependent on the concentration of the sample's ingredients and is usually not named in connection with a certain unit. Generally, methods for the normalization of intensity values have to be applied in order to achieve a reasonable comparison of samples, especially in case of biofluids. For the analysis of metabolite concentrations in urine samples as one prominent application, the concentration of the urine has to be normalized. The amount of excreted metabolites is of interest in these studies, but since a fixed volume of a urine sample is analyzed, the signal intensity emitted by a certain metabolite depends on the (usually unknown) concentration of the urine. Thus, signal strength will be denoted throughout this thesis as *(peak) intensity* without any unit terms.

changes in urine concentration require normalization

spin coupling

Up to now a single peak for chemically equivalent hydrogen nuclei has been assumed, but this is the exception rather than the rule. *Spin coupling* between neighboring magnetic nuclear dipoles leads to a weakened or amplified magnetic field B_{eff} at the nuclei and a modification of the resonance signal ν_1. This complex interaction leads to a split of the resonance signal into a multiplet signal. However, the chemical foundations of spin



(a) 600 MHz NMR spectrometer.

(b) Spectrometer structure.

Figure 2.9: (a) Picture of a 600 MHz NMR spectrometer (Bruker AVANCE 600 plus Ultrashield™) equipped with an automatic probe handler commonly used in metabonomic studies (courtesy of Dr. Christina Schreier, LipoFit Analytic GmbH). The general structure of a NMR spectrometer is shown in (b).

coupling are not relevant for the investigations pursued in this thesis and will not be discussed in detail. If any further details on spin coupling and shielding effects are of interest, these can be found in [Free 03].

2.2.5 *Industrial Measurement of NMR Spectra*

Nowadays measurement of NMR spectra is a highly automated procedure. Human interaction is required for sample preparation and spectral processing. NMR spectrometers (cf. figure 2.9a) used in industrial applications are built up of different parts as shown in figure 2.9b. The central part of the spectrometer is the persistent superconducting magnet, generating the strong magnetic field B_0. Current NMR spectrometers have a magnetic field strength up to 21 Tesla which corresponds to a Lamor frequency for hydrogen atoms up to 900 MHz, but 600 MHz is the most common frequency used in current metabonomic studies. In order to generate such a strong magnetic field that is approximately $5 * 10^5$ times stronger than the earth's magnetic field special constructions have to be used.

NMR spectrometer

superconducting magnet

The wire of the magnet coil is kept at a temperature lower than 6 K in liquid helium. Thus the resistance goes to zero and once a current is set running in the coil it will persist and no further electrical power supply is required to keep up the magnetic field. In order to reduce the need of rather expensive liquid helium for maintenance of the spectrometer a heat shield of liquid nitrogen surrounds the core of the spectrometer. For further reduction of heat flow from the environment, the whole assembly is kept in a vacuum flask. These low temperatures could negatively affect the sample, thus the sample is placed in a tube with room temperature in the center of the spectrometer – the *bore tube*.

heat shield

bore tube

In order to keep the magnetic field as homogeneous as possible within the bore tube, the sample to be analyzed is surrounded by several *shim coils*. These shim coils can generate small magnetic fields with specific spatial profiles. Thus, inhomogeneities in the magnetic field can be canceled by controlling the magnetic fields induced by the shim coils. This adjustment can be a complex task dependent on the number of shim coils, but has only to be applied once a day and not prior to each measurement procedure.

shim coils

The sample is placed by an automatic probe handler in the probehead. The probehead, located inside the bore tube, is the central part for excitation and detection of the NMR signal by a small coil. The coil has to be as close as possible to the sample in order to achieve the best results and a lot of effort is invested in the design of these probe heads. Furthermore, sample temperature can be controlled by the probehead, achieving identical conditions for each measurement. After pulse excitation by the probehead the NMR signal is recorded, amplified and converted to the FID by the digitizer. Conversion of this FID from the time domain to the frequency domain is achieved by fourier transformation as shown in figure 2.7. This spectral representation allows for the analysis of concentrations from several molecules. Each peak corresponds to a specific molecule and the peak intensity is dependent on the molecule's concentration.

probehead

receiver coil

2.3 ANALYSIS OF SPECTROSCOPIC DATA

NMR spectroscopy is a robust and reliable method for the analysis of biofluids and apart from mass spectrometry the most frequently applied measurement procedure in Metabonomics [Lenz 07]. Thereby, numerous different molecules are simultaneously measured in an automated procedure. Furthermore, only small sample volumes are required and the measurement procedure is non-destructive allowing for further measurements of the same sample. Differences in the concentration of molecules between several samples can be identified by the comparison of peak intensities between the samples' spectra. Although the un-

non-selective and non-destructive measurement procedure derlying molecules of peaks with changing concentration are not necessarily known, changes in peak intensity indicate a change in sample's composition. Systematic variations in peak intensity with respect to known properties of the sample's donor individual like gender or health status could be used to analyze new samples.

In spite of all these valuable properties of NMR spectroscopy there are some drawbacks, impeding a straightforward usage of NMR spectra for classification or data analysis purposes. Initially, the numerical range of chemical shifts used in the subsequent procedures has to be determined. If prior knowledge of relevant metabolites and their chemical shifts is present, these ranges will be analyzed and the remaining spectral information discarded. Typically this information is not available in metabonomic studies and the chemical shift range from 0 ppm to 10 ppm is used, excluding the region of the dominant and non-informative water and urea signal from 4.5 ppm to 6 ppm. These selected regions cover the main signals present in NMR spectra from urine samples and restricting the analysis to these regions is the quasi standard procedure in Metabonomics.

definition of the relevant spectral range

Furthermore, the spectra are altered by experimental or sample conditions, possibly leading to a misinterpretation. Serious perturbations of the spectra by high content of additional substances in the sample or errors in the measurement process can be detected beforehand by multivariate data analysis methods and corresponding samples are excluded from the following analysis. Further alterations caused by slight changes in physiochemical factors frequently occur, but methods to compensate these have been proposed and are still part of current research.

perturbations of spectra

2.3.1 *Peak Alignment*

In multivariate data analysis, NMR spectra are treated as high-dimensional feature vectors. Hence, corresponding peak information has to be represented by the same variable among all spectra in order to achieve a reasonable interpretation of the data. However, the electromagnetic environment of the hydrogen nucleus is mainly influenced among several physiochemical factors by ion concentration and pH value [Wu 06], leading to a change of the exact peak position in the spectrum. Thus, it is difficult to identify signals from the same molecule among several spectra according to their exact position, denoted as the *correspondence problem* [Aber 09]. These *peak shifts* have to be compensated in order to achieve a reasonable interpretation of spectra represented as vectorial data.

peak shifts

Different NMR spectra of the same sample measured after manual adjustment of the sample's pH value are shown in figure

manual adjustment of pH value

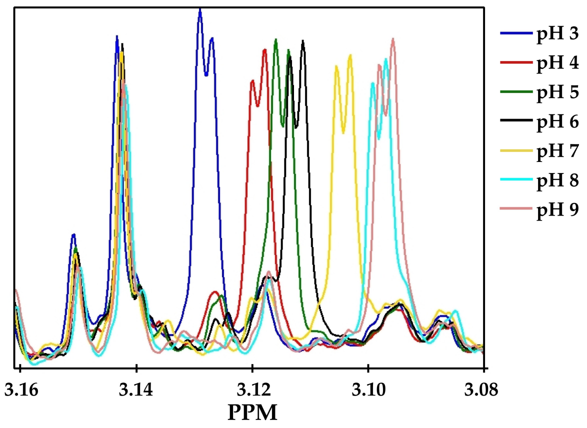


Figure 2.10: Effect of variation in pH value of a urine sample on the position of peaks in NMR spectra. Increasing the pH value leads to a shift to the right of the central peak, while other peaks nearly remain at their position.

2.10. The central peak doublet shifts to the right with increasing pH value due to the change of the electromagnetic environment of the corresponding hydrogen nucleus. Several other peaks in the same spectral region are nearly unaffected by variation of the pH value and remain at their position throughout all spectra. Such dramatic changes in pH values of urine samples would not occur naturally. However, this example illustrates the peak shifting effect of variations in physiochemical factors, and the specific sensitivity of each hydrogen nucleus to these factors.

different degrees of peak shifts

bucketing

Bucketing was presented as a first approach for minimization of peak shift effects, integrating small consecutive parts of the spectrum (buckets) to a single value located at the center of the integrated region [Holm 94, Spra 94, Holm 98b]. Thereby, exact peak positions are blurred and minor changes in the peak positions are compensated in case of adequate bucket boundaries. However, the bucket boundaries are mostly chosen independently from the underlying data and a peak shift across several buckets can generally not be avoided. A further disadvantage of the bucketing procedure is the loss of spectral resolution, complicating the reference to single peaks with changing intensity within a bucket.

trade-off between peak shift compensation and spectral resolution

The effect of different bucket boundaries on the resulting spectrum is shown in figure 2.11 based on simulated data and using a bucket width of 0.04 ppm. Using first row bucket boundaries results in a nearly full compensation of peak shifts, and intensity information from the two central peaks is still distributed over two bucket values. However, adding an offset of 0.02 ppm to the bucket boundaries results in a significant change of alignment quality. Peak shifts are still present at the right peak and information from the two central peaks is merged to a single value. Thus

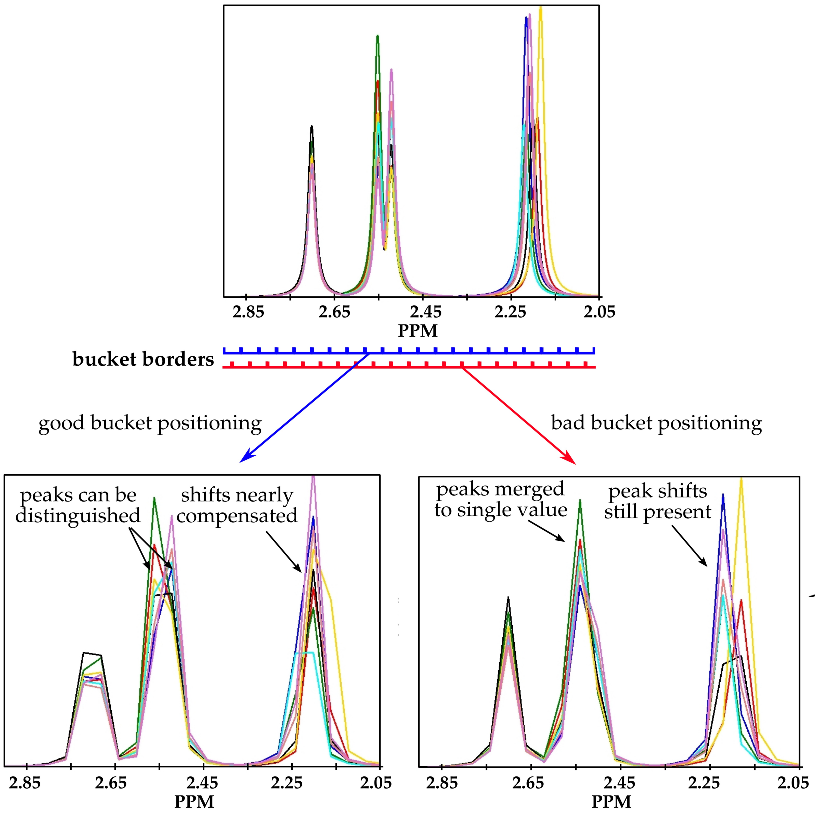


Figure 2.11: Compensation of peak shifts can be achieved by a bucketing procedure, but the success of this alignment is mainly influenced by the position of bucket boundaries.

intensity changes of the bucket value cannot be associated with one of these peaks.

peak alignment

These drawbacks of bucketing show the necessity for more sophisticated alignment procedures. Several approaches have been published [Fors 03, Stoy 04, Wu 06, Wang 06] and applied to NMR spectra [Fors 05] in recent years. Basically, alignment methods differ in their data representation, their methodology for the determination of peak shift corrections and the way of revising these shifts. Three recent alignment strategies will be shown as examples in order to clarify the problems in the identification of corresponding peaks among several spectra.

Peak Alignment Using Reduced Set Mapping

The alignment method proposed by Torgrip et al. [Torg 03], denoted as peak alignment using reduced set mapping (PARSE), aligns peaks from two different spectra by the conversion of the spectra into a peak representation and optimization of the peak-association between those spectra. Once the correspondence between peaks is determined, insertion and deletion operations

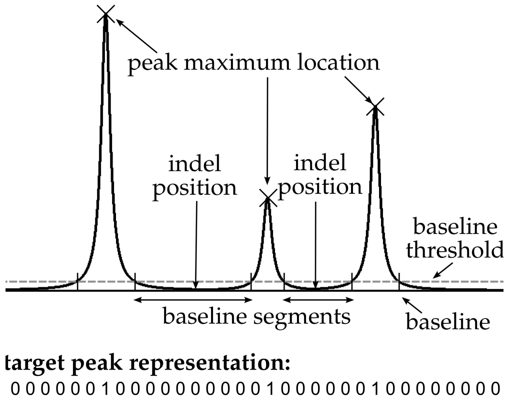


Figure 2.12: Conversion of a spectrum to peak representation and assignment of borderline regions for insertion and deletion operations (adopted from [Torg 03].

are applied in baseline regions containing non-relevant signals for the adjustment of corresponding peak positions.

peak representation

As a first step in PARSE the spectra need to be converted into a peak representations as exemplified in figure 2.12. Therefore, peak maxima and their positions are determined for each spectrum according to the Savitzky-Golay [Savi 64] first derivative of second order and transferred to the target peak representation with ones at the peak positions and zeros otherwise. The Savitzky-Golay derivatives are determined according to the higher order coefficients of a polynomial fitting for smoothing of the data. Using a polynomial of second order allows for suppression of noise signals in the data while retaining peak information similar to their original shape.

baseline regions

Subsequently, peak boundaries for putative insertion and deletion operations are identified according to the peak maxima and the estimated baseline by a running windowed median. According to this baseline and a meta-parameter α a threshold is defined indicating the transition from peak to baseline segments. Thus, peak boundaries are defined by tracing the signal from peak maxima to the left and to the right until reaching the threshold.

selection of a reference spectrum

In order to achieve an alignment of all spectra based on the peak representation and identified peak boundaries a target spectrum has to be defined. This spectrum should be as representative as possible and is usually a selected spectrum with the most common peaks or either a mean or median spectrum. The selection of a reference spectrum can be a severe drawback in heterogeneous data sets containing new or vanishing peaks. However, choosing a reference spectrum for alignment is a prerequisite in most alignment approaches. Up to now only some heuristics have been presented to achieve an automated and objective reference spectrum selection.

Final alignment is realized using the peak representation of the reference sample and each spectrum, respectively. The optimal alignment scheme is calculated by dynamic programming according to the Smith-Waterman algorithm [Smit 81]. This is algorithm is based on the efficient determination of the optimal path through a distance map **D**. This map contains the Euclidean distances between all peaks of the two spectra to be aligned (cf. [Prav 02]). An optimal path through this map is determined by backtracing the map along the minimum cumulative path. Incorporating background knowledge on the maximum distance between two shifted peaks allows for a reduction of the distance map according to [Sako 78], thereby, saving memory and reducing the computational complexity.

dynamic programming for determination of optimal insertion / deletion operations

Alternatively, representing the peak list in a directed graph allows for alignment by finding the shortest path through this graph. Each vertex is a match between two peaks, whereby, only peaks with a distance lower than a maximal value are chosen for a possible alignment. A weight is assigned to each edge according to the distance of the two peaks to be matched and the number of missed matches of peaks between the two peaks. The shortest path through this graph is determined by a breadth first search algorithm, systematically examining the entire graph and finding the optimal solution by an exhaustive search without any heuristics for the reduction of computational complexity.

alignment based on a graph representation

Both optimization methods lead to a match of corresponding peaks between the target and sample spectrum and the exact peak position of the sample spectrum is corrected according to these correspondences. Relocation of peaks is achieved by insertion and deletion operations in adjacent baseline regions. Thereby, peak shifts are corrected without modification of peak shapes. However, the matched peak lists are already a compact representation of the spectral information and can be used for further analysis as shown in [Torg 06].

alignment by insertion / deletion operations in baseline segments

Correlation Optimized Warping

Nielsen et al. first introduced *correlation optimized warping* (COW) as a methodology to align shifted signals in spectroscopic data [Niel 98]. This method is known as a piecewise or segmented data preprocessing method, aligning a sample spectrum to a previously selected reference spectrum by stretching and squeezing of spectral segments in order to optimize correlation between the two samples. Finally, the global alignment is achieved by combination of local segment modifications.

Comparable to the PARSE method, the selection of a reference spectrum influences the subsequent alignment results. Besides the selection of a reference sample according to prior knowledge Skov et. al presented a more objective approach to achieve a

reference spectrum

reasonable reference spectrum selection [Skov 06]. Although, selection of the mean spectrum or the spectrum most similar to the first principal component of a *Principal Component Analysis* (PCA) model seems reasonable, these spectra are still influenced by peak shifts. Thus, an alternative approach is presented aiming at the selection of the spectrum with the maximum similarity to all other samples in the data set $\mathbf{Z}$, consisting of N samples of size n.

similarity index

A measure reflecting this similarity is the *similarity index*, using the correlation coefficients $r(\mathbf{z}_r, \mathbf{z}_i)$ between the samples $\mathbf{z}_r$ and $\mathbf{z}_i$ $i = 1, \ldots, N$, respectively.

$$\text{Similarity index} = \prod_{i=1}^{N} |r(\mathbf{z}_r, \mathbf{z}_i)| \,,$$

where

$$r(\mathbf{z}_r, \mathbf{z}_i) = \frac{\tilde{\mathbf{z}}_r^T \mathbf{z}_i}{\|\tilde{\mathbf{z}}_r\|(\|\mathbf{z}_i\|^2 - n\bar{z}_i)^{1/2}}$$

This equation has been introduced in [Skov 06] and $\tilde{\mathbf{z}}_r$ is the centered sample $\mathbf{z}_r$ and $\bar{z}_i$ is the mean value of $\mathbf{z}_i$. The similarity index ranges from zero to one, whereby the spectrum with maximum similarity to all other samples will have the highest similarity index and is selected as reference sample.

squeezing and stretching of spectral segments for alignment

After selection of a reference spectrum, the reference spectrum $\mathbf{z}_r$ and the sample to be aligned $\mathbf{z}_s$ are divided into equally sized segments of length I. The alignment procedure basically consists of shifting the segment boundary positions of $\mathbf{z}_s$ and interpolating the segments to original size in order to achieve a maximal correlation between the spectra. This global alignment problem can be broken down by calculation of segment-wise correlations for each desired boundary shift. The globally optimal combination of shifts is determined by dynamic programming techniques [Niel 98]. The maximal position shift of segment boundaries t – denoted as *slack area* – together with the segment length I are the sole parameters of the alignment procedure to be optimized in order to achieve the best alignment results.

slack area and segment length

The effect of the different parameters and the overall principle of the alignment procedure is explained by means of a simple example shown in figure 2.13. A spectrum $\mathbf{z}_s$ of length 16 has to be aligned to the reference spectrum $\mathbf{z}_r$ of length 17. Initially, the spectra are divided into 4 segments, whereby interpolation will be used within the alignment procedure to compensate differences in the size of the last segments.

Beginning with the last segment, the left border can be a) shifted to the right, b) remain in the same position or be c) shifted to the left. According to the slack parameter t equal to one, shifts can only be a single data point to the left or to the right. Thus in this step only three modifications are possible. Dependent on the segment size of the reference sample, an interpolation

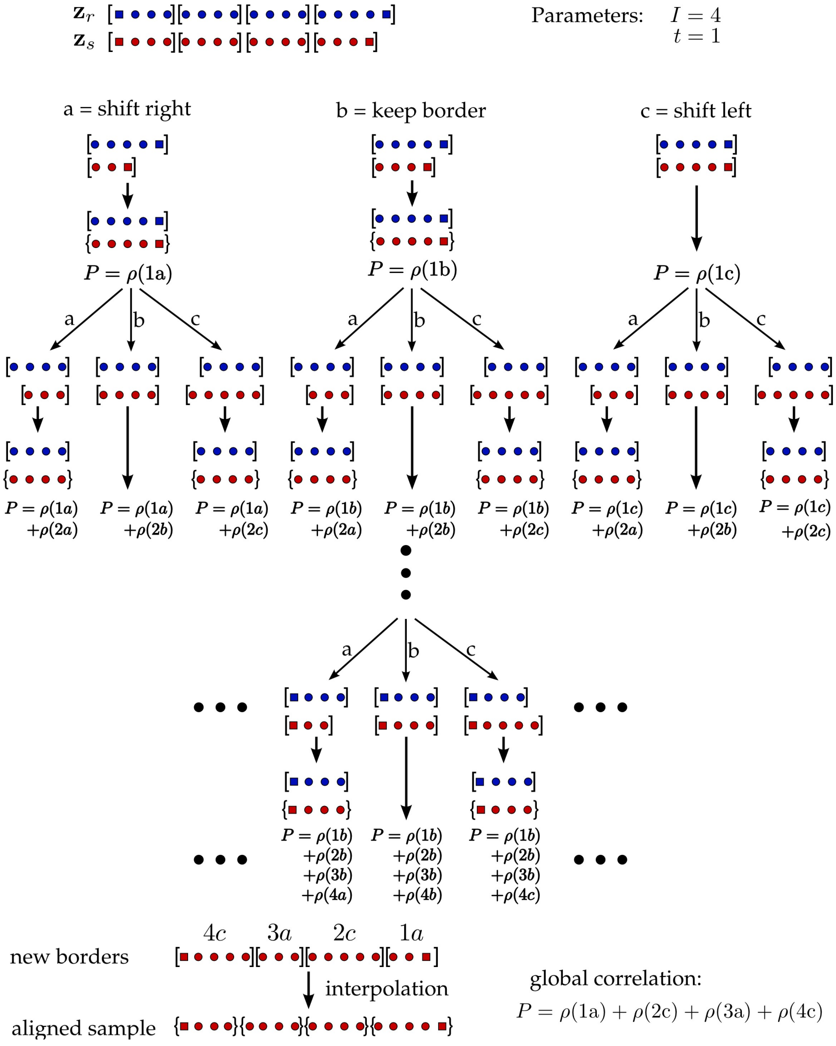

$\mathbf{z}_r$
$\mathbf{z}_s$
Parameters: $I = 4$ $t = 1$
a = shift right
b = keep border
c = shift left
$P = \rho(1a)$
$P = \rho(1b)$
$P = \rho(1c)$
a
b
c
$P = \rho(1a) + \rho(2a)$
$P = \rho(1a) + \rho(2b)$
$P = \rho(1a) + \rho(2c)$
$P = \rho(1b) + \rho(2a)$
$P = \rho(1b) + \rho(2b)$
$P = \rho(1b) + \rho(2c)$
$P = \rho(1c) + \rho(2a)$
$P = \rho(1c) + \rho(2b)$
$P = \rho(1c) + \rho(2c)$
$P = \rho(1b) + \rho(2b) + \rho(3b) + \rho(4a)$
$P = \rho(1b) + \rho(2b) + \rho(3b) + \rho(4b)$
$P = \rho(1b) + \rho(2b) + \rho(3b) + \rho(4c)$
4c 3a 2c 1a
new borders
interpolation
aligned sample
global correlation:
$P = \rho(1a) + \rho(2c) + \rho(3a) + \rho(4c)$

Figure 2.13: Example of the COW alignment procedure (see text for details, adopted from [Toma 04]).

of the sample segment has to be performed for equal segment lengths. This allows for calculation of the segment correlation $\rho(n)$ between the resulting segments at position n.

$$\rho(n) = r(\mathbf{z}_r(n), \mathbf{z}_s(n))$$

Subsequent to each of the different choices for segment border modification the process repeats for the left border of the next segment until the last segment is reached. Thereby, global correlation P can be calculated by summation of the segment correlations. In order to achieve a maximal global correlation a dynamic programming algorithm is applied, leading to the optimal combination of segment border modifications. Application of the optimal combination of segment border modifications leads to an alignment of a sample spectrum to the reference spectrum by squeezing of signal segments.

The result of the COW alignment procedure is mainly controlled by the choice of segment length and the slack parameter. These are usually determined by rules of thumb or by means of cross-validation procedure. Skov et. al presented a systematic procedure for determination of the optimal parametrization. Additionally, some modifications of the original algorithm are introduced to speed up necessary calculations allowing for a parameter optimization in reasonable time [Skov 06]. Alignment results based on different parameter settings are compared according to the newly defined *warping effect*, a summation of the *simplicity value* and *peak factor*.

warping effect

simplicity value

The *simplicity value* measures the alignment quality of a data set **Z** by analysis of the partitioning of explained variance into the singular values resulting from singular value decomposition (SVD) of **Z**. Variance is mainly explained by the first few components in case of well aligned spectra, leading to large first eigenvalues. Contrary to this, the explained variance will be distributed among several singular values in case of unaligned data. Thus, summation of the first R singular values taken to the power of four[4] will be the higher the more data variation is explained by the first components. SVD is applied on a scaled version of **Z** in the calculation of simplicity:

$$\text{simplicity} = \sum_{r=1}^{R} \left(\text{SVD}_r \left(\mathbf{Z} / \sqrt{\sum_{i=1}^{N} \sum_{j=1}^{n} z(i,j)^2} \right) \right)^4$$

where SVD_r denotes the singular value of the r$^{\text{th}}$ component from SVD. This measure ranges from zero to one, whereby a value of one indicates a perfect alignment result.

Effects of variations in physiochemical factors induce peak shifts that have to be compensated by alignment methods, but no further changes in spectral properties (e.g. peak shape) should be caused. However, modifications applied in COW could distort the peak shapes and these changes have to be taken into account in selection of a feasible slack parameter and number of segments. Changes in peak shapes are quantified in the presented approach by the *peak factor* ($0 \leq$ peak factor ≤ 1), measuring the difference between the norm of the original sample $\mathbf{z}$ and the aligned sample $\hat{\mathbf{z}}$.

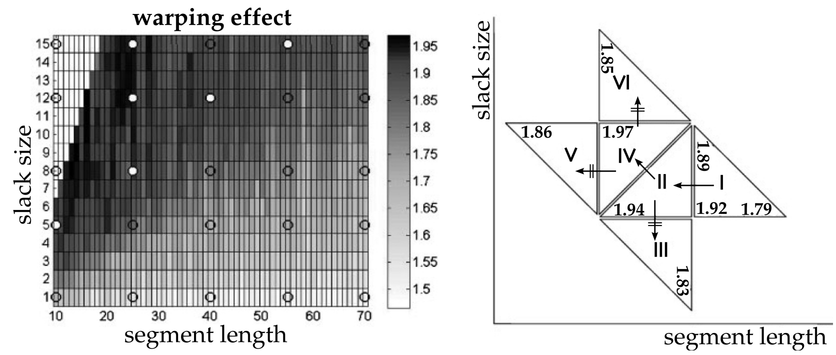
peak factor

$$\text{peak factor} = \frac{\sum_{i=1}^{N} \left(1 - \text{minimum}(c(i), 1)^2\right)}{N}$$

where

$$c(i) = \left| \frac{\|\hat{\mathbf{z}}_i\| - \|\mathbf{z}_i\|}{\|\mathbf{z}_i\|} \right|$$

[4] Using the fourth power allows for the use of this measure even if all singular values are retained. The squared sum of all singular values is by definition always equal to one [Skov 06].



(a) Warping effect values.

(b) Simplex optimization.

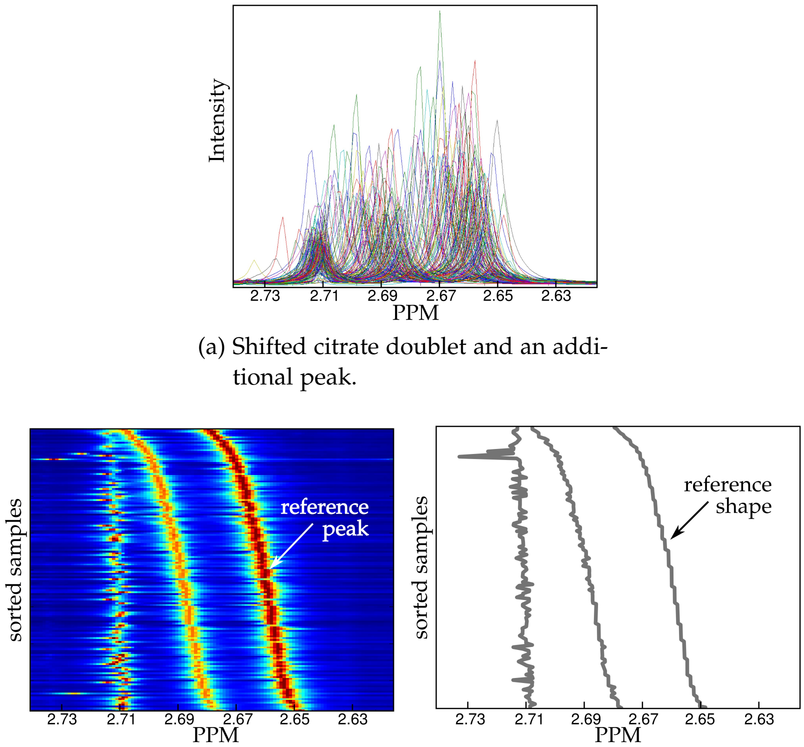
Figure 2.14: Warping effect values for different combinations of segment length and slack size are shown in (a). Filled circles are the six best parametrizations of an initial 5x5 grid evaluation. These serve as starting points for simplex optimization as illustrated in (b) (courtesy of [Skov 06]).

simplex optimization

determination of the parametrization with the maximum warping effect

Combination of the simplicity value and peak factor to the warping effect by summation facilitates the efficient comparison of alignment results achieved by different parameter settings. However, an exhaustive search for the optimal parameter setting is not feasible since the alignment procedure is still computationally demanding. Thus, a stepwise optimization procedure based on the simplex algorithm [Spen 62] has been presented, initially defining the boundaries of the optimization space according to certain spectral properties (e.g. peak width). Different parameter combinations are evaluated in this space using a 5x5 grid and the six best results serve as starting point for further search (cf. figure 2.14a). Starting at each of these points two adjacent parameter settings are evaluated, forming a triangle as shown in figure 2.14b with the starting point as base. The triangle is flipped over to the side of the line between the points in the triangle with the maximum warping effect. This process is iteratively repeated until the search ends at one configuration. The parametrization leading to the maximum warping effect among all six search procedures is regarded as the optimal configuration.

A comparison of COW to dynamic time warping [Kass 98] as alternative alignment method by Pravdova et al. has shown the capability of COW to compensate peak shifts. But limitations in baseline correction still exist [Prav 02]. Crucial aspects within this approach, such as the selection of a reference spectrum and parameter settings, have been discussed by Skov et al. and algorithmic solutions have been presented.



(a) Shifted citrate doublet and an additional peak.

(b) Sorted image representation.

(c) Indicator matrix.

Figure 2.15: Conversion of the (a) standard spectrum representation to the (b) image representation sorted according to a reference peak. The (c) indicator matrix containing peak maxima is used for detection of peak shifts by the GFHT.

Generalized Fuzzy Hough Transform

detection of shift patterns

A quite different alignment approach presented by Csenki et al. is based on the generalized fuzzy Hough transform (GFHT) [Suet 06], identifying peak shifts by detection of certain patterns in an image representation of a data set [Csen 07]. Thereby, spectra are not plotted in a typical way as shown in figure 2.15a, but stacked to form a matrix encoding the intensity by a color code (cf. figure 2.15b). Spectra are sorted within this representation according to the position of a previously defined *reference peak* with known sensitivity to physiochemical factors. pH is expected to be the major source for changes in peak positions, but for the GFHT alignment approach it is not necessary to know the exact factors leading to peak shifts. For the detection of peak shifts it is sufficient to assume a *common shift pattern* for all shifting peaks.

common shift pattern

reference peak determination

In order to achieve a reasonable representation of the spectra certain prerequisites have to be fulfilled by the reference peak. The reference peak has to be sensitive to changes in physiochemical factors in order to qualify the common shift pattern. Furthermore, the signal should be strong enough and not overlap with other signals in order to unambiguously identify the reference peak in each spectrum. The selection of the reference peak is a crucial step in this procedure and can be achieved according to the presence of strong signals with known relevance to physiochemical factors or visual investigation of the data set.

indicator matrix

Initially, the spectra are transformed into an indicator matrix having a one at peak positions detected by a preceding peak-detection method and a zero otherwise as shown in figure 2.15c. Corresponding peaks can be identified among the spectra in the indicator matrix as points of a vertical line, while shifting peaks induce slightly skewed vertical lines. Thus, the objective of peak shift detection can be reformulated by finding shapes in the indicator matrix representation similar to the shape produced by the reference peak in different positions and slight rotations. In the context of the GFHT approach the phrase *peak shape* is used for the line in the indicator matrix produced by a specific peak and should not be mixed up with the shape of a single peak in a NMR spectrum. The position of a peak shape is specific to a peak and the rotation is dependent on the particular sensitivity of the hydrogen nucleus to the shift inducing physiochemical factors.

reference shape

Differences in sensitivity to physiochemical factors can be observed in figure 2.15. The second peak of the citrate doublet has an almost equal shape like the reference peak due to the equal sensitivity to physiochemical factors. The left peak seems to be more stable than the two other peaks but large shifts can be observed in a few samples as a result of changes in physiochemical factors not affecting the citrate doublet. This demonstrates the problem of finding an optimal reference peak for sorting of the spectra, since different physiochemical factors influence the position of molecules to a different extent.

A method from the field of image analysis for the detection of shapes described by analytical functions (e.g. lines) in images is the Hough transformation [Duda 72]. In the case of shapes that are not described by an analytical function but a list of points, such as the peak positions of the reference peak forming the reference shape, the generalized Hough transform can be applied [Sama 97, Ball 81]. However, in presence of noise the generalized Hough transform would not detect every line with a shape similar to the reference shape since these are not identical in every point. Therefore, fuzzy Hough transform has been developed, allowing for slight distortions of the shape to be detected by increasing the vote of a detected shape according to the distance of a possible peak [Han 94]. Combining these two extensions of the classical Hough transform allows for the robust detection of shapes similar to a given list of points in images. This approach is applied for the detection of corresponding peaks among several spectra based on the presented indicator matrix representation.

Hough transformation

generalized Hough transformation

fuzzy Hough transformation

expansion factor

The *vote* $h(\alpha, k)$ for a shape similar to the shape of the mean-centered reference peak $\mathbf{s}$ at position k and expansion factor α in the indicator matrix $\mathbf{X}$ is defined as

$$h(\alpha, k) = H(\mathbf{X}, \mathbf{s}, \sigma) = \sum_i \sum_j x_{ij} p(i, j, \mathbf{s}, \sigma)$$

where

$$p(i, j, \mathbf{s}, \sigma) = \exp\left(-\frac{1}{2}\left(\frac{(j-k) - s_i\alpha}{\sigma}\right)^2\right).$$

*local maxima in matrix **H** indicate peak shifts*

The expansion factor denotes the relative sensitivity of the current peak to the physiochemical factors, whereby a value smaller than one indicates less sensitivity and vice versa. Calculating the vote for all spectral positions and several expansion factors results in a matrix $\mathbf{H}$ of votes $h(\alpha, k)$. Local maxima in this matrix indicate peak shapes similar to the reference peak shape. Using the information on position and expansion factor allows for the detection of corresponding peaks in the spectra. Thus, iterating over all local maxima of $\mathbf{H}$ until no further peak can be extracted results in a list of peak information and their correspondence between the spectra. Compensation of peak shifts is achieved by the modification of the exact position from all peaks corresponding to an identified shape. Alternatively, the intensity information can be directly incorporated in further analysis as already shown by PARSE.

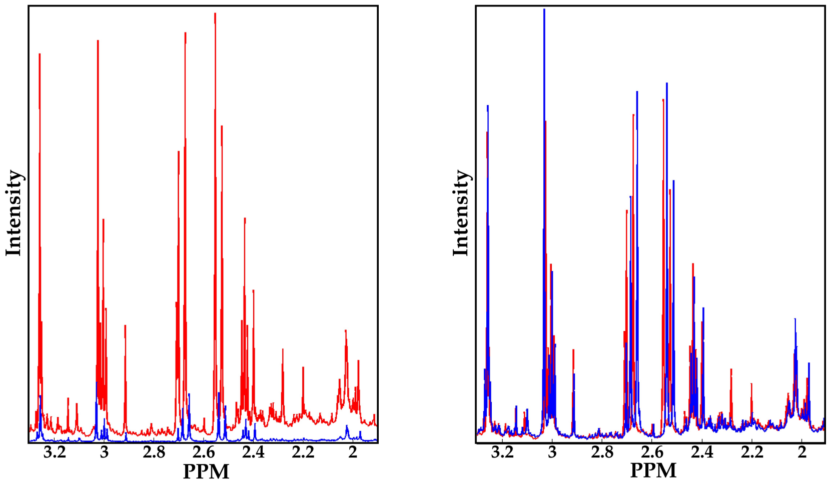
compensation of peaks shifting over each other

Assuming that every shifting peak has a common shift pattern with varying accentuation subject to the respective sensitivity to physiochemical factors, peaks can shift across each other in the spectrum. This phenomenon can solely be detected by GFHT alignment due to the incorporation of shift patterns in other spectra. This is a clear advantage of the GFHT approach, while the computational complexity could be regarded as a drawback in practical applications. However, even a computation time of several days in order to achieve an aligned data set are usually not a problem, since the organization and realization of metabonomic experiments can last several months.

2.3.2 *Compensation of Dilution Effects*

peak intensities are influenced by sample dilution

Analysis of NMR spectra is primarily based on the comparison of peak intensities, which reflect the concentrations of certain molecules. However, peak intensities are no absolute measure and can be influenced by experimental variables (e.g. sample dilution). A comparison of two spectral regions from untreated rats before and after normalization is shown in figure 2.16. Peak intensities are extremely different in the original spectra due to dilution differences, leading to an increased inter-sample variance [Crai 06, Diet 06]. In order to allow for a reasonable analysis of



(a) Original spectrum

(b) Normalized spectrum

Figure 2.16: Normalization effect of a spectral region from two untreated rats for the compensation of dilution effects.

the spectra, a change in peak intensity should only be induced by an altered molecule concentration and the influence of other variables has to be reduced.

normalization to a constant sum

Spectral normalization to a constant sum is commonly used in Metabonomics, scaling each spectrum to a previously defined total spectral area, but it has some major drawbacks. This approach is based on the assumption that the majority of a spectrum is stable and only minor changes in some peak's intensity will be present. However, even small changes in peak intensities of abundant molecules will affect the whole spectrum and change the peak intensities of molecules with an unaltered concentration. Furthermore, new peaks can occur in some spectra, especially in Metabonomics, as a metabolic product from the applied substance or caused by a seriously damaged organ, thereby decreasing the intensity of all other peaks.

normalization according to a reference substance

The spectral area of creatinine, which is assumed to be a constantly excreted substance, has been used in clinical chemistry for spectral normalization [Cons 05]. But there are serious problems from the technical and biological point of view. Creatinine peaks are sensitive to physiochemical factors and are not well separated from other peaks, impeding a reliable peak detection and peak area determination. Furthermore, creatinine excretion is not constant in case of kidney impairment and cannot be used for investigations of drugs with well-known kidney toxicity. Thus, normalization to creatinine excretion is used in only few metabonomic studies and no alternative substance with commonly accepted constant excretion has been published up to now.

probabilistic quotient normalization

An alternative scaling method, denoted as *probabilistic quotient normalization*, has been proposed by Craig et al. for spectral normalization of NMR spectra independent of a reference substance [Crai 06]. Thereby, a reference spectrum for normalization is selected and compared with corresponding data points of a sample spectrum. The median quotient is used for the scaling of the sample spectrum, thereby reducing the impact of single peaks on normalization.

histogram matching normalization

A quite different approach presented by Torgrip et al. [Torg 08] utilizes histogram matching techniques from the field of image processing [Rose 76] for spectral normalization. Each spectrum is transformed from the intensity space to *counting space* by histogram calculation. Thereby, the number of variables falling in the range of several non-overlapping intervals sorted according to their numerical range is counted. Normalization is realized by the calculation of the dilution factor, minimizing the difference between the histogram of the sample and reference spectrum. Thereby, the method is unaffected by peak shifts and the influence of each variable on the normalization procedure is independent of its value.

2.4 MULTIVARIATE DATA ANALYSIS IN METABONOMICS

data interpretation and classification

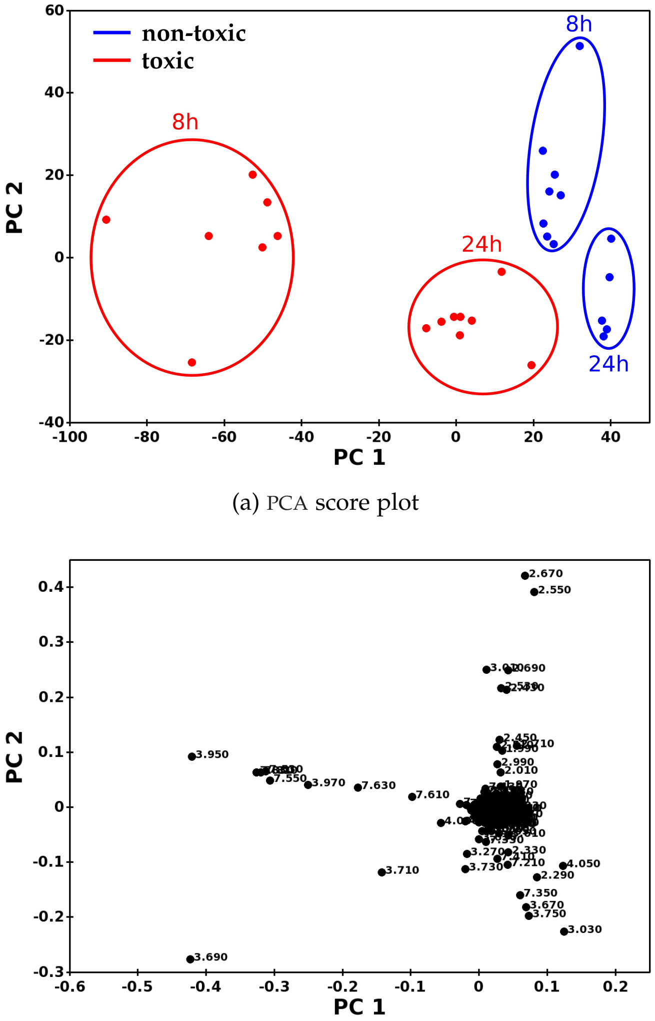
Modern spectroscopic methods used in Metabonomics produce complex and high-dimensional data sets, but due to the high cost for each measured sample their amount is usually rather limited. Therefore, methods from multivariate data analysis are applied for the interpretation of these valuable data sets, allowing for multifarious data exploration even on small data sets. *Data interpretation* and *classification* are the two major goals in Metabonomics, giving a deeper insight into the underlying biological mechanisms reflected in the data or allowing for a fast classification of unknown samples.

2.4.1 *Interpretation of NMR Data Sets*

low-dimensional data representation

Principal Component Analysis

Data analysis by projection methods allows for a low-dimensional representation and graphical interpretation of complex data sets. One of the first methods applied for data projection purposes has been *Principal Component Analysis (PCA)* [Joll 02]. Thereby, a new coordinate system is determined by the maximization of the data variance explained by the coordinate axes, which are called *Principal Components (PCs)*. These PCs are sorted in decreasing order according to the amount of explained variance. Thereby, the main variance in the data is covered by the first few PCs. It is reasonably assumed that the dimensionality of the samples can



(a) PCA score plot

(b) PCA loading plot

Figure 2.17: PCA (a) score and (b) loading plots from the first two principal components of a PCA model estimated on a set of NMR spectra derived from rat urine. Samples from rats treated with a non-toxic and toxic compound, and samples from different urine collection time-points are marked in the score plot for both compounds, respectively. Spectral positions of variables in the loadings plot are denoted in ppm.

be reduced by projection onto a subset of all PCs (cf. appendix A.1).

A projection of spectra on the first two PCs of a PCA model estimated on a data set of NMR spectra from rat urine is shown in figure 2.17a. This *score plot* allows for a graphical discrimination of samples from rats treated with a non-toxic compound (blue circles) and a nephrotoxicant (red circles) according to the projection value on the first PC. The second PC discriminates in this data set samples from different sample collection time-points after drug application.

score plot

loading plot

The contribution of each variable (or in this case spectral region) on the separation of samples in the score plot is reflected by the coefficient, denoted as *loading*, of the corresponding variable. The loadings of two PCs can be visualized by a loading plot as shown in figure 2.17b. In this plot, each point represents a specific feature of the original data set and the associated spectral regions are additionally shown. Thereby, variables relevant to sample separation are indicated by values lower or higher than the bulk of variables. Apparently, peaks in the regions of 3.9 ppm, 7.6 ppm and 3.7 ppm primarily contribute to the separation with respect to the drug-induced metabolic changes (first PC). The region around 2.6 ppm, 3 ppm and also 3.7 ppm cause a separation of samples regarding the collection time (second PC).

This straightforward graphical interpretation of a data set by PCA score and loading plots is mostly considered as first step in data analysis and has been used in numerous studies in the field of Metabonomics [Holm 98b, Fors 03, Nich 01, Wate 05]. However, this interpretation of the first PCs for data discrimination is not explicitly modeled but can be achieved due to data variance induced by differences between the samples. Using a more complex and heterogeneous data set would lead to data variance not connected with the previously mentioned sample properties and interpretation would not be that clear as presented.

Partial Least Squares

PCA as an unsupervised projection method assumes a strong connection between the variance and importance of each variable, an assumption that is not always true for spectroscopic data from biosamples. Changes in metabolite concentrations can be induced by various sources, while changes relevant for the separation of current interest can have a low variance due to the initial low substance's concentration. *Partial Least Squares (PLS)* – also known as *Projection to Latent Structures* – as supervised multivariate data analysis method is a frequently applied technique in chemometric and metabonomic studies. Based on a labeled data set correlations between the spectral data and class labels are derived [Esbe 01, Mart 89, Wold 66]. Similar to PCA a new subspace is defined by PLS but instead of maximizing the variance explained by the coordinate axes (latent variables), the covariance between the spectral variables and class labels is maximized (further details are given in A.2). Thereby, information on the class membership of each sample is integrated into the model calculation, mostly leading to a better class separation in comparison to PCA. This improved class separation and the possibility of inferring the relevant spectral regions by the analysis of the latent variables has made PLS the quasi standard analysis method in metabonomic studies [Brin 02, Keun 03, Pear 05]. Prediction of unknown samples can also be achieved by PLS discriminant analysis due to the incorporation of class information in PLS model estimation. Beside the original PLS method several modifications exist with improved model interpretation or class separation capabilities [Byle 06, Rant 07, Tryg 02].

2.4.2 *Classification of NMR Data Sets*

Not only data interpretation is of interest in metabonomic studies, but also data classification for the prediction of unknown samples. Thereby, spectra can be separated into different classes (e.g. the type of drug-induced organ toxicity) without manual interpretation of certain substances' concentrations measured by time-consuming methods and not requiring expert knowledge in the specific field for data assessment. However, according to the current literature only few attempts have been realized for an automatic classification of NMR spectra, indicating the challenging type of data and apparent problems to derive a reliable classification.

The COMET classification system

Probably the most prominent classification approach has been presented within the Consortium for Metabonomic Toxicity (COMET) [Ebbe 07, Lind 03, Lind 05], a collaboration between five major pharmaceutical companies (Bristol-Myers-Squibb, Eli Lilly & Co., Hoffman-La Roche, NovoNordisk, Pfizer Inc.) and the Imperial College London, UK. Using a data set of 12 935 NMR spectra from 1 652 rats treated from a set of 80 different compounds[5], a classification system as shown in figure 2.18 incorporating methods for data preprocessing, normalization, classification and rejection has been established.

Consortium for Metabonomic Toxicity

Initially, spectra are preprocessed in order to compensate variations induced by peak shifts or urinary concentration and a model for all control samples is built. This control model is expected to describe urinary variations of physiologically normal animals. Thus, a first discrimination of samples from control and treated animals can be achieved. Spectra regarded as samples from animals with an abnormal metabolic profile are further scaled and their similarities to known treatments are estimated by the classification of unknowns by density superimposition (CLOUDS) approach [Ebbe 03]. Finally, similarities induced by interfering signals are detected and removed for the subsequent classification based on the closest matching treatment. This multi-stage approach for the processing and classification of NMR spectra has been so far the only method validated on a large-scale database

classification of unknowns by density superimposition

[5] As most data sets used in Metabonomic studies these spectra are not available for public use.

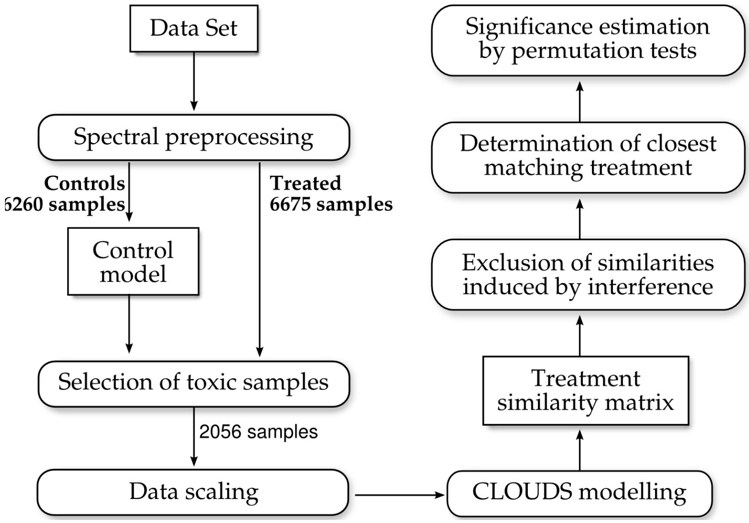

Data Set
Spectral preprocessing
Controls
6260 samples
Treated
6675 samples
Control model
Selection of toxic samples
2056 samples
Data scaling
CLOUDS modelling
Treatment similarity matrix
Exclusion of similarities induced by interference
Determination of closest matching treatment
Significance estimation by permutation tests

Figure 2.18: COMET screening method for the detection of drug-induced organ toxicities (see text for description, adopted from [Ebbe 07]).

and has proven the applicability of metabonomic methods for the detection of adverse effects in preclinical toxicology.

spectral data set

The data set used within the COMET approach contains NMR spectra from time-series samples. Therefore, urine samples are collected two times before and eight times after the application of a pharmaceutical compound. For compensation of minor peak shifts, a bucketing procedure using a bucket-width of 0.04 ppm (cf. section 2.3.1) is applied. The spectral range from 0.02 ppm to 10 ppm is used for further analysis excluding the urea and water region from 4.5 ppm to 5.98 ppm. Signals from non-endogenous compounds (e. g. drug-related compounds, DRCs) are replaced by the corresponding value of the mean control spectrum in order to concentrate on endogenous compounds in the subsequent analysis. Furthermore, variations in urine concentration are corrected by integral normalization to a value of 100.

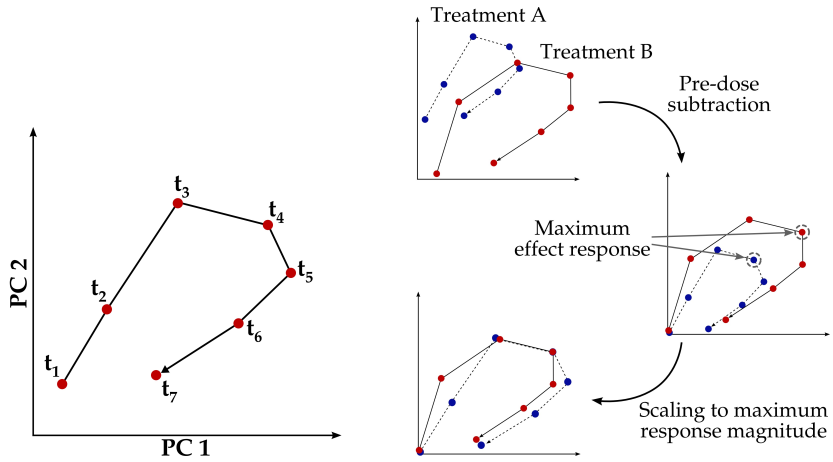
spectral preprocessing

control model

By means of an approach from multivariate data analysis, normal variations of control animals are described in a PCA model estimated on the control group of all treatments. Thereby, the number of PCs is determined using the predicted variance Q^2 on the validation set in a 7-fold cross-validation design. The predicted variance is based on the predicted residual sum of squares (PRESS), calculated as

$$\mathrm{PRESS} = \sum_{i=1}^{N} |\hat{\mathbf{z}}_i - \mathbf{z}_i|^2$$

where $\hat{\mathbf{z}}_i$ is the reconstruction of the sample $\mathbf{z}_i$ after projection into the PCA space (cf. A.1) estimated on the corresponding training



(a) Trajectory representation of a samples' time-series ($t_1 - t_7$).

(b) SMART scaling approach (adopted from [Keun 04])

Figure 2.19: Normalization of response magnitude to an applied pharmaceutical using a trajectory representation in the score space of a PCA model (see text for explanations).

set. The Q^2 value is defined with respect to the sum of squares SS of the centered data by

$$Q^2 = 1 - \frac{\text{PRESS}}{\text{SS}}$$

representing the quality of the estimated model. Subsequently, the control model is optimized by an iterative procedure restricting the selection of samples used for estimation of the model to those with a low reconstruction error after projection (cf. A.1). By means of this optimization routine control samples consistently showing abnormal spectral profiles induced by influences like stress or undiagnosed diseases can be excluded (cf. [Ebbe 07]). Finally, 4 023 of the originally 6 260 control samples are used for estimation of the control model.

exclusion of abnormal control spectra

Subsequently, the control model is used to select dosed samples for the further analysis. A sample is regarded as toxic if at least 50 % of samples corresponding to the same treatment are assumed as abnormal and if a significant toxicity of the treatment could be detected in histopathology. This selection has reduced the set of samples from treated animals showing a toxic effect to a set of size 2 056 samples. Thus, less than one third of all samples are assumed to indicate the intended reaction to the respective treatment. Animals not showing the reaction as previously assumed due to resistance, fast recovery or late onset of the toxic episode are common observations in metabonomic studies, denoted to as *non-responder*.

selection of toxic samples

Samples selected for further analysis and reduced by PCA transformation are normalized by a two-stage scaling procedure. Initially, differences in the degree of reaction to an applied treatment

data scaling

SMART procedure

trajectory representation

are compensated using the *SMART* procedure as shown in figure 2.19b. In this approach samples of the same treatment from different time-points are reduced by PCA transformation and represented as trajectory as shown in figure 2.19a [Keun 04]. The pre-dose sample of each treatment is subtracted from corresponding samples, achieving a compensation of pre-dose differences between different animals. The response magnitude to an applied pharmaceutical is normalized according to the maximum effect response, defined as the sample with the maximum squared sum of values within a time-series of samples.

VAST scaling

Subsequently, larger variations in some spectral regions are compensated by *variable stability scaling (VAST)* [Keun 03]. Thereby, each variable is scaled to unit variance and multiplied by the ratio of mean value and standard deviation. By application of this scaling method the variance of features with large variation relative to their mean value is decreased.

CLOUDS modeling

These normalized samples serve as basis for the prediction of specific organ toxicities by the CLOUDS approach, a technique based on probabilistic neural networks [Spec 90]. Each class (= treatment) is represented by a combination of several Gaussian densities. Each sample of the corresponding class serves as center point of a Gaussian density, respectively, and a predefined standard deviation σ is used. An exemplary data set and its CLOUDS representation are shown in figure 2.20. The probability for an n-dimensional sample $\mathbf{z}$ belonging to class A containing N_A samples is determined by summation over the probabilities of the sample belonging to each Gaussian density of the particular class.

$$p_A(\mathbf{z}|\mathbf{z}_{i\in A}) = \frac{1}{N_A(2\pi\sigma^2)^{n/2}} \sum_{i\in A} \exp\left(\frac{-|\mathbf{z}-\mathbf{z}_i|^2}{2\sigma^2}\right)$$

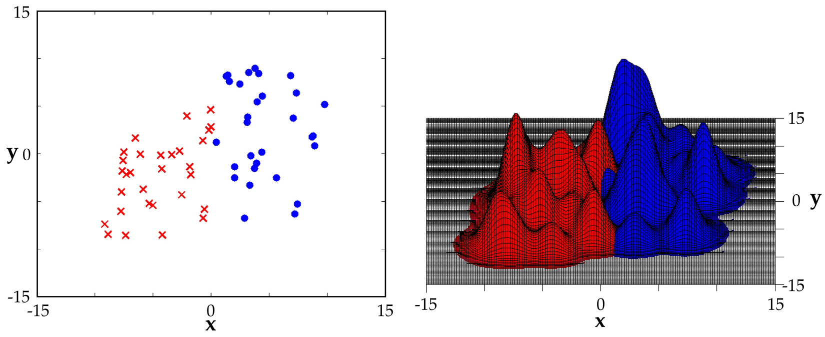
Strictly speaking, the presented classification approach is more similar to a mixture density classifier than to a neural network approach. Neither a network training nor a network structure is used in this approach. However, the approach has been derived from probabilistic neural network (PNN) theory and is applied, as stated by the author, in a non-neural architecture.

The formulation of class probability can be extended to a similarity measure between two classes A and B by assessment of the *overlap integral*, defined as

$$O_{AB} = \frac{1}{N_A N_B(4\pi\sigma^2)^{n/2}} \sum_{i\in A}\sum_{j\in B} \exp\left(\frac{-|\mathbf{z}_i-\mathbf{z}_j|^2}{4\sigma^2}\right) .$$

The *overlap similarity* between classes A and B is finally derived by a normalization of the overlap integral to $[0\ldots1]$ as follows

$$S_{AB} = \frac{O_{AB}}{\sqrt{O_{AA}O_{BB}}} .$$



(a) Exemplary two-class data set

(b) CLOUDS data representation

Figure 2.20: Representation of a data set by the CLOUDS approach by Gaussians centered at each data point with predefined standard deviation.

similarity matrix

The overlap similarity measure is used to define a similarity matrix. Rows of this matrix correspond to classes described by treatments from the training set. Columns contain groups of samples from treatments to be classified. Optimization of the Gaussian smoothing parameter σ can be achieved by computation of the similarity matrix between the treatments in the data set. The value maximizing the Shannon entropy of the matrix values is selected for the further progress. Thereby, an intermediate configuration between no overlap of classes due to a small value and great overlap induced by a high value of σ is achieved.

classification w.r.t. known treatments

Erroneously detected similarities between treatments, induced by the elimination of sample values containing DRCs (cf. [Ebbe 07]), are excluded and the three most similar treatments are determined. These similarities are allocated to one of four levels corresponding to very high, high, low and very low. Whether a treatment can be classified as toxic for a particular organ is determined according to the similarity level of the three most similar treatments with known main organ effect. Tests concerning the similarity levels of the three most similar treatments are applied. A class label is assigned in case of agreement or very high confidence. The significance of this classification is assessed by a permutation test, randomly selecting the top three hits for the classification of a treatment.

design of spectral processing and classification system

To sum up, the classification approach developed within the COMET project comprises methods for spectral preprocessing and normalization. Following this, a system for estimation of similarities between different treatments allowing for a prediction of organ toxicities is defined. By an experimental evaluation of the expert system on the given data set by leave-one-out cross-validation seven out of overall 17 substances inducing kidney toxicity could be detected. Thereby, 24 out of 31 substances inducing liver toxicity could be classified correctly [Ebbe 07].

Generally, results achieved by the COMET approach cannot be objectively compared to other approaches since the data set is not publicly available. Previous investigation of methods for the interpretation and classification of a subset of the presented data set, including hierarchical cluster analysis, principal component analysis and k-nearest-neighbor classification [Beck 03], gave first reasonable classification results. However, these alternative classification approaches evaluated on a comparable data set could not outperform the CLOUDS approach.

Alternative Classification Approaches

genetic programming

Beyond the COMET project only very few attempts for automatic classification of NMR spectra have been published. A first classification approach presented by Gray et al. in 1998 applies genetic programming [Gold 89] for the determination of the best combination of functions on the feature representation of 75 NMR spectra from human brain tumor extracts [Gray 98]. Thereby, features were extracted by projection of the spectra on the first 20 principal components from PCA (accounting for 99 % of the variance). The classes were defined as meningioma[6] and non-meningioma. Score values were combined using adapted arithmetic, trigonometric, logical and conditional functions and classified according their sign to one of the two possible classes. Evaluation of classification accuracy has been performed based on a cross-validation approach and 70 % of the samples from the test set could be predicted correctly.

SIMCA

outlier detection by PCA score plots

sparse data sets in metabonomic studies

The data set used in the previous study has been relatively small due to the complex sample acquisition process and expensive measurement methods. Holmes et al. presented in the same year *soft independent modelling of class analogy (SIMCA)* [Holm 98a] as an alternative method for classification of NMR spectra based on a set of 244 spectra from rat urine. A separate PCA model is built for the samples of each toxin type in SIMCA and new samples are assigned to the class with the minimum distance after projection. A PCA score plot revealed some samples with an unusual response to the applied pharmaceutical, and also some outlying clusters which were excluded prior to further analysis. Exclusion of these outlier samples has improved the classification accuracy from 84 % to 98 % on the test set. For the final set of 191 samples 11 compounds have been applied, which are only 14 % of the amount of samples used in the COMET project. Thus, these results are not directly comparable with those achieved in the COMET project. This is a reoccurring problem since no data sets for metabonomic studies are publicly available and usually

[6] Meningiomas are the most common tumor type in the central nervous system.

neither financial support nor the biomedical background for the acquisition of a reasonable and large-scale data set is present.

2.5 SUMMARY

analysis of biosamples

NMR spectroscopy

The improvements in spectral resolution and quality of NMR spectra achieved in the last decade allow for the non-destructive analysis of the composition of biosamples. Alternative measurement techniques from clinical chemistry can specifically quantify metabolites, enzymes or electrolytes. However, these have to be defined beforehand based on knowledge of relevant markers for the physical reaction to be detected. Furthermore, the measured concentrations have to be interpreted by an expert, which is a time-consuming process and the interpretation is dependent on the experience and knowledge of the expert. NMR spectroscopy principally measures each NMR active compound of sufficient concentration in a sample without prior preselection of putatively relevant substances. Thus, NMR spectra contain several signals and the goal of a detailed analysis is to extract the relevant information for the topic to be investigated.

physical reactions induce changes in the spectral profile

Analysis of spectroscopic data from biosamples with respect to physiological reactions induced by pathophysical stimuli is pursued in the field of Metabonomics. Information coded in spectroscopic data by peak signals is connected to known properties of the sample donor organism like gender or the physical reactions induced by an applied pharmaceutical. Peak signals, that show an intensity change induced by a relevant physiological reaction, can provide information on the underlying biological process. This information can also be used to detect this specific physiological reaction by the analysis of unknown samples.

compensation of spectrum modifications

Although NMR spectroscopy is susceptible to perturbations by physiochemical factors or sample properties, several methods have been proposed to reduce the impact of these sources allowing for a reasonable analysis of a data set. Thus, information on the health status or other physiological properties of an organism can be derived based on urine samples in a non-invasive automated analysis without the necessity of interpretation by an expert. Identification of peaks relevant for the detection of certain physiological reactions can give a deeper insight into the underlying biological mechanisms – an important aspect in safety pharmacology.

automated analysis of unknown samples

This information is usually acquired using methods from the field of multivariate data analysis. Relevant information in the given (and mostly labeled) data set is extracted and used for low-dimensional data visualization or direct analysis of the relevant variables. These correspond to spectral regions in Metabonomics. This approach requires manual interpretation but an automated analysis of unknown samples is desired. This could be achieved by methods from the field of pattern recognition, but considering the current literature only few approaches have been published.

One drawback of most pattern recognition methods is the necessity of a large-scale data set for reasonable classifier training and classification system evaluation. Collection of urine samples from experimental animals is easier than e.g. biopsy at humans. However, data sets of NMR spectra from urine samples are usually quite small due to the complex design of the experiments and the expensive measurement process. Usually, not every sample exhibits the reaction to a treatment as previously desired. Furthermore, artifacts induced by experimental conditions can alter the samples as seen within the COMET project. These samples have to be excluded from the further analysis, further reducing the size of the data set.

usually only small data sets available

Only large-scale projects like COMET can afford the buildup of huge databases by financial support of industrial partners. Up to now, this has been the only project up to now in this dimension and the data set is not publicly available. Thus, most projects aiming at the automatic classification of NMR spectra have to deal with the high data-dimensionality and rather small amount of samples in addition to the rather noisy data due to perturbations by external factors.

large data sets not available for public use

To sum up, NMR spectroscopy is a valuable method for analysis of biofluids in the context of safety pharmacology. But prior to an application of methods from the field of pattern recognition for automated classification of new samples several preprocessing steps are necessary. Subsequently, pattern recognition methods allowing for a robust classification even based on small data sets have to be applied.

3 MULTIPLE CLASSIFIER SYSTEMS

Pattern recognition methods aim at the assignment of class labels to objects, whereby each object is described by a set of features. The criteria for discrimination between different classes are automatically determined based on the given attributes. Generally, different preprocessing and feature extraction methods are available for different kinds of data and the optimal classification algorithm is usually not known beforehand. Thus, in order to derive a classification system with optimal classification accuracy, methods used within this system have to be evaluated and the combination with the best performance is usually applied for classification.

optimization of preprocessing and feature extraction methods

Interestingly, configurations with comparatively low classification accuracy are completely discarded and not further analyzed. However, samples misclassified by different classifiers do not necessarily overlap. Thus, each classifier can offer complementary information on a given data set. This distinction is induced by differences in the data set or the applied classification algorithm. The combination of different classifiers' predictions for a consensus decision could improve classification performance. Thereby, the principle of democracy is applied in order to achieve a consensus decision and not only the prediction of the best performing classifier is used. This idea motivated the still ongoing research in the field of *multiple classifier systems* - also denoted as *ensemble systems* - in the middle of the 1990s. Multiple classifier systems are based on a well-defined theoretical foundation and have successfully been applied to several complex classification problems as a result of this research.

complementary information of different classifiers

combination of classifiers in an ensemble system

This chapter will give an overview of the principles of multiple classifier systems. Although multiple classifier systems are a promising classification technique, they have not been applied to the classification of spectroscopic data in the context of Metabonomics. Thus, remarks on the application of the presented techniques in this section will be universal or focus on similar applications. Basic notation conventions and concepts consistently used during this thesis will be introduced in the following section 3.1. Section 3.2 will give an overview of properties of ensemble systems and theoretical assumptions on the advantage of combining multiple classifier in an ensemble. Methods for the construction of varying data representations for the training of different classifiers will be presented in section 3.3. A selection of frequently applied classification approaches as base classifier in an ensemble

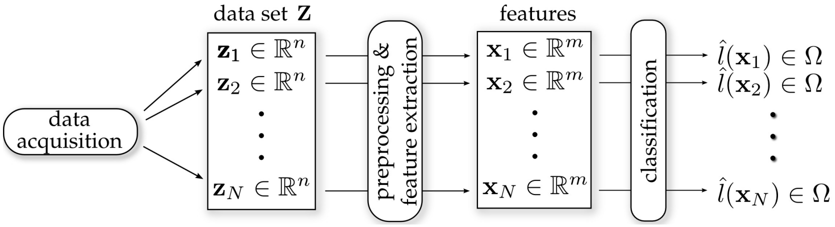

data set $\mathbf{Z}$
$\mathbf{z}_1 \in \mathbb{R}^n$
$\mathbf{z}_2 \in \mathbb{R}^n$
$\mathbf{z}_N \in \mathbb{R}^n$
data acquisition
preprocessing & feature extraction
features
$\mathbf{x}_1 \in \mathbb{R}^m$
$\mathbf{x}_2 \in \mathbb{R}^m$
$\mathbf{x}_N \in \mathbb{R}^m$
classification
$\hat{l}(\mathbf{x}_1) \in \Omega$
$\hat{l}(\mathbf{x}_2) \in \Omega$
$\hat{l}(\mathbf{x}_N) \in \Omega$

Figure 3.1: Overview of modules in a classification system.

will be presented in section 3.4, but generally any classification algorithm can be applied. Combining the predictions of all experts by aggregation methods as shown in section 3.5 is usually the last step in ensemble classification leading to a consensus decision for prediction of a sample. Boosting and random forests as extensively investigated and often applied ensemble approaches will be explained separately in section 3.6 and 3.7.

A large variety of ensemble approaches have been published for numerous applications since the first investigations on the application of multiple classifier systems. Thus, the descriptions in this section are not intended to be exhaustive and will cover the generally most important and most relevant approaches for the intended application. If no reference is explicitly given, the argumentation will be based on [Diet 00, Kitt 98, Kunc 04], which also serve as good starting point for further reading on multiple classifier systems.

3.1 NOTATION CONVENTIONS AND PATTERN RECOGNITION CONCEPTS

general structure of a classification system

The basic principle of pattern recognition systems is the classification of samples into previously defined classes in a process as shown in figure 3.1. Beginning with the data acquisition stage the process proceeds with preprocessing and feature extraction methods. The extracted features serve as data basis for the classification into distinct classes. Thereby, classes are characterized by the similarity of samples they contain, while these samples should be dissimilar to samples of other classes. Depending on the particular application, the distinction between different classes can be unambiguously defined or rather depend on the interpretation of a set of samples.

For a consistent notation throughout this thesis a set of N n-dimensional samples will be denoted as data set $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_N\}$, $\mathbf{z}_j \in \mathbb{R}^n, j = 1, \dots, N$. Each sample $\mathbf{z}_j$ in the data set $\mathbf{Z}$, has previously been assigned or has to be classified to a class $l(\mathbf{z_j}) \in \Omega$, whereby Ω is a set of c distinct classes $\Omega = \{\omega_1, \dots, \omega_c\}$.

Usually, samples will not directly be used for classification, but more characteristic representations of the samples are cal-

		$l(\mathbf{x})$	
		0	1
$\hat{l}(\mathbf{x})$	0	true negative (TN)	false negative (FN)
	1	false positive (FP)	true positive (TP)

Table 3.1: Confusion matrix of a two-class problem.

culated by feature extraction methods (e.g. projection methods, signal processing methods, etc.). Thereby, an n-dimensional sample $\mathbf{z}$ is described by a feature vector $\mathbf{x} = [x_1, \ldots, x_m]$ in an m-dimensional *feature space* with $m \ll n$. This low-dimensional feature representation describes all information relevant of the discrimination of distinct classes and improve performance of statistical pattern recognition methods by decreasing the data dimensionality while retaining the amount of samples. Generally, feature representation types can be separated into numerical and categorical features, but within this thesis only numerical features will be discussed. Based on the m-dimensional feature representation of a sample $\mathbf{x}$ a classifier D has to be defined, achieving a classification into a single class $\hat{l}(\mathbf{x}) \in \Omega$.

low-dimensional feature representation

Before some classification approaches will be presented later in this thesis, accuracy measures and training strategies will be shown in the following.

3.1.1 *Evaluation of Classification Performance*

A basic prerequisite for the successful development of a classification system is the determination of its classification performance. This information can either be used for the optimization of a single classifier or as performance measure in comparison with alternative classification systems. Different measures for estimation of classification performance exist dependent on the classification problem to be solved. In the following, performance measures specific of two-class problems will be presented, since the usual problem in metabonomic applications is to discriminate between samples labeled as non-toxic and toxic. An overview of alternative measures also applicable for multi-class problems can be found in [Bald 00].

assessment of classification performance

Given a data set $\mathbf{X}$ and labels for each sample $l(\mathbf{x_i})$, $i = 1, \ldots, N$, the performance of a given classifier D can be evaluated by the comparison of the original labels with the predicted labels $\hat{l}(\mathbf{x_i})$ achieved by classification. Therefore, a *confusion matrix* as shown in table 3.1 is built counting the amount of negative samples ($l(\mathbf{x_i}) = 0$) labeled as negative (true negative) and positive (false positive), and positive samples ($l(\mathbf{x_i}) = 1$) labeled as positive (true positive) and negative (false negative).

confusion matrix

classification accuracy

The overall classification performance of D is expressed by the *classification accuracy* Λ, comparing the amount of correctly labeled samples with the overall amount of samples.

$$\Lambda = \frac{TP + TN}{FN + FP + TN + TP}$$

sensitivity

specificity

In binary classifications (e.g. classification as negative or positive with respect to a drug-induced organ toxicity in drug design) two additional measures are usually given in order to indicate the performance of the classification system regarding different aspects. The *sensitivity* measures the percentage of positive samples which have been correctly identified as positive, while the *specificity* measures the proportion of negative samples classified as negative.

$$\text{sensitivity} = \frac{TP}{TP + FN} \qquad \text{specificity} = \frac{TN}{TN + FP}$$

confidence interval

An experimental evaluation of a classification system on a finite test set can generally indicate the capabilities of the system. However, this performance cannot be guaranteed for the classification of alternative data sets, since only a fraction of all possible samples generated by the underlying statistical process are investigated. A confidence interval $[\Lambda_l, \Lambda_u]$ can be used in order to assess the variability of an evaluated classification performance achieved on a data set of size N (cf. e.g. [Schl 03]). The lower bound of this interval is defined by

$$\Lambda_l = \frac{N}{N + q^2}\left(\Lambda + \frac{q^2}{2N} - q\sqrt{\frac{\Lambda(1-\Lambda)}{N} + \frac{q^2}{4N^2}}\right)$$

and the upper bound by

$$\Lambda_u = \frac{N}{N + q^2}\left(\Lambda + \frac{q^2}{2N} + q\sqrt{\frac{\Lambda(1-\Lambda)}{N} + \frac{q^2}{4N^2}}\right) .$$

rating of changes in classification accuracy as statistically significant

Within these calculations q represents the $(1 - \frac{\alpha}{2})$ quantile of the standard normal distribution, whose values can be found in literature. A commonly used confidence level is $1 - \alpha = 95\,\%$, indicating that the real classification accuracy lies in the given interval with a probability of 95 percent. Furthermore, the confidence interval can be used in order to rate an achieved improvement or degradation of classification performance caused by modifications of the classification system as statistically significant. If a classification performance outside of the confidence interval is achieved, the change is designated as statistically significant and has not only been obtained due to the finite test set.

Two-class data sets employed for the evaluation of classification performance are usually designed to have a comparable amount of negative and positive samples. Thereby, the classification performance, which is mostly used as the primary performance measure, is not mainly influenced by the predictions for the samples of the prevalent class. However, an almost equal amount of positive and negative samples cannot always be achieved in case of sparse data sets. Thus, a robust measure for performance assessment even in case of imbalanced data sets is needed. One of these measures is the *Matthews Correlation Coefficient (MC)* [Bald 00], incorporating all four entries of the confusion matrix for the determination of classification performance. This measure ranges from -1 to $+1$, whereby -1 indicates full disagreement and $+1$ full agreement between true and predicted class labels. A random prediction is indicated by a value of 0.

imbalanced data sets

matthews correlation coefficient

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

According to literature, a multiplicity of further measures for classification performance dependent on the application or objectives of evaluation have been defined. Measures presented in this section are relevant to this thesis and the interested reader is referred to the literature for further information (cf. e.g. [Bald 00]).

3.1.2 *Strategies for Classifier Training and Evaluation*

Besides the large variety of measures for the estimation of classification performance described in the literature (cf. e.g. [Bald 00]), experimental design for *training* and *testing* of a data set has been discussed for a long time [Tous 74]. Although, using the same data set for training and testing has been proposed as *resubstitution method* [Smit 47], this is generally not a valid procedure for the estimation of the classification performance Λ. The resulting classifier is expected to be *overtrained* on the data and will fail on unseen data. The *generalizability* of a classification system, achieving a reliable prediction for unseen data, has become an important prerequisite of classification systems for their successful application to real-world classification tasks.

design of experimental evaluations

generalizability

A method for training and evaluation of a classifier allowing for a reliable estimate of Λ is the *K-fold cross-validation* [Koha 95]. Thereby, the data set is randomly split into K parts of comparable size and classifier evaluation is carried out on one part, whereby the union of all other parts is used for classifier training. This training and evaluation procedure is repeated K times. Each time a different part is used for testing and the final Λ is calculated by averaging over all test results, respectively. In case of K equal to N the method is called *leave-one-out (LOO)* cross-validation, as in each of the K folds a single sample is used for testing.

k-fold cross-validation

leave-one-out cross-validation

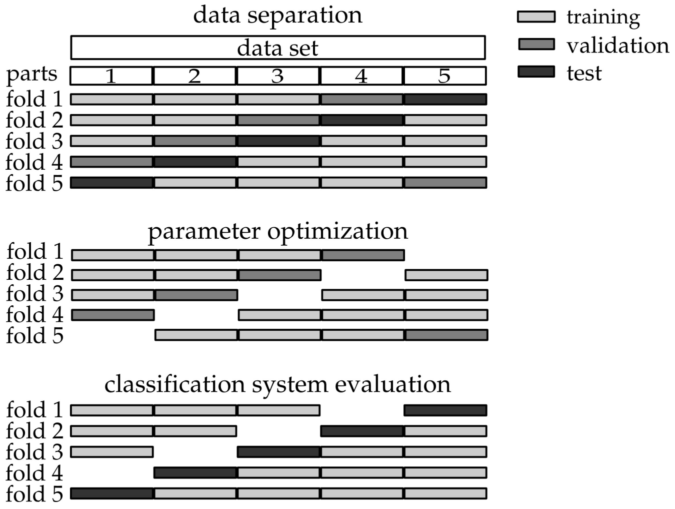


Figure 3.2: Structure of data set separation and combination for five-fold cross-validation and test.

parameter optimization

independence of the test set

Most classification algorithms require the optimization of specific parameters in order to adapt the classification decision to the particular problem. If these parameters cannot be determined according to data properties like the amount of samples or dimensionality, these are usually optimized according to the classification performance on the test set under parameter variation. Thereby, overfitting on the training set can be avoided by the selection of the configuration optimally performing on the test set. However, incorporating knowledge in the final performance of the classification system for parameter selection violates the necessity of independence of the test set.

k-fold cross-validation and test

Parameter optimization and an independent test set can be achieved by applying the cross-validation principle, but additionally using one of the $K - 1$ training parts for validation as shown in figure 3.2. Thereby, in case of a five-fold cross-validation and test approach three fifths of the five data set parts are used for the training of the classification algorithm, one fifth for parameter optimization, and the remaining fifth for final testing. This partitioning is rotated K times and the final results are averaged over all respective classifications. However, in case of small data sets this approach can reduce classification performance of the final classifier due to the decreased amount of samples incorporated in classifier evaluation and training.

If multiple evaluations based on different parts for testing have been used for the optimization of a classifier, the question arises which classifier should be used for final classification. A straightforward solution would be to use the best performing classification approach combined with the optimized parameters for the training of a classifier on the whole data. However, an alternative way of combining classifiers trained and optimized on individual cross-validation sets is the aggregation in an ensemble system.

3.2 CONCEPTS OF MULTIPLE CLASSIFIER SYSTEMS

The design of a classification system achieving a reasonable performance on a certain classification task generally consists of defining an adequate feature extraction and classification method. Even though this sounds quite straightforward, methods achieving good results in other fields may not be suitable for the current application. In order to determine the optimal system design several combinations of preprocessing and feature extraction methods have to be evaluated. Classification systems evaluated within this optimization process usually differ in their data representation or focus on different data aspects for classification. Thus, each classification approach performs well for a certain subset of the test samples and misclassifies the remaining samples. If the sets of misclassified samples from each evaluated classification approach do not strongly overlap, each of the classification approaches provides complementary information on the samples.

different approaches for classification system optimization

disjunct false classifications

Multiple classifier systems aim at taking advantage of this property. Therefore, predictions of several *base classifiers*, each giving diverse prediction results for the same samples, are combined instead of relying on a single classifier. Thereby, an improved classification performance and generalizability is expected. This basic principle of ensemble systems is graphically illustrated in figure 3.3. In this example a set of three classifiers D_1, D_2, D_3 are given, misclassifying partially different regions of samples from the data space $\mathbf{U}$ (cf. figure 3.3a). The combination of the classifiers by assigning the label predicted by at least two of the three classifiers leads to a smaller region of misclassified samples in comparison with the single classifiers as shown in figure 3.3b.

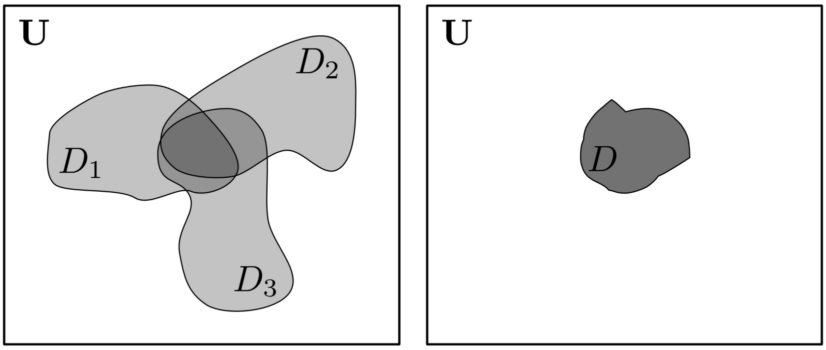
combination of base classifiers

This example illustrates two basic prerequisites that have to be fulfilled by multiple classifier systems according to Hansen et al. [Hans 90]. As a first prerequisite, the base classifiers' regions of misclassified samples should not strongly overlap. Otherwise, no improved classification performance can be achieved. Furthermore, the classification performance of each individual classifier has to be better than random guessing in order to decrease the region of misclassified samples by increasing the number of experts in the ensemble.

diverse and accurate base classifiers

3.2.1 *General Structure of Multiple Classifier Systems*

The general architecture of an ensemble system (cf. figure 3.4) is almost identical to the structure of common classification systems as previously shown in figure 3.1. Differences solely exist



(a) Regions of misclassified samples for classifiers D_1, D_2, D_3.

(b) Region of misclassified samples after combination of the three classifiers to a single classifier D.

Figure 3.3: Exemplary graphical illustration of the ensemble effect. The region of misclassified samples (gray area) in the data space $\mathbf{U}$ is reduced by combining the three classifiers to a single classifier (adopted from ([Kim 03]).

regarding the classifier training and determination of a final classification for each sample. This implies two important steps which are specific for each multiple classifier system and are also crucial for the classification performance. First of all an *ensemble creation* method has to be applied in order to train L diverse classifiers. Each of these *base classifiers* D_i, $i = 1, \dots, L$ predicts the label of a given sample $\mathbf{x}$ and these predictions $\mathbf{s} = [s_1, \dots, s_L] \in \Omega^L$ have to be combined to a final prediction $\hat{l}(x)$ by aggregation methods.

ensemble creation

base classifier

ensemble aggregation

Variations concerning the ensemble creation method, the base classifier algorithm selection or the ensemble aggregation method are usually evaluated in order to achieve the best classification performance for the current classification problem. Although some methods tend to perform well on a variety of applications, they are not expected to be optimal for every application.

3.2.2 *Why Do Ensembles Work?*

no theoretical but empirical proof

Generating an ensemble of accurate and diverse classifiers combined via a voting scheme is expected to increase the classification performance in comparison with each individual classifier. But is it always possible to construct good ensembles in practice? Although no theoretical proof can be given to affirm this question, Dietterich presented three fundamental reasons that demonstrate the advantages of multiple classifier systems [Diet 00].

statistical reason

Classifier training utilizing a learning algorithm involves the selection of a classifier from the space of available classifiers as shown in figure 3.5. If only small data sets are available no robust estimation of the underlying statistical models can be achieved. In this case several classifiers D_i can be determined, each giving a

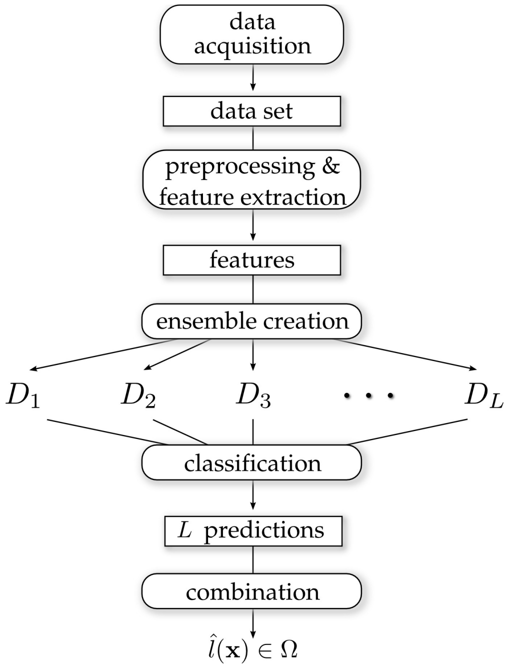
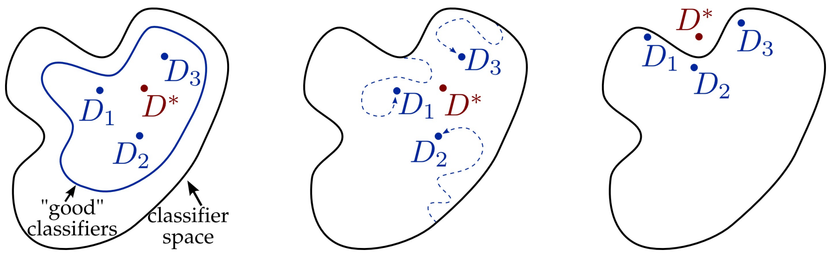


Figure 3.4: General structure of an ensemble system.

reasonable classification accuracy as indicated by the inner region in figure 3.5a. These classifier perform worse in comparison to the optimal classifier D^*, but averaging over their predictions is expected to compensate individual false classifications and improve classification performance.

Classifier training usually involves the optimization of specific parameters in order to obtain the optimal classifier that can be achieved in the classifier space. However, these algorithms can get stuck in local optima and dependent on the starting point of optimization different classifiers close to the optimal classifier are generated (cf. figure 3.5b). Again, by averaging over the predictions of the individual classifiers, classification

computational reason



(a) Statistical

(b) Computational

(c) Representational

Figure 3.5: Different examples for explanations why an ensemble classification can improve classification performance compared with a single classifier in practical applications (adopted from [Diet 00] and [Kunc 04]).

performance closer to the optimal classifier than each individual classifier can be achieved.

representational reason

The third reason for an improved classification performance of the ensemble is based on the classifier space spanned by the applied classification algorithms as shown in figure 3.5c. If for example a non-linear classifier is needed for optimal classification of a data set but only linear classifiers are applied, the non-linear decision can be approximated by averaging over the decisions of linear classifications from base classifiers. Defining a different classifier space including D^* can generally achieve a high classification accuracy, but it is more straightforward to train an ensemble of simple classifiers than a single complex classifier. Furthermore, the single complex classifier is again endangered to get stuck in local optima during optimization, which can be avoided by averaging over several simple classifiers.

Although the given examples are no theoretical foundation proving the general predominance of multiple classifier systems over single classifier approaches, they demonstrate the suspected benefit by averaging over several simple classifiers. The success of multiple classifier systems in quite a number of practical applications and the demonstrated suitability for special cases further supports this assumption.

3.3 ENSEMBLE CREATION TECHNIQUES

Training a classifier based on a given data set is everyday practice in the field of pattern recognition. But achieving an ensemble of multiple accurate and diverse classifiers requires additional *ensemble creation* methods. In some applications the problem arises to combine different classifiers due to different input sources or the use of different classification approaches. In these cases the strategy for the creation of the ensemble is already defined. Notwithstanding, this is rather the exception than the rule. A selection of the most important generic ensemble creation methods that can be used in practical applications will be presented in the following.

3.3.1 *Data Set Modification*

classifier performance dependent on the training set

Classifier training using a learning algorithm aims at deriving regularities of the data or estimating the underlying statistical model for discrimination between distinct classes. These discrimination criteria are learned based on a given data set. Thus, the data set is crucial for the performance and predictions made by the classifier. Consequently, diverse classifiers can be achieved by using different subsets of the data set for each of the classifier training procedures. This results in a set of classifiers predicting

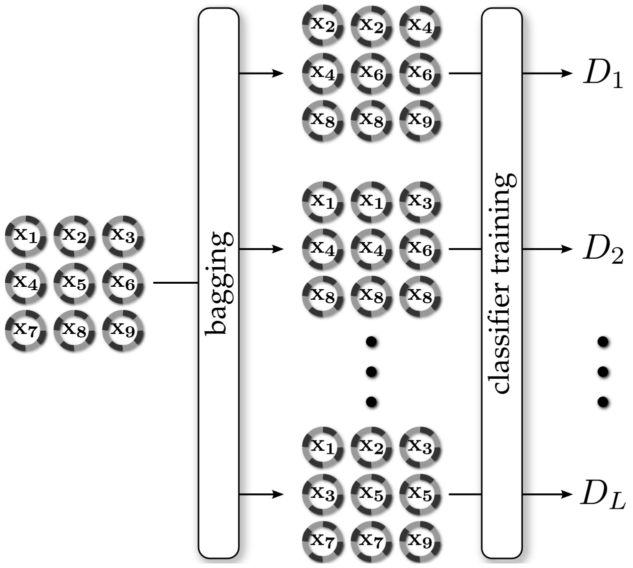


Figure 3.6: Bagging procedure for creation of L different classifiers for an exemplary data set. For each bootstrap replicate nine samples are randomly selected with replacement from the data set of size nine.

the label of a new sample based on a particular data subset used for training.

In order to achieve diverse classifiers by this procedure, *unstable* learning algorithms [Brei 96a] are required, which show a strong dependence between their prediction and samples used for training. If even small modifications of the data set do not lead to changes in classification results, predictions from each classifier in the ensemble will be identical. No improvement in classification performance of the ensemble system can be achieved in this case. Neural networks and classification trees are regarded as unstable classifiers, while the k nearest neighbor is stable [Brei 96b].

unstable classifiers

A quite straightforward method for the selection of data subsets was presented by Breiman, denoted as *Bagging* (abbreviation for *b*ootstrap *agg*regat*ing*) [Brei 96a]. Subsets are determined by a random selection of N samples with replacement from the original data set as shown in figure 3.6. The resulting subsets are denoted as *bootstrap replicates*, containing on an average 63.2 % of the original samples and several replicates [Brei 96a]. Bagging is combined according to the originally proposed approach with decision or regression trees as base classifiers. Ensemble aggregation is achieved by the assignment of the class with the maximum number of votes. Several experiments have shown an improved classification performance by using the bagging procedure. Bagging can also be combined with alternative classification algorithms and ensemble aggregation techniques as long as unstable classification algorithms are used.

bagging

Partitioning of a data set is also a quite common method used for the creation of cross-validation sets as shown in section 3.1.2. Thereby, the whole data set is split into K parts and classifiers are trained on the union of all parts each time excluding one part. This can also be applied for the creation of an ensemble of classifiers, whereby background knowledge on similarities of samples can be introduced in order to improve diversity of the resulting classifier. Due to the similarity of this procedure to the cross-validation procedure ensembles built according to this method are denoted as *cross-validated committees* [Parm 96].

cross-validated committees

3.3.2 *Input Feature Modification*

Not only the selection of training samples but also the definition and selection of features heavily influences the classification decisions and performance of each classifier. Thus, achieving variability in the selection or the type of features used for the learning algorithm would cause diverse classifiers by the integration of different information on the samples in the discrimination decision.

random subspace sampling

The *Random Subspace Sampling (RSS)* method originally proposed by Ho for the construction of an ensemble of decision trees selects a subset of features randomly [Ho 98]. Thereby, each classifier uses one out of multiple feature subset combinations as shown in figure 3.7, leading to a set of diverse predictions. This method is expected to perform well especially in the case of redundant features by distributing these among several subsets rather than combining them for a single classification. But as shown by Tumer and Ghosh, RSS could lead to worse results if all features are essential in order to derive a robust classification [Tume 96].

variation of preprocessing and feature extraction methods

The general structure of a classification system (cf. figure 3.1) includes the application of preprocessing and feature extraction methods prior to the final classification. Several combinations of these methods are possible and the optimal combination is usually determined by an experimental evaluation. However, each combination usually emphasizes different aspects of the data which seems to be advantageous to classification purposes. This variability in data representation can again be used for ensemble creation by training a classifier on each combination, achieving a reasonable classification performance and integration into an ensemble.

3.3.3 *Assessment of Ensemble Diversity*

Modifications regarding the selection of samples or features used for the training of a classifier leads to diverse classifiers, which

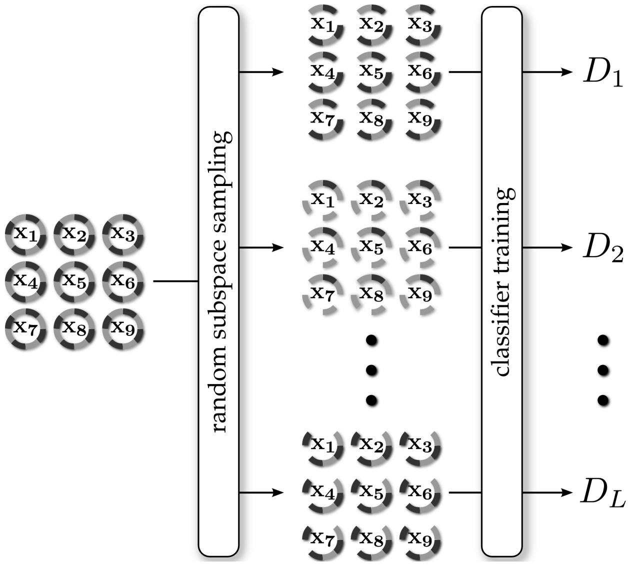


Figure 3.7: Exemplary random subspace sampling procedure for creation of L different classifiers by random selection of a subset of five features from an eight-dimensional data set.

is a major prerequisite for the successful application of multiple classifier systems. But no commonly accepted definition of ensemble diversity can be found in the literature. Kuncheva and Whitaker presented a comprehensive overview and comparison of diversity measures and discuss the question whether the classification performance of an ensemble is strongly connected with classifier diversity [Kunc 03].

Q statistic

The estimation of diversity between a pair of classifiers can be achieved by direct comparison of their predictions for the samples $\mathbf{x}_j, j = 1, \ldots, N$ from the labeled data set $\mathbf{X}$. Coding the predictions of a single classifier by a binary vector containing a 1 at position j if the classifier predicts the sample $\mathbf{x}_j$ in the correct class and 0 otherwise, allows to compare the similarity of classifier D_a and D_b by Yule's Q statistic [Yule 00] as follows

pairwise similarity of classifiers

$$Q_{a,b} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}} .$$

Thereby, N^{cd} represents the amount of positions with value c in the coded prediction vector of D_a and a value of d in the vector of D_b. The pairwise Q statistic is estimated for each combination and averaged over the number of combinations in order to achieve a final Q statistic for a set of L classifiers.

averaged Q statistic

$$Q_{av} = \frac{2}{L(L-1)} \sum_{i=1}^{L-1} \sum_{k=i+1}^{L} Q_{i,k}$$

Independence among the classifiers of an ensemble is indicated by a value of Q_{av} close to zero, while negative values are achieved if different samples are misclassified by the base classifiers. The latter is expected to have a positive effect on classification performance and to be a possible criterion to appraise the quality of ensemble creation techniques.

Entropy Measure

High diversity can intuitively be recognized in case of two-class classification problems if fifty percent of the base classifiers assign the correct label to a given sample, while the remaining classifiers assign the incorrect label. The other extreme is full agreement among all classifiers, indicating no diversity in the ensemble. This intuitive understanding of diversity is reflected by the entropy measure E, indicating maximum diversity by a value of 1, and 0 if full agreement among all classifiers is present.

entropy of predictions

$$E = \frac{1}{N}\frac{2}{L-1}\sum_{j=1}^{N}\min\left\{\left(\sum_{i=1}^{L} y_{j,i}\right), L - \left(\sum_{i=1}^{L} y_{j,i}\right)\right\}$$

where

$$y_{i,j} = \begin{cases} 1, & D_i \text{ classifies } \mathbf{x}_j \text{ correctly} \\ 0, & \text{otherwise.} \end{cases}$$

The entropy measure E is not a standard entropy function since no logarithm function is used. Cunningham and Carney presented in [Cunn 00] a more classical formulation of the entropy measure. But these two measures are similar related to the ensemble accuracy as mentioned by Kuncheva [Kunc 04].

Kohavi-Wolpert Variance

Considering different training sets for a single classifier, Kohavi and Wolpert defined an estimation for the variance of the prediction y for a single sample $\mathbf{x}$ by

$$\text{variance}_x = \frac{1}{2}\left(1 - \sum_{i=1}^{c} P(y = \omega_i|\mathbf{x})^2\right) .$$

variance of predictions for each sample

Transferring this formulation to predictions achieved by different classifiers $D_1, \dots, D_L$ allows for an estimation of ensemble diversity. Simplifying the formula by restricting classifier outputs to correct and incorrect instead of being an element of Ω leads to the following definition of the *Kohavi-Wolpert variance (KWV)* according to [Kunc 04]:

$$\text{KWV} = \frac{1}{NL^2}\sum_{j=1}^{N} Y(\mathbf{x}_j)(L - Y(\mathbf{x}_j))$$

where

$$Y(\mathbf{x}_j) = \sum_{i=1}^{L} y_{i,j}\,.$$

Comparing KWV values of different ensemble classifications allows to determine the ensemble with the largest value as the ensemble with the most diverse classifiers.

Measure of difficulty

The last diversity measure that will be presented in this section has been inspired by a study of Hansen and Salamon [Hans 90], denoted as the *measure of "difficulty"* θ. Thereby, a discrete random variable Z taking values from $\{0, 1/L, 2/L, \ldots, (L-1)/L, 1\}$ is assumed, denoting the proportion of classifiers that correctly classify a given sample $\mathbf{x}$. The L classifiers are used to classify the samples in $\mathbf{X}$, resulting in an estimated probability mass function of Z.

Figure 3.8 depicts exemplary histograms of Z using seven classifiers, each classifying 60 % of the samples from $\mathbf{X}$ in the correct class. In case of independence between the classifiers the histogram is similar to a binomial distribution as shown in figure 3.8a. In this case the ensemble classification accuracy would be increased to approximately 75 % by the prediction of samples into the class with the majority of votes for ensemble aggregation. If the ensemble consists of identical classifiers a full agreement in all decisions is achieved (cf. figure 3.8b). Either 100 % or 0 % of all classifiers predict the correct label and classification performance remains at 60 %. The optimal case of negatively dependent classifiers is shown in figure 3.8c, whereby each classifier correctly predicts different subsets of the data. Thus, a perfect classification is achieved by the combination of diverse classifiers, further supporting the assumed connection between ensemble diversity and accuracy. Generally, an ensemble with negatively dependent classifiers is desired, but due to the generally low diversity of classifiers in most applications, rather positive dependence is usually the case.

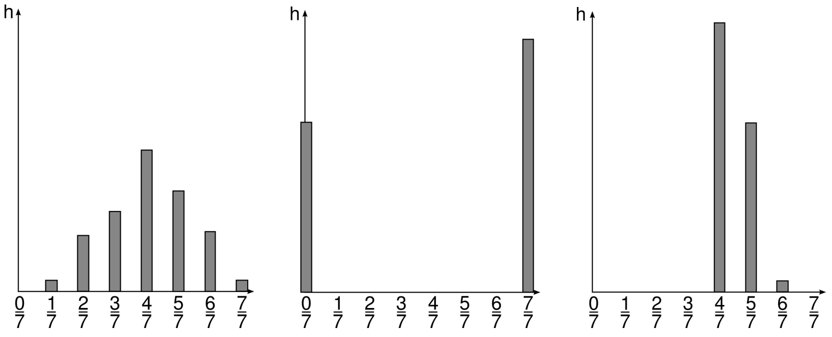
independence of base classifiers

positive dependence of base classifiers

negative dependence of base classifiers

In order to derive a diversity measure according to this representation, the distribution of correctly classified proportions has to be evaluated. The histogram with diverse classifiers shown in the right plot has a small variance, while the ensemble with identical classifiers produces a histogram with large variance as shown in the middle plot. Thus, the variance of the histogram serves as indicator on the ensemble diversity, whereby a variance close to zero corresponds to a high ensemble diversity.

variance of histogram values



(a) Independent (b) Positive dependent (c) Negative dependent

Figure 3.8: Exemplary histograms of values assigned to Z in case of three different dependency situations of classifiers within an ensemble. The x-axis corresponds to the proportion of classifiers assigning the correct label (adopted from [Kunc 04]).

Summary

Beyond the presented diversity measures several other are proposed in literature [Akse 03, Banf 03, Kunc 03, Ruta 03], further illustrating the variety of diversity measures. A comparison of these methods was presented by Kuncheva et al., also addressing the question whether the assumed relation between classification performance and diversity can be confirmed [Kunc 03, Kunc 04]. Within their experiments a strong relation between different diversity measures could be shown, but no measure with strongest relation to improvement in classification accuracy could be denoted. However, Q_{av} is recommended due to its easy calculation and indicating independence by a value of zero and negative dependence by negative values.

no diversity measure with the strongest relation to accuracy

Although a relation between ensemble diversity and classification accuracy could be observed in experiments pursued in [Kunc 04], most practical applications have to deal with quite similar base classifiers. Thus minor changes in diversity would not lead to a significant improvement in classification performance. Nevertheless, diversity in an ensemble has to be enforced in order to achieve a classification performance superior to the single best classifier.

diversity has to be enforced

3.4 BASE CLASSIFIERS

In the progress of ensemble generation, classifiers are trained on data sets generated by ensemble creation techniques. These form the basis for the prediction of new samples by individually assigning class labels to the samples. Only few prerequisites have to be fulfilled by the base classifiers, namely diversity and accuracy. While diversity is mainly induced by ensemble creation methods, unstable classification algorithms can further increase diversity for an improved ensemble performance. The required accuracy of base classifiers is not intended to achieve a perfect classification, but an accuracy better than random guessing [Hans 90]. Thus, they are usually referred to as *weak classifiers* and the desired *strong* classification of the ensemble is achieved by the combination of their predictions.

unstable classifier increase diversity in the ensemble

base classifiers are usually weak learner

However, also strong classification approaches such as support vector machines or neural networks can be used as base classifiers. Even though single strong classifiers can achieve reasonable classification results, the combination of diverse strong classifiers is expected to slightly increase classification performance and – as the most important aspect – improve generalizability.

ensemble of strong classifiers increases generalizability

The choice of the learning algorithm for base classifiers is not restricted to a certain algorithm and the suitability of different approaches has to be evaluated for the problem at hand. Nevertheless, some algorithms have shown good results in several applications and are generally considered to be a good starting point for further investigations of ensemble variations. An overview of commonly used base classifier algorithms is presented in the following.

3.4.1 *k Nearest Neighbor Classifier*

The basic assumption of pattern classification approaches is the similarity of samples from the same class, while being different to samples from other classes. Thus, similarity between samples can be used as criterion for the classification of new samples. Similarity is usually defined in a geometrical sense, indicating similarity between two samples if their distance in the feature space is smaller than to other samples.

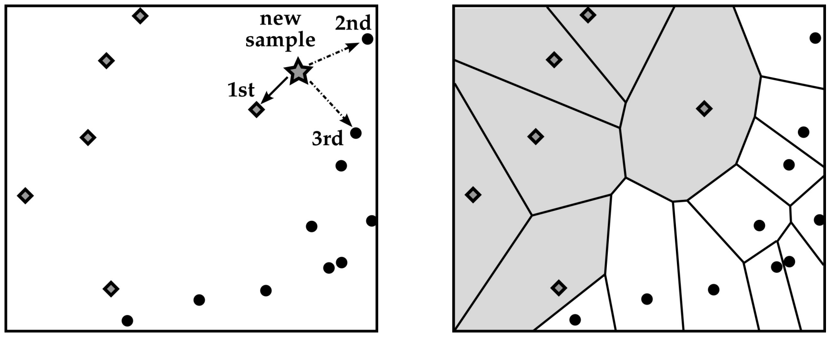
classification according to similar samples

These considerations are directly used in the *nearest neighbor (NN)* classification rule by assuming a labeled set of samples as representatives of their respective classes. New samples are classified by calculation of the distances to all representatives and assignment of the closest representative's label. Figure 3.9a shows an exemplary two-class data set and a new sample to be classified. According to the NN rule the new sample is classified into the class of the diamonds. Classification regions of the data set based on the euclidean distance are shown by a Voronoi diagram in figure 3.9b. The partitioning of the data space is dependent on the chosen metric for distance calculations and according to the current classification problem different metrics can be evaluated.

different metrics

Extending the NN rule by classification to the class most represented among the, e.g., three closest neighbors would lead in the example to a classification into the class of the circles. Generally, extending the classification rule to the k closest samples is applied

extension to the k nearest neighbors



(a) 1-NN and 3-NN classification

(b) Voronoi tessellation

Figure 3.9: Classification of a new sample via the 1-NN and 3-NN classification rule and the corresponding Voronoi diagram.

in real-world applications in order to prevent false classifications in the presence of noisy or falsely labeled samples in the training set. In consistency with NN classification this approach is denoted as *k nearest neighbor (kNN)* classification.

prototype selection

Assuming the whole training set as representatives is straightforward but with the currently available large data sets in mind this procedure can produce problems. Two of these problems are the large memory usage and computational complexity for determination of the distances to all representatives. Thus, common practice is to determine prototypes for a given data set that serve as representatives of the classes, thereby reducing memory usage and computational complexity. Furthermore, the determination of representative prototypes allows for the exclusion of noisy objects, and improvements in classification performance can be achieved [Ferr 99].

One method of the reduction of the set of representatives is the selection of a subset of the original samples (*prototype selection*), thereby optimizing the subset selection with respect to the classification performance on an independent test set (cf. e.g. [Hart 68, Skal 94, Wils 72]). Alternatively, the original data set can be used to derive new samples as representatives (*prototype extraction*) by competitive learning, gradient descent optimization, or bootstrap random methods (cf. e.g. [Bezd 98, Deca 97, Hama 97]).

application in metabonomic investigations

The *k*NN classification approach has been applied for the detection of drug-induced organ toxicities based on NMR spectra by Beckonert et al. in [Beck 03] and has later been reviewed by Keun [Keun 06]. Spectra are preprocessed in this approach using a bucketing procedure with a bucket-width of 0.04 ppm, spectral normalization to a constant sum and exclusion of the water and urea signal (cf. section 2.3). PCA, PLS and hierarchical cluster analysis are first applied in order to visualize groupings of spectra related to the same toxin type by score plot and dendogram analysis, respectively. A first grouping of spectra could be observed by these methods, whereby the groups correspond to spectra of urine samples from experimental animals treated with pharmaceuticals inducing liver or kidney toxicity, or having no toxic effect. Samples that are regarded to indicate a toxic effect but have no significant difference to control animals in the dendogram analysis are excluded from the further analysis.

Classification performance of the *k*NN approach is determined by means of a LOO and two-fold cross-validation design. In the final evaluation the two different types of organ toxicity, and control samples could be detected by an individual accuracy of approximately 90 %. Another important outcome of the experimental evaluation is the drop of classification performance of about 1–4 % and the increased robustness of classification models when using the two-fold instead of LOO cross-validation design. This change in classification performance demonstrates the *optimistic* results achieved by LOO cross-validation and the importance of a valid experiment design. This is a critical aspect in metabonomic investigations where multiple samples from the same animal are collected at different time points.

kNN outperforms PCA and SIMCA approaches

These first experiments using pattern recognition methods for the detection of drug-induced organ toxicities have shown promising results. In comparison to similar studies using PCA, SIMCA [Robe 00, Holm 98a] or probabilistic neural networks [Holm 01] the *k*NN classification approach achieves improved classification results [Beck 03]. Thus, the *k*NN classification approach is a straightforward but competitive classification approach for metabonomic applications.

3.4.2 *Decision Trees*

graph structure

nodes and branches

The most intuitive approach for the discrimination of samples from different classes is to decide according to discriminating properties of the samples. This approach of deducing the class membership by decisions regarding a certain property of the sample is used by *decision trees*, representing the decision process by a graph terminology. *Internal nodes* and the *root* of the graph contain certain features of the samples. Decisions regarding the value of the features are specified by the *branches* to the child nodes. Each *leaf node* represents a class label, whereby the same class label can be represented by multiple leaf nodes. The class relationship of a sample can be traced by following a path fulfilling all conditions at the branches until a leaf node is reached. An exemplary decision tree for discrimination between different fruits is shown in figure 3.10. Initial discrimination is achieved according to the color as nominal feature and final specification of the fruit is based on size, shape or weight of the possible fruits, respectively.

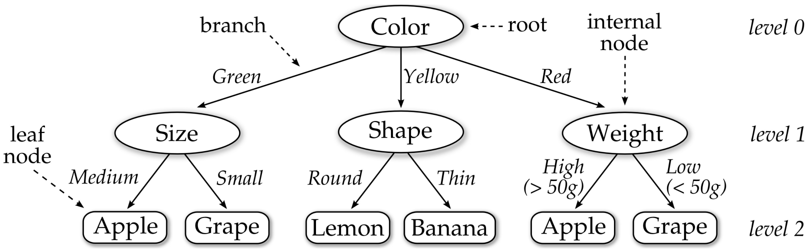


Figure 3.10: Decision tree for discrimination between apple, grape, cherry, lemon and banana (adopted from [Duda 01], p. 397).

tree construction

branching factor

The essential step in order to achieve a classification by decision trees is the tree construction procedure. Thereby, the feature and criterion of each node and branch have to be defined, allowing for the best classification of a given labeled data set. Initially, all samples are assigned to the root and the best feature for the separation of the sample set into subparts and assignment to the child nodes has to be determined. Basically, these *splits* can be binary or non-binary depending on the tree design or the user-defined *branching factor*. Since each non-binary split into K parts can be represented by $K-1$ binary splits only binary splits are assumed in the following. After the separation of samples to the child nodes this progress is iteratively repeated until all samples assigned to a node have the same class label or only few differently classified samples are present. The construction of a decision tree is heavily influenced by the training samples. Therefore, tree classifiers are regarded as unstable classifiers and are often used as base classifier algorithm in ensemble classification systems.

minimization of impurity at nodes

The objective in selection of the best splitting feature and value is the classification accuracy and simplicity of the final tree. Therefore, the impurity of samples at descendant nodes can be measured according to their corresponding class labels and the feature minimizing this impurity is selected for splitting at the current node. In case of nominal features, this splitting results in child nodes according to single categories or subsets of categories, while for real-valued features a split in two child nodes is usually achieved according to a specific threshold.

entropy-based impurity

Assuming a node t in a decision tree and a set of samples with known classes assigned to that node, the probability P_j of class $\omega_j \in \Omega = \omega_1, \ldots, \omega_c$ at that node can be estimated according to the proportion of samples from class ω_j. The entropy of probabilities of all possible classes can be used as one criterion for the definition of the impurity at node t by

$$i_E(t) = -\sum_{j=1}^{c} P_j \log P_j$$

whereby a minimum value $i_E(t) = 0$ is achieved if samples of only a single class are assigned to node t.

A further measure for impurity is based on the misclassification achieved if a class label for a sample at node t is selected randomly according to the probability of the different classes, denoted as the *Gini impurity*.

Gini impurity

$$i_G(t) = 1 - \sum_{j=1}^{c} P_j^2$$

The *misclassification impurity* measure has an even stronger relation to the classification performance. For this measure the current internal node is treated as a leaf node and the expected error by assigning the label of the class with the maximum percentage of samples at the node is measured.

misclassification impurity

$$i_M(t) = 1 - \max_{j=1,\dots,c} P_j$$

An improvement in splitting according to a certain feature is indicated by comparison of the impurity before splitting and the averaged impurity of the child nodes. The feature leading to the maximum improvement is selected for the splitting of a sample subset at node t.

Construction of the tree is stopped if nodes contain only samples of the same class and the node becomes a leaf with the corresponding class. Thereby, a perfect classification of the training set can be achieved if no identical samples with different class labels are in the data set. However, large trees tend to overfit on the training data and early stopping of the tree construction procedure is proposed, leading to impure nodes but achieving an increased generalizability of the decision tree (cf. [Duda 01]). If early stopping is applied, probably beneficial splits beyond the stopping point will not be used for classification, a phenomenon denoted as the *horizon effect* [Duda 01].

overfitting on the training data

horizon effect

Pruning methods have been proposed in order to circumvent this problem by building the tree to its full size and retrospectively pruning superfluous parts of the tree by merging descendant nodes into a leaf node. Esposito et al. gave a comprehensive overview of pruning methods in [Espo 97] to which the interested reader is referred for further information. An extensive comparative study presented in the work of Esposito et al. investigated the purpose of pruning methods on different classification problems. Although a general statistically significant improvement could not be achieved by pruning methods for all problems, pruning methods were regarded as advantageous to most applications. However, pruning is usually not applied for decision trees as base classifier in ensemble methods in order to increase the diversity of the different decision trees. An improved generalizability

pruning methods

is achieved by averaging over the predictions of decision trees. Furthermore, decision stumps are often applied in ensemble methods, thus stopping tree construction after splitting at the root.

CART

The theory of decision trees presented up to now has been mainly oriented according to the *classification and regression tree (CART)* approach proposed by Breiman et al. [Brei 84]. Besides the CART approach two alternative tree designs are commonly used for classification, namely ID3 and C4.5. The major difference between CART and ID3[1] proposed by Quinlan [Quin 86] is the type of features used for the separation of data sets. While nominal and real-valued features can be used within CART, ID3 restricts the type of data to nominal features. In order to handle real-valued features these have to be discretized in attribute bins. The number of descendants at each node is equal to the number of categories for the selected variable. Nodes are split until all features have been used for splitting or all nodes are pure, leading to a depth of ID3 trees equal to the number of input variables.

ID3 tree

C4.5 tree

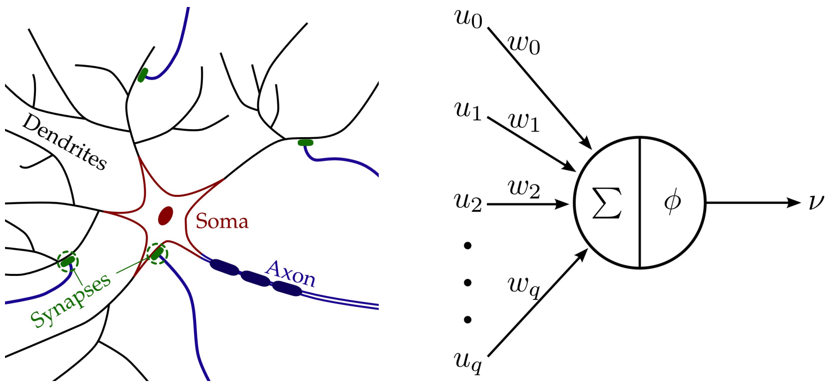
According to [Duda 01] the C4.5 algorithm [Quin 93] is the most popular decision tree approach and a successor of the ID3 algorithm. Real-valued features are treated as in CART and nominal features like in ID3. The major difference between the C4.5 algorithm and CART, besides the possibility of non-binary splits in case of nominal data, is the treatment of missing features. If a test sample has a missing value at node L used for the splitting of the sample set, all succeeding branches are followed to the leaf nodes. A final label is assigned according to the labels of these leaf nodes and the probabilities of each at branch at the node L. Furthermore, the C4.5 algorithm applies pruning methods in order to delete unnecessary splits and create leaf nodes at the corresponding node.

application in metabonomic studies

The application of decision trees in metabonomic applications is generally a promising approach, since a classification of spectra is achieved according to metabolite concentrations above or below an optimized classification threshold. This classification approach reflects the common practice in safety pharmacology by analyzing changes of specific metabolites. The non-selectivity of NMR spectroscopy allows for an automated selection of these metabolites and new biomarkers could be detected. A comparison of six different multivariate methods including CART for identification of biomarkers based on an artificial set of NMR spectra carried out by Rousseau et al. has shown good detection results of CART in the absence of noise [Rous 08]. But as soon as the noise level in the data is increased, no reliable identification of biomarkers could be achieved. These experiments demonstrate the sensitivity

sensitivity to noise

[1] ID3 is the third algorithm of a series of *active dichotomizer* algorithms.



(a) Biological neuron

(b) Artificial neuron

Figure 3.11: Structure of a (a) biological and (b) artificial neuron receiving signals from several inputs and transmitting a signal.

of decision trees to noise and as proposed by Rousseau more robust related models have to be applied.

3.4.3 *Neural Networks*

mathematical model of the human brain

An approach to model human intellectual abilities are *artifical neural networks (ANNs)* (or simply *neural networks*). These are similarly designed as the human brain by several interconnected functional units working in parallel. ANNs are regarded as a valuable classification tool for pattern recognition problems and are also often used as base classifier in ensemble approaches [Hans 90, Opit 99, Schw 00, Zhou 02]. ANNs are like decision trees unstable classifiers, leading to changes in the network parameters in case of modified training sets.

functional units of the human brain

The functional units of the human brain are the neurons and their connections, allowing for admission and transmission of electric signals. Neurons consist of different functional parts as shown in figure 3.11a. The central part of the neuron is the *soma*, containing cytoplasm and the nucleus like every other human cell. Electric signals are transmitted through the *axon* and *synapses* via *dendrites* to other neurons. The transition of an electric signal from an axon to a dendrite is realized at the synapses. These serve as connection of neurons by ion exchanges, which induce spiked electric signals in the dendrite. These signals are further transmitted to other neurons.

input and output neurons

Inspired by the biological model, artificial neurons consist of several inputs and a single output as shown in figure 3.11b. *Synaptic weights* $\mathbf{w} = [w_0, \ldots, w_q] \in \mathbb{R}^{q+1}$ are assigned to each variable of the input vector $\mathbf{u} = [u_0, \ldots, u_q] \in \mathbb{R}^{q+1}$ and the *activation function*

output $\nu \in \mathbb{R}$ is generated according to an *activation function* ϕ of the weighted summation of input features.

$$\nu = \phi \left(\sum_{i=1}^{q} w_i u_i + w_0) \right)$$

Thereby, u_0 is set equal to one and w_0 serves as bias. Different activation functions ϕ can be applied but the sigmoid function is the most common one. The sigmoid function is differentiable and incorporates characteristics from linear functions close to zero and from threshold functions for larger values.

$$\phi(\xi) = \frac{1}{1 + \exp(-\xi)}$$

Rosenblatt's perceptron

Rosenblatt applied a threshold activation function, defining a separation hyperplane in $\mathbb{R}^q$, in order to achieve a classification of samples by the so-called *perceptron* [Rose 62].

$$\phi(\xi) = \begin{cases} 1, & \text{if } \xi \geq 0, \\ -1. & \text{otherwise} \end{cases}$$

perceptron training

A modification of the hyperplane's position is achieved by the adjustment of the weights $\mathbf{w}$ aiming at an optimal classification of a labeled data set $\mathbf{X}$. The training procedure initializes the weights randomly and samples of $\mathbf{X}$ are successively classified. Weight modifications are applied in case of false predictions by

$$\mathbf{w} \longleftarrow \mathbf{w} - \nu \eta \mathbf{x_j}\,,$$

learning rate

where ν is the classification result for $\mathbf{x_i}$ and η specifies the *learning rate* of the training procedure. In case of a linearly separable data set in $\mathbb{R}^m$ this training procedure will converge to the definition of a linear separating function allowing for perfect classification of samples from $\mathbf{X}$. But if $\mathbf{X}$ is not linearly separable the training procedure will not converge and no perfect classification can be achieved.

multilayer perceptron

By the connection of multiple perceptrons to a network of neurons in accordance with the biological model more complex and even non-linear discriminant functions can be defined. The most famous type of ANNs is the *multilayer perceptron (MLP)* as shown in figure 3.12, a *feedforward* network with directed connections from one *layer* to all perceptrons of the subsequent layer. Perceptrons are not connected within a layer or to nonadjacent layers. Thus, the output of perceptrons from one layer serves as input for the subsequent layer and three different layer types can be distinguished.

input layer

Each variable x_i of an m-dimensional input-vector is the sole input of corresponding perceptrons of the *input layer* and the

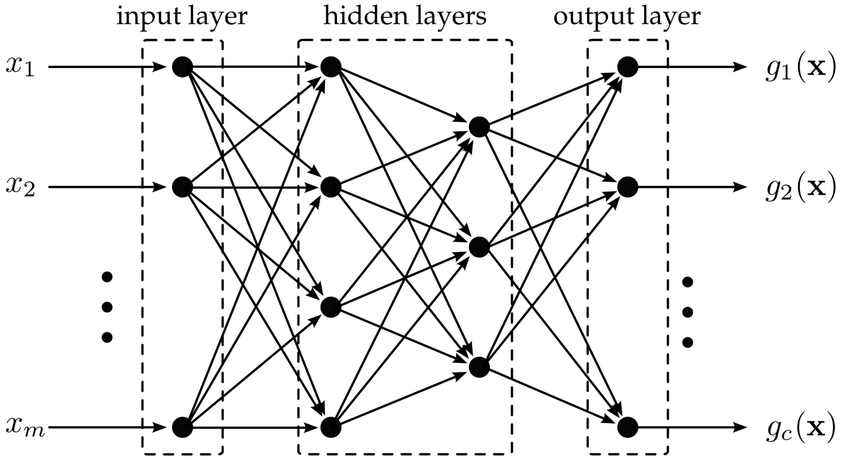


Figure 3.12: Structure of a MLP with an input layer for the *m*-dimensional sample **x**, two hidden layers and an output layer with decision functions *g* for each of the *c* possible classes (adopted from [Kunc 04]).

activation function is the identity function. The output of the input layer is connected to all perceptrons of the subsequent *hidden layer* and their output *v* is further transmitted to subsequent hidden layers. Outputs of the last hidden layer are submitted to perceptrons of the *output layer*, whereby each perceptron defines a decision function for each of the *c* classes according to the sample **x**. The index *k* of the class assigned to the sample **x** is determined by the output neuron with maximum value of the corresponding decision function.

hidden layer

output layer

$$k = \underset{j=1,\dots,c}{\operatorname{argmax}}\, g_k(\mathbf{x})$$

While the number of perceptrons in the input and output layer is defined according to the sample dimensionality *m* and number of classes *c*, respectively, the number of hidden layers and corresponding perceptrons is usually not restricted. However, it has theoretically been shown that a MLP with a single hidden layer can approximate any discrimination function with a predefined precision [Bish 95, Scar 98]. Despite these theoretical results the question arises how to train the MLP.

a single hidden layer is sufficient

Given an ANN with a defined number of hidden layers, perceptrons, and a differentiable activation function the network parameters are determined according to the *backpropagation algorithm* (cf. e.g. [Bish 95]). Thereby errors at the output layer are propagated backwards through the net to inner nodes. Thus, the gradient of the error of the network is estimated with respect to the network parameters and weights minimizing this error are determined by a gradient descent algorithm. Backpropagation training facilitates a fast determination of reasonable network parameters allowing for optimization of network designs in prac-

backpropagation training

tical applications. Several modifications of the backpropagation algorithm have been presented up to now aiming at an improved training procedure regarding stability and convergence (cf. e.g.. [Loon 97]). Details on backpropagation training is beyond the scope of this thesis and the interested reader is referred to one of the publications on ANNs (e.g. [Bish 95, Duda 01]) for further reading.

local minima

Backpropagation training is an optimization algorithm for adjustment of network weights in order to achieve the best classification performance on a cross-validation set. However, this optimization procedure could get stuck in local minima. The finally optimized network weights are dependent on factors such as the weight initialization, value of the learning rate, or the set of samples used for training. Determination of the global optimum in the presence of many local optima is the problem of most parameter optimization algorithms as discussed in section 3.2.2.

neural network ensembles

The basic principle of multiple classifier systems is to combine a set of non-optimal classifiers in order to approximate the optimal but unknown classifier. This approach has been applied for an ensemble of neural networks in several studies [Hans 90, Opit 99, Schw 00]. Hansen et al. induced differences between neural networks by variations in the selection of initial weights or training samples, and the sequence these samples are used within backpropagation training. Final prediction of a sample is achieved by assigning the class label with the maximum number of votes from the individual networks. Experiments presented by Hansen et al. demonstrate the improved classification performance of the ensemble in comparison to the best individual neural network [Hans 90].

application of ANNs to spectroscopic data

Some of the first approaches for the classification of NMR spectra by ANNs has been presented by Anthony et al. [Anth 95] and later on by Lisboa et al. [Lisb 98]. Results achieved in these experiments are based on concentration information of several molecules extracted from a spectral profile. Furthermore, only a very limited data set is used. Holmes et al. applied ANNs for the prediction of drug-induced organ toxicities on a significant larger data set and an increased spectral range is used, as it is common practice in current metabonomic studies [Holm 01]. In these experiments *PNNs* [Spec 90], as a special type of ANNs, are used for distinction between different toxicity classes. Thereby, an exponential function instead of a sigmoidal one is used as activation function at the nodes, allowing for the definition of non-linear decision boundaries. In fact, class regions are described by a sum of Gaussians with a defined variance centered at all training patterns of the classes, respectively.

probabilistic neural network

The objective in the experimental evaluation of the PNN approach is to differentiate between five different toxicity classes and two different rat strains. A PNN is trained on a single training set (583 samples) and the classification performance is determined on a test set (727 samples). Different types of organ toxicity in combination with the rat strain could be classified with an accuracy of 94 %. In comparison with alternative classifications by a MLP or SIMCA approach, PNNs achieved the best overall classification performance. The MLP has shown the worst classification performance assigning only in 52 % of the cases the correct class label. This performance could be caused by noise in the data from environmental factors as stated by Ott et al. [Ott 03]. These perturbations of the samples seem to be compensated by PNNs.

PNNs outperform MLP and SIMCA approaches

Generally, ANN approaches seem to be promising for metabonomic applications, but in order to achieve stable classification results large data sets are required. Furthermore, network optimization by backpropagation training is sensitive to noise in the data. Noisy data sets are a common problem in metabonomic studies, thus either advanced preprocessing methods have to be applied in order to reduce this noise or alternative ANNs types such as PNNs have to be used.

3.4.4 *Support Vector Machines*

The *k*NN classification rule demonstrated the idea for classification of new samples according to their similarity to known samples. Classification by *support vector machines (SVMs)* [Vapn 95] aims not at the description of class regions but at a discrimination between samples from two classes with respect to their relative position to a separating hyperplane[2]. The position of the separating hyperplane is optimized in a training procedure in order to discriminate samples of two classes and to achieve a maximum distance to a subset of training samples. These samples specifying the position of the hyperplane are denoted as *support vectors (SVs)* and maximization of the distance between the hyperplane and SVs improves generalizability of the classification. The book of Schölkopf and Smola [Scho 02] is a comprehensive work on the theory and application of SVMs and serves as basis for the following argumentation. Further details on SVM classification and training can be found in [Cort 95, Chri 00, Plat 99a].

separating hyperplane

support vector

Separation of two classes by a linear function in the original data space is commonly used as a toy example, but this is hardly ever the case in real-world problems. But the application of a suitable transformation can allow for a linear discrimination in the higher dimensional feature space, representing a non-linear separation in the data space. Thus, in the SVM approach the

[2] Generally, every n-class problem can be reformulated by several two-class problems, thus explanations regarding the classifications by SVMs are restricted to two-class problems for simplification.

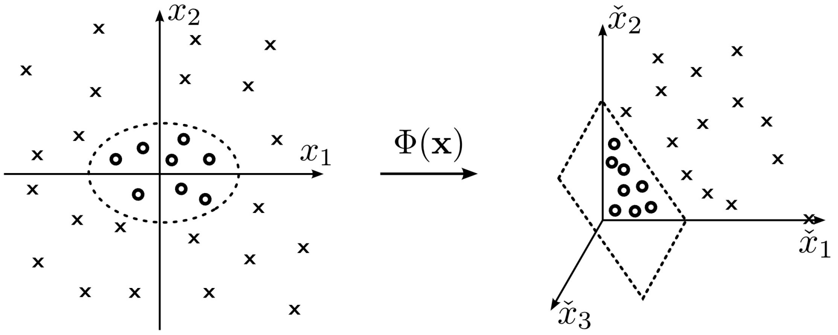


Figure 3.13: Transformation from the data space into the feature space of higher dimensionality by $\Phi(\mathbf{x}) = (\check{x}_1, \check{x}_2, \check{x}_3) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$ (adopted from [Scho 02]).

data transformation in a feature space

transformation Φ is applied for mapping of samples from the data space $\mathcal{X}$ into a feature space $\mathcal{H}$ of higher dimensionality.

$$\begin{aligned}\Phi : \mathcal{X} &\longrightarrow \mathcal{H}\\ \mathbf{x} &\longmapsto \check{\mathbf{x}} := \Phi(\mathbf{x})\end{aligned}$$

Transformation of an exemplary non-linearly separable data set into a higher-dimensional feature space as shown in figure 3.13 allows for separation by a hyperplane.

kernel function

Distance estimations in the feature space can be achieved by sample transformation and calculation of the scalar product. However, this can be computational demanding and as it is an essential step in optimization of the hyperplane position an effective distance estimation approach is required. Kernel functions k are usually applied, allowing for distance estimation between two samples $\mathbf{x}_a$ and $\mathbf{x}_b$ in feature space without prior transformation.

$$k(\mathbf{x}_a, \mathbf{x}_b) := \langle \check{\mathbf{x}}_a, \check{\mathbf{x}}_b \rangle = \langle \Phi(\mathbf{x}_a), \Phi(\mathbf{x}_b) \rangle$$

common kernel functions

Although kernel functions can specifically be defined for the current application, some kernel functions have frequently been used in practical applications showing reasonable classification results. Besides the linear kernel, representing the distance in the original space, prevalent kernel functions are the polynomial, sigmoidal and radial basis function (RBF) kernel defined as follows (cf. [Chan 01]):

$$\begin{aligned} k(\mathbf{x}_a, \mathbf{x}_b) &= (\gamma \langle \mathbf{x}_a, \mathbf{x}_b \rangle + \Theta)^d && \textbf{polynomial kernel}\\ k(\mathbf{x}_a, \mathbf{x}_b) &= \exp\left(-\gamma \|\mathbf{x}_a - \mathbf{x}_b\|^2\right) && \textbf{RBF kernel}\\ k(\mathbf{x}_a, \mathbf{x}_b) &= \tanh\left(\gamma \langle \mathbf{x}_a, \mathbf{x}_b \rangle + \Theta\right) && \textbf{sigmoidal kernel}\end{aligned}$$

Kernel-specific parameters $d \in \mathbb{N}$ and γ, $\Theta \in \mathbb{R}$ have to be adjusted for each classification problem by an optimization procedure. The RBF kernel is recommended for most applications, achieving reasonable classification results and only a single kernel parameter γ defining the width of Gaussians used for transformation has to be optimized.

In order to achieve a discrimination between two classes a separating hyperplane H and the corresponding normal vector $\mathbf{w}$ have to be defined.

$$H = \{\,\mathbf{x} \,|\, \langle \mathbf{w}, \mathbf{x} \rangle + b = 0\} \qquad \text{with} \quad \mathbf{w} \in \mathcal{H},\ b \in \mathbb{R}$$

decision function

The hyperplane H allows for the formulation of a decision function f for the classification of a sample $\mathbf{x}$ in one of the two possible classes $\omega \in \{\pm 1\}$.

$$f(\mathbf{x}) = \operatorname{sgn}\left(\langle \mathbf{w}, \mathbf{x} \rangle + b\right)$$

generalizability of the classification decision

Basically, in the linearly separable case as shown in figure 3.14a, several hyperplanes can be defined allowing for a perfect discrimination based on the current data set used for SVM training. The generalizability of a classification system is besides the classification performance the main criterion for assessment of classification quality. Thus, a hyperplane achieving good results also on unseen samples has to be determined.

margin maximization

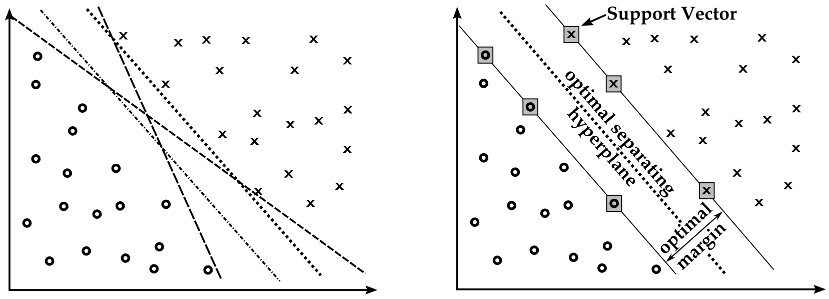
Considering a large *margin* in class separation it is quite intuitive to assume that the classification performance on unseen samples will be comparable to the performance on the training set. This intuitive understanding of generalizability can further be supported by assuming the same dependence for the generation of training and test samples, which is the basic assumption in pattern classification. Thereby, differences between training and test patterns are induced by class-specific variability of the samples. If a separation with a margin larger than the maximum variability is achieved, the separating hyperplane will generalize well on unseen samples. Thus, the optimal hyperplane should maximize the minimal distance to the SVs, thereby maximizing the distance between two class regions (cf. figure 3.14b) by:

$$\underset{\mathbf{w}\in\mathcal{H},\, b\in\mathbb{R}}{\text{maximize}} \quad \min\{\, \|\mathbf{x} - \mathbf{x}_i\| \mid \mathbf{x} \in \mathcal{H}, \langle \mathbf{w}, \mathbf{x} \rangle + b = 0,\ i = 1, \ldots, m\}\,.$$

In order to construct the optimal hyperplane with a normal vector leading to the largest margin the following optimization problem has to be solved:

$$\begin{aligned} \underset{\mathbf{w}\in\mathcal{H},\, b\in\mathbb{R}}{\text{minimize}} \quad & \tau(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \\ \text{where} \quad & \omega_i \left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\right) \geq 1 \quad \text{for } i = 1, \ldots, N\,. \end{aligned}$$

This constrained optimization problem, comprising the objective function τ and inequality constraints, can be solved by intro-



(a) Hyperplanes separating a linearly separable data set.

(b) Optimal hyperplane maximizing the margin.

Figure 3.14: Data set to be classified by a SVM. Several hyperplanes can be used for class discrimination but only a single one maximizes the margin, thus being optimal for classification.

dual optimization problem

ducing *Lagrange multipliers* α_i and a *Lagrangian* to form the *dual optimization problem*

$$\begin{aligned} \underset{\boldsymbol{\alpha}\in\mathbb{R}^N}{\text{maximize}} \quad & W(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j \omega_i \omega_j \, k\left(\mathbf{x}_i, \mathbf{x}_j\right) \\ \text{where} \quad & \alpha_i \geq 0, \quad i = 1, \ldots, N \\ \text{and} \quad & \sum_{i=1}^{N} \alpha_i \omega_i = 0. \end{aligned}$$

The set of Lagrangian multipliers $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$ characterizes the influence of each training sample for construction of the hyperplane. Samples with a Lagrangian multiplier α_i equal to zero are irrelevant for classification of new samples and the subset of samples with $\alpha_i > 0$ are the SVs, influencing the classification decision. This can be visualized by a sheet lying along the hyperplane. Each SV exerts a force relative to α_i in direction of the normal vector on the sheet. Thereby, forces of SVs from each side of the sheet sum up to zero and stabilize the position of the sheet, thus *supporting* the plane sheet. Classification of an unknown sample $\mathbf{x}$ is realized by the decision function $f(\mathbf{x})$ as follows:

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{N} \omega_i \, \alpha_i \, k\left(\mathbf{x}, \mathbf{x}_i\right) + b\right) .$$

Although transformation of a data set in a space of higher dimensionality is expected to increase separability of two data sets, a linear separation in the kernel space cannot be guaranteed in case of still overlapping class regions. In order to account for a non-linearly separable data set *slack variables* are introduced allowing for SVs on the wrong side of the hyperplane or a distance

slack variable

smaller than one. Thus, the condition of the samples' distance to the hyperplane is alleviated as follows:

$$\begin{aligned} &\omega_i \left(\langle \vec{\mathbf{w}}, \mathbf{x}_i \rangle + b \right) \geq 1 - \xi_i \qquad \text{for } i = 1, \ldots, N \\ &\text{where} \quad \xi_i \geq 0\,. \end{aligned}$$

A well-performing classifier has to determine an optimal combination of a normal vector **w** and the amount of slack variables used for determination of the separating hyperplane. Thereby, the trade-off between maximization of the margin and classification performance is controlled. The number of slack variables can be controlled by a slack parameter C, which is included in the minimization problem as follows:

slack parameter C

$$\underset{\mathbf{w}\in\mathcal{H},\,\boldsymbol{\xi}\in\mathbb{R}^N}{\text{minimize}} \quad \tau\left(\mathbf{w}, \boldsymbol{\xi}\right) = \frac{1}{2}\left\|\mathbf{w}\right\|^2 + \frac{C}{N}\sum_{i=1}^{N}\xi_i\,.$$

This minimization problem can be reformulated as dual problem by a Lagrangian as follows

$$\begin{aligned} \underset{\boldsymbol{\alpha}\in\mathbb{R}^N}{\text{maximize}} \quad & W\left(\boldsymbol{\alpha}\right) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j\omega_i\omega_j\, k\left(\mathbf{x}_i, \mathbf{x}_j\right) \\ \text{with} \quad & 0 \leq \alpha_i \leq \frac{C}{N}, \quad i = 1, \ldots, N \\ \text{and} \quad & \sum_{i=1}^{N}\alpha_i\omega_i = 0\,. \end{aligned}$$

Thereby, a minimization of the training error and a maximization of the margin is achieved. However, the absolute term C is dependent on the range of data values. The incorporation of the alternative parameter ν as a data-independent optimization parameter is presented by Schölkopf et al. in [Scho 00]. Besides the regulation of slack variables, the amount of SVs used for the definition of the separating hyperplane, representing the complexity of the data separation, is controlled via the value of ν. The optimization problem can be redefined incorporating this parameter by

slack parameter ν

$$\begin{aligned} \underset{\mathbf{w}\in\mathcal{H},\,\boldsymbol{\xi}\in\mathbb{R}^N,\rho,b\in\mathbb{R}}{\text{minimize}} \quad & \tau\left(\mathbf{w}, \boldsymbol{\xi}, \rho\right) = \frac{1}{2}\left\|\mathbf{w}\right\|^2 - \nu\rho + \frac{1}{N}\sum_{i=1}^{N}\xi_i \\ \text{where} \quad & \omega_i\left(\langle \mathbf{x}_i, \mathbf{w} \rangle + b\right) \geq \rho - \xi_i \\ \text{and} \quad & \xi_i \geq 0, \quad \rho \geq 0\,. \end{aligned}$$

The dual problem is defined based on this optimization problem using a Lagrangian as follows

$$\begin{aligned} \underset{\boldsymbol{\alpha}\in\mathbb{R}^N}{\text{maximize}} \quad & W(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i,j=1}^{N} \alpha_i\alpha_j\omega_i\omega_j k\left(\mathbf{x}_i, \mathbf{x}_j\right) \\ \text{with} \quad & 0 \le \alpha_i \le \frac{1}{N}, \quad i = 1, \ldots, N \\ \text{and} \quad & \sum_{i=1}^{N} \alpha_i\omega_i = 0 \\ \text{and} \quad & \sum_{i=1}^{N} \alpha_i \ge \nu\,. \end{aligned}$$

In comparison with the regulation by the slack parameter C the influence of support vectors is controlled by an additional condition and is not included in the primary optimization problem. Further remarks on the relation of C and ν can be found in [Chan 01].

SVMs are regarded as a powerful classification approach, which has achieved reasonable classification results in numerous applications (cf. e.g. [Hsu 02]). ANNs have also achieved reasonable classification results in various fields, but certain aspects support the application of rather SVMs than ANNs in metabonomic applications. As stated by Burbidge et al. [Burb 01], and later supported by Byvatov et al. [Byva 03], SVMs also achieve improved classification results when only few but high-dimensional samples are available. The definition of a separating hyperplane can already be achieved by a subset of the training samples and no statistical models have to be estimated. Discrepancy between high data dimensionality and low number of samples is a common problem in metabonomic studies, thus SVMs are expected to achieve good classification results even on this complex data sets.

reasonable results on sparse and high-dimensional data sets

strong classifier

Similar to NNs, SVMs are regarded as strong learning algorithms and even a single SVM can achieve a reasonable classification performance. The combination of multiple SVMs generated by bagging or AdaBoost methods [Kim 03, Li 05] is expected to slightly increase the ensemble performance in comparison with single SVMs and also improve generalizability. This suggestion is supported by an empirical study of Wang et al. [Wang 09] using different ensemble construction techniques for SVM ensembles. These ensembles are evaluated on 20 sets from the UCI machine learning repository [Asun 07] and an industrial study of gear defect detection. SVM ensembles could not outperform single SVM classification on every data set, but on an average an improved classification performance could be achieved.

improved performance and generalizability of SVM ensemble approach

The application of SVMs for the classification of NMR spectra for metabonomic applications has not been extensively investigated up to now. While SVMs are already known in the related field of chemometrics [Belo 02b, Belo 02a], one of the first classification approaches based on a data set of NMR spectra using SVMs has been presented by Masoum et al. [Maso 07]. The goal in this study was to discriminate farmed and wild salmon, and different countries of origin. The whole data set was divided for an experimental evaluation into a training (74 samples), validation (45 samples) and test set (22 samples). Data reduction has been applied by averaging every two spectral values, leading to a data set of only 141 samples but 11 501 dimensions. Spectral preprocessing comprises COW alignment for peak shift reduction and standard normal variate scaling [Barn 89] for compensation of dilution effects. An evaluation of different kernel functions has proven the RBF kernel to be the optimal choice due to the low number of kernel parameters and the ability to model a reasonable class separation even in case of this complex data set. A perfect discrimination between wild and farmed salmon could be achieved on the test set, while the origin was falsely predicted for a single sample. These classification results are especially impressive under the consideration of the data dimensionality.

classification of NMR spectra by SVMs

nearly perfect classification of all samples

To sum up, SVMs allow for a reasonable classification even in case of high-dimensional and sparse data sets. Competitive classification results could be achieved in several applications. In contrast to ANNs, SVMs can compensate outlier samples by slack variables and show a decreased sensitivity to noisy samples.

3.5 ensemble aggregation techniques

Ensemble creation methods form the foundation to achieve a set of L diverse base classifiers D_i, $i = 1, \dots, L$ and to predict the label of a test sample $\mathbf{x}$ by each of them. By means of base classifier outputs a final ensemble classification $\hat{l}(\mathbf{x})$ is realized using *ensemble aggregation techniques*. This combination of base classifier outputs can basically be achieved by two different types of methods, *fusion* and *selection*. In fusion each classifier has an influence on the final decision, while in selection the prediction of the classifier regarded as most suitable for the current sample is selected.

classifier fusion and selection

Different aggregation strategies can be applied according to [Xu 92] dependent on the type of classifier outputs. The most common and universal output type is the assignment of a class label $s_i \in \Omega$, $i = 1, \dots, c$ to each of the test samples by the base classifiers, respectively, denoted as the *abstract level*. Thus, the output of the ensemble is a vector of predictions $\mathbf{s} = [s_1, \dots, s_L] \in \Omega^L$. By this output a classification into a single class without specification of alternative but less plausible predictions or definition of classification certainty is achieved. An alternative output type is

abstract level

rank level

measurement level

oracle level

the *rank level* by assigning not a single class label to a test sample but a subset of Ω ranked according to their grade of confidence to be the correct label. A further possible output of a base classifier is the definition of a vector containing values reflecting the support for a sample to be classified to each of the individual classes, denoted as the *measurement level*. Finally, an *oracle level* output can be used in case of a labeled data set as shown in section 3.3.3, assigning a 1 if prediction of the base classifier is correct and 0 otherwise.

Several aggregation approaches have been proposed up to now. The approaches presented in the following will not be comprehensive but restricted to the relevant ones for this thesis. Classification problems from the field of Metabonomics usually aim at discrimination between very few classes. Thus outputs from the rank level will not be discussed. Since the correct label of new samples is unknown, outputs from the oracle level will not be useful for the considered problem. A discussion of methods presented in this section and further ensemble aggregation approaches can be found in [Kunc 04, Xu 92].

3.5.1 *Majority Voting*

Determination of a consensus decision by *majority voting* based on outputs from the abstract level has a long history as for example in democratic votings in politics. Majority voting is defined by assigning the class with the maximum number of votes among a set of classifiers' predictions to a test sample.

This understanding of majority voting can be mathematically formulated by the definition of a c-dimensional vector $[d_{i,1}, \ldots, d_{i,c}] \in [0,1]^c$ containing a value of one at the position of the class predicted by the classifier D_i and zero otherwise. The index k of the class assigned to a sample $\mathbf{x}$ is defined by

$$k = \underset{j=1,\ldots,c}{\operatorname{argmax}} \sum_{i=1}^{L} d_{i,j}\,.$$

An alternative formulation is to assign a label only if the percentage of votes for the majority class is above a certain threshold and reject it otherwise in order to guarantee a certain confidence in the classification [Xu 92].

classification performance of majority voting

The majority vote is one of the most common fusion schemes due to its simplicity and effectiveness, and its theoretical properties have been investigated in several studies (cf. e.g. [Lam 97, Lin 03, Kunc 04]). In order to estimate the classification performance achieved by majority voting, assume a set of L independent classifiers, whereby L should be odd. Each classifier assigns the correct label to a sample $\mathbf{x} \in \mathbb{R}^m$ with known probability p. Thus, the classification accuracy of the ensemble using the

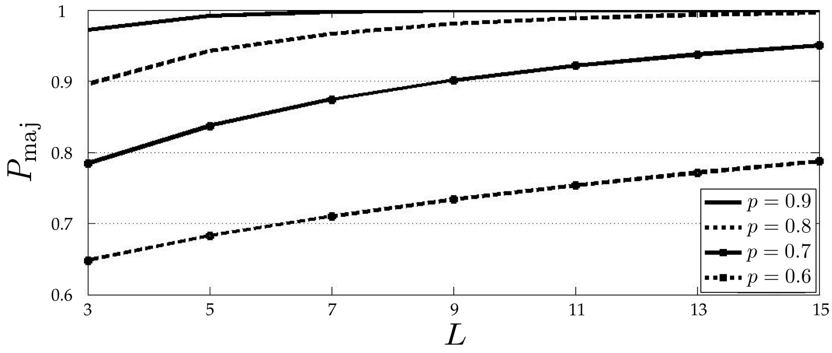


Figure 3.15: Majority vote accuracy of independent classifiers with specific probability p to assign the correct label under variation of the ensemble size L.

majority vote for output combination can be estimated according to [Nitz 82] as

$$P_{\text{maj}} = \sum_{m=\lfloor L/2 \rfloor + 1}^{L} \binom{L}{m} p^m (1-p)^{L-m} \,.$$

Some values of P_{maj} are shown in figure 3.15 for demonstration of the effect of p and L on the expected ensemble classification performance. According to the calculation of the expected classification performance, different results of the ensemble accuracy under variation of L can be achieved, known as the Condorcet Jury Theorem [Nitz 82]:

condorcet jury theorem

1. If $p > 0.5$, then P_{maj} increases up to 1 if $L \longrightarrow \infty$
2. If $p < 0.5$, then P_{maj} decreases down to 0 if $L \longrightarrow \infty$
3. If $p = 0.5$, then P_{maj} is equal to 0.5 for any L.

These results further support the general intuition to achieve at least a classification performance of 0.5 by the base classifiers in order to increase classification performance by ensemble combination. However, identical performance of each base classifier as assumed above is highly unrealistic. Shapley and Grofman have proven the validity of these results also in case of unequal probabilities symmetrically distributed around the mean p with a value greater than 0.5 [Shap 84].

An example of three classifiers having each a probability of 0.6 to assign the correct label to one of 10 different examples illustrates the influence of dependence between classifiers' predictions on the ensemble performance. In the optimal case, when maximal two of the three classifiers agree on the correct label, a classification performance of 0.9 can be achieved. But in the worst case the classification performance can decrease down to 0.4 (cf. [Kunc 04], p. 117). Thus, an improved classification performance cannot be guaranteed by using majority voting, but the probability to achieve good classification results increases with the number of base classifiers and their mean performance.

3.5.2 *Accuracy-Based Weighting*

improved influence of the best experts' predictions

The influence of each expert on the final ensemble classification is equally distributed among all L classifiers in majority voting. But it is more intuitive to control the influence of each classifier with respect to its individual competence to classify a given sample. Thus, by assigning a weight w_i to each classifier the *weighted majority voting* can be formulated by

$$k = \operatorname*{argmax}_{j=1,\dots,c} \sum_{i=1}^{L} w_i d_{i,j} \,,$$

where k reflects the index of the class assigned to the sample $\mathbf{x}$. Thereby, not only the amount of classifiers but also their joined influence on the classification decision is decisive for the final prediction.

An example originally shown by Shapley and Grofman [Shap 84] is given in the following in order to demonstrate the effectiveness of weighted majority voting. Given a set of five independent classifiers $D_1, \dots, D_5$ with individual classification performance of 0.9, 0.9, 0.6, 0.6 and 0.6. Majority voting would lead in this example to an expected classification accuracy of $P_{\mathrm{maj}} = 0.877$ by

$$P_{\mathrm{maj}} = 3 \times 0.9^2 \times 0.4 \times 0.6 + 0.6^3 + 6 \times 0.9 \times 0.1 \times 0.6^2 \times 0.4$$

as the probability that at least three classifiers assign the correct label. Thus, majority voting would lead to a worse classification accuracy compared to the single best classifier. But if the power of the most competent classifiers is increased by assigning weights $\frac{1}{3}$, $\frac{1}{3}$, $\frac{1}{9}$, $\frac{1}{9}$ and $\frac{1}{9}$, respectively, the expected classification performance is increased to $P^{w}_{\mathrm{maj}} = 0.927$ by

$$P^{w}_{\mathrm{maj}} = 0.9^2 + 2 \times 3 \times 0.9 \times 0.1 \times 0.6^2 + 2 \times 0.9 \times 0.1 \times 0.6^3 \,.$$

Due to the increased influence of the two best performing classifiers a correct classification can be achieved if these two assign the correct class label, or at least one and the majority of the three remaining classifiers.

Actually, several other weightings would achieve an improved classification performance. But in case of L conditionally independent classifiers, such that

$$P(\mathbf{s}|\omega_j) = \prod_{i=1}^{L} P(s_i|\omega_j)$$

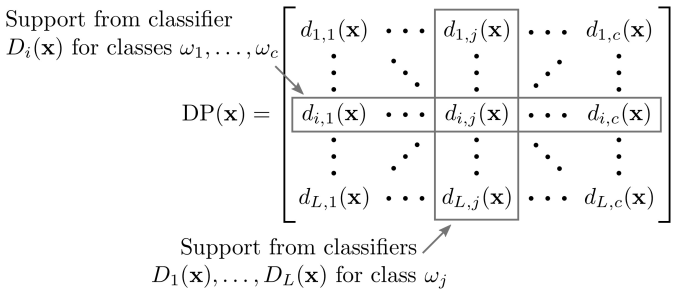

Support from classifier
$D_i(\mathbf{x})$ for classes $\omega_1, \dots, \omega_c$
$\mathrm{DP}(\mathbf{x}) =$
$d_{1,1}(\mathbf{x})$ $\cdots$ $d_{1,j}(\mathbf{x})$ $\cdots$ $d_{1,c}(\mathbf{x})$
$d_{i,1}(\mathbf{x})$ $\cdots$ $d_{i,j}(\mathbf{x})$ $\cdots$ $d_{i,c}(\mathbf{x})$
$d_{L,1}(\mathbf{x})$ $\cdots$ $d_{L,j}(\mathbf{x})$ $\cdots$ $d_{L,c}(\mathbf{x})$
Support from classifiers
$D_1(\mathbf{x}), \dots, D_L(\mathbf{x})$ for class ω_j

Figure 3.16: Decision profile for a sample $\mathbf{x}$; $d_{i,j}$ corresponds to the support of D_i for class ω_j (adopted from [Step 06]).

is fulfilled, with accuracies $p_1, \dots, p_L$ the maximum ensemble classification performance is achieved according to [Shap 84] if weights are assigned by

$$w_i \propto \log \frac{p_i}{1 - p_i} .$$

Taking the prior probabilities of the respective classes into account, the index of the class assigned to a sample $\mathbf{x}$ is determined by

$$k = \operatorname*{argmax}_{j=1,\dots,c} \left[\log P(\omega_j) + \sum_{i=1}^{L} d_{i,j} \log \frac{p_i}{1 - p_i} \right] .$$

Majority voting and accuracy-based weighting are combination methods for output types of the abstract level, thereby every classifier assigns a unique class label to a test sample. Several other straightforward combination schemes even for outputs from measurement level have been proposed up to now. An overview and an experimental comparison of a selection of these can be found in [Kitt 98, Kunc 04].

3.5.3 *Class-Conscious Combiners*

While outputs from the abstract level assign only a single class to a sample, outputs from the measurement level indicate their support for classification in each of the c different classes, respectively. Thus, each expert defines a c-dimensional real-valued vector and the outputs of L classifiers can be organized in a matrix as shown in figure 3.16, denoted as *decision profile* (DP).

decision profile

The fusion approach realized by majority voting can also be applied to real-valued outputs and not only for label outputs. Thereby, the support μ_j for the class ω_j is determined according to the support given to the class ω_j by the L classifiers as denoted in column j of the DP. The index k of the finally predicted class is determined according to the maximum support achieved.

$$k = \operatorname*{argmax}_{j=1,\dots,c} \mu_j = \operatorname*{argmax}_{j=1,\dots,c} S[d_{1,j}(\mathbf{x}), \dots, d_{L,j}(\mathbf{x})]$$

combination function

Different combination functions S can be applied and common rules are the *average*

$$\mu_j = \frac{1}{L} \sum_{i=1}^{L} d_{i,j}(\mathbf{x})$$

and *product* combination

$$\mu_j = \prod_{i=1}^{L} d_{i,j}(\mathbf{x}) \,.$$

Apart from these, alternative combination rules have been proposed, but the suitability of each method is usually dependent on the chosen base classifier algorithm and the problem at hand [Kitt 02].

3.5.4 *Class-Indifferent Combiners*

Decision profiles as described in the previous section can also be used for the classification of new samples without restrictions on the column of the DP to be used. Thereby, DPs of samples from each class are transferred into an *intermediate feature space* and dependent on this representation the final class for a given test sample is determined.

Decision Templates

class specific decision templates

A classifier fusion approach proposed by Kuncheva et al. aims at the definition of a typical DP_j for each class ω_j, called the *decision template (DT)* [Kunc 01]. The DT_j for class ω_j is calculated according to the DPs of samples from the data subset $\mathbf{X}_j$ containing N_j samples from the class ω_j by

$$\text{DT}^j = \frac{1}{N_j} \sum_{x_k \in \mathbf{X}_j} \text{DP}(\mathbf{x}_k) \,.$$

similarity of a sample to all DTs

The classification of a test sample $\mathbf{x}$ is achieved by the calculation of the similarity between the DP for the sample and the DTs of each of the c classes (cf. figure 3.17). The index k of the class with the maximum similarity determines the final classification result and is calculated according to

$$k = \underset{j=1,\dots,c}{\text{argmax}} \left[1 - \frac{1}{L \times c} \sum_{m=1}^{L} \sum_{n=1}^{c} (\text{DT}^j_{m,n} - d_{m,n}(\mathbf{x}))^2 \right] ,$$

whereby $\text{DT}^j_{m,n}$ is the entry of the DT for class ω_j in the row m and column n. The presented similarity criterion is the euclidean distance between the two vectors in the $L \times c$-dimensional space. But also alternative similarity measures can be applied as presented in [Kunc 01].

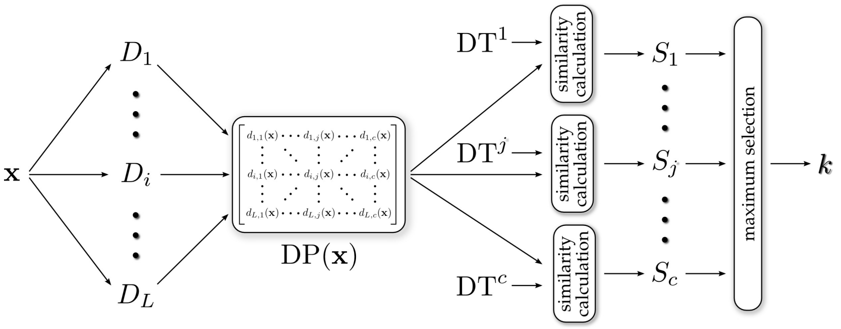


Figure 3.17: Decision template classification procedure.

The advantage of the decision template approach is its simplicity and its application has been presented in several applications (cf. e.g. [Diet 03, Kitt 02, Step 06]). Like most classifier fusion methods the decision templates approach is not superior to all alternative fusion schemes, but shows comparable classification performance. Additionally, it is more elegant due to the incorporation of support for each of the c classes from all L classifiers in the final classification.

Behavior Knowledge Space

Huang et al. proposed a fusion scheme named *behavior knowledge space* jointly considering the support of all classifiers to all classes for final classification [Huan 95]. Thereby, each DP is transformed into an L-dimensional feature vector $\boldsymbol{\delta}(\mathbf{x}) = [\delta_1(\mathbf{x}), \ldots, \delta_L(\mathbf{x})]$ by assigning the index of the class with the maximum support from classifier D_j for the sample $\mathbf{x}$ to feature $\delta_j(\mathbf{x})$. The resulting space of transformed DPs is denoted as the behavior knowledge space. Since only values from the interval $[1, \ldots, c]$ can be assigned to each feature, each point in this space corresponds to a certain bin.

conversion of DPs in feature vectors

The classification of a new sample $\mathbf{x}$ is achieved by the transformation of DP$(\mathbf{x})$ in the feature space. The amount of samples from class ω_j in this bin is determined according to the histogram function $h_j(\boldsymbol{\delta}(\mathbf{x}))$ and sample $\mathbf{x}$ is classified into the most representative class as follows

amount of samples in bins in feature space

$$k = \begin{cases} \underset{j=1,\ldots,c}{\operatorname{argmax}} h_j(\boldsymbol{\delta}(\mathbf{x})) & \text{if } \sum\limits_{i=1}^{c} h_i(\boldsymbol{\delta}(\mathbf{x})) > 0 \text{ and } \dfrac{h_k(\boldsymbol{\delta}(\mathbf{x}))}{\sum\limits_{i=1}^{c} h_i(\boldsymbol{\delta}(\mathbf{x}))} \geq \lambda \\ 0 & \text{otherwise.} \end{cases}$$

The label 0 corresponds to a rejection of the sample if either no sample from the training set is at the specific position, or the proportion of patterns from class ω_k is below a threshold λ.

In summary, new samples are assigned to the class with the majority of samples at the corresponding position of the feature

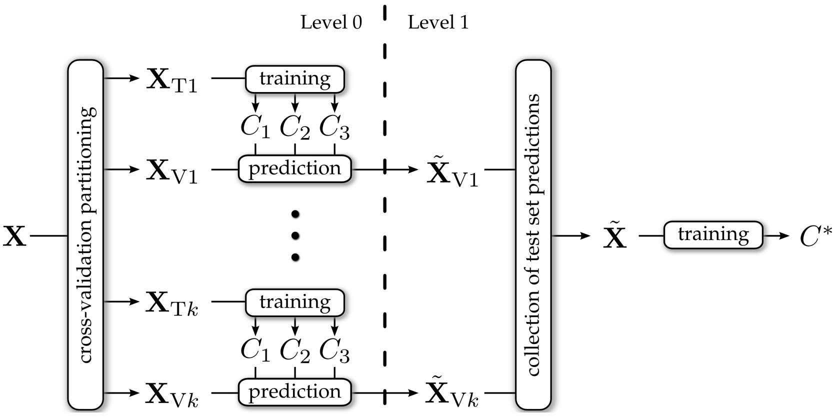


Figure 3.18: Generalized stacking procedure.

compensation of different class sizes

space. Kittler et al. propose to introduce class-specific weights with respect to the class-specific prior probabilities in the determination of the maximum proportion of samples from each class [Kitt 02]. Thereby, classes of different size in the training set can be taken into account. An experimental comparison demonstrates consistently better results achieved by the behavior knowledge space approach in comparison with the decision templates approach [Kitt 02]. Even increasing the number of experts has lead to improved classification results. Thus, no expert selection has to be applied in order to increase classification performance.

3.5.5 *Generalized Stacking*

stacked generalization

Wolpert presented *stacked generalization* as an ensemble aggregation strategy focusing on a high generalizability in ensemble aggregation [Wolp 92]. Thereby, ensemble aggregation is achieved by the training of an additional classifier on the predictions of multiple classifiers for a given data set by means of a cross-validation approach. Thus, an additional layer of abstraction is introduced by interpretation of predictions from base classifiers as features for a given sample.

As a first step in the stacked generalization approach (cf. figure 3.18), a given data set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is split into k parts. With respect to these parts, k different training $\mathbf{X}_{T1}, \ldots, \mathbf{X}_{Tk}$ and corresponding test sets $\mathbf{X}_{V1}, \ldots, \mathbf{X}_{Vk}$ are created according to the cross-validation principle (cf. section 3.1.2). L different classifiers $C_1, \ldots, C_L$ are trained on each of the k training sets. These *level-0 models* are applied for prediction of class labels from samples in the respective test sets.

level-0 models

Predictions of level-0 models serve as features of the new data set representation $\tilde{\mathbf{X}} = \{\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_N\}$, by

$$\tilde{\mathbf{x}}_n = [\hat{l}_1(\mathbf{x}_n), \ldots, \hat{l}_L(\mathbf{x}_n)] \, , n = 1, \ldots, N \, ,$$

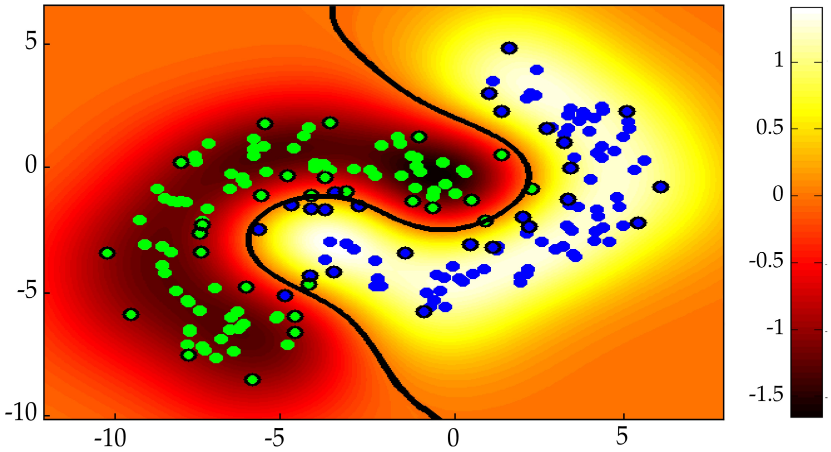


Figure 3.19: Two-class problem and the separating hyperplane of an RBF SVM. Distances in the RBF space are coded by color and the SVs are marked by black circles.

where $\hat{l}_i(\mathbf{x}_n)$, $i = 1, \ldots, L$ is the prediction of the class label from classifier C_i for sample $\mathbf{x}_n$. A *level-*1 *generalizer* C^* is trained on this newly created *level-*1 *data* $\tilde{\mathbf{X}}$ for classification of unknown samples.

*level-*1 *generalizer*

Investigations by Ting et al. have shown that class probabilities rather than class predictions from level-0 models should be used to form the level-1 data [Ting 99]. These cause an improved classification by the stacked classifier as shown by an experimental evaluation. However, class probabilities are not estimated by each classification approach as exemplary presented for SVMs. Classification is achieved by SVMs with respect to the relative position of samples to a separating hyperplane as outlined in section 3.4.4 . Thus, SVMs produce generally no (pseudo-)probabilities that could be used for a level-1 data representation.

*class probabilities from level-*0 *models*

SVMs do not estimate class probabilities

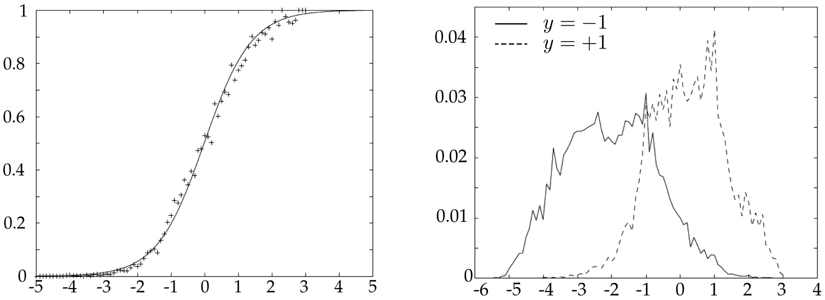
An exemplary classification of two banana shaped classes[3] by a RBF SVM is shown in figure 3.19 . Differences can be noticed in the confidence for the classification of samples with respect to their relative position to the separating hyperplane. Samples close to the class boundaries could be moved to the other side of the hyperplane by added noise of low intensity. Thus, the larger the distance to the separating hyperplane is, the higher is the confidence in the correct classification.

confidence in classification w.r.t. to the distance to the separating hyperplane

These observations have been used by Platt for the definition of a transformation of SVM outputs to class pseudo-probabilities [Plat 99a]. Evaluation of different transformation functions by Platt on three sets from the UCI data set [Asun 07] have finally led to the choice of a parametrized sigmoid function. Thereby,

SVM class pseudo-probabilities

[3] The data set was generated using the `gendatb` command from the PRTOOLS toolbox (Version 4.1.4) in Matlab. Software available at `http://www.prtools.org`.



(a) Posterior probabilities from Bayes' rule (+) and a sigmoidal fitted to SVM outputs (-)

(b) Class-conditional histogram of SVM outputs.

Figure 3.20: (a) Comparison of posterior probabilities generated for the UCI Adult data set by Bayes' rule (plus marks) and the sigmoidal fitted by the algorithm of Platt on SVM outputs. The class-conditional histograms of SVM outputs are shown in (b) (adopted from [Plat 99b]).

the probability for a sample $\mathbf{x}$ to be labeled as 1 (and not as -1) with respect to its distance $f(\mathbf{x})$ to the separating hyperplane is defined by

$$P(l(\mathbf{x})=1|\mathbf{x}) \approx P_{a,b}(f(\mathbf{x})) = \frac{1}{1+\exp(af(\mathbf{x})+b)}\,.$$

optimization of transformation parameters

For a given data set $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$ the optimal parameters a and b are determined by the solution of the following regularized maximum likelihood problem, which is a cross-entropy error function:

$$\underset{a,b\in\mathcal{R}}{\text{minimize}} \left[-\sum_{n=1}^{N}\left(t_n \log\left(P_{a,b}(\mathbf{x}_n)\right) + (1-t_n)\log\left(1-P_{a,b}(\mathbf{x}_n)\right)\right)\right],$$

$$\text{with } t_n = \begin{cases} \frac{N_+ +1}{N_+ +2} & \text{if } l(\mathbf{x}_n) = +1 \\ \frac{1}{N_- +2} & \text{if } l(\mathbf{x}_n) = -1 \end{cases},$$

where N_+ and N_- corresponds to the amount of samples with $l(\mathbf{x})=+1$ and $l(\mathbf{x})=-1$, respectively. Lin et al. presented in [Lin 07] an optimization approach for the determination of a and b. Thereby, improvements with respect to the algorithm proposed by Platt could be achieved.

This transformation is exemplary shown in figure 3.20a for the outputs of a linear SVM trained by a 3-fold cross-validation procedure on the Adult data set as performed by Platt in [Plat 99a]. The sigmoid function is fitted to the SVM outputs, and posterior probabilities for all samples falling into the same bin of size 0.1 of a histogram estimated on the SVM outputs are shown (cf. figure 3.20b). These posterior probabilities are determined according to

Bayes′ rule on the histogram estimates of class-conditional densities from the SVMs trained via cross-validation (cf. [Plat 99a] for further details). As can be seen by this illustration, the determined sigmoid function achieves a good estimate of the posterior probabilities and is used as an alternative level-0 data representation besides binary predictions achieved by SVMs as base classifiers.

3.6 boosting methods

An ensemble system incorporating methods for ensemble creation, base classifier prediction, and ensemble aggregation is the *boosting* approach. The boosting approach is based on a weighted combination of multiple classifiers trained in a cascade. Each classifier in the cascade is weighted and trained in order to compensate errors made by previously trained classifiers. Thus, the base classifiers are boosted by their weighted combination for final classification.

AdaBoost

weighting of samples

In the *AdaBoost* approach, which is derived from *ada*ptive *boost*ing [Freu 97], the training of classifiers in the cascade is controlled by weights assigned to each sample. These weights reflect the influence of each sample on the training and choice of the base classifier. These weights are modified within an iterative procedure as shown in algorithm 1, increasing the weights of falsely classified samples and vice versa. Weights of samples can be used in two different ways within classifier training. In case of the resampling procedure weights regulate the probability of each sample to be selected for the training set. Alternatively, weights can be incorporated in the determination of the training error. In the following the resampling procedure will be presented as shown in algorithm 1 for two-class problems.

equal relevance of all samples

Given a data set $\mathbf{X}$ of size N and the corresponding class labels $l(\mathbf{x}_i)$, $i = 1, \ldots, N$, sample weights w_i are initially selected to be uniform. In each of the $k_{\max}$ iterations a classifier D_k, $k = 1, \ldots, k_{\max}$ is trained using a data set sampled according to $\mathbf{w}^k$. Thus, the influence of each sample on the classifier is controlled by the samples′ weights $\mathbf{w}^k$. The classification error of D_k for $\mathbf{X}$ is calculated as weighted sum of misclassifications, whereby $\mathbf{w}^k$ determines the influence of each sample on the error. Each classifier D_k has to achieve at least a classification error smaller than 0.5 and greater than zero in order to achieve an exponentially decreasing classification error [Freu 01]. Otherwise the weights are set to be uniform and the next best classifier is trained. If $\epsilon_k > 0$ and $\epsilon_k < 0.5$ the weight α_k for the classifier D_k is calculated. Finally, the sample weights are updated and normalized to sum up to one, thereby increasing the weight of wrongly classified samples leading to a stronger influence on the classification decision in the following iteration.

focus on falsely classified samples

Algorithm 1 AdaBoost algorithm (adopted from [Freu 97])

Input: data matrix $\mathbf{X}$ and corresponding labels $l(\mathbf{x}_i)$, $i = 1, \ldots, N$, maximum number of iterations $k_{\max}$

Output: set of classifiers D_k and corresponding weights α_k

```
$\forall i : w_i^1 \leftarrow \frac{1}{N}$
for $k \leftarrow 1, \ldots, k_{\max}$ do
    train $D_k$ using $\mathbf{w}^k$; get a hypothesis $h_k = X \rightarrow [0, 1]$
    Calculate the error $\epsilon^k$ of $D_k$
        $\epsilon^k \leftarrow \sum_{j=1}^{N} w_j^k |h_t(\mathbf{x}_i) - l(\mathbf{x}_i)|$
    calculate the classifier weight $\alpha_k$
        $\alpha_k \leftarrow \frac{\epsilon}{1-\epsilon_k}$
    update the sample weights
        $w_j^{k+1} \leftarrow \frac{w_j^k \alpha_k^{(1-|h_t(\mathbf{x}_i)-l(\mathbf{x}_i)|)}}{\sum_{i=1}^{N} w_i^k \alpha_k^{(1-|h_t(\mathbf{x}_i)-l(\mathbf{x}_i)|)}}$
end for
return $D_1, \ldots, D_k$ and $\alpha_1, \ldots, \alpha_k$
```

The classification of a new sample is achieved by a weighted summation of the hypotheses h_k produced by each of the $k_{\max}$ classifiers.

$$h(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum\limits_{k=1}^{k_{\max}} (\log \frac{1}{\alpha_k})\, h_k(\mathbf{x}) \geq \frac{1}{2} \sum\limits_{k=1}^{k_{\max}} \log \frac{1}{\alpha_t} \\ 0 & \text{otherwise.} \end{cases}$$

The presented algorithm achieves a binary classification but also multi-class classifications can be realized by modifications of the original algorithm denoted as AdaBoost.M1 and AdaBoost.M2 (cf. e.g. [Freu 97]).

An exemplary two-class data set to be classified by the AdaBoost algorithm using threshold classifiers as base classifiers is shown in figure 3.21. In the first round each sample has the same weight (visualized by the size of the sample) and a hypothesis h_1 is determined achieving the best classification accuracy. Weights of samples misclassified by h_1 are increased while the remaining ones are decreased and h_2 is determined in the second boosting round. Finally, h_3 is determined and the overall classification decision h can be formulated according to α_1, α_2, α_3 achieving a perfect discrimination between the two classes on the training set.

sensitivity to outliers and noise

Although the focusing on samples hard to predict is the basic principle of AdaBoost, this approach can have severe drawbacks in case of noisy samples and outliers present in the data. In

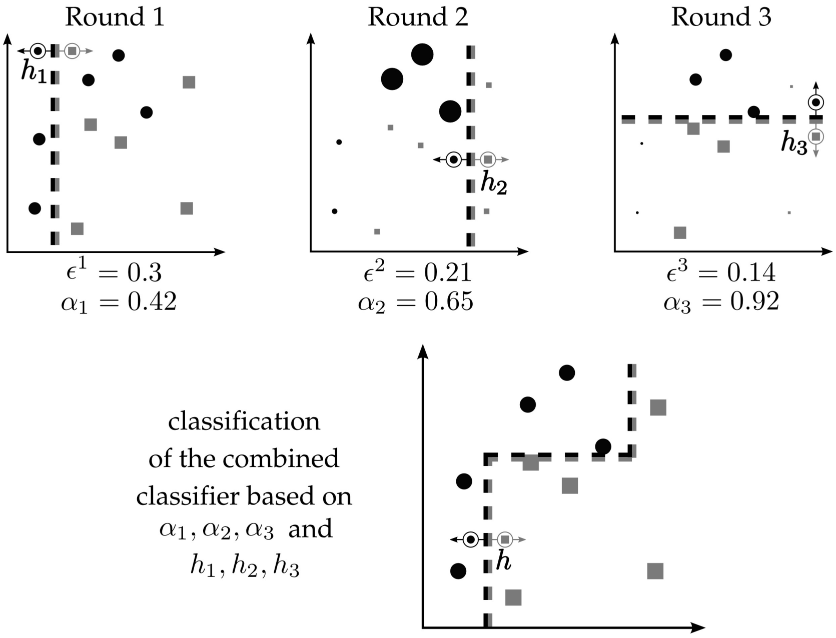


Figure 3.21: Toy example of the discrimination of a two-class data set with threshold classifiers in each boosting round and the final classification. The size of samples corresponds to the weight assigned to the respective sample in each boosting round.

these cases AdaBoost tends to overfit on the data trying to correctly classify the complete data set. Several modifications of the original approach have been presented in order to achieve a more robust classification. Basically, these modified approaches limit the maximum weight that can be assigned to a sample (cf. e.g. [Freu 01, Meir 03]) or modify the calculation of sample weights (cf. e.g. [Frie 00, Jin 03]) in order to reduce the influence of potential outliers or noisy samples on the training procedure. Approaches aiming at fast training even in case of large data sets have been proposed [Brad 07, Hall 07], but in applications in the field of Metabonomics rather high-dimensional and sparse data sets are the usual case. Three different modifications of the original AdaBoost approach, which are expected to achieve improved classification results on sparse and high-dimensional data sets, will be presented in the following.

modified boosting approaches

3.6.0.1 *LogitBoost*

The *LogitBoost* approach introduced by Friedman et al. is expected to achieve stable classification results even on small and high-dimensional data sets [Frie 00]. Thereby, boosting is reformulated as additive logistic regression. A classification model is estimated by a stepwise optimization of the binomial log-likelihood function, which seems to be more suitable in classification than the

additive logistic regression

exponential criterion applied in the original AdaBoost. By this modification, weights are increased linearly instead of exponentially.

Investigations presented by Dettling et al. have shown increased performance of LogitBoost on small and high-dimensional gene expression data sets [Dett 03]. Within these experiments one-level decision trees, denoted as decision stumps, have been used as base classifiers. Further improvements in classification performance could be achieved by the selection of features with high significance for class discrimination as stated by the Wilcoxon signed rank test [Wilc 45]. This statistical test estimates the significance of features for class discrimination according to a labeled two-class data set.

3.6.0.2 *BagBoost*

LogitBoost combined with bagging

An extension of the LogitBoost approach for large data sets has been proposed by Dettling by the inclusion of the bagging method (cf. section 3.3.1) within the LogitBoost approach, denoted as *BagBoost* [Dett 04]. Thereby, not just a single classifier is estimated within each boosting round but a set of classifiers each trained on a bootstrap replicate of the original data set.

reduced bias and variance

The rationale for this approach is the combination of the boosting committee with lower bias and slightly increased variance with the nearly identical bias but lower variance introduced by the bagging procedure. Experiments on gene expression data supported the assumed classification improvements due to both reduced bias and variance [Dett 04]. Classification results were not only competitive to bagging and boosting, but also to discriminant analysis and SVMs.

3.6.0.3 *L_2 Boost*

gradient descent technique

Interpretation of the boosting approach as functional gradient descent technique has motivated the formulation of *L_2 Boost* by Bühlmann et al. [Buhl 03]. Thereby, a hypothesis h_k is determined according to the gradient descent with respect to the loss function $\frac{l(\mathbf{x})-h_k(\mathbf{x})}{2}$.

An experimental evaluation of this approach has shown comparable results to LogitBoost in case of data sets of low dimensionality, while the latter achieves the best results on data sets of high dimensionality. Generally, the L_2 Boost approach seems to work quite well with small decision trees as base classifiers or even decision stumps.

3.6.0.4 *Summary*

Apart from the presented boosting approaches, several other methods based on the AdaBoost algorithm have been presented up to now with modifications advantageous to a particular problem (cf. e.g. [Meir 03]). However, a systematic comparison of the most promising boosting approaches on a variety of data sets is still missing. But differences between the approaches seem to be rather small in published evaluations of different boosting methods.

Generally, boosting approaches show reasonable results in several applications although it is a quite straightforward approach taking advantage of the combination of several weak classifiers – the general idea of multiple classifier systems. Additionally, parameters of most boosting approaches are very limited or can even be selected according to data properties like data set size or data dimensionality. Thus no complex parameter optimization by cross-validation has to be applied. Although most boosting approaches tend to overfit on the data in case of noisy data sets, modifications of the original boosting algorithm have been proposed, also achieving a reasonable classification in these cases.

no application on NMR data

Although boosting methods are regarded as a valuable classification tool for several classification problems, they have not been applied for the detection of drug-induced organ toxicities based on NMR spectra up to now.

3.7 random forests

combination of multiple decision trees

Bagging has been presented in section 3.3.1 as a generic ensemble creation method in order to achieve multiple data subsets as a basis for an ensemble system. Breiman presented a variant of the bagging approach in combination with decision trees as base classifier algorithm, denoted as *random forest* [Brei 01]. Generally, a random forest is defined as a set of tree-structured classifiers grown with respect to a set of independent and identically distributed random variables. Each tree in the forest classifies a sample in a single class. The random variables control the ensemble creation technique regarding the subset of samples or features, or variations on tree parameters in order to achieve multiple diverse classifiers. Also a combination of these techniques would lead to a random forest by selection of a subset of samples and features for classifier training (cf. [Lati 00]).

random input selection

decision stumps

The most successful method of ensemble creation is the random input selection as shown in figure 3.22. Thereby, bootstrap replicates of the original data set are sampled and a random feature subset of size m' ($m' < m$) is selected randomly for the determination of the best splitting feature at each node (cf. section 3.4.2). Breiman proposes the application of full CART trees without pruning, but in several experiments decision stumps have been used within the random forest approach. The performance of the trees within a random forest is controlled via the param-

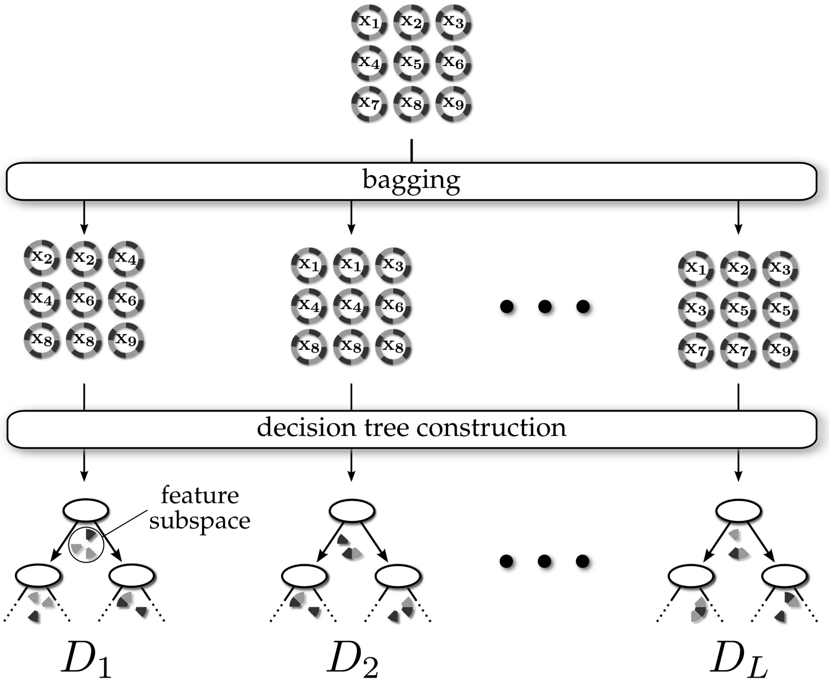


Figure 3.22: Scheme for construction of a random forest using random input selection. Different sample selections are used for the creation of random forests by random selection of a subset of features ($m' = 3$) at each node for determination of the best splitting feature.

eter m'. Large subspaces used for determination of the optimal splitting criterion will increase the individual performance of the decision trees but also increase their correlation. Assignment of small values to m' will result in a diverse set of low performing decision trees and the overall classification performance is achieved by their combination in the ensemble. While the number of trees has nearly no influence on the ensemble performance (common default values are 1 000 – 5 000) the value of m' can have a great influence on the ensemble if not all features are useful for classification. In these cases large values for m' have to be selected, otherwise there is no significant influence on classification performance (cf. e.g. [Brei 01, Diaz 06]).

validation by out-of-bag samples

assessment of features' relevance

Most classification approaches require an independent validation set for the optimization of their parametrization and estimation of classification performance. But within the random forest approach this can be achieved by the classification of samples not used within the decision trees, respectively, denoted as *out-of-bag samples*. An additional useful property of the random forest approach is the possibility to estimate the relevance of each feature for classification. This is achieved by adding noise to each variable separately and according to the amount of changing predictions each feature's relevance to the classification decision is estimated. However, this works only if no dependent features are present in the data since each feature is changed independently of each other. An approach to reveal these dependent variables for estimating their impact on the classification has been presented by Bureau et al. [Bure 05]. But a full evaluation of all possible variable combinations usually fails due to the exponentially growing computational complexity.

The random forest approach using random input selection has several advantages in comparison with alternative classification approaches as shown in [Brei 01]. It is comparable with AdaBoost, as the state-of-the-art classification approach in the field of multiple classifier systems in the late 90s, achieving sometimes worse but also better results. In contrast to AdaBoost random forests are relatively insensitive to outliers and noise in the samples since the influence of falsely classified samples on the classification decision is not increased within the forest construction procedure. Training a random forest is computationally efficient since only simple comparisons have to be evaluated at each node. Furthermore, each tree in the forest can be trained independently of each other, thus allowing for the parallelization of the forest construction.

insensitive to outliers and noise

3.8 SUMMARY

Like all classification approaches ensemble systems aim at the improvement of classification performance for a certain application. However, the major difference to other classification systems is the incorporation of multiple classifiers in order to improve classification performance by final ensemble combination. Experimental evaluations clearly indicate the advantages of multiple classifier systems, being a promising approach from the field of pattern classification.

boosting classification performance by multiple classifiers

All ensemble systems have a common overall structure but differ in the methods applied. In order to take advantage of multiple classifiers, diversity in the classifications from the set of base classifiers has to be achieved. Otherwise, the classification performance of the individual classifiers and the ensemble will be identical. Different generic ensemble creation methods have been proposed based on modifications of the data set by means of varying samples or feature selections. Thus, each base classifier is trained on a different data basis. Base classifiers have to be sensitive to these variations in order to achieve diverse classifications. Besides their instability only a classification performance better than 50 % is a prerequisite for an adequate classification algorithm. If these conditions are fulfilled, predictions from the diverse classifiers are finally aggregated to an ensemble classification which is expected to be better than every individual

diverse base classifiers

ensemble creation

base classifier

ensemble aggregation

classifier. The applied fusion method depends on the output type of the base classifiers and several aggregation methods have been proposed.

optimization of ensemble methods

The optimal combination of ensemble creation, classification and ensemble aggregation methods can be determined according to an experimental evaluation. Furthermore, out-of-bag samples can be used in ensemble approaches like boosting or random forests in order to determine the best ensemble design. However, the ensemble design is not restricted to the proposed methods and classification performance can usually be improved by the adaptation of the specific methods to the problem at hand. The identification of the major problems for a successful classification of the present data set and design of a suitable ensemble system is the major challenge but also represents the flexibility of ensemble systems.

4 MATERIALS AND EVALUATION METHODS

Before new ensemble systems for the robust classification of NMR spectra will be presented in the next chapter, the data sets and evaluation strategies used for the development of the new approaches are presented in this chapter. Data acquisition is an important aspect in metabonomic applications due to the complex experiment design and the expensive sample measurement by NMR spectroscopy. Two different data sets will be presented in section 4.1, followed by a detailed description of the evaluation strategy used in this thesis.

4.1 DATA SETS

The design of a data set that is used for training and evaluation is an important step in the development of new classification systems. The data set should contain a reasonable amount of representative examples from the particular domain in order to achieve a good performance of the classification system on unseen samples.

prerequisites for metabonomic data sets

In order to fulfill this prerequisite in the presented metabonomic application, a multitude of realistic experiments from studies pursued in safety pharmacology have to be performed. First of all, a representative amount of pharmaceuticals with different effects on the organism have to be applied, including harmless substances in order to produce control samples. Furthermore, the same pharmaceutical has to be applied to multiple experimental animals in order to compensate variations in the response induced by different genders or genetic variations of the particular individuals. Since the point in time of the maximum response to a pharmaceutical is usually unknown, it is not sufficient to collect a single sample. Thus, samples from the same individual have to be collected at multiple time points.

reasonable experiment design for acceptance of results

A thoughtful design of experiments is the key for acceptance of results achieved in the study. Especially in drug design, each newly developed pharmaceutical has to be accepted by particular federal institutions before it can be sold in the respective country. For example, the U.S. Food and Drug Administration (FDA) has to control the efficacy and possible adverse effects of drugs and foods in the United States of America. Only particular studies are accepted by this institution in order to investigate possible adverse effect. Thus, a valid design of experiments has to be defined in order to achieve reliable results.

The investigation of the toxic effect of even a small set of compounds requires the generation of hundreds of samples due to the complex design of experiments. This acquisition is a complex and laborious procedure, including the following steps:

1. specification of compounds to be applied
2. design of the animal study and sample collection
3. sample analysis by clinical chemistry
4. measurement by NMR spectroscopy.

no data sets publicly available

Therefore, the buildup of a large data set of NMR requires considerable amounts of financial expenditures, specific equipments, and expert knowledge in the respective fields. Due to these facts, metabonomic data sets are usually strictly confidential and are not available for public use. Hence, the sets of spectra used in this thesis are not publicly available and partially blinded. However, all details are given that are necessary in order to follow results presented from the experimental evaluation.

In the following, two data sets as they are used for experimental evaluations presented in this thesis are presented. Following this, the combination of predictions from several samples is presented, which allows for a classification with respect to a compound applied in a particular dose. Finally a detailed description of the evaluation procedure by cross-validation is given.

4.1.1 *Real Data Set from Safety Pharmacology (`REAL.NMR`)*

47 compounds

The primary data set used in this thesis for the development of new classification systems is a set of ^{1}H NMR spectra of urine samples from rats treated with one of 47 different compounds. Throughout this thesis this data set is denoted as `REAL.NMR`. The urine samples originate from a safety pharmacology study carried out by the pharmaceutical company Boehringer Ingelheim Pharma GmbH & Co KG (BI). In this study, each compound was applied to a group of eight experimental animals, composed of 4 male and 4 female rats, either orally (p.o.), intravenously (i.v.), intraperitoneally (i.p.) or inhalatively (i.h.). These rats were housed in metabolic cages and urine samples were collected 8 and 24 hours after the application of the compound. The use of urine instead of blood allows the collection of biofluid samples at different points in time, while animals would have to be euthanized before blood samples can be taken. As a further important aspect, the number of experimental animals used for this study is reduced.

applied to 8 rats

sample collection after 8 and 24 hours

The main response is expected to be detectable in the first 24 hours and no further samples have been collected. Thus, 16 samples are collected for each compound. Certain experiments have been repeated using a different dose of the particular compound in order to investigate dose dependent effects. Some compounds have been applied multiple times for the generation of an increased amount of samples for relevant compounds.

The NMR measurements were performed at LipoFit Analytic GmbH (Regensburg, Germany) by order of BI. Samples were measured on a Bruker AVANCE 600 plus Ultrashield™ NMR spectrometer, followed by automatic and manual spectroscopic postprocessing procedures. Further details on the experimental methods (animal study, probe handling, clinical chemistry, NMR spectroscopy) can be found in [Lien 08b]. The full set of processed NMR spectra was provided by BI for the evaluation of classification systems.

semi-automatic measurement and postprocessing procedure

A small set of samples could not be analyzed by clinical chemistry or NMR spectroscopy due to insufficient sample volumes and was, hence, excluded from the data set. Furthermore, some samples have shown large differences to the remaining samples, indicated by outliers in PCA score plots, and are also excluded. These differences are induced by strong NMR signals from drug related compounds or by perturbing artifacts from the measurement process.

Each of the 896 NMR spectra from the final data set is labeled according to literature references as being *non-toxic* (651 samples) or *toxic* (245 samples) with respect to the proximal tubule (cf. [Lien 08b])[1]. The applied compounds in combination with their dose, amount of samples and the percentage of samples indicating a toxic effect according to the analysis of results from safety pharmacology are shown in tables 4.1 and 4.2. Internal substances of BI are not listed with their real name due to nondisclosure agreements and are referred to as *internal compound*.

896 samples (651 non-toxic; 245 toxic)

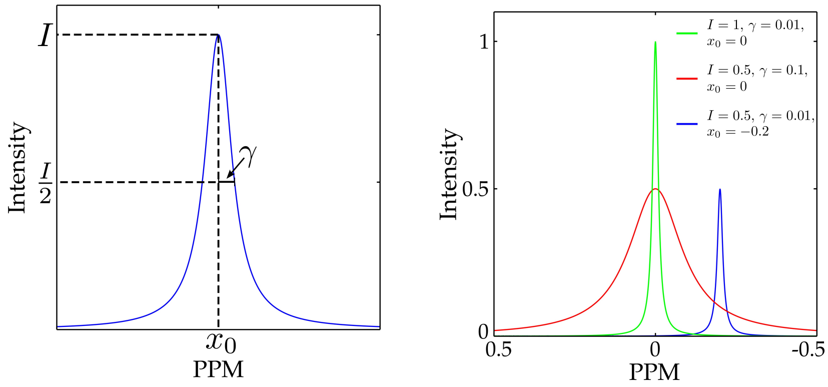
The majority of compounds judged as non-toxic are not inducing any physiological reaction in the organism, but rather show no toxic effect with respect to the proximal tubule. Thus, different physiological reactions are induced by the applied pharmaceuticals and the objective of classification methods is to determine variations in molecule concentrations indicating an organ toxicity of a particular organ. The small amount of real control samples is the major difference to the data set generated in the COMET project.

only few real control samples

The low percentage of samples, which are expected to induce a toxic effect but are judged as non-toxic according to results from clinical chemistry, demonstrates the low sensitivity of clinical chemistry. This is a known problem, especially in the analysis of samples collected only few hours after single administration of a

low sensitivity by clinical chemistry

[1] Chloroquine is judged as non-toxic contrary to the literature reference given in [Lien 08b] due to differences in the applied dose and analysis results by clinical chemistry.



(a) Properties of a Lorentzian peak (b) Different peak parametrizations

Figure 4.1: Characteristics of a parametrized Lorentzian line and examples of peaks resulting from different parametrizations.

compound. Thus, the objective of new classification methods is to achieve a robust detection of drug-induced organ-toxicities even for this particular type of samples. Therefore, literature references are used for labeling as non-toxic or toxic in order to achieve a reliable detection of organ toxicities where clinical chemistry is not sensitive enough to detect them.

4.1.2 *Artificial Data Set (`SIM.NMR`)*

data set of spectra with known properties

Even though a set of 896 samples is not regarded as a large data set in comparison with other sets used in pattern classification applications, the `REAL.NMR` data covers a reasonable range of applied compounds and toxic samples. In consideration of the complex sample collection and expensive measurement procedure this data set is a good basis for the development and evaluation of classification methods in the field of Metabonomics. The presence of peak signals allowing for classification of spectra is expected due to results from experimental evaluations, but the amount, relevance and positions of these signals are not known. Thus, in order to model particular aspects of the data, which should be identified by the classification methods, a data set of simulated spectra is used in addition to the `REAL.NMR` data set for evaluation and classification interpretation.

Lorentzian peak shape

NMR spectra are mainly composed of noise in baseline regions and several peak signals with variation in their exact position and intensity. Baseline regions are modeled by random values in a small range around zero, but the shape of simulated peaks should be comparable with those in real NMR spectra. Following [Koh 08], peaks in NMR spectra can be described by Lorentzian lines, which are described by a parametrized Lorentzian function. Given a peak at position x_0, the peak intensity I and the half-

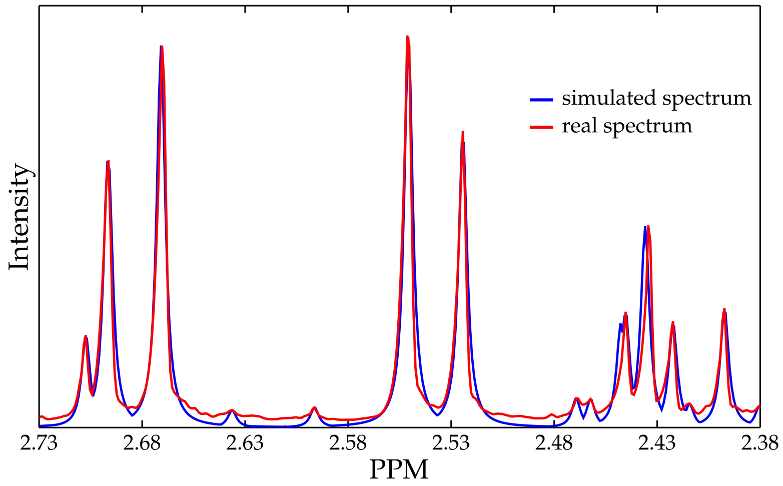


Figure 4.2: Comparison of the spectral region around the citrate peak of a simulated and a real spectrum.

width at half-maximum (HWHM) γ (cf. figure 4.1a), the spectral value $s(x)$ at the position x is determined by

$$s(x) = I * \frac{\gamma}{(x - x_0)^2 + \gamma^2}.$$

These three parameters influence the position, intensity and shape of the peak signal as shown in figure 4.1b. A spectrum is now defined as the sum of a noisy baseline and a set of parametrized peaks.

spectrum as a sum of parametrized peaks

In order to achieve a peak representation similar to ^{1}H-NMR spectra from urine samples, the 156 most dominant peaks from a control spectrum are determined. The control spectrum is measured from a urine sample after the application of Natrosol™. Peak detection is performed by the algorithm used in the PARSE procedure (cf. section 2.3.1) and only peaks that could be robustly detected are used. Thus, a representative non-toxic NMR spectrum can be defined according to the identified peak positions and intensities. A HWHM of 0.0025 ppm is used for peak simulation, which is determined by visual investigation of real NMR spectra. A comparison of a simulated and a real spectrum in the region of the citrate peak is shown in figure 4.2. Major differences can be observed with respect to the baseline, but the peak shapes are well simulated.

156 peaks from a control spectrum

HWHM of 0.0025 ppm

A single peak is fully defined by its position, intensity and HWHM, but changes in position and intensity have to be induced in order to simulate different spectral profiles. A statistical model of changes specific to each peak based on training samples can not be determined due to the problems in peak detection and alignment. Thus, random changes in position and intensity up to a maximum value are applied to each peak. These maximum values are specifically defined for each peak.

random changes in position and intensity

Four multiplets have been identified by visual investigation of the REAL.NMR spectra and modeled with the same peak shifts and intensity changes. This leads to the definition of a set of 148 signals for the simulation of the spectra. Peak shifts of up to 0.005 ppm to the left and to the right are randomly applied to 30 % of the spectra. Intensity changes are applied in different ranges dependent on the peak height. 129 groups show random intensity changes from about 50 % of their intensity defined by the peak detection algorithm. In every spectrum a single peak has the same intensity in order to identify scaling artifacts. The remaining signals vary in their intensity between 20 % and 80 %. Additionally, 35 Random peaks with varying intensity and position are added to the spectra in order to induce randomness in the simulated spectra. Furthermore, samples originating from the same compound are simulated by the addition of up to five random peaks to the spectra of the particular compound.

35 random peaks

substance-specific random peaks

Peaks defined up to now have shown purely random changes in intensity and contain no information for classification purposes. In order to define spectral profiles specific to non-toxic and toxic samples, five different biomarker peaks are added to the spectra. The position, intensity and maximum variation of these biomarker patterns are shown in table 4.3. The class-specific range of the intensity values are chosen in order not to allow for a correct classification using a single peaks. Thus, no classification system is expected to achieve a perfect classification, but different classification approaches can be compared with respect to their opportunity to combine and localize relevant spectral regions.

5 biomarker peaks

The simulation of random peaks is illustrated in figure 4.3 for the biomarker peak BP_2. The mean peak intensity for non-toxic and toxic samples is defined as 0.5 and 0.7, respectively. Intensity changes are induced from -0.2 to 0.2 . Furthermore, the peak position randomly varies between 4.49 ppm and 4.51 ppm. Different peak maxima defined by class-specific changes in intensity and position define two overlapping regions, whereby, no linear discrimination between non-toxic and toxic samples can be achieved. This also holds for the remaining biomarker peaks defined in table 4.3.

960 spectra (480 non-toxic; 480 toxic

A final data set consisting of 960 spectra is defined, using these definitions of normal, random and class-specific peaks, whereby the sets of spectra labeled as non-toxic and toxic each contain 480 spectra. For data set creation it is assumed that samples originate from the application of 20 different compounds. Thus, samples originating from the application of a particular compound contain up to five compound specific peaks, which other samples do not contain. 10 compounds are used to define 16 spectra for each compound and for every of the remaining compounds 32 samples are simulated. Furthermore random peaks and class-

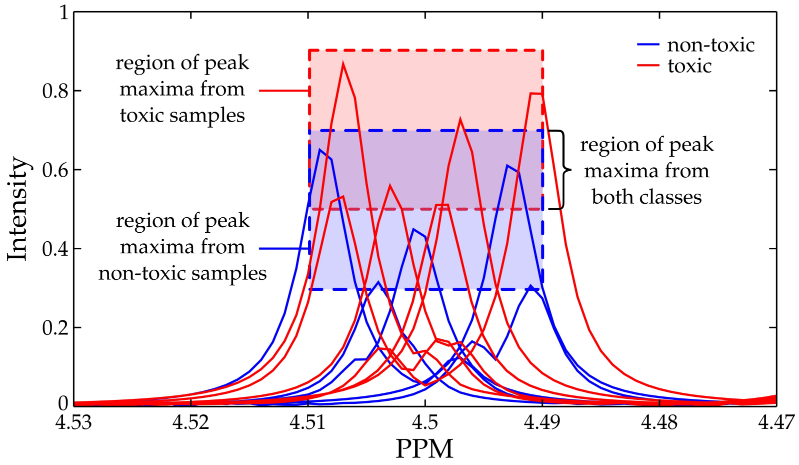


Figure 4.3: Demonstration of the regions of non-toxic and toxic peaks for an exemplary simulated biomarker signal with simulated changes in peak intensity and position.

specific peaks are included. Thereby, an amount of samples and compounds comparable to the REAL.NMR data set is achieved.

The simulated data set SIM.NMR is expected to model only a part of the full complexity present in the REAL.NMR data set. Results achieved on this data set are used for the comparison of classification systems, but they do not reflect their true performance on real metabonomic data sets. However, the evaluation of newly developed ensemble systems on the simulated data allows for a comparison of their interpretability. The presence of biomarker patterns in the REAL.NMR data set is not proven and can only be assumed with respect to the evaluation results. In contrast to this, the position of biomarker patterns is known for the simulated data set and can be used for the validation of the final results.

evaluation of interpretability

4.2 Evaluation of Classification Results

An experimental evaluation has to be properly designed in order to determine the real performance of a classification system. A selection of performance measures and strategies for classifier evaluation are described in section 3.1.1. Methods applied for the evaluation of the proposed classification approaches are outlined in this section.

performance measure and evaluation strategies

4.2.1 *Performance Measures*

Classification accuracy (acc), specificity (spec), and sensitivity (sens) are commonly used performance measures in metabonomic applications as shown in section 3.1.1. Additionally , the

MC as main optimization criterion

Matthews Correlation Coefficient (MC) is applied in order to evaluate the classification performance even for classifications of the imbalanced REAL.NMR data set[2]. Thus, the MC serves as primary optimization criterion in the following evaluations and the remaining measure are denoted in addition.

4.2.2 *Cross-Validation and Test*

optimization of classification parameters

Most classification methods require the optimization of particular parameters. Thus, in order to allow for parameter optimization on a validation set and determination of the final performance on an independent test set, a stratified[3] five-fold cross-validation and test approach is applied (cf. section 3.1.2). Initially, the whole data set is split into five parts of comparable size and five different cross-validation folds are defined. Three fifths are used for classifier training and parameters specific to the classifier are optimized on one fifth, denoted as validation set. The final performance of the optimized classifier is determined on the remaining fifth.

presence of almost identical samples

Precautions have to be taken if almost identical samples from the same class are present in a data set. An example of an application where nearly identical samples can lead to false evaluation results is the image-based face-detection. If samples from the same person exist in the training and test set, a classifier using characteristics specific for this single individual would lead to good results for samples in the test set. However, the objective is to determine the performance of the system for unknown samples. Thus, very similar samples have to be contained in the same fold in order to avoid their use for training and evaluation.

retention of groups of samples

The presented REAL.NMR data set contains similarities between spectra due to samples collected from the same animal or after application of the same compound. Furthermore, compounds can be grouped with respect to their target and indications as shown in table 4.4. Compounds with the same indication are expected to have a similar effect on the organism. Thus, groups of samples associated with compounds of the same indication are retained in the creation of the folds, avoiding similarities of samples in training and validation or test sets induced by the same indication. The low number of indications limits the number of folds used in this thesis to five. Otherwise, very small folds, or folds containing only negative or positive samples would be defined.

[2] Classifications for imbalanced data sets, containing unequal amounts of samples labeled as non-toxic and toxic, would lead to good results even by assigning the label of the prevalent class to all samples. Therefore, alternative evaluation measures have to be applied.

[3] Stratified cross-validation aims at creation of folds with an almost equal proportion of samples from the different classes as the original data set.

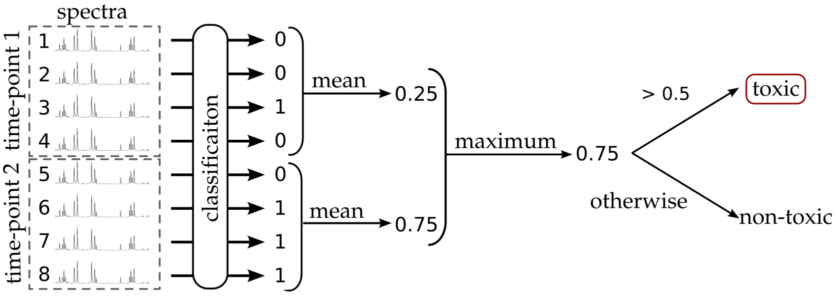


Figure 4.4: Combination of predictions for samples associated with the same compound. The mean value of all samples from the same collection time is determined and the final prediction for the compound is achieved with respect to the maximum of predictions for all time points.

4.2.3 *Combination of Predictions for Compounds*

Predictions achieved by a classification system are always assigned to individual samples. However, the main objective is the detection of compounds, which induce a particular organ toxicity. Thus, the predictions for samples of the same compound have to be combined in order to achieve a classification as non-toxic or toxic with respect to the particular dose.

combination of sample predictions

A first combination of the samples associated with the same substance-dose combination, and collected at the same time after application can be achieved by averaging over their predictions. Samples classified as non-toxic are labeled as "0" and in case of a predicted toxic reaction the label "1" is assigned. Toxic effects are expected to be detectable at specific collection times and not at each of them. Since, even a detected organ toxicity at a single point of time is sufficient to classify a substance-dose combination as inducing an organ toxicity, the maximum over predictions for the respective points in time is used. If the final value is larger than 0.5, the compound is classified as toxic. The principle of this *group-classification* is shown for an experiment with two collection time-points in figure 4.4.

IDENTIFIER	SUBSTANCE	DOSE	♯ SAMPLES	TOXIC
Cpd04	BEA 2108 BR	10 μg/kg i.h.	15	13 %
Cpd05	BI internal 01	10 mg / kg p.o.	31	0 %
Cpd06	BI internal 02	30 mg / kg p.o.	15	0 %
Cpd07	BI internal 02	100 mg / kg p.o.	16	6 %
Cpd08	BI internal 02	300 mg / kg p.o.	16	25 %
Cpd10	BI internal 04	10 mg / kg p.o.	31	6 %
Cpd11	BI internal 05	25 mg / kg p.o.	32	9 %
Cpd12	BI internal 06	10 mg / kg p.o.	16	25 %
Cpd13	BI internal 06	100 mg / kg p.o.	32	78 %
Cpd15	BI internal 07	100 mg / kg p.o.	13	33 %
Cpd16	BI internal 08	5 mg / kg p.o.	16	0 %
Cpd17	BI internal 08	15 mg / kg p.o.	16	13 %
Cpd18	BI internal 09	0.5 mg / kg p.o.	16	6 %
Cpd19	BI internal 10	30 mg / kg p.o.	16	0 %
Cpd20	BI internal 11	100 mg / kg p.o.	15	20 %
Cpd21	BI internal 12	50 mg / kg p.o.	15	13 %
Cpd23	BI internal 14	200 mg / kg p.o.	16	25 %
Cpd24	BI internal 15	100 mg / kg p.o.	16	31 %
Cpd25	BI internal 16	50 mg / kg p.o.	16	0 %
Cpd26	BI internal 17	54 mg / kg p.o.	16	19 %
Cpd27	BI internal 18	300 mg / kg p.o.	16	6 %
Cpd28	BI internal 19	54 mg / kg p.o.	16	0 %
Cpd29	BI internal 20	30 mg / kg p.o.	16	6 %
Cpd30	BI internal 21	300 mg / kg p.o.	16	13 %
Cpd31	Chloroquine	30 mg / kg p.o.	14	0 %
Cpd33	BI internal 22	100 mg / kg p.o.	16	6 %
Cpd34	BI internal 23	25 mg / kg p.o.	16	0 %
Cpd35	BI internal 24	120 mg / kg p.o.	16	6 %
Cpd40	HP-β-CD	30 mg / kg i.v.	14	7 %
Cpd41	Hydrochlorothiazide	20 mg / kg p.o.	16	25 %
Cpd43	Imipramine	100 mg / kg p.o.	16	19 %
Cpd45	NaCl	0.9 %i.v.	16	19 %
Cpd46	NaCl / Glucose	0.9 % / 5 % isoosm. 1:1 i.p.	14	0 %
Cpd49	Natrosol	0.5 % p.o.	60	0 %
Cpd50	Netilmicin	20 mg / kg i.p.	16	6 %

Table 4.1: Substances regarded as non-toxic with respect to the proximal tubule according to literature references (cf. [Lien 08b]). The percentage of samples labeled as toxic according to the results from clinical chemistry are additionally given.

IDENTIFIER	SUBSTANCE	DOSE	♯ SAMPLES	TOXIC
Cpd01	2-BEA	100 mg / kg i.p.	16	69 %
Cpd02	Amiodarone	200 mg / kg p.o.	15	7 %
Cpd03	Ampothericin B	4 mg / kg i.p.	16	6 %
Cpd09	BI internal 03	100 mg / kg	13	80 %
Cpd14	BI internal 06	300 mg / kg p.o.	15	100 %
Cpd22	BI internal 13	10 mg / kg	16	0 %
Cpd32	Cisplatin	10 mg / kg i.p.	16	25 %
Cpd36	Folic acid, unbuffered	200 mg / kg i.p.	16	13 %
Cpd37	Gentamicin	100 mg / kg s.c.	14	7 %
Cpd38	HCBD	100 mg / kg i.p.	16	56 %
Cpd39	$HgCl_2$	1 mg / kg i.p.	16	44 %
Cpd42	Hydroquinone	100 mg / kg p.o.	8	75 %
Cpd44	Indomethacin	30 mg / kg p.o.	13	15 %
Cpd47	$NaCrO_4$	30 mg / kg i.p.	16	44 %
Cpd48	NaF	20 mg / kg i.p.	16	13 %
Cpd51	N-phenyl-2-naphthylamine	250 mg / kg p.o.	16	0 %
Cpd52	Pluronic F-68	500 mg / kg i.v.	7	43 %

Table 4.2: Substances regarded as toxic with respect to the proximal tubule according to literature references (cf. [Lien 08b]). The percentage of samples labeled as toxic according to the results from clinical chemistry are additionally given.

NAME	POSITION [ppm]	INTENSITY NON-TOXIC	INTENSITY TOXIC
BP_1	1.0 (±0.01)	0.1 (±0.1)	0.2 (±0.2)
BP_2	4.5 (±0.01)	0.5 (±0.2)	0.7 (±0.2)
BP_3	4.23 (±0.005)	0.5 (±0.2)	0.3 (±0.2)
BP_4	6.8 (±0.005)	0.2 (±0.1)	0.1 (±0.1)
BP_5	8.11 (±0.005)	0.3 (±0.1)	0.4 (±0.1)

Table 4.3: Position and intensity of biomarker patterns added to the simulated data set. Positions and intensities are randomly varied in the denoted ranges.

SUBSTANCE	INDICATION
2-BEA	Experimental
Amiodarone	Anti-Arrhytmic
Ampothericin B	Antibiotic
BEA 2108 BR	Anticholinergic Drug
BI internal 01	PDE IV
BI internal 02	DPP IV / Diabetes
BI internal 03	CGRP / Megrim
BI internal 04	PDE IV
BI internal 05	MCH
BI internal 06	Factor 10a
BI internal 07	CGRP / Megrim
BI internal 08	PDE IV
BI internal 09	Squalen-Cyclase-Inhibitor
BI internal 10	Squalen-Cyclase-Inhibitor
BI internal 11	CGRP / Megrim
BI internal 12	EGFR
BI internal 13	Fibrinogen-R-Antagonist
BI internal 14	Fibrinogen-R-Antagonist
BI internal 15	EGFR
BI internal 16	NK1-Antagonist
BI internal 17	NK1-Antagonist
BI internal 18	LTB4-Antagonist
BI internal 19	NK1-Antagonist
BI internal 20	NK1-Antagonist
BI internal 21	LTB4-Antagonist
Chloroquine	Antimalaria Agent
Cisplatin	Cytostatic Drug
BI internal 22	5HT3R-Antagonist
BI internal 23	MCH
BI internal 24	EGFR
Folic acid, unbuffered	Vitamin
Gentamicin	Antibiotic Aminoglycosid
HCBD	Experimental
$HgCl_2$	Experimental
HP-β-CD	Excipient
Hydrochlorothiazide	Diuretikum
Hydroquinone	Experimental
Imipramine	Tricyclic Antidepressant
Indomethacin	NSAID
NaCl	Vehicel
NaCl / Glucose	Vehicel
$NaCrO_4$	Experimental
NaF	Experimental
Natrosol	Zellulose / Vehicel
Netilmicin	Antibiotic Aminoglycosid
N-phenyl-2-naphthylamine	Experimental
Pluronic F-68	Excipient

Table 4.4: Compounds and their division into different indication classes.

5 ADVANCED ENSEMBLE APPROACHES FOR CLASSIFICATION OF NMR SPECTRA

The reliable detection of drug-induced organ toxicities is an important prerequisite for efficient drug design in pharmaceutical industry. Histopathology or analysis of biofluids by clinical chemistry allow for the identification of organ toxicities, but these are complex procedures requiring expert knowledge in the respective fields. Thus, an automatic detection of organ toxicities is desired for a fast evaluation of animal experiments in order to give support to the determination of possible adverse effects induced by a pharmaceutical. Sample analysis can be achieved by NMR spectroscopy in a non-destructive and non-selective way, leading to spectral profiles specific to the organism's health status. Changes in the metabolism of an organism can be detected in these spectra as variation of peak intensities.

Metabonomics

Methods from the field of pattern recognition have been applied to analyze these complex NMR samples in order to automatically detect organ toxicities based on patterns of spectral changes. However, these patterns are not known beforehand and their identification is a complex task. Spectral signals are altered by different factors such as noise and peak shifts, whereby the information on the concentration of specific molecules is not easily accessible. Furthermore, available NMR data sets are rather small with only hundreds to thousands of samples due to the expensive measurement procedure.

complex and sparse data sets

As presented in chapter 3, ensemble methods have led in several applications to more accurate and robust classification models in comparison with single classifier approaches. The basic design of ensemble methods aims at the creation of base classifiers with different *views* on the data and their combination by ensemble aggregation methods. Different views can either be induced by the modification of the data or variation in the training of the base classifier. Generally, ensemble methods are not restricted to certain types of algorithms for the creation of base classifiers, and any approach leading to a set of diverse classifiers can be applied. This flexibility in the ensemble design allows for the definition of views specific to the classification of NMR spectra as either indicating an organ toxicity or not. Thus, robust and accurate classifications can be achieved even for noisy and sparse data sets.

ensemble of classifiers

adaptive ensemble design

NMR spectra are rich of information on the concentration of multiple molecules contained in a sample. However, the identification of signals, which show intensity changes in case of organ toxicities, is an important aspect in order to achieve a robust classification. Since the measurement procedure is non-selective, only a small fraction of signals present in a spectrum is expected to be relevant to classification purposes. The flexibility of ensemble systems in the creation of base classifiers can be used to overcome this problem of many spectral regions not being relevant to classification purposes. By estimating the relevance of each spectral region in a data-driven way, this information can be included in the ensemble creation, finally leading to an improved ensemble performance. Furthermore, the information on relevant spectral regions can be used in safety pharmacology for the determination of biomarker patterns. The identification of molecules corresponding to the biomarker patterns is expected to allow for a more robust and reliable detection of organ toxicities in clinical chemistry.

few relevant signals

focus on particular spectral regions

Results presented in the following description of developed ensemble methods are used for the illustration and justification of particular decisions in the design of the ensemble creation procedures. An extensive experimental evaluation of the presented approaches and a comparison with alternative classification methods is given in chapter 6.

5.1 VARIATION OF SPECTRAL PREPROCESSING METHODS FOR ENSEMBLE CREATION

Raw NMR spectra are initially described by 250 000 spectral variables. This spectral range is larger than the region containing peak signals. Thus, a commonly accepted preprocessing procedure is the exclusion of spectral regions below 0.02 ppm and above 10 ppm for a metabonomic analysis. Furthermore, the region around the water and urea peaks (4.5 ppm - 6 ppm) is excluded in order to reduce disturbing artifacts induced by the water and urea peak.

This restriction on particular spectral regions is a standard procedure, used in several publications. However, this is the only standardized preprocessing procedure. A variety of other preprocessing approaches is proposed in the literature, showing improvements in metabonomic investigations of the respective data sets. Although the application of a bucketing procedure using a bucket width of 0.04 ppm is proposed in several publications, this bucket width is not well motivated[1] and can further

no standardized preprocessing procedure

[1] In fact, this bucket width has first been motivated in the early days of Metabonomics. Bucketing was used as a method to reduce the number of variables below 256 as the maximal number of columns in an Excel sheet. Using the spectral range from 0.02 ppm to 4.5 ppm and 6 ppm to 10 ppm has led to the choice of 0.04 ppm.

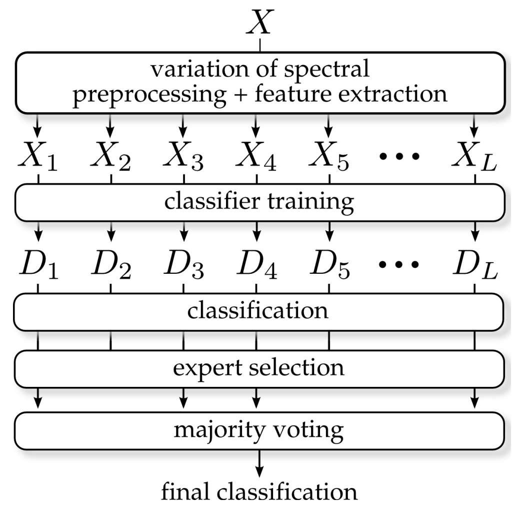

X
variation of spectral preprocessing + feature extraction
X_1 X_2 X_3 X_4 X_5 ••• X_L
classifier training
D_1 D_2 D_3 D_4 D_5 ••• D_L
classification
expert selection
majority voting
final classification

Figure 5.1: General structure of the ensemble system based on variation of spectral preprocessing and feature extraction methods.

be optimized for some particular data set. Due to the lack of a standard preprocessing protocol, available preprocessing methods have to be evaluated and the best performing one is used for final application.

selection of the best single preprocessing procedure

This is a classical situation, where multiple classifiers with varying performances are generated, but only a single one is used due to the restriction to a single classifier system. Furthermore, the best performance is always limited to the particular data set used during optimization. Thus, no general statement about the applicability of different preprocessing methods can be achieved. Ensemble systems allow to overcome this restriction and combine the set of classifiers. This ensemble creation procedure based on the variation of preprocessing methods is a first general proof of concept of the suitability of ensembles for metabonomic applications. This ensemble system has been presented in [Lien 08a, Lien 08b] and its general structure is shown in figure 5.1.

combination of different preprocessing procedures

A first variation is achieved in this approach by the selection of 0.01 ppm, 0.02 ppm or 0.04 ppm as bucket width. Large buckets improve the compensation of peak shifting effects but also decrease the spectral resolution and increase the number of peaks combined to a single variable. The selected bucket width either allows for an increased spectral resolution or for a compensation of peak shifts.

variation of bucket width

In addition to the bucketing operation, standard normal variate (SNV) transformation is used for scaling spectroscopic data sets, normalizing the variables of each spectrum to a mean-value of zero and standard deviation of one [Barn 89]. This transformation has been developed for the normalization of near infrared spec-

optional SNV scaling

troscopic data, but has also shown good results for NMR spectra [Jank 09, Maso 07]. Variation of bucket width and the optional application of the SNV transformation generates six different data sets.

variation in the number of PLS components

Common practice is also the projection of the spectral data set by PLS transformation (cf. section A.2) into a feature space of lower dimensionality. The PLS transformation focuses on variables relevant to class discrimination, allowing for an improved performance of a subsequent classification procedure. The number of PLS components is usually determined according to the classification performance on a validation set. However, this is only a valid procedure if the estimated PLS model achieves for each feature a correct estimation of its relevance to class discrimination. Otherwise, features generated by additional PLS components contain information that lead to an improved performance of a subsequent classification procedure. Thus, the number of PLS components for the transformation of the six preprocessed data sets is also evaluated by building models using one to five PLS components and increasing the number up to fifty in steps of five.

ensemble combination

This variation of preprocessing and feature extraction finally leads to a set of 90 differently preprocessed data sets, which serve as data basis for the training of base classifiers. Each classifier predicts a sample dependent on the respective preprocessing procedure, leading to different sets of misclassified samples. Combination of these predictions by majority voting is expected to lead to a reduced set of misclassified samples as introduced in section 3.2.

5.2 MODIFICATION OF RSS FOR IMPROVED ENSEMBLE CLASSIFICATION

The ensemble approach presented in the previous section serves as a proof of concept for the applicability of ensemble systems in metabonomic applications. Ensemble systems allow for the design of ensemble creation and aggregation procedures in consideration of characteristics of some particular data set. The classification system described in this section is based on a prominent ensemble creation procedure, which is adapted for the domain of NMR classification.

Reconsidering the representation of molecule concentrations in an NMR spectrum (cf. section 2.2), it becomes clear that certain variables, which correspond to specific regions within the spectrum, reflect the concentration of specific molecules. Only a small fraction of the molecules measured by NMR spectroscopy is expected to show changes in concentration caused by an induced organ toxicity. Thus, a selection of variables relevant to the classification of samples as non-toxic or toxic has to be determined.

selection of spectral regions

The RSS procedure, as introduced in section 3.3.2, randomly selects a subspace of variables for the creation of a new data set Φ. Using this ensemble creation procedure, a subspace selection and improved generalizability can be achieved [Ho 98]. However, according to the originally proposed RSS procedure, all dimensions are selected with equal probability for inclusion in each of the subspaces. In order to include differences in the relevance of particular variables for classification purposes in the subspace selection procedure, weights are assigned to each variable. High weights are assigned to the most relevant variables and increase their probability to be selected for a subspace. However, in order to derive this improved RSS procedure, a reasonable weight distribution has to be estimated.

RSS

weighted RSS

An iterative optimization procedure has first been presented in [Lien 07] in order to determine a weight distribution for a particular data set and a detailed description is given in this section. The optimization procedure aims at increasing the weights of features regarded as most relevant to classification purposes in an automated way. Due to the correspondence of single variables to particular spectral regions, an automated selection of relevant regions of the spectra is achieved. By this newly developed enhancement of the RSS procedure, the mean classification performance of base classifiers is increased and finally leads to an improved ensemble performance. Due to the modification of a weight distribution used within the RSS procedure, this approach is denoted as *adapted random subspace sampling (ARSS)*.

automatic weight optimization

Optimization of Weighted RSS

Prior to the optimization of variable weights, a bucketing procedure using a bucket width of 0.01 ppm is applied to the spectral data set. Peak shifts are a common problem in metabonomic investigation and influence the success of subsequent classification approaches. Using a bucket width smaller than the quasi-standard of 0.04 ppm reduces the effect of peak-shift compensation, but improves the spectral resolution. Thus, each bucket value contains intensity information of a smaller number of peaks and the selection of particular variables is comparable with the selection of specific peak signals.

preprocessing by a bucketing procedure

The optimization procedure, as shown in figure 5.2 and algorithm 2, starts with uniform weights w_i, $i = 1, \dots, m$ assigned to each dimension. According to these weights, a set of L subspaces is extracted. Thereby, n dimensions are drawn randomly with replacement from the originally n-dimensional data set. This leads to L data sets $\Lambda_j \in \mathcal{R}^n$, $j = 1, \dots, L$ with in average 63.2 %

uniform weight distribution as starting point

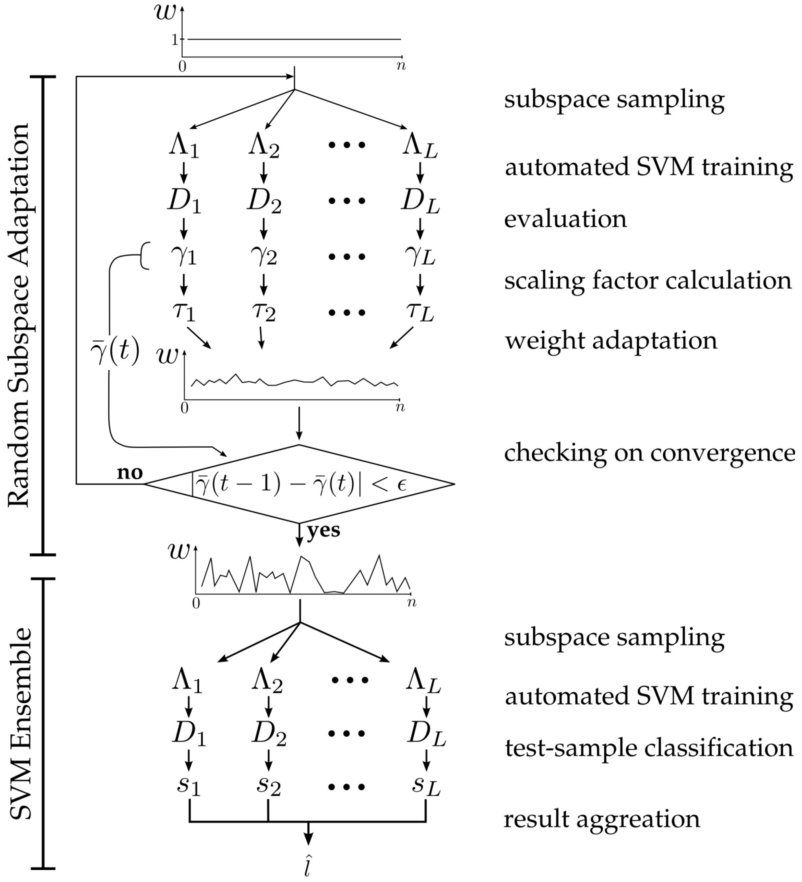


Figure 5.2: Weight adaptation procedure.

unique features from the original data set [Brei 96c]. This is the general procedure applied in RSS and serves as starting point for the optimization procedure.

The random selection of variables used for the definition of the subspaces $\Lambda_1, \ldots, \Lambda_L$ is the basis for the training of diverse classifiers $D_1, \ldots, D_L$ with classification performances $\gamma_1, \ldots, \gamma_L$. In this ensemble, the estimated model for each classifier and its classification performance are dependent on the selection of variables contained in the particular subspace. Thus, variables used by the classifier with the best performance γ_B allow for an improved classification of a validation set in comparison to variables used for estimation of the classifier model with the worst performance γ_W. This information on the variables used by each classifier and the resulting classification performance is the basis for the modification of weights used in the subspace sampling procedure.

classifier performance dependent on the subspace

The features selected for the subspace of the best base classifier are necessary in order to achieve a robust classification. Therefore, weights of features used in this subspace are upscaled by multiplication with the learning rate ν ($\nu > 1$). However, if a feature

upscaling of weights

Algorithm 2 Weight optimization procedure for RSS

Input: m-dimensional data set $\mathbf{X}$, convergence threshold ϵ, learning rate ν, ensemble size L, number of repetitions K

Output: optimized weight distribution $\mathbf{w}$

```
$w_i = 1, i = 1, \dots, m$                                ● *initialize weight distribution* $\mathbf{w}$
$\bar{\gamma}(0) = 0$
$t = 0$
**repeat**
  **for** $k \leftarrow 1, \dots, K$ **do**
    **for** $l \leftarrow 1, \dots, L$ **do**
      extract subspace $\Lambda_l^k$ given $\mathbf{w}$
      train classifier $D_l^k$ on $\Lambda_l^k$
      determine classifier performance $\gamma_l^k$
    **end for**
    determine the best $\gamma_B$ and worst classification performance $\gamma_W$ of $D_1, \dots, D_L$
    **for** $l \leftarrow 1, \dots, L$ **do**
      calculate scaling factor $\tau_l^k$ for $\Lambda_l$ by
      $\tau_l^k = \begin{cases} \frac{1}{\nu} + \left(\frac{2(\gamma_l^k - \gamma_W)}{\gamma_B - \gamma_W}\right)(1 - \frac{1}{\nu}) & \text{if } \frac{\gamma_l^k - \gamma_W}{\gamma_B - \gamma_W} < 0.5 \\ (2 - \nu) + \frac{(2\nu - 2)(\gamma_l^k - \gamma_W)}{\gamma_B - \gamma_W} & \text{otherwise} \end{cases}$
    **end for**
  **end for**
  **for** $l \leftarrow 1, \dots, L$ **do**
    $\tau_l = \sqrt[K]{\prod_{k=1}^{K} \tau_l^k}$                  ● *normalize scaling factors*
  **end for**
  **for** $i \leftarrow 1, \dots, m$ **do**
    $c = 0$
    $s = 1$                                               ● *dimension scaling*
    **for** $l \leftarrow 1, \dots, L$ **do**
      **if** $\Lambda_l \cap \{\lambda_i\} \neq \emptyset$ **then**
        $c = c + 1$
        $s = s * \tau_l$                                  ● *rescale scaling value*
      **end if**
    **end for**
    **if** $c \neq 0$ **then**
      $s = \sqrt[c]{s}$                                   ● *normalize scaling value*
    **end if**
    $w_i = w_i * s$                                       ● *scale variable weight*
  **end for**
  $t = t + 1$
**until** $|\bar{\gamma}(t) - \bar{\gamma}(t-1)| < \epsilon$          ● *check on convergence*
**return** optimized weights $\mathbf{w}$
```

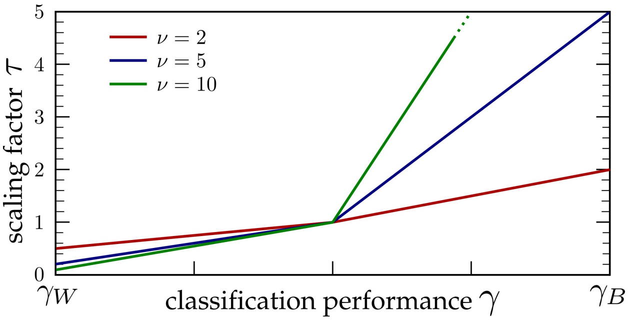


Figure 5.3: Scaling factors τ for different learning rates ν.

contained in the subspace of the best classifier is also used by the worst classifier, no inference of the feature's relevance can be determined. Thus, the weight of this feature is multiplied by $1/\nu$, thereby revoking the previous modification and the feature's weight remains at its value. Furthermore, weights of features used by the classifier with mean classification performance are multiplied by one since their relevance for classification cannot be determined in relation to the features used by the best and worst classifiers.

downscaling of weights

determination of scaling factors

Based on these principles a scaling factor τ_j has to be defined for each subspace. Therefore, scaling factors defined in the previous paragraph are used as fix points and linear interpolation is applied between them, resulting in the following formulation:

$$\tau_j = \begin{cases} \frac{1}{\nu} + \left(\frac{2(\gamma_j - \gamma_W)}{\gamma_B - \gamma_W}\right)\left(1 - \frac{1}{\nu}\right) & \text{if } \frac{\gamma_j - \gamma_W}{\gamma_B - \gamma_W} < 0.5 \\ (2 - \nu) + \frac{(2\nu - 2)(\gamma_j - \gamma_W)}{\gamma_B - \gamma_W} & \text{otherwise} \end{cases} \quad \text{for } j = 1, \ldots, L\,.$$

Generally, alternative scaling functions can also be applied in this stage, but the exact formulation is expected to be of minor importance. The proposed calculation fulfills the previously mentioned requirements and the learning rate can be adjusted by a single parameter. Thus, this function is a reasonable choice for this application.

The main step in the optimization procedure is the modification of the weight distribution. Scaling factors calculated for each of the subspaces are the basis for determination of the weight distribution used in the next iteration. A single dimension can be used in several subspaces and the factor for scaling of the dimension's weight is calculated as the geometric mean of all corresponding scaling factors. For example, if the feature with the index $k = 1, \ldots, n$ is used in the subspaces Λ_2, Λ_5 and Λ_{12}, w_k is changed by

normalization of scaling factors

$$w_k = w_k * \sqrt[3]{\tau_2 * \tau_5 * \tau_{12}}\,.$$

In order to derive $\mathbf{p} = \{p_1, \ldots, p_n\}$, which controls the probability for each dimension to be chosen in the subspace selection procedure, the percentage of each weight from the sum of all weights is determined by

selection probability

$$p_i = \frac{w_i}{\sum\limits_{j=1}^{n} w_j} .$$

With respect to these modified probabilities, a new ensemble is created and the process is repeated until convergence is achieved. In this context, convergence is defined by a change in the mean classification performance $\bar{\gamma}$ of the individual base classifiers below a predefined threshold ϵ. Variables with upscaled weights due to a relatively high classification performance achieved by classifiers using this variable are favored in the next iteration and are again modified in the next iteration.

convergence criterion

Scaling factors determined in each iteration are dependent on the subspace selection. In order to avoid putative statistical artifacts induced by the subspace sampling procedure, multiple ensembles are created using the same weight distribution. Combination of scaling factors from several repetitions is achieved by the calculation of their geometric mean.

creation of multiple ensembles in each iteration

Results of an exemplary weight optimization procedure for the `REAL.NMR` data set, as presented in section 4.1.1, are shown in figure 5.4. SVMs using a linear kernel function are the base classifier algorithm in this approach, allowing for a reasonable classification performance and fast parameter optimization. The mean performance of single classifiers increases in the optimization procedure due to the preference of variables with high weights. As an effect of the improved single classifier performance, the ensemble performance on the cross-validation set also increases. The ensemble diversity, estimated by the entropy measure, decreases during this optimization procedure by about 80 %, however the ensemble performance on the test set is increased. Thus, despite the decreased ensemble diversity, the ensemble performance on the cross-validation and test set is improved by the optimization of the weight distribution used in the ensemble creation. The learning rate ν controls the speed of convergence in this approach. Values close to one will lead to only small changes of variable weights, but increasing the learning rate could speed up the optimization procedure.

improved mean classifiers

improved ensemble performance

As can be seen in figure 5.4, the iterative optimization of weights used in a weighted RSS procedure leads to an improvement of the base classifier performance. As an effect of the improved base classifiers, the ensemble performance on the cross-validation and test set also increases. Thus, it is reasonable to focus in classification on only specific parts of the spectra, which contain the most relevant information for classification purposes.

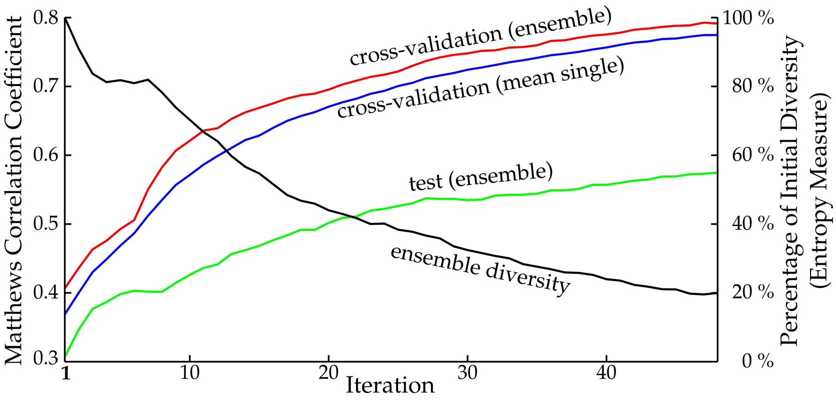


Figure 5.4: Change of diversity, mean single classifier performance and ensemble performance on the cross-validation and test set throughout the proposed weight adaptation procedure.

Thereby, more accurate and robust classification models can be trained and combined to an ensemble system.

5.3 ENSEMBLE OF LOCAL EXPERTS

In the previous section, an approach for guiding an RSS procedure to favor relevant features in the selection of feature subspaces was presented. This restriction on certain spectral regions for classification is the central aspect in the adaptation of ensemble methods for the automatic classification of NMR spectra. Although this has been implicitly achieved by the optimization of RSS weights, features with low weights can still influence the final ensemble. Thus, no clear separation between non-informative and informative regions can be achieved. The focus on the most relevant spectral regions and their automatic determination is presented in this section by the *ensemble of local experts* approach.

ensemble of local experts

The overall structure of the new ensemble approach for classification of NMR spectra is shown in figure 5.5. Contrary to the previously presented approach, a bucketing procedure is applied using a small bucket width of 0.001 ppm. The objective of this bucketing procedure is not the compensation of peak shifts but the equal discretization of all spectra. Thereby, changes in peak intensity of single peaks are noticeable and can be used for classification.

high spectral resolution

general structure of the ensemble system

Initially, multiple spectral regions Ψ_i, $i = 1, \ldots, L$ are selected for further progress, denoted as the *spectral regions of interest* (SROIs). An alignment procedure for the compensation of peak-shift effects is applied to these SROIs and L base classifiers are trained. These classifiers serve as *local experts* with a focused view on the data including signals from a limited amount of molecules. Since these local experts are the basis for the definition of the

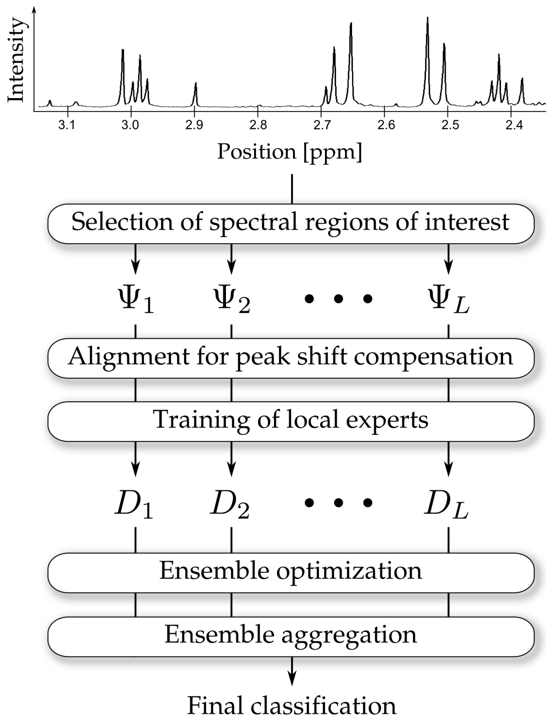

Intensity
3.1
3.0
2.9
2.8
2.7
2.6
2.5
2.4
Position [ppm]
Selection of spectral regions of interest
Ψ_1 Ψ_2 • • • Ψ_L
Alignment for peak shift compensation
Training of local experts
D_1 D_2 • • • D_L
Ensemble optimization
Ensemble aggregation
Final classification

Figure 5.5: Overview of the ensemble of local experts approach.

ensemble system, it is denoted as ensemble of local experts and has first been presented in [Lien 08c]. Finally, an ensemble optimization procedure is applied for the improvement of ensemble classification performance by the subsequent ensemble aggregation method. An optimization can be achieved by selection of a subset of local experts used for the final ensemble classification or application of alternative ensemble aggregation strategies.

stepwise ensemble creation procedure

The proposed procedure mainly consists of four different steps:

1. Selection of SROIs

2. Alignment of SROIs

3. Training of base classifiers

4. Ensemble optimization

application to new samples

While the training of base classifiers can be achieved by supervised learning, the remaining steps require methods suitable for the current application. The optimization of these methods is achieved via cross-validation as presented in section 4.2. Finally, the relevant SROIs, alignment models, parameters used for the training of base classifiers and the final ensemble aggregation method are obtained and can be used for the construction of the final classification system. Therefore, all samples are partitioned into the final SROIs, which have been used in the ensemble, and the alignment procedure is applied. Parameters specific for the particular base classifier are used to train the local experts and the final ensemble combination is achieved. Details on the methods used in the respective steps are presented in the following sections.

5.3.1 *Determination of Spectral Regions of Interest*

The separation of a full spectrum into several segments allows for the training of classifiers using only a limited amount of information (represented by peak intensity changes) for their classification decision. Since usually concentrations of several molecules are necessary to identify a toxic effect, it is unrealistic to assume that the use of a single local expert is sufficient to allow for a perfect classification. But their combination in an ensemble can achieve a higher classification than each individual classifier. Thus, spectral regions used by the local experts should contain information relevant to class discrimination. However, in this step the question arises how these regions can be reliably identified since no information on the relevance of spectral signals for classification is available.

focus of each local expert on a short spectral region

combination of multiple views in an ensemble

Generally, peaks represent the main signals in an NMR spectrum. However, peak overlaps and signals with an intensity similar to the noise intensity complicate approaches for robust peak detection. Furthermore, peak detection is usually applied to each spectrum separately. In order to determine spectral regions independently of the current spectrum, detected peaks have to be matched against each other. This is the basic problem in peak alignment. Thus, the robust detection of SROIs is a complex task and usually no background knowledge of relevant signals can be applied.

definition of SROIs independent of a specific spectrum

Considering the problems in peak detection, an approach for the automated extraction of SROIs independent of peak detection results is developed (cf. figure 5.6). This approach is based on sliding windows, defining SROIs by windows of fixed size and an overlap of 50 % on the whole spectrum. Moving the sliding window by 50 % of the window size to the next position is used for the compensation of problems caused by disadvantageous border positions.

sliding window approach

In order to achieve a local view on the spectra, each SROI should include only a limited amount of signals. However, peaks in the same spectral region are expected to originate from the same chemical group. Combination of these signals can improve classification performance. In order to allow for sliding windows with a very local view and also larger SROIs containing multiple signals, sliding windows are applied in multiple scales. The smallest window size of 0.025 ppm is equal to the width of dominant

windows in different scales

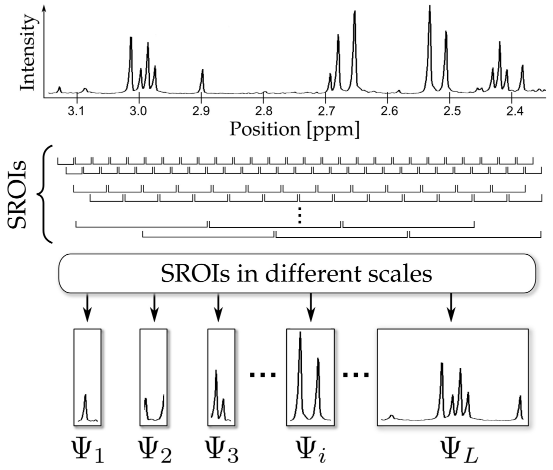


Figure 5.6: Sliding window with overlap approach in multiple scales for separation of a spectrum into multiple SROIs $\Psi_1, \dots, \Psi_L$.

peaks and doubled each time up to a size of 0.4 ppm, allowing for the combination of neighboring peak signals.

Application of this approach on the spectral range from 0.2 to 4.5 ppm and 6 to 10 ppm[2] in five scales using window sizes of 0.025, 0.05, 0.1, 0.2 and 0.4 ppm leads to the definition of 1241 SROIs. These spectral regions form the basis for the further process of peak alignment and base classifier training.

1241 initial SROIs

In fact, no explicit selection of a subset of the full spectral range but a separation of the spectrum into multiple parts is achieved. In this early stage of the classification procedure, no information on the relevance of spectral regions is present or can be derived from the spectral profile. Even the presence of strong signals with significantly changing intensities does not indicate a relevant spectral region. These intensity changes could also be induced by other sources independently of the investigated organ toxicity.

Thus, a collection of SROIs is defined in this step and the restriction on particular spectral regions is achieved by the selection of a subset of all available local experts. A first exclusion of local experts can be achieved by the estimation of their classification performance by a cross-validation procedure. In case of a low classification performance, signals in the respective SROI are not regarded as relevant to the current classification problem and are excluded from further progress.

initial selection of local experts w.r.t. their classification performance

However, this first exclusion of local experts does not respect their combined performance in the final ensemble. Thus, a further ensemble optimization procedure is applied to an explicit selec-

[2] The spectral region from 0.02 to 10 ppm contains the main peak signals and is analyzed in metabonomic approaches. The spectral range from 4.5 to 6 ppm contains the dominant peak signals from water and urea, and is usually excluded.

optimization of the expert selection w.r.t. ensemble performance

tion of local experts with respect to the classification performance achieved by ensemble aggregation methods. Neither the most relevant spectral regions nor the optimal number of local experts is known beforehand. Thus, the final selection of local experts has to be determined in a data-driven way.

SROIs are the basis for the training of local experts, but in order to achieve a robust classification of SROIs peak shifts have to be compensated. Subsequent to the application of an alignment procedure, base classifiers are trained and ensemble optimization procedures are applied.

5.3.2 *Compensation of Peak Shifts*

compensation of peak shifts

Differences in sample properties, such as pH, temperature or ion concentration, induce a slight modification of the peak position. These perturbations have to be compensated in order to allow for a reasonable classification (cf. 2.3.1). Each hydrogen atom is influenced to a different extent by changes in pH, temperature or ion concentration dependent on the electromagnetic environment of the nucleus. Thus, a correction of peak shifts cannot be achieved by assigning a global shift factor moving the whole spectrum to the correct position. However, the separation of the spectrum into several SROIs allows to correct peak shifts present in quite small spectral regions containing only few peaks. Since these short spectral regions will be processed independently from each other in the subsequent process, the problem of aligning the whole spectrum is reduced by the alignment of each SROI.

separate alignment of each SROI

Stoyanova et al. presented in [Stoy 04] an approach to the automatic detection of peak shifts based on a principal component analysis of a spectral data set. If no mean-centering is applied prior to the estimation of the PCA model, peak shifts are indicated in the second PC by a shape similar to the first derivative of a peak signal. This characteristic can either be used for the detection of peak shifts or for the quantization of peak shifts in the current spectrum according to the influence of the distinct model components.

identification of peak shifts

The first two PCs for a simulated data set of an unaligned and aligned peak, and the corresponding normalized eigenvalues are shown in figure 5.7. In the unaligned case, the second PC has the shape of a peak's first derivative and merely 69.2 % of the variance in the data is described by the first PC. This explained variance increases to 99.8 % in the aligned case and the second PC has no derivative shape. Thus, the quality of an alignment procedure can be quantized by the determination of the explained variance of the first PCs. Actually, the distribution of the explained variance among the first PCs has also been applied in the COW alignment procedure as presented by Skov et al. [Skov 06]. In this

high explained variance of first PC indicates aligned peaks

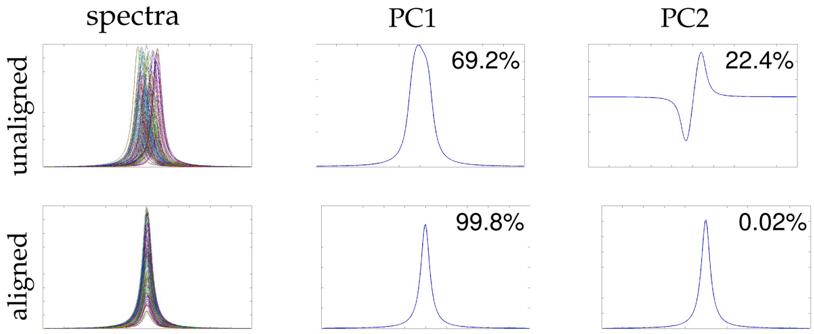


Figure 5.7: First two PCs of an unaligned and aligned simulated data set. The percentage of explained variance by each individual PC according to their eigenvalue is denoted.

context, the alignment quality is determined with respect to the simplicity value (cf. section 2.3.1). Therefore, the maximization of the explained variance of the first PC represents the main objective of the new SROI alignment procedure.

maximization of the first PC's explained variance

The iterative alignment scheme is based on the evaluation of several positive and negative alignment shifts separately applied to each spectrum. The shift leading to the maximum similarity of the current spectrum to the remaining spectra is assigned as the *shift factor* δ_i, $i = 1, \ldots, N$ (cf. figure 5.8). Before the alignment procedure can be applied, the maximum shift δ_{max}, that is expected to occur in a given spectral data set, has to be defined. Given p as the number of shifts to be evaluated at each iteration, discrete evaluation shifts $\delta_{\text{-max}}, \ldots, 0, \ldots, \delta_{\text{max}}$ are defined from the negative maximum shift up to the positive maximum shift in equally sized steps. Increasing p leads to a finer grid for the search of an adequate alignment shift of the current spectrum in each iteration, but also increases the computational complexity of the alignment procedure. The similarity of each shifted version of $\mathbf{x}_i$ to the remaining data set is determined by calculation of the reconstruction error ϵ using the first PC θ_1 of a PCA model estimated on the remaining data.

determination of shift factor δ for each spectrum

reconstruction error ϵ

$$\epsilon_i = \|\mathbf{x_i} - ((\mathbf{x_i}\theta_1^T)\theta_1)\|$$

Peak shifts are modeled by the second PC and if the first PC is sufficient for the description of the current spectrum, peak shift effects are reduced. Thus, maximum similarity is achieved if the reconstruction error becomes minimal. Therefore, the shift value corresponding to the minimal reconstruction error is assigned as shift factor δ_i to the sample $\mathbf{x}_i$. This procedure is applied to each sample in the data set and repeated until the changes in shift factors converge to a value close to zero.

alignment shift: shift minimizing ϵ

This procedure requires the estimation of N PCA models, where N is equal to the amount of samples in the data set. In order to

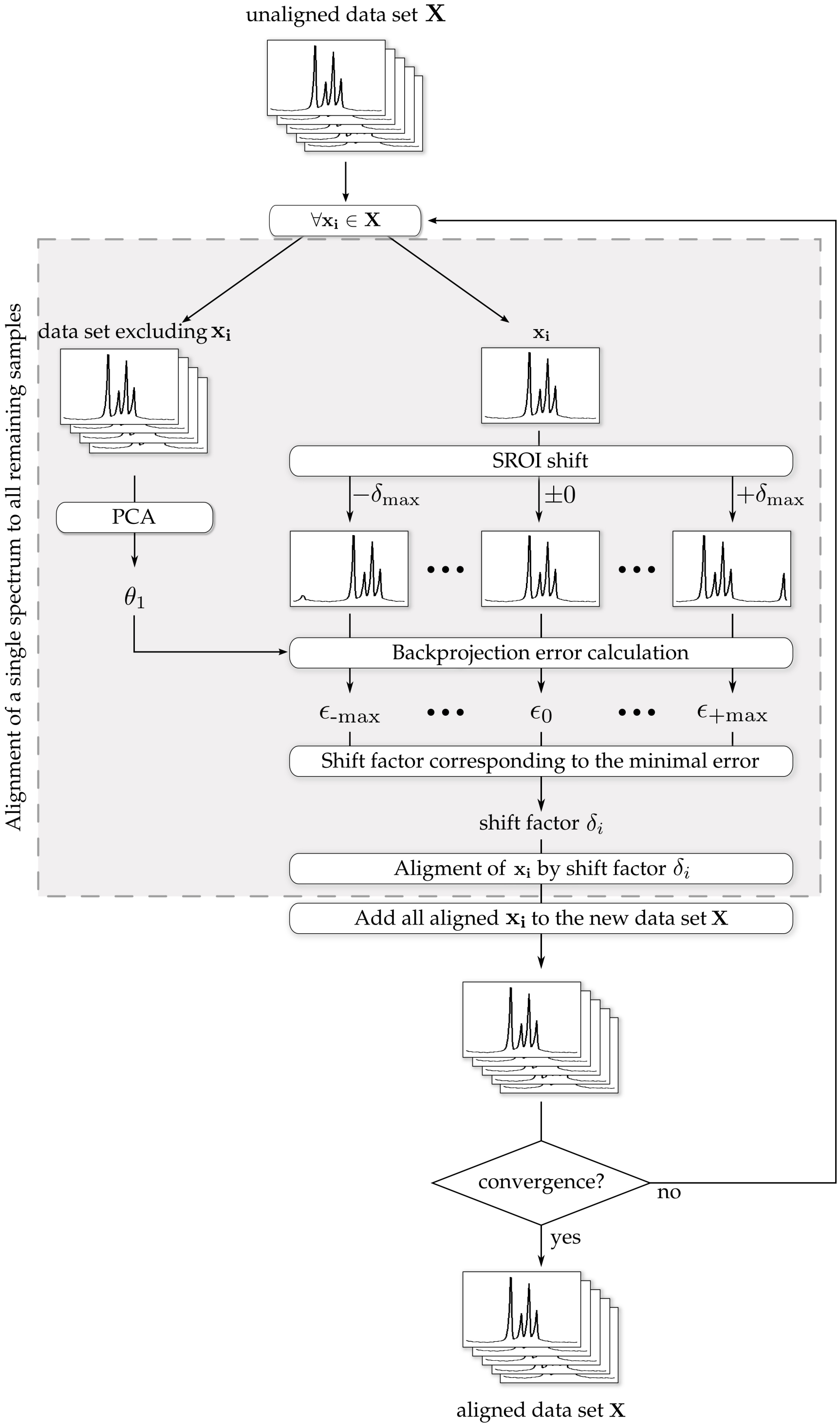

unaligned data set $\mathbf{X}$
$\forall \mathbf{x_i} \in \mathbf{X}$
Alignment of a single spectrum to all remaining samples
data set excluding $\mathbf{x_i}$
$\mathbf{x_i}$
PCA
θ_1
SROI shift
$-\delta_{\text{max}}$
± 0
$+\delta_{\text{max}}$
Backprojection error calculation
$\epsilon_{\text{-max}}$
ϵ_0
$\epsilon_{+\text{max}}$
Shift factor corresponding to the minimal error
shift factor δ_i
Aligment of $\mathbf{x_i}$ by shift factor δ_i
Add all aligned $\mathbf{x_i}$ to the new data set $\mathbf{X}$
convergence?
no
yes
aligned data set $\mathbf{X}$

Figure 5.8: Scheme for the alignment of a data set $\mathbf{X}$. Each sample $\mathbf{x_i}$ is aligned to the remaining data set by means of the first PC θ_1 of a PCA model. The shift factor δ_i leading to the minimal reconstruction error is assigned to the spectrum $\mathbf{x_i}$ and the aligned spectrum is added to the next data set. This procedure is iterated until shift factors converge close to zero.

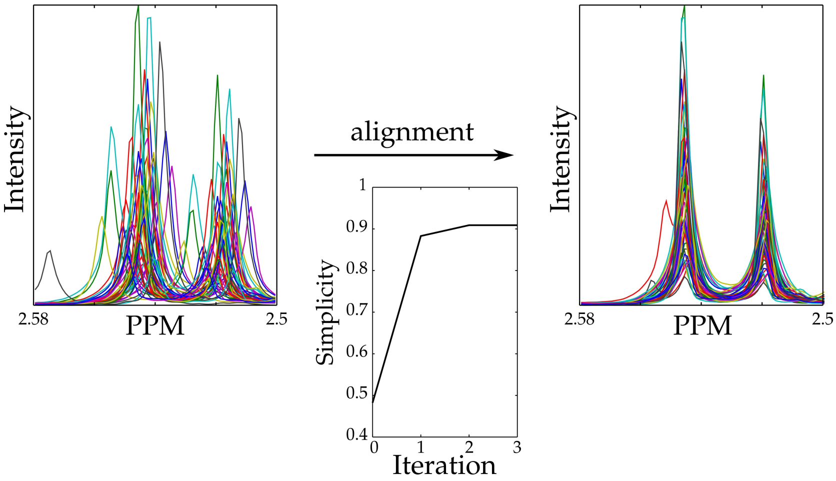


Figure 5.9: Alignment results for one of the citrate doublets and the change of simplicity during the iterative optimization.

reduce the computational complexity of the alignment procedure, blocks of samples can be defined. These blocks are excluded from the whole data set, shifted by the evaluation shifts and the minimal reconstruction error is determined with respect to the PCA model determined on the remaining data set. Randomization of the selection of samples contained in each block after each iteration allows for an alignment independent of a particular partition of samples into these blocks.

alignment of blocks of spectra by a single PCA model

An exemplary alignment result for one of the citrate doublets from 50 randomly selected spectra is shown in figure 5.9. Citrate peaks are present in every spectrum from urine samples and show a high sensitivity to changing physiochemical factors. Thus, these peaks are selected for the demonstration of the alignment achieved by the procedure described in this section. The set of spectra is randomly split into 20 sample subsets and aligned against the remaining ones. The simplicity value of the unaligned data set is equal to 0.48, whereby the values range from zero to one and a well aligned data set has a simplicity of one. An improvement of the alignment quality is achieved in the first iteration and finally converges to a value of 0.91 in the following iterations. The speed of convergence is mainly dependent on the data, but the main improvement is achieved in the first iterations.

increased simplicity value and alignment quality

The proposed alignment procedure has several advantages in the application for alignment of SROIs. Contrary to other approaches, no reference spectrum is required in order to achieve a well-aligned data set. The alignment of new samples to this data set is also straightforward by estimation of a final PCA model for the whole data set and alignment of the new samples by means of the first PC of this model. Furthermore, well aligned spectra can be obtained in a reasonable time independently from

no reference spectrum required

the results of a peak detection algorithm. Thus, this alignment scheme is applied on the selected SROIs by the multi-scale sliding window approach and used for the training of base classifiers by supervised learning.

5.3.3 *Ensemble Optimization and Aggregation*

The alignment of SROIs allows for the training of classifiers with a limited view on the spectra and their integration into an ensemble of local experts. SROIs containing information relevant for discrimination between non-toxic and toxic samples are determined in the ensemble of local experts approach by the selection of a subset of all available experts. The problem of determining an optimal subset of local experts for ensemble classification can be compared to *feature selection* as a common problem in pattern classification tasks. Thereby, given a set of L features, the subset of l $(l < L)$ features leading to the optimal classification performance of a subsequent classification system has to be determined.

ensemble optimization comparable to feature selection

An exhaustive search by the evaluation of all possible feature combinations is usually not practical due to the computational complexity of this procedure. For example, the determination of the optimal subset of 10 features from 100 available ones by an exhaustive search requires the evaluation of

$$\binom{100}{10} = \frac{100!}{10!(100-10)!} = 1.73 * 10^{13}$$

different subsets. Assuming that 100 different combinations could be evaluated within a second, the whole evaluation would last nearly 5 500 years. This demonstrates the necessity of alternative approaches, which achieve good selections in a reasonable time even though these could lead to suboptimal solutions. The determination of a reasonable selection of local experts in a computationally feasible way is denoted in this thesis as *ensemble optimization*.

ensemble optimization

An overview of different ensemble aggregation strategies has been presented in section 3.5, but these consider all available classifiers for determination of the final classification. Methods optimizing the selection of local experts to be used for ensemble aggregation by majority voting, and an aggregation method focusing on generalizability of the aggregation method will be presented in the following.

Sequential Optimization

Marill and Green introduced a feature selection algorithm, denoted as *sequential backward selection (SBS)* [Mari 63], and its bottom up version *sequential forward selection (SFS)* is proposed by

Whitney [Whit 71]. Thereby, a feature selection is iteratively optimized by stepwise exclusion / inclusion of features. Once a feature is excluded / included in the subset, this decision cannot be revoked during optimization, which is denoted as the *nesting effect* [Pudi 94].

iterative inclusion or exclusion of features

nesting effect

In order to overcome the nesting effect, the *sequential backward floating selection (SBFS)* and *sequential forward floating selection (SFFS)* approaches have been presented by Pudil et al. [Pudi 94]. These methods check after each selection of a feature to be excluded / included, whether the inclusion / exclusion of an already evaluated feature improves ensemble quality. Feature selections optimized by floating methods are not expected to be optimal but show reasonable results by preventing the nesting effect. Furthermore, the optimization procedure has a low computational complexity (cf. [Pudi 94]).

application for ensemble optimization

The application of these floating methods for the problem of ensemble optimization can be realized by using predictions from local experts in the ensemble as features to be selected. The quality of a given selection is determined by the classification performance achieved by an ensemble aggregation method (e.g. majority voting).

Before presenting details on selection procedures, some formal definitions have to be introduced. The objective of the ensemble optimization is to determine a selection $\Phi_k = \{\phi_1, \ldots, \phi_k\}$, $\phi_i \in \{1, \ldots, L\}$, $i = 1, \ldots, k$ of k, $1 \le k \le L$, local experts from the set of all classifiers $\tilde{\Phi} = \{D_1, \ldots, D_L\}$. The quality of an expert selection Φ_k is determined according to the classification performance achieved by majority voting, denoted as $J(\Phi_k)$. In accordance with [Pudi 94], the significance $S_{k-1}(\phi_i)$ of a classifier ϕ_i from the selection Φ_k is determined by

significance S_{k-1} of an included classifier

$$S_{k-1}(\phi_i) = J(\Phi_k) - J(\Phi_k \setminus \{\phi_i\}) \,,$$

where $\Phi_k \setminus \{\phi_i\}$ is the set Φ_k excluding the classifier ϕ_i. The significance $S_{k+1}(\hat{\phi}_j)$, $j = 1, \ldots\ L - k$ of a classifier $\hat{\phi}_j$ from the set $\tilde{\Phi} \setminus \Phi_k$ of unselected classifiers is calculated by

significance S_{k+1} of an excluded classifier

$$S_{k+1}(\hat{\phi}_j) = J(\Phi_k \cup \{\hat{\phi}_j\}) - J(\Phi_k) \,.$$

iterative optimization of the expert selection

The SFFS procedure mainly consists of three different steps that are repeated until a maximum number of iterations or convergence to a final selection is achieved. This procedure is initialized with an empty set of classifiers Φ_0 and the first two experts are selected according to the general SFS procedure. After this initialization, the SFFS procedure works as follows (cf. [Pudi 94]):

STEP 1: INCLUSION OF THE BEST LOCAL EXPERT

Determination of the most significant classifier ϕ_b not already included in the ensemble according to

$$\phi_b = \operatorname*{argmax}_{\hat{\phi}_i \in \tilde{\Phi} \backslash \Phi_k} J(\Phi_k \cup \{\hat{\phi}_i\})$$

and inclusion in Φ_k forming the new selection Φ_{k+1}.

$$\Phi_{k+1} = \Phi_k \cup \{\phi_b\}$$

STEP 2: CONDITIONAL EXCLUSION OF THE WORST EXPERT

Determination of the least significant classifier ϕ_w in the set Φ_{k+1} by

$$\phi_w = \operatorname*{argmax}_{\phi_j \in \Phi_{k+1}} J(\Phi_{k+1} \setminus \{\phi_j\})\,.$$

If $\phi_w = \phi_b$ then set $k = k+1$ and proceed with step 1. Otherwise exclude classifier ϕ_w from Φ_{k+1} forming the new set Φ'_k by

$$\Phi'_k = \Phi_{k+1} \setminus \{\phi_w\}\,.$$

If $k > 2$ store the new set of classifiers Φ'_k and the corresponding classification performance $J(\Phi'_k)$ and proceed with step 3, else return to step 1.

STEP 3: CONTINUE CONDITIONAL EXCLUSIONS

Determination of the least significant classifier ϕ_s in the set Φ'_k by

$$\phi_s = \operatorname*{argmax}_{\phi_j \in \Phi'_k} J(\Phi'_k \setminus \{\phi_j\})\,.$$

If $J(\Phi'_k \setminus \{\phi_s\}) \leq J(\Phi_{k-1})$ then set

$$\Phi_k = \Phi'_k \quad \text{and} \quad J(\Phi_k) = J(\Phi'_k)$$

and continue with step 1. Otherwise set

$$\Phi'_{k-1} = \Phi'_k \setminus \{\phi_s\} \quad \text{and} \quad k = k-1\,.$$

If $k = 2$, set

$$\Phi_k = \Phi'_k \quad \text{and} \quad J(\Phi_k) = J(\Phi'_k)$$

and return to step 1. Otherwise repeat step 3.

floating ensemble cardinality

The cardinality of the set of classifiers will *float* up and down throughout this process in order to include the best unselected expert or exclude the least significant expert. After this process has stopped, the optimal selection achieved within SFFS is selected. While the SFFS procedure is a bottom-up procedure starting with an empty set of classifiers, the SBFS procedure works by means of a top-down strategy. The full set of classifiers is used at the beginning and the least significant expert is excluded or a previously

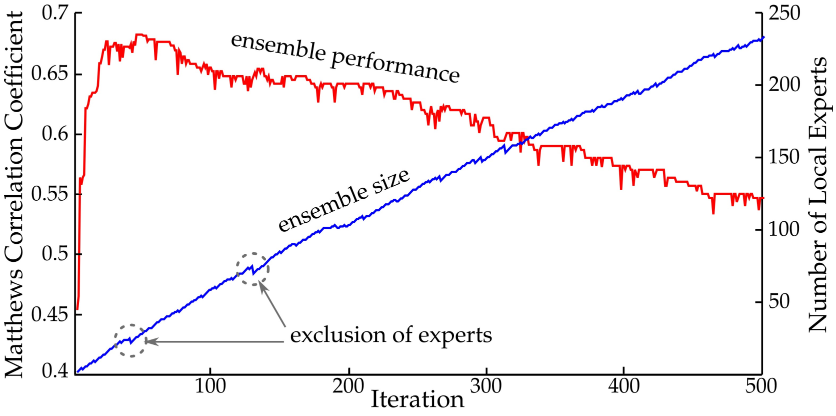


Figure 5.10: Change of diversity and ensemble size in the first 500 iterations of an SFFS optimization on an exemplary set of 554 local experts.

excluded expert is integrated in the ensemble. Due to its similarity to the SFFS procedure, this method will not be presented in this work. The interested reader is referred to [Pudi 94] for a detailed description.

Results of an exemplary ensemble optimization using the SFFS procedure on a set of 554 local experts for the `REAL.NMR` data set (cf. section 4.1.1) is shown in figure 5.10. The ensemble performance increases up to an MC of approximately 0.68 in the first iterations using an ensemble of 25 experts. The inclusion of local experts is continued after this solution has been found, but beyond this point the ensemble performance is decreasing. Previously included experts are removed from the ensemble in several steps during the optimization procedure as shown by the number of local experts included in the ensemble in each step. Thereby, the nesting effect can be avoided, which is a clear advantage of the SFFS procedure in comparison to SFS. A classification performance below a previously defined threshold or a constantly decreasing performance can be used as stopping criterion for this optimization. However, results from the first 500 iterations are shown for demonstration of changes in ensemble performance and size.

improvement of ensemble performance in the first iterations

avoidance of the nesting effect

To sum up, floating methods determine an optimized subset of local experts and improve classification performance of the final ensemble. Since a pool of numerous classifiers is initially created due to the segmentation of the spectrum into multiple SROIs, an efficient optimization procedure is required. Floating methods allow to determine an optimized classifier selection in a reasonable time and are a promising ensemble optimization approach investigated in this thesis for the selection of local experts.

fast and robust ensemble optimization

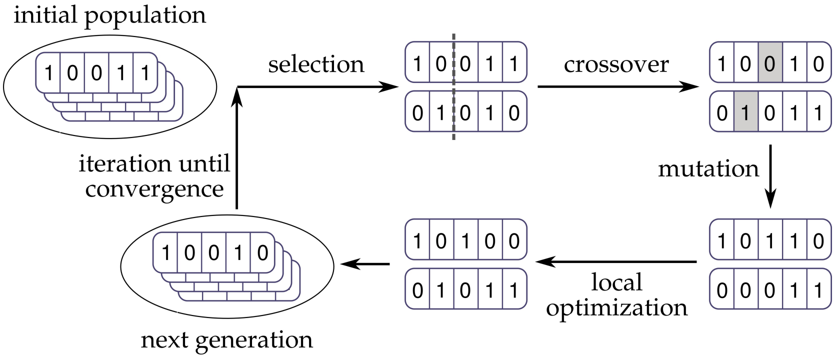


Figure 5.11: Basic structure of hybrid genetic algorithm optimization. Individuals are randomly selected from the initial population, modified by crossover and mutation operations, and optimized by local search operations until the next generation is created. In each iteration a new generation is created until convergence is achieved.

Hybrid Genetic Algorithm Optimization

A comparison of several feature selection algorithms presented by Oh et al. [Oh 04] has shown good results for SFFS but further promising feature selection approaches are genetic algorithms (GAs). These have also been investigated in earlier studies for their application in feature selection and their effectiveness has been proven [Bril 92, Raym 00, Sied 89, Yang 98]. An extension of GAs by local search operations are *hybrid genetic algorithm (HGA)*. These have shown superior results in [Oh 04] for data sets with more than fifty features. The preselection of local experts is expected to result in a set of typically a few hundred classifiers, wherefore HGAs are a promising approach for the determination of an optimized selection of local experts.

extension of genetic algorithm approaches

HGAs are non-sequential approaches evaluating multiple selections at the same time. The distribution of solutions in the parameter space is optimized by interaction of these solutions using evolutionary concepts and local optimization methods (cf. figure 5.11). HGAs have been presented by Oh et al. [Oh 04] as an extension of the GA approach introduced by Goldberg [Gold 89]. By combination of the GA and local search operations, individuals are improved in order to find local optima. Thus, no further iterations are required to search in the region close to the best individuals.

combination of GA principles and local search operations

The general structure of HGAs for expert selection optimization is shown in algorithm 3. In an initial step, expert selections have to be encoded as chromosomes. This is achieved by the encoding of each selection by a string of L binary digits, where L is the number of local experts initially trained. The inclusion of expert D_i, $i = 1, \ldots, L$ is encoded by assignment of “1” to the *gene* at position i

chromosome encoding

Algorithm 3 HGA for local experts selection

Input: ripple factor r, population size P, crossover probability p_c and mutation probability p_m

Output: optimized expert selection Φ

```
initialize population $\mathbf{\Phi} = \{\Phi_1, \dots, \Phi_P\}$
repeat
  for $g \leftarrow 1, \dots, \frac{P}{2}$ do
    select offsprings $\Phi_a$ and $\Phi_b$ w.r.t. their fitness $J$
    one-point crossover of $\Phi_a$ and $\Phi_b$ w.r.t. $p_c$
    mutation of $\Phi_a$ and $\Phi_b$ w.r.t. $p_m$

    for $i \leftarrow 1, \dots, r$ do
      remove the least significant expert from $\Phi_a$ and $\Phi_b$
    end for

    for $j \leftarrow 1, \dots, 2 * r$ do
      add the most significant expert to $\Phi_a$ and $\Phi_b$
    end for

    for $k \leftarrow 1, \dots, r$ do
      remove the least significant expert from $\Phi_a$ and $\Phi_b$
    end for

    add optimized offsprings to the next population $\mathbf{\Phi}'$
  end for
  replace $\mathbf{\Phi}$ with $\mathbf{\Phi}'$
until convergence criterion fulfilled
return best individual of $\mathbf{\Phi}$
```

in this string, and the exclusion by a "0". For example, the string "10001010" encodes the expert selection $\Phi = \{1, 5, 7\}$, including predictions from D_1, D_5 and D_7 for majority voting. The initial population of P individuals is created by assigning randomly a zero or one to the genes of each individual's chromosome.

selection of the initial population

The selection of two individuals for the application of crossover and mutation operations is controlled by the fitness of each individual. This fitness corresponds to the classification performance $J(\Phi_k)$, $k = 1, \dots, P$ achieved by majority voting on the particular selection Φ_k of predictions from local experts. According to this fitness, the probability p_i of the ith individual to be selected for the population of the next generation is determined by fitness-proportionate selection, also denoted as roulette-wheel scheme [Mitc 96], as follows:

fitness of each individual

$$p_i = \frac{J(\Phi_i)}{\sum\limits_{j=1}^{P} J(\Phi_j)} .$$

This principle can be visualized by a circular "roulette-wheel" and each individual has a slice of size proportional to its fitness

roulette-wheel

$J(\Phi_i)$ on this wheel. In order to determine a population of size P, the wheel is spun P times and the individual under the marker of the wheel is selected for the next generation in each spin.

crossover operation

A crossover operation is applied with respect to the crossover probability p_c to two individuals, which are selected according to the roulette-wheel scheme (cf. figure 5.11). The crossover operation is the main procedure to create new individuals in GAs by exchanging all genes from two individuals behind a randomly determined crossing point. Generally, multiple crossover points can be selected, but since these can be achieved by multiple one-point cross-overs, single crossover points are considered in this thesis.

mutation operation

Each gene of the two offsprings is modified with respect to the mutation probability p_m by changing a zero into a one and vice versa. Mutation operations are of minor importance in GAs but can allow for explorations of regions in the parameter space that could not be reached by using only the recombination scheme. Thus, mutations cause a broader evaluation of the parameter space even if the algorithm has converged to a local optimum.

local search operations

After recombination and mutation, the offsprings are further optimized by local search operations in order to find local optima that can be determined by simple inclusion and exclusion of few experts. These local search operations are comparable with the optimization procedure of the SFFS. The grade of optimization is controlled by the ripple factor r.

In the first optimization stage, the r least significant experts are excluded from the current selection Φ. The least significant expert ϕ_e is defined by maximizing the individuals fitness after exclusion.

$$\phi_e = \operatorname*{argmax}_{\phi_j \in \Phi} J(\Phi \setminus \{\phi_j\}) \,.$$

Subsequently, $2r$ experts are added to the ensemble, whereby the expert ϕ_a is significant if its inclusion leads to a maximization of the individual's fitness.

$$\phi_a = \operatorname*{argmax}_{\phi_j \in \Phi} J(\Phi \cup \{\phi_j\}) \,.$$

Finally, r experts with the least significance are again excluded from the ensemble. These inclusion and exclusion operations aim at an improvement of each individual in order to find local optima close to the individual in the solution space.

repetition of selection, recombination, mutation, and optimization

The two offsprings are finally added to the next generation and the selection, recombination, mutation and optimization procedures are repeated until a population size equal to the previous population is achieved. This newly created population replaces the previous one and the process is repeated until a convergence criterion such as a minimal change in fitness or a

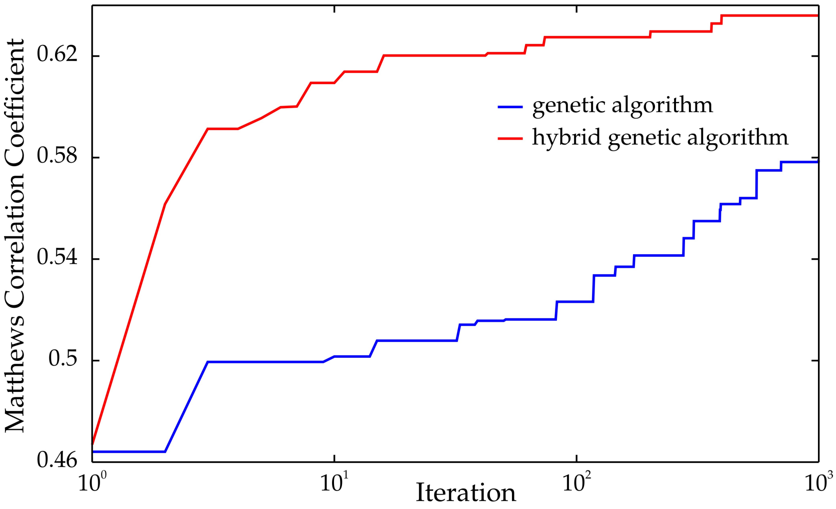


Figure 5.12: Change of the MC of the best individual during GA and HGA optimization. The number of iterations is shown on a logarithmic scale for a better visualization of changes in the first iterations.

maximum number of iterations is fulfilled. In order to avoid the loss of previously found good solutions, the individuals with the best fitness, which are denoted as the *elite*, remain unchanged and are added to the next population.

Although GA and HGA approaches are identical with respect to the genetic operations applied for the exploration of the solution space, both approaches differ in their speed of convergence. While GAs need several iterations for the determination of the local optimum close to an already found solution, this local optimum is directly searched in the HGA approach by local optimization procedures. Thus, the HGA approach requires fewer iterations to find local maxima close to solutions contained in the population and finally achieves better results as shown in figure 5.12.

HGAs require fewer iterations

HGA and SFFS are regarded as the best methods for feature selection. Therefore, the application of these methods for the determination of the optimal subset of local experts for ensemble classifications seems promising. Due to the multi-scale approach for the determination of SROIs, a rather large set of local experts is initially trained. Thus, optimization algorithms such as the HGA are required, allowing for the determination of a reasonable expert selection even in case of a multitude of available experts.

Stacking of Classifiers

SFFS and GA approaches aim at the explicit selection of experts for the final ensemble. A final classification is achieved by majority voting with respect to the predictions of the selected experts. Although both approaches are expected to increase the ensemble

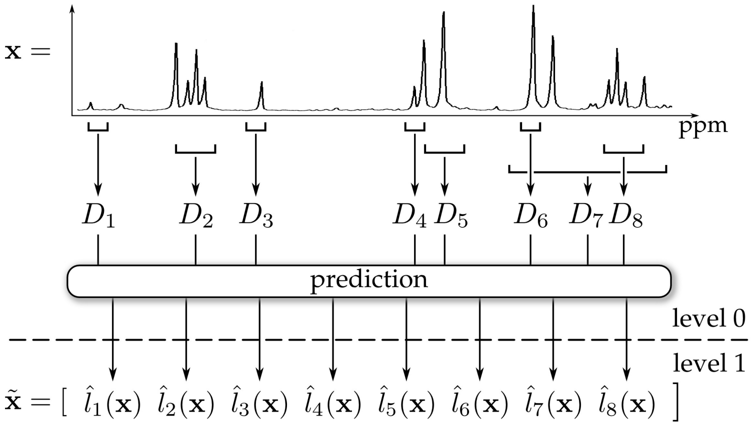

$\mathbf{x} =$
ppm
D_1
D_2
D_3
D_4
D_5
D_6
D_7
D_8
prediction
level 0
level 1
$\tilde{\mathbf{x}} = [\ \hat{l}_1(\mathbf{x})\ \hat{l}_2(\mathbf{x})\ \hat{l}_3(\mathbf{x})\ \hat{l}_4(\mathbf{x})\ \hat{l}_5(\mathbf{x})\ \hat{l}_6(\mathbf{x})\ \hat{l}_7(\mathbf{x})\ \hat{l}_8(\mathbf{x})\ ]$

Figure 5.13: Exemplary creation of level-1 data by prediction of local experts $D_1, \ldots, D_8$ based on a preselected subset of SROIs.

overfitting of ensemble optimization methods

classification performance, optimization is performed on only one data set. Thus, the optimization of the non-trainable majority voting procedure by selection of local experts may lead to an overfitting effect and perform worse in the classification of unseen samples.

An ensemble of local experts aggregation approach presented in [Lien 09] aiming at an improved generalizability is inspired by the stacked generalization procedure described in section 3.5.5. This stacking approach is achieved by using the predictions of local experts as a new data representation as shown in figure 5.13. Thus, each local expert $D_1, \ldots, D_L$ predicts a sample with respect to the specific SROI as either being non-toxic or toxic, and predictions from all local experts serve as features. As stated in section 3.5.5, probabilistic outputs rather than class labels should be used for the creation of the level-1 data representation. Thus, a stacked classification algorithm is trained using this data representation and is used for the classification of new samples.

predictions of classifiers are used as features for a stacked classifier

No further selection of local experts is achieved if outputs from all local experts are used for the training of a stacked classifier. Thus, a further optimization step is required in order to achieve a new data representation, thereby, focusing on features regarded as most relevant to class discrimination. The PLS transformation, as introduced in section 2.4.1, is a projection method in order to achieve a data representation of low dimensionality containing the main information relevant to classification. Thereby, the influence of each feature on the projection is determined based on a labeled data set (cf. section A.2). Thus, the projection of the level-1 data set by PLS transformation is expected to implicitly emphasize relevant features and improve classification performance of the stacked classifier.

focus on relevant features by PLS transformation

To sum up, stacking allows for more complex ensemble aggregations than those achieved by simple majority voting, and an improved performance is expected. The influence of local experts on the data representation used by the combiner is controlled by the application of a PLS transformation. Incorporation of class probabilities rather than class predictions is applied due to the expected improvement of the classification results.

5.3.4 *System Overview*

Three essential steps of the proposed ensemble of local experts approach have been presented in this section, namely

1. determination of SROIs
2. alignment of SROIs
3. ensemble aggregation.

determination of SROIs

The combination of these methods leads to the final ensemble approach as shown in figure 5.14 . In the first step, SROIs are determined by a sliding window approach with overlap in different scales. These short spectral regions restrict the view of the local experts using only a subset of the whole information contained in the spectra.

alignment of SROIs

In order to achieve a reasonable classification of SROIs, peak shifts have to be compensated. This is achieved by an appearance-based alignment procedure, maximizing the similarity of all spectra by an optimization of sample-specific SROI shifts. These aligned SROIs serve as basis for the training of local experts, representing the base classifiers in the proposed ensemble approach.

selection of local experts for ensemble aggregation

Ensemble aggregation by majority voting is a commonly used method, but the explicit selection of local experts used for the final combination of predictions is expected to increase ensemble classification performance. This represents the general assumption, that signals from few molecules are expected to give information for the reliable detection of organ toxicities. A first exclusion of local experts can be realized with respect to their individual classification performance. Local experts with a low classification performance are expected to use spectral regions containing no relevant information for class separation. These local experts could increase the ensemble diversity but will also decrease the mean classification performance of all local experts, which finally leads to a decreased ensemble performance. Ensemble optimization allows for the explicit selection of local experts for final ensemble voting. Alternatively, a stacked classifier can be used for aggregation of predictions from local experts.

stacking approach

The optimization of several parameters for the ensemble of local experts approach is performed by a five-fold cross-validation

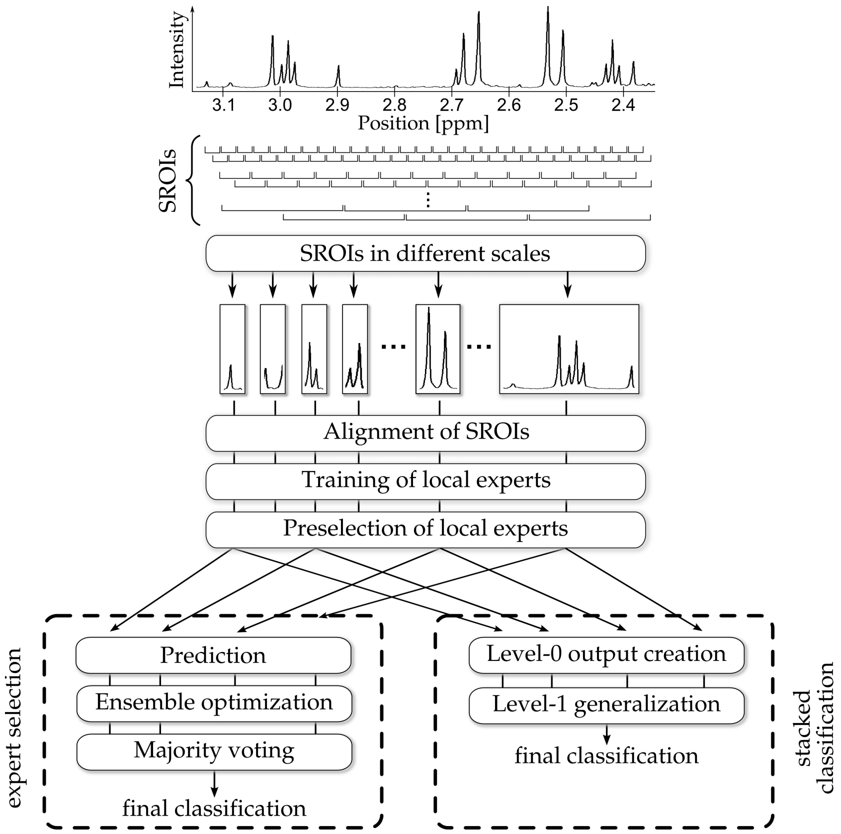


Figure 5.14: Overview of the ensemble of local experts approach. SROIs are initially selected by a multi-scale sliding window approach and are aligned by an appearance-based alignment procedure. A preselected set of local experts trained on aligned SROIs can either be used for expert selection procedures or for a stacked classification approach.

and test procedure. The final classification performance is averaged over the performances achieved on the test set of each fold. However, for the final application in metabonomic studies, a single classification system instead of five systems trained for each cross-validation fold has to be trained. Therefore, the selection of SROIs, alignment models and classifier parameters for each SROI, and the weights for local experts or the stacked classifier parameters are used to build a final ensemble on all available samples.

5.4 INTERPRETABILITY OF ENSEMBLE CLASSIFICATION

relevant spectral regions and confidence in classification

In addition to the classification of new samples as non-toxic or toxic, ensemble approaches allow for further interpretation. The two main objectives are the determination of spectral regions containing relevant information for class discrimination and assessment of confidence in the classification.

Identification of Relevant Spectral Regions

Inference on spectral regions relevant to the distinction between non-toxic and toxic samples is an important prerequisite for the identification of new biomarkers. A fast and reliable detection of organ toxicities can be achieved by using methods from clinical chemistry for the quantization of these biomarkers.

new biomarkers

The identification of spectral regions containing peak signals of biomarker candidates is achieved in the ensemble of local experts approach by the analysis of SROIs used for the training of local experts contained in the final ensemble after optimization. Counting for each point in the spectrum the number of local experts using this point indicates its relevance to classification purposes. Thereby, an implicit weighting of spectral regions is achieved by using overlapping SROIs in different scales. The higher the number of local experts for a spectral region is, the more relevant signals are contained in this region.

SROIs of local experts in final ensemble

This weighting of spectral regions with respect to their relevance for classification purposes is the main objective in the modified RSS procedure and directly optimized by an iterative procedure. Thus, the final weight distribution allows for the discriminant between relevant and non-relevant variables. Spectral regions with the highest weights can be used as a starting point for further investigations.

optimized weight distribution in the ARSS approach

The determination of relevant spectral regions does not automatically allow for the unique identification of molecules corresponding to peaks in these regions. Although databases containing NMR signals for a multitude of molecules are available, not every molecule contained in urine samples from rats is contained in these databases. Furthermore, identification with respect to a single peak position may lead to several possible matches. The amount of possible matches can be reduced by the incorporation of the position of additional peaks induced by the same molecule, but this information is not always available. Thus, the unique and automatic identification of substances contained in the urine samples being relevant to class discrimination is still a challenging task and will not be further investigated in this thesis.

identification of molecules corresponding to peak signals

Assessment of the Degree of Toxicity

Organ toxicities are induced in a rather gradual progress and the binary classification as non-toxic or toxic is not common practice, rather different degrees of toxicity are assigned. Information on this grade of toxicity is useful for the interpretation of a prediction *confidence in the classification decision*

and can be used as a value of confidence in the classification. In case of severe organ damages, several spectral regions will show a spectral profile different from the profile of control samples. Thus, the percentage of base classifiers in the ensemble approaches predicting a sample as toxic can be regarded as an indicator on the grade of changes in the spectral profile and the degree of induced organ toxicity.

dose-dependent effect

The grade of an induced organ toxicity is usually dependent on the dose of the applied compound. The determination of a dose allowing for the desired effect of the pharmaceutical and not inducing severe organ toxicities is an important aspect in safety pharmacology. Assessment of the degree of toxicity with respect to the percentage of base classifiers assigning the label toxic to a sample allows for the observation of dose-dependent reactions. Thereby, changes in the degree of toxicity can be observed by increasing the dose up to the dose inducing a toxic effect.

5.5 SUMMARY

sparse and complex data sets

The analysis of biofluids by NMR spectroscopy leads to the representation of concentration information for several signals in a spectrum. In order to determine particular changes in the metabolism, which indicate an organ toxicity, classification approaches have to be applied, achieving even for these complex and sparse data sets robust classification results.

ensemble by variation of preprocessing and feature extraction methods

Ensemble approaches presented in this chapter are designed for the classification of NMR spectra respecting particular characteristics of this type of data. The ensemble based on the variation of preprocessing and feature extraction methods aims at the inclusion of multiple classifiers trained on differently processed data sets in an ensemble system. This incorporation of different "views" on the data improves classification performance and serves as proof of concept for the applicability of ensemble methods for the detection of organ toxicities.

focus on relevant spectral regions

Although multiple spectral signals are present in an NMR spectrum, only a small fraction is expected to contain information necessary for a robust classification. Thus, ensemble systems are designed to focus on spectral regions relevant to classification, leading to an improved ensemble performance. This is the main principle proposed in this thesis and realized by two different ensemble approaches.

optimized weight distribution for RSS ensemble

The relevance of spectral regions is determined in the ARSS approach by the iterative optimization of a weight distribution used in a weighted RSS procedure. Features with the highest weights are favored in subspace selection, thereby increasing the final ensemble performance. Views on short spectral regions are used in the ensemble of local experts approach for the training of classifiers and their combination in an ensemble system. The optimization of the selection of local experts finally used in the ensemble improves the ensemble performance by an emphasis on the spectral regions relevant to classification. This ensemble not only improves classification performance but also allow for an interpretation of the ensemble decision. The percentage of votes for a classification as toxic is used as indicator on the degree of the toxic effect. Furthermore relevant spectral regions, which are identified by high RSS weights or views of local experts, serve as starting points for the identification of biomarker signals.

ensemble of local experts

interpretability of ensemble approaches

6 EVALUATION

The automatic classification of NMR spectra from urine samples for automatic detection of drug-induced organ toxicities is a rarely investigated field. One of the most advanced approaches is the classification system developed within the COMET project as presented in section 2.4.2. However, neither the software nor the data set used within the COMET project is available for public use. This lack of benchmark data sets and systems is a common problem in the comparison of newly developed classification systems with alternative approaches in the field of Metabonomics.

lack of public data sets and reference systems

Thus, the developed ensemble approaches presented in the previous chapter are compared in an experimental evaluation with approaches from pattern classification and ensemble theory. The bases for these evaluations are the data sets and evaluation methods presented in chapter 4.

experimental evaluation

After the presentation of evaluation results for a selection of reference classification approaches from single and multiple classifier theory, the parametrization of ensemble approaches is optimized in section 6.2. The final ensemble design is chosen with respect to the evaluation results on the validation set of the `REAL.NMR` data set. Classification and interpretation of these optimized ensembles is investigated in section 6.3. In order to validate results obtained on real NMR spectra, the ensemble approaches are applied to a simulated data set with known properties. Results concerning the classification performance and interpretability are presented in section 6.4. Finally, the ensemble approaches are applied for the detection of alternative toxicity types, and this chapter concludes with a summarizing section.

comparison to reference systems

ensemble optimization

final evaluation and interpretation

simulated spectra

alternative toxicity types

6.1 EFFECTIVENESS OF REFERENCE APPROACHES

The first classification results presented in this section are achieved by the evaluation of selected reference classification systems for the detection of drug-induced organ toxicities for the `REAL.NMR` data set. These approaches are from the field of single and multiple classifier systems. The classification performances of these methods serve as basis for the comparison of results achieved by the proposed advanced ensemble methods.

performance of reference classification systems

NN, *k*NN, and SVMs using a linear or RBF kernel have shown reasonable classification performances in previous informal experiments. Thus, these are used as representatives for single classifier approaches. Preprocessing of NMR spectra is optimized

single classifiers: NN, kNN, linear SVM, RBF SVM

METHOD	MC	ACC [%]	SPEC [%]	SENS [%]
NN	0.363	75.8 (±2.8)	86.2	48.2
*k*NN ($k = 15$)	0.374	77.7 (±2.7)	92.7	38.0
Linear SVM	0.449	77.2 (±2.7)	82.0	64.5
RBF SVM	**0.472**	77.8 (±2.8)	81.6	**67.8**
RSS + Linear SVM	0.427	79.1 (±2.7)	92.2	44.5
RSS + RBF SVM	0.424	79.4 (±2.6)	94.3	39.6
Random Forest	0.429	**79.6** (±2.6)	**95.2**	37.9
LogitBoost	0.402	78.1 (±2.7)	90.8	44.5
BagBoost	0.424	78.9 (±2.7)	91.4	45.7
L_2Boost	0.381	77.3 (±2.7)	90.0	43.6

Table 6.1: Classification performances on the validation set of the `REAL.NMR` data set for different approaches from single classifier and ensemble theory. Ensemble methods show a low sensitivity, leading to worse MCs in comparison with single SVM classifications.

optimization of preprocessing and feature extraction methods

by the application of a bucketing procedure using a bucket width of 0.01, 0.02 or 0.04 ppm. The chosen bucket widths either achieve a higher spectral resolution or an improved peak shift compensation. The scaling of spectra for the correction of variations in the sample concentration is achieved either by SNV or integral scaling. Furthermore, a PLS transformation is optionally applied. The dimensionality of the final representation is varied from one to five, and from five to fifty in steps of five. Thereby, classifier-specific parameters, preprocessing procedures and the number of PLS components are optimized for each classifier on the validation set.

ensemble methods: RSS SVM, random forest, LogitBoost, BagBoost, L_2Boost

Random forests, three different boosting approaches, and RSS approaches using either linear or RBF SVMs as base classifiers are applied as representative classification algorithms from the field of multiple classifier systems. The particular boosting variants are chosen due to their proposed ability to achieve reasonable classification results even in case of data sets with few samples of high dimensionality (cf. section 3.6). The data set used for the evaluation of ensemble approaches is preprocessed by integral scaling and a bucketing procedure using a bucket width of 0.01 ppm. In this step a bucket width smaller than the quasi-standard of 0.04 ppm is applied in order to allow for the selection of single relevant signals in the ensemble approaches. Parameters of these ensemble methods are optimized by a grid-search algorithm and the best performing parametrization is finally chosen.

The classification performances of the reference systems on the validation set of the `REAL.NMR` data set are shown in table 6.1. The best values for each of the performance measures are marked as

METHOD	MC	ACC [%]	SPEC [%]	SENS [%]
NN	0.302	73.1 (±2.9)	83.4	45.7
kNN ($k = 15$)	0.283	75.0 (±2.8)	**91.7**	30.6
Linear SVM	**0.364**	73.2 (±2.9)	78.2	**60.0**
RBF SVM	0.311	72.4 (±2.9)	80.0	52.2
RSS + Linear SVM	0.355	**76.6** (±2.8)	90.0	40.8
RSS + RBF SVM	0.292	74.9 (±2.8)	90.3	33.9
Random Forest	0.292	74.6 (±2.9)	89.1	35.9
LogitBoost	0.301	74.3 (±2.9)	87.4	39.6
BagBoost	0.339	75.9 (±2.8)	89.1	40.8
L$_2$Boost	0.311	74.9 (±2.8)	88.3	39.2

Table 6.2: Classification performances on the test set of the `REAL.NMR` data set for different approaches from single classifier and ensemble theory. Random forest and boosting methods show a low sensitivity, leading to worse MCs in comparison with single SVM classification.

bold numbers and the confidence range (level of confidence: 95 %) is denoted for each classification accuracy. SVMs using either a linear or RBF kernel perform best with respect to the MC as main optimization criterion in this thesis. Although the random forest approach achieves the best classification accuracy, the single RBF SVM achieves a higher MC value due to its high sensitivity. This is caused by the imbalanced data set and demonstrates the preference of the MC for optimization rather than the overall classification performance. Generally, a decreased sensitivity of the investigated ensemble systems in comparison with the single SVMs can be observed.

RBF SVM best approach on validation set

Parameters specific for each classifier are optimized with respect to the classification performance on the validation set. Thus, the final performance has to be determined on a test set in accordance to the cross-validation and test procedure described in section 4.2.2. Classification results on the test set presented in table 6.2 demonstrate the low performance of nearest neighbor approaches. Ensemble methods also have on the test set a low sensitivity, resulting in a decreased MC value. Single SVM approaches achieve reasonable classification results on the test set, and the linear SVM shows the best MC value among all evaluated classification approaches. The low generalizability of the RBF SVM is expected to be caused by a complex discrimination function, which is too adapted for the training data. The presence of false positive and negative training samples decreases the generalizability and an improved performance on the test set is achieved by the linear SVM with a rather simple discrimination function.

linear SVM best approach on test set

All ensemble approaches investigated as reference systems perform worse in comparison with the single RBF and linear SVM as the best classifier on the validation and test set, respectively. Therefore, the linear and RBF SVM performances are used in following evaluations as reference for comparison of newly developed ensemble systems.

decreased sensitivity of ensemble methods

The main difference between the ensemble and single classifier approaches is their decreased sensitivity, which leads to a lower MC and a higher overall classification performance. This reduced sensitivity is expected to be caused by the combination of multiple base classifiers achieving a sensitivity below fifty percent. The combination of these base classifiers in an ensemble finally leads to a reduced sensitivity and MC value.

6.2 OPTIMIZATION OF ADVANCED ENSEMBLE METHODS

The general ensemble design for advanced ensemble methods was presented in chapter 5, but details on the optimal base classifier algorithm, ensemble optimization or ensemble aggregation strategy were not specified. These have to be investigated for the particular data set with respect to classification results achieved by a cross-validation procedure. This section presents evaluation results for the optimization of the different ensemble approaches and the final performances on the test set will be shown in the next section.

6.2.1 *Variation of Spectral Preprocessing Methods*

definition of different views on the data by variation of preprocessing methods

SVMs serve as base classifier algorithm

The single classifier results presented in the previous section are based on the optimization of preprocessing and feature extraction methods. Instead of selecting the configuration with the best performance, the ensemble approach presented in section 5.1 combines classifiers trained during this optimization. Each preprocessing and feature extraction combination is expected to model different aspects of the data. Linear or RBF SVM models, which have shown the best performance in the previous evaluation, are trained on differently preprocessed data sets and combined in a multiple classifier system.

The optimization of the set of base classifiers used for the final ensemble was motivated for the ensemble of local experts approach. While this leads to a selection of spectral regions in the ensemble of local experts, this principle is applied in the variation of spectral preprocessing methods for the improvement of the final ensemble performance. Not every spectral preprocessing procedure is expected to achieve a performance better than random guessing, or that their combination leads to an ensemble of

METHOD	MC	ACC [%]	SPEC [%]	SENS [%]
Single Linear SVM	0.449	77.2 (±2.7)	82.0	**64.5**
Ensemble Linear SVMs	**0.504**	**81.2** (±2.6)	**90.6**	56.3
Single RBF SVM	0.472	77.8 (±2.7)	81.6	**67.8**
Ensemble RBF SVMs	**0.585**	**83.8** (±2.4)	**90.0**	67.3

Table 6.3: Comparison of the single best SVM using either a linear or RBF kernel by variation of preprocessing methods and an ensemble classification. The subset of classifiers combined in an ensemble is optimized on the validation set of the REAL.NMR data set.

ensemble optimization by an HGA approach

diverse classifiers. Thus, hybrid genetic algorithm optimization using a population size of 100, $p_c = 0.6$, $p_m = 0.05$ and $r = 3$ (cf. section 5.3.3) is applied for optimization of the classifier selection for the final ensemble. HGAs have shown on this ensemble type the best results in comparison with alternative base classifier selection approaches presented in section 5.3.3 and are used in this evaluation.

improved performance by combination of different views

The ensemble performance is significantly increased in comparison with the single SVM results by the combination of an optimized selection of classifiers as shown in table 6.3. While the ensemble of linear SVMs leads to a reduced sensitivity, the combination of RBF SVMs results in an improved overall performance and only a slight change in sensitivity. Thus, RBF SVMs are used in the following investigations as base classifier algorithm for the preprocessing ensemble.

proof of concept for ensemble methods

This performance outperforms all classification approaches for the REAL.NMR validation set presented up to now. Finally, these results serve as the proof of concept for the general applicability of advanced ensemble approaches for the automatic detection of drug-induced organ toxicities by classification of NMR spectra.

6.2.2 *Adapted Random Subspace Sampling*

Evaluation results presented in the previous section demonstrate improvements in classification performance by the combination of multiple classifiers instead of using only the single best classifier. This approach follows the general ensemble idea of combining multiple diverse classifiers. The adapted random subspace sampling (ARSS) procedure presented in section 5.2 is designed to respect particular characteristics of NMR spectra for ensemble creation. The final ensemble performance is increased by the optimization of a weight distribution used in a weighted RSS procedure, which leads to the exclusion of non-relevant spectral signals.

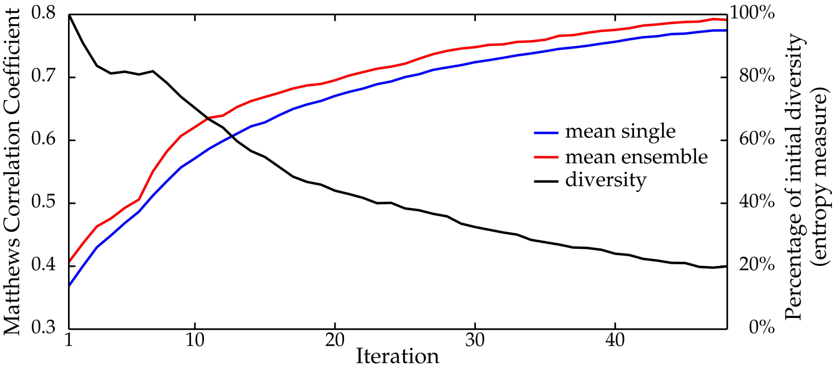


Figure 6.1: Change of diversity, mean single classifier performance and ensemble performance throughout the iterative weight optimization procedure on the validation set of the `REAL.NMR` data set.

linear SVM as base classifier in optimization

SVMs using a linear kernel function serve as base classifiers in the iterative weight optimization procedure due to their reasonable performance shown in section 6.1. Furthermore, only a single parameter has to be optimized, which allows for the determination of an optimized SVM model for each subspace. The ensemble creation is repeated 20 times in each iteration and each ensemble consists of 25 base classifiers. The repetition of the ensemble creation aims at the compensation of artifacts induced by the random ensemble creation procedure. The ensemble size is chosen in order to achieve an optimization in a reasonable time. Each iteration requires the parameter optimization and training of 500 SVMs. Thus, increasing the ensemble size or the number of repetitions would lead to a higher computational effort.

20 RSS SVM ensembles of size 25 in each iteration

increasing base classifier and ensemble performance

The weight optimization leads to an improved mean base classifier performance as shown in figure 6.1. As an effect of the improved base classifiers, the ensemble performance increases and finally converges after 48 iterations. The weight distribution used in the last iteration is used for the construction of the final SVM ensemble. Diversity decreases during the iterative optimization due to the focusing on the most relevant spectral regions. However, this does not lead to a decreased ensemble performance. Thus, the performance of base classifiers has a greater influence on the ensemble performance than the diversity.

Evaluation results of an SVM ensemble using either linear or RBF kernels before and after optimization of the weight distribution are shown in table 6.4. The ensemble size is varied for each configuration from five to 500 in steps of five and the best performing ensemble is finally selected. SVM ensembles using a uniform weight distribution perform worse than the best SVM presented in section 6.1. This can be explained by the low sensitivity of the base classifiers, leading to a low sensitivity of the

BASE CLASSIFIER	WEIGHTS	MC	ACC [%]	SPEC [%]	SENS [%]
Linear SVM	uniform	0.427	79.1 (±2.7)	92.2	44.5
Linear SVM	optimized	**0.790**	**91.9** (±1.8)	**96.5**	**79.6**
RBF SVM	uniform	0.424	79.4 (±2.6)	94.3	39.6
RBF SVM	optimized	**0.790**	**91.9** (±1.8)	**96.3**	**80.0**

Table 6.4: Comparison of RSS SVM ensembles using either uniform or optimized weight distributions for the subspace selection procedure on the validation set of the `REAL.NMR` data set.

final ensemble. However, using the optimized weight distribution for the ensemble creation procedure leads to a significant improvement of the final ensemble performance for both SVM types. The final ensemble shows a very high specificity and sensitivity, which has not been achieved by any of the single classifier or ensemble systems presented in the previous sections.

final ensemble performance is significantly improved

The two investigated types of SVMs show only minor differences by using the same weight distribution. This can be explained by the representational reason for the good performance of ensemble systems (cf. section 3.2.2). Thereby, a non-linear classification is achieved by combination of multiple linear SVMs in an RSS ensemble.

To sum up, classification performance is not improved by the application of the original RSS procedure due to differences in the relevance of variables and the low sensitivity of the base classifiers. By focusing on the most relevant features, the MC value could be significantly increased from 0.424 to 0.790 using RBF SVMs as base classifier. These results further illustrate the possibility ofdesigning ensemble classification systems with respect to particular characteristics of a data set. Furthermore, an important result of these investigations is the different importance of spectral signals for classification purposes.

focus on relevant spectral regions

particular ensemble design for the classification problem

6.2.3 *Ensemble of Local Experts*

Following the proof of concept for the general applicability of ensemble methods for the classification of NMR spectra and improvements achieved by focusing on particular spectral regions, the ensemble of local experts approach is evaluated on the `REAL.NMR` data set. This approach was introduced in section 5.3 and aims at a more explicit selection of spectral regions than the one achieved in the ARSS procedure. Therefore, base classifiers are trained on short spectral regions and combined in an ensemble system.

The initial separation into SROIs is achieved in a multi-scale sliding window approach using windows of size 0.025, 0.05, 0.1,

five scales of sliding windows

0.2 or 0.4 ppm and an overlap of 50 %. The minimal window size is determined according to the peak width of a peak of average intensity in an NMR spectrum. This window size is doubled until windows containing multiple signals are obtained. This separation on spectra defined from 0.2 to 10 ppm excluding the spectral region from 4.5 to 6 ppm leads to the initial definition of 1241 SROIs.

1241 initial SROIs

Peak shifts in these short spectral regions are corrected by means of the appearance-based alignment procedure presented in section 5.3.2. The maximum peak shift of 0.045 ppm used in this alignment approach is the maximum shift observed for the citrate peak in the `REAL.NMR` data set. This shift is expected to be an upper bound for peak shifts occurring in the data set. The alignment quality is determined with respect to the simplicity value as presented in section 2.3.1 and large values indicate well aligned spectra.

peak shift compensation by an alignment procedure

Application of the alignment procedure on the training, validation and test sets created by the five-fold cross-validation and test procedure increases the mean simplicity value from 0.846 to 0.903. Thereby, the alignment quality is increased for each SROI and is expected to lead to more robust local experts. The focus of this thesis is on the development of classification systems with a high classification performance and not the development of new alignment methods or data representations. Thus, the appearance based alignment scheme is applied on each SROI and differences to alternative alignment strategies are not further investigated.

increased simplicity value for each SROI

RBF SVMs are trained on the aligned SROIs and their classification performance on the validation set is used as criterion for an initial exclusion of local experts. Thereby, SROIs not leading to an MC value greater than 0.2 are not used in the next steps, leading to a set of 554 SROIs. The threshold of 0.2 is chosen in order to exclude regions where a classification performance only slightly better than random guessing or even worse is achieved. Nevertheless, several SROIs are still retained, which can be combined in the final ensemble optimization procedure.

preselection of local experts ⇒ 554 SROIs

The sliding window approach is applied in five scales using an overlap of 50 %, wherefore each spectral point can be contained in up to ten different SROIs. Figure 6.2 illustrates the number of SROIs located on each spectral point after the applied preselection procedure. The count of SROIs is flipped over the horizontal axis for visualization purposes. This set of SROIs serves as basis for the determination of the best base classifier, ensemble optimization and ensemble aggregation approach in the following.

Choice of the Base Classifier Algorithm

Generally, different base classifier algorithms can be applied for sample classification according to ensemble theory. However, the

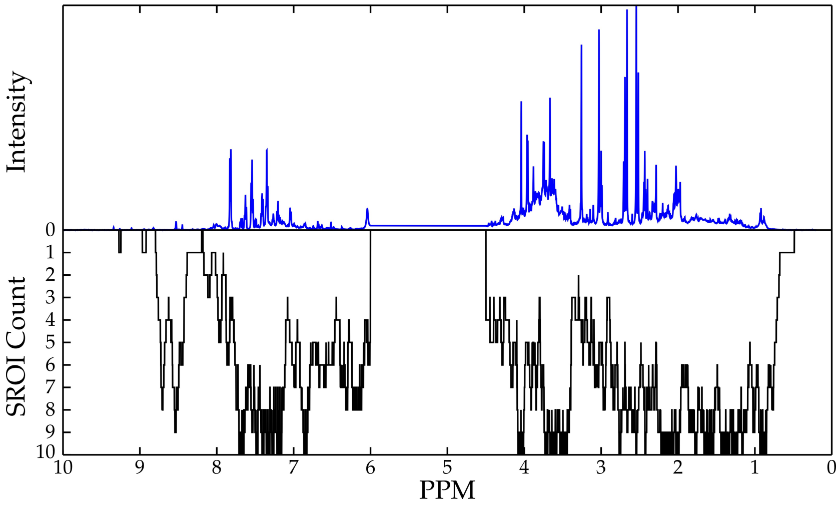


Figure 6.2: Visualization of a control spectrum and the amount of SROIs located on the particular spectral regions. SROIs not leading to an MC value above 0.2 in RBF SVM classification on the validation set of the REAL.NMR data set are excluded. SROI counts are flipped over the horizontal axis for visualization purposes.

BASE CLASSIFIER	MC	ACC [%]	SPEC [%]	SENS [%]
NN	0.525	82.5 (±2.5)	98.2	40.8
*k*NN	0.525	82.5 (±2.5)	98.2	40.8
Random Forest	0.484	81.1 (±2.6)	98.5	35.1
Linear SVM	0.580	84.0 (±2.4)	99.5	42.9
RBF SVM	**0.683**	**87.6** (±2.2)	**99.8**	**55.1**

Table 6.5: Comparison of different base classifier algorithms for the ensemble of local experts. SFFS is used as ensemble optimization method on the validation set of the REAL.NMR data set.

expected amount of local experts leading to robust classifications is rather limited due to the putative low number of relevant spectral signals. Thus, local experts should already allow for a high classification accuracy in order to improve the final ensemble performance.

strong local experts required

In section 6.1, selected classification algorithms are evaluated for the classification of NMR spectra. A comparison of the ensemble performance achieved by their use as base classifiers and optimization of the expert selection by SFFS is shown in table 6.5. NN and *k*NN achieve equal classification results as base classifiers and the performance of random forests is low due to their decreased sensitivity. As in case of the single classifier evaluation, SVMs show the best classification performance and the best results are achieved using the RBF kernel. In this case, the non-linear

RBF SVMs used as local experts

METHOD	MC	ACC [%]	SPEC [%]	SENS [%]
Majority	0.233	74.7 (±2.8)	**100**	7.3
Accuracy Based	0.259	75.1 (±2.8)	**100**	9.0
DT	0.578	83.7 (±2.4)	90.6	65.3
SFS	0.666	87.1 (±2.2)	99.5	53.9
SBS	0.633	86.0 (±2.3)	98.6	52.7
SFFS	**0.683**	**87.6** (±2.2)	99.8	55.1
GA	0.589	84.4 (±2.4)	99.4	44.5
HGA	0.633	85.8 (±2.3)	99.7	49.0
Stacking	0.576	83.3 (±2.4)	88.9	**68.2**
PLS Stacking	0.611	85.2 (±2.3)	92.8	64.9

Table 6.6: Classification performance after application of different ensemble optimization procedures on the validation set of the `REAL.NMR` data set.

classification performance of RBF SVMs cannot be achieved by a combination of multiple linear SVMs since for each SROI only a single expert is trained. Therefore, RBF SVMs serve as base classifier algorithm used in the following evaluations.

Determination of the Final Ensemble Aggregation

optimization of ensemble aggregation

The initial exclusion of local experts with respect to their classification performance on the validation set leads to the focusing on spectral regions, which allow for at least a minimal discrimination between toxic and non-toxic samples. But this modification of the ensemble composition does not respect the final performance of local experts after their aggregation. Therefore, ensemble optimization methods presented in section 5.3.3 are applied, aiming at the determination of an expert selection with a high performance.

low sensitivity of majority and accuracy-based voting

Evaluation results of common ensemble combination methods, optimized ensemble decisions and the stacked classifier approach are shown in table 6.6. Majority voting and accuracy-based weighting achieve the lowest classification performances, mainly due to the very low sensitivity. The initial selection of local experts allows for base classifiers in the ensemble with low sensitivity. Thus, their combination leads to a very low sensitivity and MC value. Even though all classifiers are used in the decision template approach for the determination of DTs, these templates can model classifications with low confidences and improve the final ensemble performance.

Ensemble optimization methods are applied to the predictions achieved on the validation set and all optimized selections perform better than the non-selective ensemble aggregation techniques. Sequential optimization methods show an improved ensemble performance by using the floating variants, whereby the nesting effect is avoided. Also the combination of GAs with local search operations in the HGA approach increases the ensemble performance in comparison with the original GA approach. Although these differences are not statistically relevant, an improved expert selection could be defined. The final ensemble size varies between 15 and 67, which further supports the expectation, that only a small fraction of the originally defined 1241 SROIs are relevant to classification.

SFFS is the best ensemble optimization method

Ensemble combination by a stacked classifier is realized using an RBF SVM on the probabilistic outputs of all local experts. PLS transformation is optionally applied on the probabilistic SVM outputs using three PLS components. The algorithm used for stacked classification, the choice of probabilistic classifier outputs and the number of PLS components is optimized by cross-validation. The classification performance of the stacked classifier is worse than most alternative approaches, but it is increased by using a PLS transformation on the level-0 data before the stacked classifier is applied. Furthermore, the trainable combiner is expected to avoid overfitting effects that can occur by using optimization methods based on majority voting. The final performance of the ensemble approaches on the test set has to be used in order to determine their performance on unknown data.

PLS transformation improves the stacking approach

6.2.4 *Comparison of Ensemble Approaches*

Before presenting results on the interpretability of ensemble systems and their performance on the test set, a summarizing comparison of evaluation results achieved by the different approaches is shown in table 6.7. All ensemble approaches achieve significantly higher classification accuracies than the single classifier approaches. Among the ensemble approaches, the ARSS approach leads to the best classification results. High specificity and even sensitivity rates are achieved by the ARSS procedure. The sensitivity of the local experts approaches is lower than for most of the other classifiers. This can be caused by the low sensitivity of local experts due to the absence of strong biomarker signals allowing for a clear discrimination between non-toxic and toxic signals.

ARSS best approach on the validation set

6.3 effectiveness and interpretability of advanced ensemble approaches

Evaluation results presented in the previous section lead to the final design of the ensemble approaches. This section will compare the evaluation results achieved on the test set, representing

METHOD	MC	ACC [%]	SPEC [%]	SENS [%]
Linear SVM	0.449	77.2 (±2.7)	82.0	64.5
Single RBF SVM	0.472	77.8 (±2.8)	81.6	67.8
Preprocessing Ensemble	0.585	83.8 (±2.4)	90.0	67.3
ARSS ensemble	**0.790**	**91.9** (±1.8)	96.3	**80.0**
Local Experts + Selection	0.683	87.6 (±2.2)	**99.8**	55.1
Local Experts + Stacking	0.611	85.2 (±2.3)	92.8	64.9

Table 6.7: Comparison of the best single classifier approaches and advanced ensemble methods on the validation set of the REAL.NMR data set. All ensemble systems outperform the single classifiers, and the modified RSS procedure achieves the best results.

METHOD	MC	ACC [%]	SPEC [%]	SENS [%]
Single Linear SVM	0.364	73.2 (±2.9)	78.2	60.0
Single RBF SVM	0.318	72.4 (±2.9)	80.0	52.2
Preprocessing ensemble	0.435	77.3 (±2.7)	83.9	60.0
ARSS ensemble	0.549	83.0 (±2.5)	92.5	58.0
Local Experts + Selection	0.420	78.5 (±2.6)	**95.7**	35.9
Local Experts + Stacking	**0.562**	**83.1** (±2.5)	90.6	**63.3**

Table 6.8: Comparison of evaluation results on the test set of the REAL.NMR data set. Generally, performances decrease in comparison with validation results but all ensemble systems still significantly outperform the single classifier approaches.

the real performance of the investigated ensemble approaches. Besides the sample-wise classification results, predictions for applied pharmaceuticals and further interpretation of the ensemble decisions can be achieved. These results are presented in the following for the test set of the REAL.NMR data set.

6.3.1 *Test Classification Performance*

results on the test set reflect the real performance

Classification performances on the validation set are used during the ensemble design for the identification of the best parametrization of base classifiers, the ensemble creation procedure and ensemble aggregation. Thereby, the ensemble performance is influenced and the real performance of the classification system has to be determined on an independent test set. Due to the limited amount of samples in metabonomic applications, a five-fold cross-validation and test procedure (cf. section 4.2.2) is applied.

The final ensemble performance is determined on the test set of the REAL.NMR data set and shown in table 6.8. Ensemble approaches still significantly outperform single classifier approaches. Generally, a decreased classification performance can be observed in comparison with results achieved on the validation set. Each ensemble method is optimized with respect to the particular validation set, which leads to overfitting effects resulting in a decreased performance on the test set.

ensemble approaches achieve the best results

The local experts approach in combination with expert selection by ensemble optimization leads to relatively low classification performances. This overfitting on the validation data is induced by the ensemble optimization procedure. Application of the stacking approach for ensemble aggregation improves the classification performance on the test set. Thus, the generalizability of the ensemble of local experts approach is improved by the stacking procedure. Finally, a better classification performance is achieved than in the ARSS procedure.

stacking improves the generalizability of the ensemble of local experts

Even a classification accuracy of 80 % represents a reasonable classification result for this kind of data. Several non-responders are expected due to the low number of collection times and the single administration of the pharmaceuticals. Class labels are assigned with respect to literature references since the real labels cannot be robustly determined by clinical chemistry due to its low sensitivity. Thus, false positive and also false negative samples are expected to be contained in the data set, thereby, decreasing the classification performance of the whole system. However, the presented ensemble approaches achieve a reasonable classification performance on the test set and significantly outperform single classifier approaches.

6.3.2 *Compound Classification*

The presented classification systems achieve a classification of samples as non-toxic or toxic. However, the objective in safety pharmacology is to detect organ toxicities induced by a particular compound. Therefore, classifications of samples corresponding to the same compound are combined to a final classification for each base classifier by the procedure presented in section 4.2.3. If a toxic reaction is detected at any collection time in more than half of the samples, the compound is classified as toxic. Thereby, non-responders or samples from animals with kidney problems induced by other sources (e.g. stress, genetic factors) can be compensated by averaging over several samples.

combination of sample predictions for each compound

Compound classification results achieved by the best single classifiers and the ensemble approaches are shown in table 6.9. These results indicate a general improvement in classification performance. The best performing classification systems are the ARSS ensemble and the local experts approach in combination with the stacking procedure. Both methods misclassify only two

ARSS and the ensemble of local experts achieve the best results

METHOD	MC	ACC [%]	SPEC [%]	SENS [%]
Single Linear SVM	0.617	80.8 (±10.6)	77.1	**88.2**
Single RBF SVM	0.506	75.0 (±11.5)	71.4	82.4
Preprocessing ensemble	0.578	76.5 (±10.6)	82.9	76.5
ARSS ensemble	**0.779**	**90.4** (±8.2)	94.3	82.4
Local Experts + Selection	0.734	88.5 (±8.8)	**97.1**	70.6
Local Experts + Stacking	**0.779**	**90.4** (±8.2)	94.3	82.4

Table 6.9: Comparison of compound classification results on the test set for single classifiers and ensemble methods. Combination of several samples for prediction of each compound increases the classification performance of all ensemble methods.

non-toxic and three toxic compounds. These results are promising taking the noisy, sparse and complex data sets into account.

Note that the amount of classifications is reduced from 896 to 52 by combination of sample predictions to compound classifications. Thereby, the significance levels are about four times larger than for sample classifications. Furthermore, high changes in MC values can be observed by misclassifying only few compounds.

47 out of 52 compounds are correctly classified

The classification performance of most approaches is reduced on the test set, but the percentage of detected organ toxicities is still reasonable for this kind of data. Each compound has only been applied once and urine samples are collected in the first 24 hours. Thus, sensitivity is expected to be very low. However, 82 % of all compounds labeled as toxic could be identified by the ARSS procedure and the local experts approach in combination with the stacking procedure. Overall, 47 of 52 compounds could be correctly classified, which proves the effectiveness of the new approach combining ensemble methods and metabonomic applications. A perfect classification is not expected since not every organism reacts in the first 24 hours as it should according to literature references. Thus, the presented classification systems achieve very good results for the identification of toxic reactions at the proximal tubule as a possible method applied in the early stages in drug development.

6.3.3 *Assessment of the Grade of Toxicity*

differences in the grade of a toxic effect

The combination of sample predictions to compound classifications allows for a classification decision for each compound. Furthermore, the consensus in the ensemble for a toxic classification can be used as indicator on the degree of the toxic effect. Severe organ damages in the proximal tubule are expected to lead to significant changes in the urine composition, thereby inducing several modifications of peak intensities. The more spectral

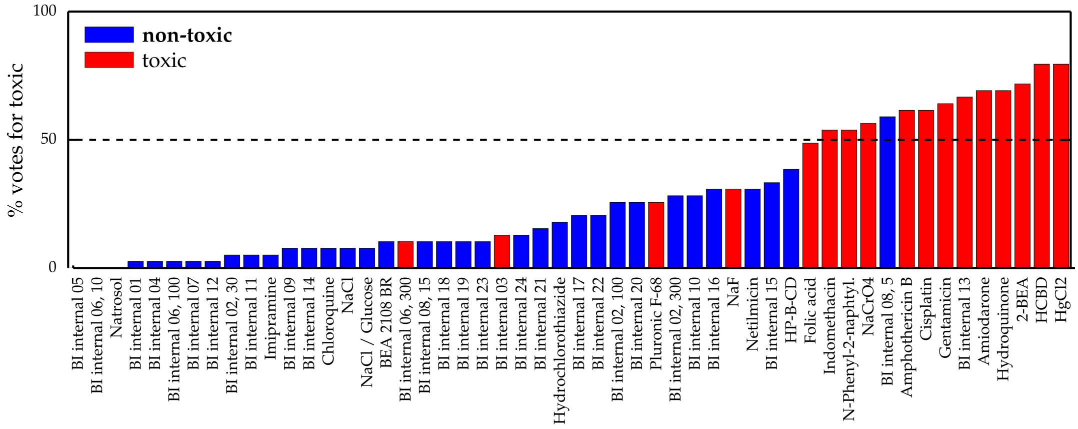


Figure 6.3: Percentage of votes for a toxic classification of each compound from an optimized ensemble of local experts for the test set of the REAL.NMR data set.

ensemble consensus as indicator

changes occur, the higher the percentage of votes as toxic in the ensemble approach should be, since even classifiers with a low sensitivity are expected to detect them.

This information on the percentage of base classifiers voting for a classification as toxic after combination to compound classifications can be visualized as exemplary shown in figure 6.3 for the ensemble of local experts using an optimized selection of base classifiers. According to this plot, several compounds can be clearly classified as non-toxic or toxic. Few samples are close to the threshold of 50 % votes for the toxic class, representing the discrimination criterion between the non-toxic and toxic class. Compounds close to this threshold are expected to be neither a clear non-toxic nor toxic compound.

grade of toxicity of two exemplary compounds

The correspondence of the expected toxic effect and the determined degree will not be discussed for each compound, but two prominent representatives will be mentioned. Most applied compounds have at least some effect on the organism, although not every one is known to induce a toxic effect at the proximal tubule. Natrosol® is the only compound, which is expected to be harmless to the organism. As can be seen in figure 6.3, none of the local experts classifies Natrosol® as toxic. In contrast, mercury chloride ($HgCl_2$) is a strong nephrotoxin, inducing severe damages in the kidney. Approximately 80 % of all local experts from the final ensemble assign the label toxic to the compound, which indicates a strong physiological reaction reflected in the spectral profile.

dose-dependent change of the ensemble consensus

Three samples are applied in different doses and a change of the toxic effect can be observed in figure 6.3. The ensemble agreement is shown for each dose of the three compounds in table 6.10. An increase of the expected degree of toxicity can

COMPOUND	DOSE	% VOTES FOR TOXIC	CLASS
BI internal 02	30 mg / kg	5.1	non-toxic
BI internal 02	100 mg / kg	25.6	non-toxic
BI internal 02	300 mg / kg	28.2	non-toxic
BI internal 06	10 mg / kg	0.0	non-toxic
BI internal 06	100 mg / kg	2.6	non-toxic
BI internal 06	300 mg / kg	10.3	toxic
BI internal 08	5 mg / kg	59.0	non-toxic
BI internal 08	15 mg / kg	7.7	non-toxic

Table 6.10: Percentage of votes for the toxic class from an optimized selection of local experts for compounds applied in multiple doses.

be observed for compounds *BI internal 02* and *BI internal 06* by comparing low and high dose classifications. The low dose of *BI internal 08* has a higher value in this experiments than the high dose. These low dose samples have led to peculiar results in previous classification experiments. Thus, it is expected that some other sources (e.g. stress) influenced the urine composition of these samples.

The same approach can be applied for the visualization of the ensemble decision for each compound achieved by the preprocessing ensemble or the ARSS ensemble. However, the majority of base classifiers in these approaches achieve the same classification due to their low diversity. Thus, the interpretation of ensemble decisions achieved in these approaches for the assessment of the grade of an induced organ toxicity leads to worse results than those from the ensemble of local experts approach. Furthermore, the probabilistic outputs of a stacked SVM can be used for this visualization. However, for the illustration of the general principle, investigations in this context are shown in an exemplary way for the ensemble of local experts approach.

6.3.4 *Determination of Relevant Spectral Regions*

The determination of the degree of an organ toxicity is of interest for compounds contained in the current data set. In addition to this information, the determination of spectral regions relevant for robust classification can allow for an improved classification in future investigations. The identification of relevant spectral signals is the starting point for the definition of biomarker molecules, which can be used by alternative measurement methods for the detection of organ toxicities. Although the identification of spectral regions is necessary for a robust classification, it does not automatically lead to the underlying biomarker molecules. How-

identification of biomarker signals

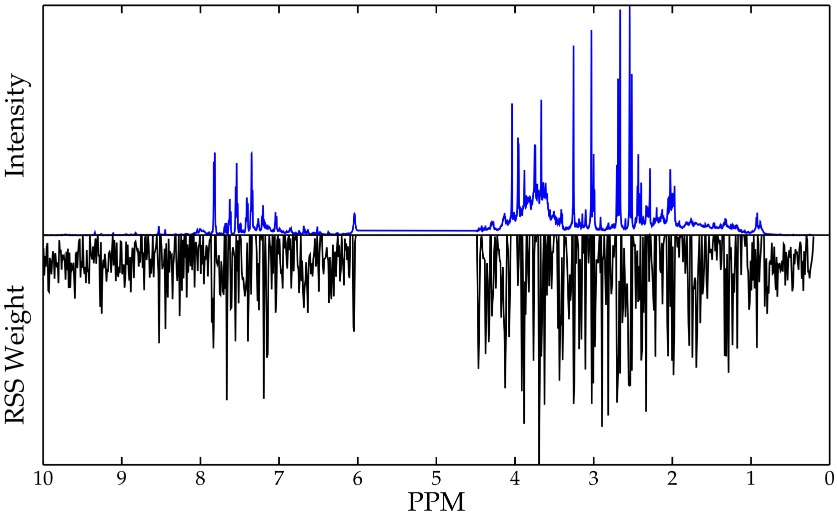


Figure 6.4: Weight distribution determined in the ARSS procedure and an exemplary non-toxic spectrum. The plot of the RSS weights is flipped over the horizontal axis for visualization purposes.

ever, it is an important information to start the search for relevant peak signals.

The ensemble approach based on variation of preprocessing methods does not allow for the unique identification of spectral regions mainly influencing the ensemble decision. PLS models can be interpreted with respect to the weights assigned to each variable for feature transformation. However, the problem of combining information from multiple experts remains, whereby some experts do not use a PLS transformation for feature extraction.

Inference on relevant spectral regions can directly be achieved in the ARSS approach with respect to the weight distribution determined in the iterative optimization procedure. High weights are assigned in this optimization to the most relevant variables. The final weight distribution and a typical non-toxic spectrum are shown in figure 6.4. Several spectral points with high weights can be noticed, which are mainly in regions of peak signals. In contrast, regions with nearly no noticeable signals have low weights.

high weights in ARSS indicate relevant signals

The small bucket-width of 0.01 ppm used for preprocessing the spectra improves the spectral resolution. Thereby, correspondences between single weights and peak signals can be achieved. However, there is no obvious separation between non-relevant and relevant spectral regions, but can be achieved by the definition of a threshold. An alternative interpretation is the ranking of signals. Thereby, identification of underlying molecules starts using the signal with the highest weights and proceeds in descending order.

threshold for detection of relevant signals needed

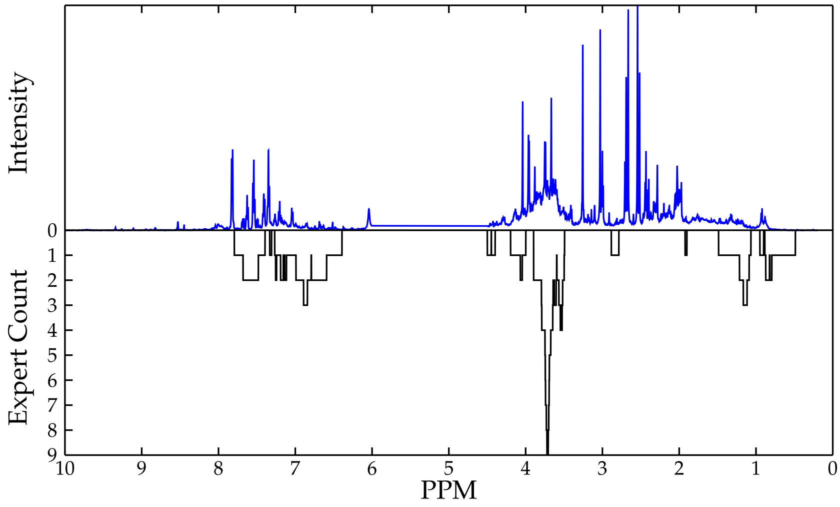


Figure 6.5: Visualization of the amount of base classifiers using each spectral point in the optimized ensemble of local experts. The most relevant spectral regions are expected to be used by the most local experts.

final selection of local experts defines relevant spectral regions

The ARSS procedure requires the definition of a threshold for discrimination between non-relevant and relevant spectral regions. In contrast, the ensemble of local experts approach contains this information in the spectral regions used by local experts. The ensemble composition is optimized in order to maximize classification performance, thereby, the focus on particular peak signal is achieved by inclusion or exclusion of local experts. The amount of local experts from the final ensemble using each particular spectral point is shown in figure 6.5.

Up to ten different local experts can be localized on the same spectral point due to the multi-scale sliding window approach with overlap for initial SROI determination. One spectral signal at approx. 3.7 ppm, is used by nearly all local experts trained on this point, which indicates the relevance of this signal for class discrimination. Comparing this result to the weights determined in the ARSS procedure shows that the same spectral point is also identified as the most relevant signal. The remaining regions used in the local experts procedure are also located on variables with high weights, but these variables are not that clearly identifiable as in the ARSS approach due to the larger window sizes. However, even larger windows focus the view of local experts on the regions containing the most relevant signals for class discrimination. Based on these results, further investigations for identification of underlying molecules can be performed.

6.3.5 *Summary*

Classification and interpretation results presented in this section demonstrate the effectiveness of the developed ensemble methods for the detection of an organ toxicity at the proximal tubule by classification of NMR spectra from urine samples. The main characteristic of these ensemble methods is the focusing on particular spectral regions, which are regarded as relevant for classification purposes. This relevance is determined in an automated approach, finally leading to a significant improvement of classification performance with respect to other investigated classification approaches. In particular, the generalizability of the ensemble of local experts is improved by the application of a stacked classifier, finally leading to the best classification performance on the test set.

general idea: focus on automatically determined relevant spectral regions

Classifications for all samples corresponding to a particular compound are combined in order to derive a final compound classification. Thereby, classification performance is increased by compensation of non-responders or samples not showing the expected toxic reaction due to resistance of the organism or a late onset of the toxic effect. The consensus of classifiers is used as indicator on the induced toxic effect, whereby a high percentage of votes for classification as toxic is regarded as a strong toxic effect.

grade of toxic effect for each compound

Further interpretation capabilities are presented for the ensemble of local experts and the ARSS procedure by deriving the spectral regions mainly relevant for classification purposes. The optimized weight distribution or the spectral regions used by local experts identify spectral points or regions as most relevant. These can be used as starting point for the identification of molecules corresponding to the relevant signals.

determination of relevant spectral regions

6.4 CLASSIFICATION AND INTERPRETATION OF THE SIM.NMR DATA SET

The REAL.NMR data set serves as basis for evaluation and interpretation of newly developed ensemble approaches. The presence of signals in this data set, allowing for the discrimination between non-toxic and toxic compounds, is derived from evaluation results. However, their exact spectral position and individual significance for classification purposes is not known. Thus, a simulated data set SIM.NMR, as described in section 4.1.2, is used for further evaluation and interpretation of the different ensemble approaches.

biomarker signals not known for real spectra

6.4.1 *Evaluation Results*

Classification results on the SIM.NMR data set are not intended to represent the performance that can be achieved in real metabonomic applications, but they allow for a comparison of the presented classification approaches. The classification results on the REAL.NMR data set are dependent on the design of experiments, the amount of samples, and the quality of the spectra. The relation of changes in spectral profiles to the health status of the particular organ is probably the most important aspect. This dependence is modeled in the SIM.NMR data set for five spectral signals, but none of them allows for a perfect discrimination between non-toxic and toxic samples.

simulation of five biomarker signals

Evaluation of the approaches investigated in the previous section is performed by a five-fold cross-validation and test procedure. Due to equally sized non-toxic and toxic classes in the SIM.NMR data set, each fold contains an equal amount of non-toxic and toxic samples. Parameter optimization regarding the best single SVM or ensemble configurations is performed as described in the previous sections. Note that the stacking procedure achieves the best results for this data set without PLS transformation. This is caused by the reasonable performance of all experts determined after the initial experts selection procedure.

The sample classification performances on the validation set and test set shown in table 6.11 support the results obtained for the REAL.NMR data set. All ensemble approaches outperform the single SVMs on the validation set due to their possibility of adapting to the current data set. Furthermore, these approaches are able to focus on particular spectral regions and exclude regions with random intensity changes. The best performance on the validation set is achieved by the ARSS procedure, followed by the local expert approaches.

ensemble approaches achieve the best results

The classification performance on the test set indicates a low generalizability of the preprocessing ensemble. Also the local experts approach using the majority voting technique on an optimized selection of base classifiers leads to a reduced performance on the test set. The remaining approaches show a minor decreasing classification performance and the ARSS approach is still the best performing classification procedure. Compound classifications are not shown for this data set, since a perfect classification is achieved for all classification approaches on the validation and test set.

high generalizability of the ARSS and ensemble of local experts approach

Sample classification results support the expectation that no perfect classification of the simulated data set can be achieved. Thus, it is a reasonable data set for the evaluation of the ensemble approaches developed in this thesis. Generally, ensemble

METHOD	MC	ACC [%]	SPEC [%]	SENS [%]
VALIDATION SET				
Single Linear SVM	0.777	88.9 (±2.0)	88.9	88.6
Single RBF SVM	0.777	88.9 (±2.0)	88.1	89.6
Preprocessing Ensemble	0.840	92.0 (±1.7)	91.5	92.5
ARSS Ensemble	**0.952**	**97.6** (±1.0)	**97.7**	97.5
Local Experts + Selection	0.933	96.7 (±1.1)	96.9	96.5
Local Experts + Stacking	0.924	96.1 (±1.2)	94.2	**98.1**
TEST SET				
Single Linear SVM	0.735	86.8 (±2.1)	86.7	86.9
Single RBF SVM	0.746	87.3 (±2.1)	86.9	87.7
Preprocessing Ensemble	0.607	80.2 (±2.5)	75.6	84.8
ARSS Ensemble	**0.927**	**96.4** (±1.2)	**96.5**	96.3
Local Experts + Selection	0.871	93.5 (±1.6)	94.8	92.3
Local Experts + Stacking	0.908	95.4 (±1.3)	93.9	**96.9**

Table 6.11: Comparison of sample classifications on the validation and test set of the SIM.NMR data set. Ensemble approaches outperform the single classifier approaches, but the preprocessing ensemble and the optimized local expert selection indicate a decreased generalizability.

approaches show an improved classification performance in comparison with the single classifier approaches.

6.4.2 *Quality of Biomarker Identification*

evaluation of interpretability

The SIM.NMR data set can be used for the comparison of different classification approaches, but the main advantage is the known position of biomarker peaks. Thus, the interpretability of the ARSS procedure and the ensemble of local experts approach is compared with respect to their capability to identify the five biomarker positions described in table 4.3.

biomarker signals have high weights in ARSS

A comparison of the weight distribution achieved in the ARSS procedure is shown in figure 6.6 together with an exemplary simulated control spectrum. The positions of the biomarker signals are marked by red crosses. Eight spectral positions can be identified with a weight higher than half of the maximum weight at position 4.5 ppm, whereby the four highest weights are located on biomarker peaks. The last biomarker at the spectral position of 1.0 ppm is weighted with the sixth largest weight. This can be explained by an additional peak at the same position with randomly changing intensity. Thus, the last biomarker is supposed to be the worst of the five biomarkers, which is reflected by the weight assigned in the ARSS approach.

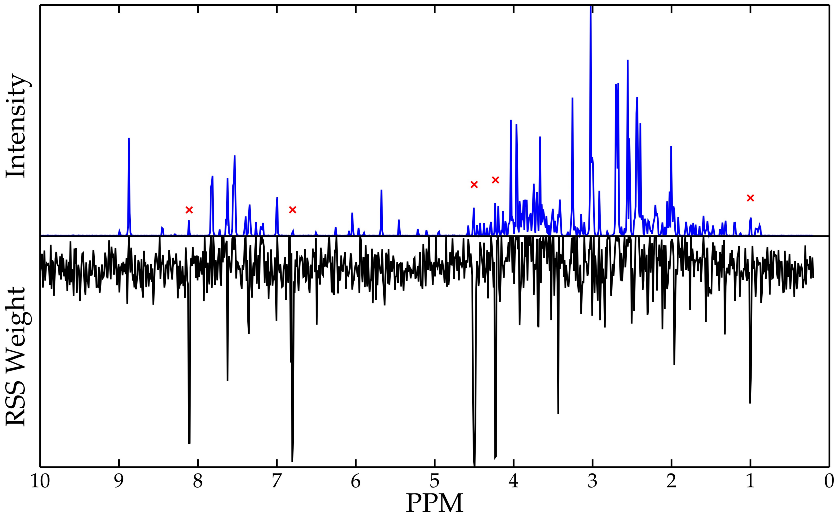


Figure 6.6: Comparison of the optimized weight distribution for the SIM.NMR data set and known biomarker peaks marked by crosses in an exemplary simulated spectrum.

all biomarker signals are used by local experts

Identification of relevant spectral regions can be achieved in the ensemble of local experts approach by counting for each spectral position how often it is used by local experts in the optimized ensemble. According to the results shown in figure 6.7, four spectral regions are of major relevance to classification and two further regions are used by one or two local experts. Nearly all local experts in the final ensemble are located on spectral regions containing simulated biomarker peaks. However, only their combination leads to an MC value of 0.871 on the test set, whereby the best single local expert merely achieves a value of 0.569. The biomarker at position 1.0 ppm is again regarded as a peak of minor importance according to the number of local experts since it is overlapping with a peak of randomly changing intensity.

differences in the interpretability of both approaches

This example demonstrates the two major differences in the identification of relevant spectral regions based on the analysis of the ensemble decisions. The interpretation of ARSS weights leads to well located spectral points which are expected to be relevant for classification purposes. However, the definition of possible biomarker patterns can only be achieved with respect to a threshold, that has to be defined for each experiment. Spectral regions used in the final ensemble of local experts are expected to be relevant, since otherwise their exclusion would have led to an improved classification performance. Although the view of local experts is focused on the relevant spectral regions, the identification of single relevant spectral signals cannot be achieved as in the ARSS approach.

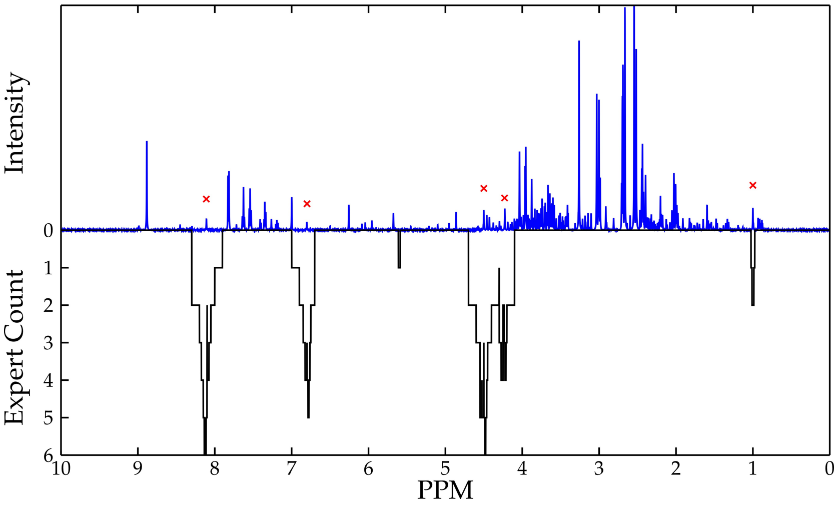


Figure 6.7: Count of local experts using each spectral point in the optimized ensemble for the SIM.NMR data set. Biomarker peaks are marked as crosses in an exemplary simulated spectrum.

Thus, each of the two approaches has advantages and disadvantages. However, both ensemble systems allow for a reasonable classification and their results can be combined for interpretation. Thereby, relevant spectral regions identified in the ensemble of local experts procedure are used to determine variables with high weights assigned by the ARSS in these regions. This would result in the exact positions of the biomarker peaks in the SIM.NMR data set and no additional signals would be detected. Thus, following the general ensemble idea, the drawbacks of each approach for identification of biomarker signals can be compensated by the combination of their results.

combination of identified relevant spectral signals

6.5 detection of alternative toxicity types

The proximal tubule is used in this thesis as the main organ region for the detection of toxic adverse effects. This region of the kidney is among other regions responsible for the regulation of the urine composition. Thus, cell damage in this region is expected to result in a changed concentration of particular molecules in the urine.

Compounds applied in this study do not only induce toxic effects at the proximal tubule, but also alternative toxicity types can be observed. Alternative toxicity types can either be characterized by cell damages in other organ regions than the proximal tubule or are not specific for a certain organ. Three alternative toxicity types investigated in this section are the renal papillary necrosis, acute liver toxicity and phospholipidosis. Note, that compounds do not always induce a single type of toxicity, but dif-

differences in the toxicity type or location of the toxic effect ferent combinations of induced organ toxicities can be observed for different compounds.

A subset of the spectra used in the previous evaluations can be classified as non-toxic or toxic for the three alternative toxicity types according to literature references. Furthermore, a few spectra, which are not used in the previous investigations due to the absence of literature references for the proximal tubule, are used for the particular toxicity type if literature references are available. The amount of samples and class sizes of the data sets for the alternative toxicity types are given in table 6.12.

TOXICITY TYPE	SAMPLES	NON-TOXIC	TOXIC
Renal Papillary Necrosis	583	481	102
Acute Liver Toxicity	500	392	108
Phospholipidosis	733	403	330

Table 6.12: Data set and class sizes for alternative types of organ toxicities investigated in this thesis.

In general, cell damages lead to changes in the cell metabolism and an increased permeability of the cell membrane. The effect on the urine composition is dependent on the intensity of the induced damage and the specific organ. In the following, details on the particular metabolic changes and evaluation results for the detection of three alternative toxicity types based on NMR spectra from urine samples will be presented.

Renal Papillary Necrosis

transportation of urine

The renal papilla is a part of the kidney responsible for the transportation of the urine in the collecting duct after passing the nephron (cf. figure 2.1). Some pharmaceutical compounds induce toxic reactions at this particular part of the kidney, leading to cell damages. Thereby, the cell content may leak in the urine and other substances can pass the renal papilla due to the damaged cells.

reasonable classification results but low sensitivity

The fraction of samples, which are expected to show a toxic reaction at the renal papilla according to literature references, is rather small. Thus the sensitivity of the classification systems is the main problem as shown in table 6.13. However, ensemble methods achieve a reasonable MC value and the amount of detected organ toxicities is quite high with respect to the low number of toxic samples. Thus, a relation between the health status of the renal papilla, and the urine composition is expected.

METHOD	MC	ACC [%]	SPEC [%]	SENS [%]
		SAMPLE CLASSIFICATION		
Single Linear SVM	0.364	85.2 (±2.9)	**100**	15.7
Single RBF SVM	0.194	78.6 (±3.3)	89.2	28.4
Preprocessing Ensemble	0.116	80.4 (±3.2)	94.8	12.7
ARSS Ensemble	**0.557**	88.0 (±2.6)	94.6	**56.9**
Local Experts + Selection	0.435	86.5 (±2.8)	99.4	25.5
Local Experts + Stacking	**0.557**	**88.3** (±2.6)	95.8	52.9
		COMPOUND CLASSIFICATION		
Single Linear SVM	0.339	81.3 (±13.2)	**100**	14.3
Single RBF SVM	0.189	75.0 (±14.4)	88.0	28.6
Preprocessing Ensemble	0.448	84.4 (±12.4)	**100**	28.6
ARSS Ensemble	**0.714**	**90.6** (±10.5)	96.0	**71.4**
Local Experts + Selection	0.486	84.4 (±12.5)	96.0	42.9
Local Experts + Stacking	0.605	87.5 (±11.5)	96.0	57.1

Table 6.13: Evaluation results on the test set for the detection of renal papillary necrosis.

Acute Liver Toxicity

regularization of blood composition

The liver plays a key role in the metabolism of vertebrates and severe liver damages can be lethal for the organism. Although the liver is able to regenerate, cell damages lead to a reduced functionality and changes in the metabolism. Thus, by a changed metabolic activity of the liver, the blood composition can change and thereby indirectly affect the urine composition.

Evaluation results for sample and compound classification on the test set for the detection of liver toxicities with respect to literature references are shown in table 6.14. Generally, a liver toxicity can be induced in different regions of the liver, leading to different modifications of the blood composition. For this application, a compound is defined as being toxic if an organ toxicity in any region of the liver is expected to occur in the first 24 hours. The pattern of spectral changes is expected to be more specific for the acute than for the general liver toxicity.

heterogeneous damage patterns lead to a low sensitivity

The main problem is the robust detection of positive samples or compounds. This leads to low MC values for most classification approaches except for the ARSS ensemble and the local experts approach in combination with the stacking procedure. Both approaches achieve a reasonable classification performance under consideration of the single drug administration and the few collection times. Furthermore, metabolic changes induced by toxic effects in other organs further modify the urine compositions and

METHOD	MC	ACC [%]	SPEC [%]	SENS [%]
SAMPLE CLASSIFICATION				
Single Linear SVM	0.158	73.8 (±3.8)	86.5	27.8
Single RBF SVM	0.161	74.0 (±3.8)	86.7	27.8
Preprocessing Ensemble	0.181	79.2 (±3.6)	98.5	9.3
ARSS Ensemble	0.509	**85.0** (±3.1)	95.4	47.2
Local Experts + Selection	0.446	83.8 (±3.2)	**99.0**	28.7
Local Experts + Stacking	**0.511**	83.2 (±3.3)	88.8	**63.0**
COMPOUND CLASSIFICATION				
Single Linear SVM	0.222	73.1 (±16.2)	89.5	28.6
Single RBF SVM	0.222	73.1 (±16.2)	89.5	28.6
Preprocessing Ensemble	0.052	69.2 (±16.7)	89.5	14.3
ARSS Ensemble	**0.703**	**88.5** (±12.5)	**100**	57.1
Local Experts + Selection	0.330	76.9 (±15.5)	**100**	14.3
Local Experts + Stacking	0.470	76.9 (±15.5)	78.9	**71.4**

Table 6.14: Evaluation results on the test set for the detection of liver necrosis toxicity.

reduce the ability to detect changes induced by the acute liver toxicity.

Phospholipidosis

The intracellular accumulation of phospholipids with lamellar bodies is the main characteristic of drug-induced phospholipidosis. The affection of organs with phospholipidosis leads to inflammatory reactions and changes in histology [Ande 06]. Thus, phospholipidosis is not a type of toxicity for a particular organ, but different organs show similar damages as in case of a drug-induced organ toxicity. The target organ may vary for phospholipidosis inducing compounds, but the toxic effect is supposed to be present for each of them. If particular "damage patterns" independent of the affected organ are recognizable in the urine composition, a detection of phospholipidosis can be achieved.

similar toxic effects can occur at different organs

Sample and compound classification results on the test set shown in table 6.15 clearly demonstrate the low classification performance for the detection of drug-induced phospholipidosis. The ARSS ensemble is the sole approach achieving an MC value above 0.2 for sample classification. Good results of the local experts approach for classification of compounds is expected to be caused by an advantageous distribution of falsely classified samples. Although an MC value of 0.372 is achieved, the sample classification performance of 0.110 demonstrates the low perfor-

low classification accuracies for all approaches

METHOD	MC	ACC [%]	SPEC [%]	SENS [%]
SAMPLE CLASSIFICATION				
Single Linear SVM	0.047	53.1 (±3.6)	61.8	42.4
Single RBF SVM	0.063	54.6 (±3.6)	68.7	37.3
Preprocessing Ensemble	0.089	56.3 (±3.6)	**79.2**	28.5
ARSS Ensemble	**0.335**	**67.3** (±3.4)	72.7	60.6
Local Experts + Selection	0.110	56.2 (±3.6)	62.8	48.2
Local Experts + Stacking	0.038	50.1 (±3.6)	72.7	**65.8**
COMPOUND CLASSIFICATION				
Single Linear SVM	−0.022	48.8 (±14.6)	42.8	55.0
Single RBF SVM	0.120	56.1 (±14.5)	**61.9**	50.0
Preprocessing Ensemble	−0.024	48.8 (±14.6)	47.6	50.0
ARSS Ensemble	**0.372**	**68.3** (±13.7)	**61.9**	75.0
Local Experts + Selection	**0.372**	**68.3** (±13.7)	**61.9**	75.0
Local Experts + Stacking	−0.012	48.8 (±14.6)	14.6	**80.0**

Table 6.15: Evaluation results on the test set for the detection of phospholipidosis.

mance of this approach for the detection of phospholipidosis. Also the classification accuracy of nearly 70 % achieved by the ARSS procedure in sample classification is not high enough for applications in safety pharmacology. Thus, a robust detection of phospholipidosis based on the analysis of NMR spectra from urine samples is not expected.

Conclusions

Results for the detection of three toxicity types as alternatives to the detection of organ damages at the proximal tubule further demonstrate the improved classification performance achieved by ensemble methods. However, a reasonable classification performance cannot be achieved for every toxicity type. The main problem in the detection of renal papillary necrosis and liver toxicity is the low number of positive samples in the data set. Thus, a robust model for toxic samples can hardly be estimated, leading to low sensitivity values. However, the ARSS procedure and the ensemble of local experts stacking procedure achieve even for these data sets a good classification performance. Whether this performance is sufficient for the final application in safety pharmacology studies has to determined with respect to the performance of alternative methods and the required classification accuracies.

ensemble approaches outperform single SVMs

In contrast to the renal papillary necrosis and liver toxicity, a robust detection of phospholipidosis could not be achieved.

no reasonable classification results for phospholipidosis

Although the ARSS procedure could classify about 67 % of all samples, this is not expected to be robust enough for application in safety pharmacology. Phospholipidosis can be observed in different organs, and induced changes in urine composition are expected to be not specific enough in order to achieve a robust detection.

The proximal tubule has a great influence on the urine compositions and is used as the main application in this thesis. Promising results could be achieved in the detection of renal papillary necrosis and liver toxicity, but the collection of further positive samples is expected to lead to an increased sensitivity.

6.6 SUMMARY

evaluation of advanced ensemble approaches

The effectiveness of the new ensemble approaches presented in chapter 5 was determined in this section on a real data set from safety pharmacology and a simulated data set for the validation of identified relevant signals. Furthermore, the performance of classification methods for the detection of alternative toxicity types was investigated.

Evaluation results for the `REAL.NMR` data set demonstrate the low performance of classical pattern recognition approaches from single and multiple classifier theory. The best classification results are achieved by SVM approaches, which serve as reference for comparison of newly developed systems. Even though the reference systems are chosen due to their ability to achieve a reasonable classification even in case of sparse and complex data sets, classification performance needs to be improved for application in safety pharmacology.

newly developed ensemble methods outperform reference systems

Advanced ensemble methods are optimized with respect to the classification results achieved on the validation set of the `REAL.NMR` data set and finally applied on the test set. All ensemble approaches outperform SVMs as reference classification systems. Reasonable classification results are achieved by the ARSS procedure and the ensemble of local experts stacking approach. The stacking procedure improves the generalizability of the ensemble of local experts approach, which results in the best evaluation results among the investigated methods. Combination of sample classifications to compound predictions further improves the classification performance of the classification approaches.

general idea: focus on relevant spectral regions

To focus on the spectral regions relevant to the discrimination between non-toxic and toxic samples represents the main characteristic of newly developed ensemble methods and the central hypothesis of this thesis. The ARSS procedure and the ensemble of local experts represent two different approaches for the determination of relevant spectral regions and their incorporation in the classification decision.

Ensemble systems not only improve the classification performance, but also allow for an interpretation of their ensemble decision. The percentage of base classifiers detecting an organ toxicity is an indicator of the degree of the toxic effect. Furthermore, the spectral signals used by the ARSS procedure and the ensemble of local experts approach can be determined and used as starting point for the identification of biomarker signals.

interpretability of ensemble methods

The promising results obtained in these evaluations are validated on a simulated data set. The ARSS approach and the ensemble of local experts outperform the reference systems in the classification of simulated NMR spectra and all biomarker signals in the spectra are identified.

improved performance of ensemble methods validated on a simulated data set

The improved classification performance of ensemble systems is further validated in the classification of other toxicity types than the proximal tubule as the main organ region investigated in this thesis. Although classification results are below those achieved in the detection of organ toxicity with respect to the proximal tubule, ensemble approaches focusing on relevant spectral regions outperform remaining classification approaches.

7 CONCLUSION

Industrial drug design aims at the development of new pharmaceuticals having a benefit for the patient. While this design starts with the creation or combination of molecules having the intended effect on the organism, other side effects such as organ toxicities are usually unknown. Only if these adverse effects are known and can influence the design, a safe and efficacious drug can be developed.

design of efficacious and safe drugs

In order to give support for a robust detection of organ toxicities, new classification approaches for the automated classification of urine samples based on the measurements from NMR spectroscopy were developed in this thesis. This inference on specific changes in the metabolism indicating a toxic effect based on the analysis of spectroscopic data has been investigated in the last decade(s) by different institutions. Thereby, mainly classical pattern recognition approaches were used on rather limited data sets.

automatic classification of NMR spectra

Classification systems developed in this thesis are based on ensemble methods, which achieved competitive classification results in different applications even in case of sparse and complex data sets. However, their suitability for metabonomic applications has not been investigated up to now. Consequently, the basic concepts of ensemble theory are used for the design of a classification system respecting specific characteristics of NMR spectra.

ensemble methods for robust classification of sparse data sets

SUMMARY

Approaches for the identification of information contained in NMR data relevant to the discrimination between induced toxic effects are mainly based on multivariate data analysis methods, such as PCA or PLS. Only few studies investigated the application of alternative methods from the field of pattern recognition theory on rather limited data sets.

only few approaches applied for classification in metabonomic applications

The detection of characteristic changes in the spectral profiles from urine samples is the main principle for the classification of new samples as being non-toxic or toxic with respect to a particular organ. The reliable detection of the relevant peaks in the spectrum is the most challenging task due to the usually rather small data sets. Therefore, robust classification methods that achieve an automated determination and extraction of spectral

analysis of urine samples for inference on the health status

reliable detection of relevant spectral signals

signals or regions relevant to the discrimination between non-toxic and toxic samples are presented in this thesis.

ensemble systems show competitive results in many applications

definition of a particular ensemble design for classification of NMR spectra

definition of different views by variation of preprocessing and feature extraction methods

Ensemble systems are one prominent example of classification methods which have shown competitive classification results in several applications. The general ensemble principle is used in this thesis for the design of ensemble classification systems, which incorporate characteristics of the particular data set. Thus, the effectiveness and flexibility of ensemble systems is combined for the design of a robust classification system. The general concept of combining multiple classifiers is achieved in a first ensemble approach by the integration of classifiers trained on differently preprocessed data sets in a multiple classifier system. Combination of these different "views" on the data in an ensemble leads to an improved classification performance. This result serves as proof of concept for the applicability of ensemble methods for the classification of NMR spectra. However, the main improvement is expected to be achieved by respecting particular characteristics of the data in the ensemble design.

emphasis on relevant spectral signals

weighted subspace selection procedure

The emphasis on the relevant peaks in the classification system is the fundamental approach for the improvement of classification results presented in this thesis. The relevance of each bucket value is determined in the ARSS approach by the iterative optimization of a weight distribution, which is used for the subspace selection. Thereby, weights are increased if their selection in a subspace leads to higher classification results than those achieved by other base classifiers. Finally, the ensemble performance is increased in each iteration due to the improved base classifiers.

training of local experts on short spectral regions

The initial step in the ensemble of local experts approach is the definition of spectral regions by a sliding windows approach in different scales. Local experts are trained on the short spectral regions, thereby, focusing the view of each base classifier on specific spectral signals. The optimization of the selection of local experts for the final ensemble aggregation is achieved by approaches originally used for feature selection problems. Thereby, local experts using the most relevant parts of the spectrum are used for construction of the final ensemble decision. This selection of local experts is prone to overfitting and an improved generalizability is achieved by a stacking approach for ensemble aggregation.

automatic detection of relevant spectral signals

determination of the grade of a toxic effect

The identification of the relevant spectral regions is achieved by the presented ensemble approaches in an automated way. Thereby, no background knowledge has to be incorporated. Thus, sample ingredients, which have not been regarded as relevant to the detection of organ toxicities, can be used for classification and new biomarkers can be identified. The location of relevant spectral regions is indicated either by high weights in the ARSS procedure or by the specific regions used by local experts contained in the final ensemble. Furthermore, the percentage of base classifiers assigning the label toxic to an applied compound is used as indicator on the degree of the toxic effect.

The focus of the ensemble classification is concentrated on particular spectral regions in the presented ensemble approaches which leads to an improved accuracy in the classification of spectra. However, the primary objective is the determination of adverse effects for each applied compound and not for every sample. Thus, the classifications of spectra corresponding to each compound are combined for each sample collection time. A compound is classified as being toxic if the combination of predictions leads to a toxic label at one of the collection times. Thereby, samples showing a non-toxic profile due to an increased resistance of the organism or a late onset of the toxic effect are compensated and a classification of each applied compound is achieved.

combination of sample predictions for each compound

The capabilities of the newly developed advanced ensemble methods for automatic classification of NMR spectra are investigated in an experimental evaluation based on a real set of NMR spectra from safety pharmacology. A valid optimization and test strategy is developed by the use of a five fold cross-validation and test procedure, which allows for the optimization and evaluation of the ensemble methods. Evaluation results have clearly shown a significant improvement in classification performance of ensemble methods in comparison with single SVM approaches, which outperformed a selection of alternative ensemble approaches. The main improvements are achieved by ensemble approaches focusing on relevant spectral regions determined in an automated way. While the base classifier selection procedure of the ensemble of local experts approach leads to overfitting effects, an improved generalizability is achieved by the alternative stacking approach. The ARSS procedure and the ensemble of local experts show also in compound classification the best evaluation results among the investigated approaches. These results are confirmed for the classification of alternative toxicity types, although a reasonable classification performance cannot be achieved for each type.

determination of ensemble performance by an experimental evaluation

ensemble approaches outperform reference systems

classification of alternative toxicity types

Inference on the most relevant spectral regions is shown for the `REAL.NMR` data set, but these results cannot be validated since no information on the location of real biomarker signals is available. Thus, a data set of simulated spectra containing mainly peaks signals with randomly changing intensity and five biomarker peaks are used for evaluation of the interpretation results. The combination of results from both methods leads to the identification of the exact signals of all biomarker signals, thereby demonstrating the effectiveness of the ensemble methods.

interpretability evaluated on a set of simulated spectra

OUTLOOK

classification performance and interpretability is the key for acceptance

larger data sets in an alternative data representation

Metabonomics has a large potential to influence the developments of techniques used in drug design for the assessment of toxic adverse effect after drug application. Ensemble methods presented in this thesis achieve superior results in comparison with alternative investigated approaches. However, the accuracy has to be improved in order to achieve acceptance of these automated detection methods. The success of these methods has been reduced up to now by the relatively small data sets and the spectral data representation. Generally, developments of new instruments with stronger magnetic fields reduce the overlap of peak signals and can improve description of alternative data representations such as peak lists. Furthermore, 2D-NMR is an alternative measurement procedure, which further reduces the overlap of signals by their positioning on a 2D plane rather than on a 1D ppm scale. These improvements combined with steadily decreasing measurement costs, and advances in the data analysis techniques are expected to be the key for achieving improved results in the classification and interpretation of future spectroscopic data sets.

classification of time-series data

A further promising alternative to the presented classification approaches is the incorporation of the changes in the spectral profile of the same animal over time. The collection of samples before the drug application allows for the determination of a normal spectral profile for each animal. With respect to these spectra, changes in the spectral profile can be qualified for further collection points and used as information for an applied classifier. Therefore, a data representation as shown for the SMART scaling [Keun 04] can be created, or classifiers respecting time-series data such as Hidden Markov Models (HMMs) can be applied. However, in order to use samples from several collection points and to train robust classification models, a very large data set of high quality has to be acquired.

A APPENDIX

A.1 PRINCIPAL COMPONENT ANALYSIS

The basic concept of Principal Component Analysis (PCA) is to transform a high-dimensional data set into a representation with lower dimensionality while retaining most of the variation present in the data. This transformation is achieved by projection into a new coordinate system, whereby the axes, the so-called Principal Components (PCs), are uncorrelated and sorted according to the amount of explained data variation so that the main variation in the data is described by the first few PCs.

data transformation in space of lower dimensionality

Variation present in a data set to be analyzed is used for the estimation of PCs. Variation is characterized by the scatter-matrix **S** of a data set $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ defined as:

retention of variation in the data

$$S = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}})^T,$$

where $\bar{\mathbf{x}}$ is the mean vector of the data set defined as

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i\,.$$

For decorrelation of the data a transformation $\mathbf{\Theta}$ is required, which transforms **S** to a diagonal matrix $\tilde{\mathbf{S}}$

$$\tilde{\mathbf{S}} = \frac{1}{N}\sum_{i=1}^{N}\Theta(\mathbf{x}_i - \bar{\mathbf{x}})[\Theta(\mathbf{x}_i - \bar{\mathbf{x}})]^T = \mathbf{\Theta}\mathbf{S}\mathbf{\Theta}^T$$

In order to retain the Euclidean distance between samples an orthonormal transformation must be applied. The transposed matrix $\mathbf{\Theta}^T$ of eigenvectors $\boldsymbol{\theta}_i$ of **S** fullfills the constraints of orthonormality and is used for the decorrelation of **S**. Application of the transformation to a data set with zero mean

orthonormal transformation

$$y = \mathbf{\Theta}^T(\mathbf{x} - \bar{\mathbf{x}})$$

leads to the transformed scatter matrix by

$$\tilde{\mathbf{S}} = \mathbf{\Theta}^T\mathbf{S}\mathbf{\Theta} = \mathbf{\Theta}^T\mathbf{\Theta}\mathbf{\Lambda}\mathbf{\Theta}^T\mathbf{\Theta} = \mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_N \end{pmatrix},$$

where $\lambda_1, \lambda_2, \ldots, \lambda_N$ are the eigenvalues of $\mathbf{S}$ expressing the variance of the data described by the PCs.

use of the m first PCs for transformation

reconstruction error ϵ

Information on the amount of each PC's explained variance can be used for dimension reduction of a n-dimensional data set by using only the m first PCs ($m < n$) with the largest eigenvalues for the construction of a feature space. Projecting the original samples onto this new subspace reduces the sample dimensionality while retaining the majority of data variance. The loss of information by this projection can be expressed by the reconstruction error ϵ, which is basically the amount of variance explained by the discarded PCs

$$\epsilon = E\{\|\mathbf{x} - \mathbf{x}'\|\} = E\{\| \sum_{i=m+1}^{n} \mathbf{y}_i\theta_i\|^2\} = \sum_{i=m+1}^{n} \lambda_i$$

where

$$\mathbf{x}' = \sum_{i=1}^{m} \mathbf{y}_i\theta_i \, .$$

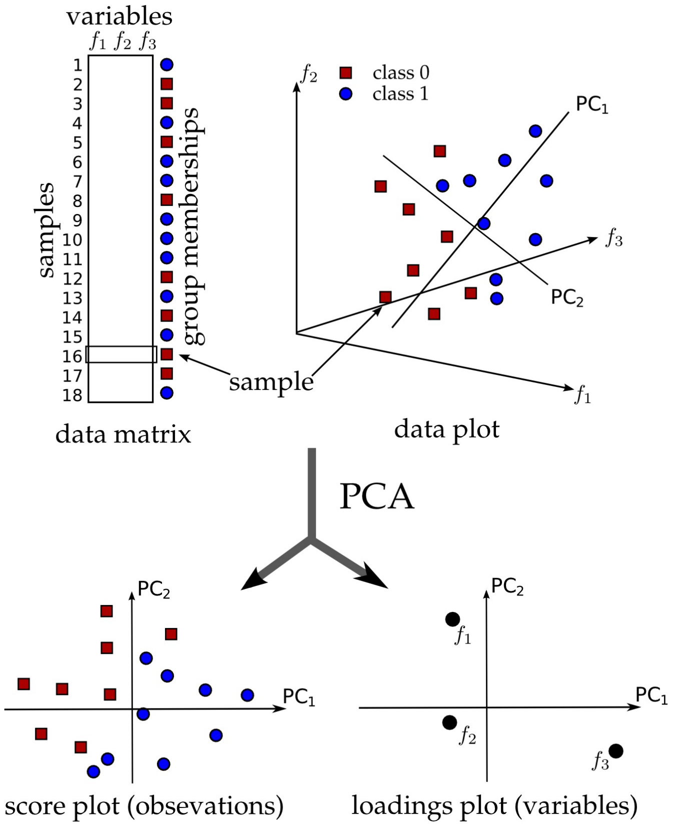


Figure A.1: A PCA model approximates the main variation in a given data set. Interpretation can be pursued by low dimensional representation of the samples in a score plot or describing the influence of each input variable on the PCs (adopted from [Tryg 07]).

The projected samples are usually referred to as *scores* (cf. figure A.1). These can be used for visualization purposes in a score plot or even as feature representations of a data set in a classification system. Using only the first few PCs reduces the data dimensionality but the main variance present in the data is retained.

scores

score plot

PCs mainly reweight the dimensions of the input space according to their relevance for description of the currently major variance in the data. Thus, further knowledge on the variation of each dimension in a data set can be achieved by analysis of these weights, which are also referred to as *loadings*. If a separation between classes is visible in the score plot, a loading plot can be used to determine the features relevant to the class separation. Each point in the loading plot corresponds to a particular feature. Important variables are located in the periphery of the loading plot and variables with low influence on the model are close to the origin. This interpretation is usually applied in fields, where each variable has a certain meaning and the most relevant features have to be determined (e.g. chemometrics, metabonomics).

loading plot

Further details of approaches for the estimation of PCA models, and applications are described in [Fink 08] and [Joll 02], which are the basis for previous explanations and are a good starting point for further reading on PCA.

A.2 PARTIAL LEAST SQUARES REGRESSION

Probably the best-known method in metabonomics is the Partial Least Squares (PLS) regression, also referred to as projection to latent structures. PLS has been introduced by the Swedish statistican Herman O. A. Wold in 1966 [Wold 66]. Comparable with PCA, a new coordinate system for the projection of the independent variables (samples) is estimated within PLS. The optimization criterion for this estimation is not the explained variation of the new coordinates. Instead, the covariance between the independent variables and a given set of dependent variables (e.g. class labels) is maximized. Thus, PLS is a supervised method.

data transformation w.r.t. to variance in the data and class labels

Due to the incorporation of additional information on the samples regarding their class membership, an improved discrimination between different groups can be achieved by PLS in comparison with PCA. Furthermore, new samples can be classified according to the estimated PLS model. Due to its convincing results in mutlivariate data analysis and its integration into several commercial and non-commercial programs, PLS has become increasingly popular in the field of metabonomics and chemometrics.

matrix of independent variables

The set of n k-dimensional independent variables is denoted in the following as $n \times k$ matrix $\mathbf{X}$, and the n l-dimensional independent variables as $n \times l$ matrix $\mathbf{Y}$. The $\mathbf{Y}$ matrix is in case of discriminant analysis formed as a "dummy matrix" of zeros and ones, whereby each column corresponds to one of the different classes. Each sample has a one in the column of the class it belongs to and a zero in the column of classes it does not belong to. Vectors $\mathbf{t}$, $\mathbf{u}$, $\mathbf{w}$, $\mathbf{c}$ and $\mathbf{p}$ are columns of the matrices $\mathbf{T}$, $\mathbf{U}$, $\mathbf{W}$, $\mathbf{C}$ and $\mathbf{P}$, and b are diagonal elements of the diagonal matrix $\mathbf{B}$.

matrices of orthogonal factors, and weights

Decomposition of the matrices $\mathbf{X}$ and $\mathbf{Y}$ (cf. figure A.2) is achieved by a common set of orthogonal factors formed as matrix $\mathbf{T}$ and specific weights $\mathbf{P}$ and $\mathbf{C}$. Thereby, the matrix $\mathbf{Y}$ is only approximated and $\mathbf{Y} \neq \hat{\mathbf{Y}}$.

$$\begin{aligned}
&\mathbf{X} = \mathbf{T}\mathbf{P}^T \\
&\hat{\mathbf{Y}} = \mathbf{T}\mathbf{B}\mathbf{C}^T\,, \\
&\text{where} \quad \mathbf{T}^T\mathbf{T} = \mathbf{1}
\end{aligned}$$

For the determination of the matrix $\mathbf{T}$ a maximzation problem according the covariance between $\mathbf{X}$ and $\mathbf{Y}$ is formulated in order to achieve the desired dependency between the two matrices. Thereby, a set of factors $\mathbf{w}$ and $\mathbf{c}$ has to be determined for a

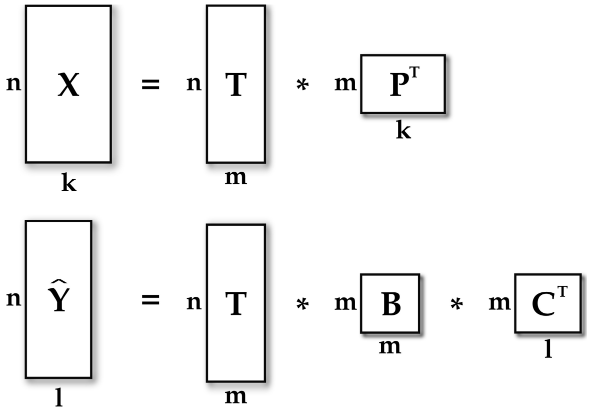


Figure A.2: Decomposition of matrices X and Y.

linear combination of columns from **X** and **Y**. This maximization problem can be formulated according to the following:

$$\begin{aligned} \underset{\mathbf{t},\mathbf{u}\in\mathbb{R}^N}{\text{maximize}} \quad & \mathbf{t}^T\mathbf{u} \\ \text{with} \quad & \mathbf{t} = \mathbf{X}\mathbf{w} \\ & \mathbf{u} = \mathbf{Y}\mathbf{c} \\ \text{where} \quad & \mathbf{w}^T\mathbf{w} = \mathbf{1} \\ & \mathbf{t}^T\mathbf{t} = \mathbf{1}\,. \end{aligned}$$

Solving this maximization problem for **t** and **u** allows for the calculation of **p** and **b** by

$$\begin{aligned} \mathbf{p} &= \mathbf{E}^T\mathbf{t} \\ \mathbf{b} &= \mathbf{t}^T\mathbf{u}\,, \end{aligned}$$

where **E** is the matrix **X** after mean centering and variance normalization. The influence of **t** can now be subtracted from **X** and **Y** , and the process is iteratively repeated until the desired number of components is calculated or **X** equals a zero matrix.

interpretation by scores and loadings plots

The interpretation of the estimated model parameters is comparable with the scores and loadings plots explained for PCA in section A.1, while PLS has shown improved interpretation results compared with PCA due to the incorporation of class labels in the estimation of model components. One of the possible algorithms for the calculation of a PLS model is outlined in the following section.

NIPALS for Estimation of PLS Components

A commonly used method for the estimation of PLS components is the nonlinear iterative partial least squares (NIPALS) algorithm, which will be outlined based on [Abdi 03]. Mean-centered and variance normalized matrices **X** (independent variables) and **Y** (dependent variables) are the input parameters for the PLS estimation method shown in algorithm 4. The final model components are determined in an iterative approach up to a predefined number of iterations. Increasing the number of model components improves the description of the data, while simultaneously decreasing the generalization capability of the model for the hihiprediction of unknown samples.

Algorithm 4 Calculation of a PLS model according to the NIPALS algorithm (cf [Abdi 03, Wold 04])

```
Input: mean centered and variance normed matrices X and Y
    and amount of factors $m$ to be calculated
Output: Components of the PLS model: T, U, W, C, P and B
$i = 0$
while $(\mathbf{X} \neq$ zero matrix   AND   $i \leq m)$ do
    initialize u randomly
    $i = i + 1$
    $\mathbf{t}_{\text{old}} = \mathbf{0}$
    $\mathbf{t}_{\text{new}} = \mathbf{1}$
    while $\mathbf{t}_{\text{old}} \neq \mathbf{t}_{\text{new}}$ do
        $\mathbf{t}_{\text{old}} = \mathbf{t}_{\text{new}}$
        $\mathbf{w} = \frac{\mathbf{X}^T\mathbf{u}}{\mathbf{u}^T\mathbf{u}}$                • weights for X
        $\mathbf{t} = \frac{\mathbf{X}\mathbf{w}}{\mathbf{w}^T\mathbf{w}}$                • factors for X
        $\mathbf{c} = \frac{\mathbf{Y}^T\mathbf{t}}{\mathbf{t}^T\mathbf{t}}$                • weights for Y
        $\mathbf{u} = \frac{\mathbf{Y}\mathbf{c}}{\mathbf{c}^T\mathbf{c}}$                • factors for Y
        $\mathbf{t}_{\text{new}} = \mathbf{t}$
    end while
    $\mathbf{p} = \frac{\mathbf{X}^T\mathbf{t}}{\mathbf{t}^T\mathbf{t}}$                    • subtract the influence of t ...
    $\mathbf{X} = \mathbf{X} - \mathbf{t}^T\mathbf{p}$                • ...from X
    $\mathbf{Y} = \mathbf{Y} - \mathbf{t}^T\mathbf{c}$                • ...from Y
    $b = \mathbf{t}^T\mathbf{u}$                        • regression weight
    store t, u, w, c, p and $b$ in the corresponding
    matrices T, U, W, C, P und B
end while
return T, U, W, C, P und B
```

A.3 Example of PC and PLS Regression

Principal Component Regression (PCR) and Partial Least Squares Regression (PLSR) are based on the analysis methods presented in the previous two sections. A regression and not a classification problem is assumed. This section[1] demonstrates the application of both methods on a set of 60 near infrared (NIR) spectra at 401 wavelengths, whereby the analyzed gasoline samples have different octane ratings (cf. [Kali 97]). The spectra of the data set are shown in figure A.3.

data set of 60 NIR spectra at 401 wavelengths and given octane ratings

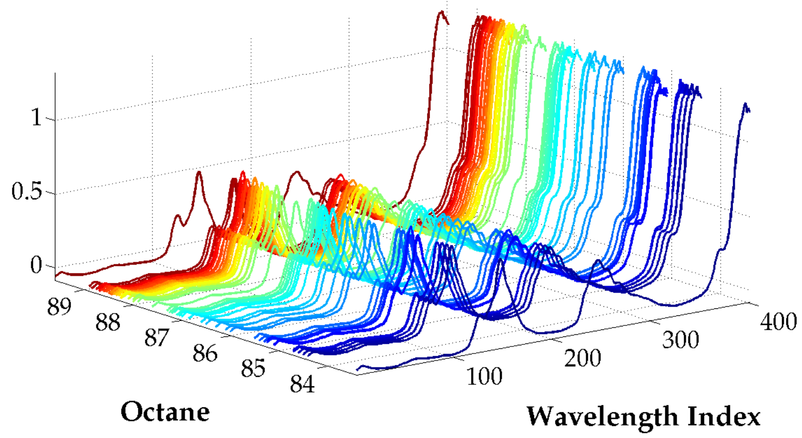


Figure A.3: Data set of 60 near infrared spectra of gasoline samples at 401 wavelengths and their octane ratings.

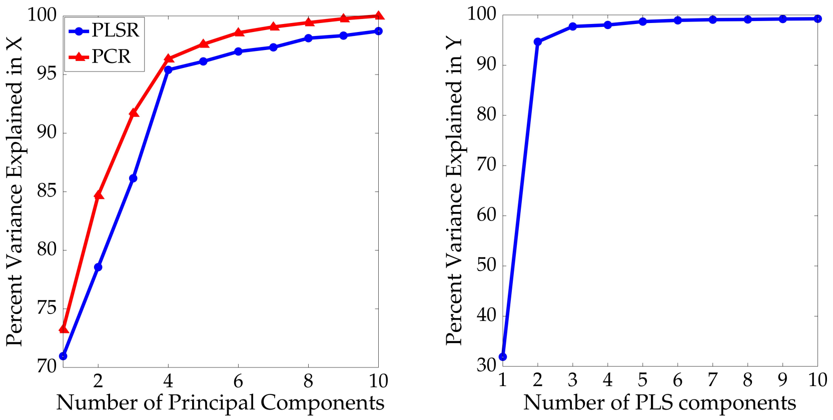
PCR and PLSR models are estimated under variation of the number of components. Thereby, the percentage of explained variance in the data space increases with each new component as shown in figure A.4a. PCR components are estimated in order to maximize the explained variance in the data space X. As a result, the explained variance of PCR components is higher than those of PLS components, since these are estimated by taking also the response variable and not only the predictor variables into account. The percentage of variance in the response variable explained by the PLSR components is shown in figure A.4b.

percentage of explained variance in X and Y

More than 95 % of variance in the data space is explained by the first four components of both regression approaches. However, two comoponents of the PLSR model are already sufficient to fit the data in the Y space as shown in figure A.4b. In practice, methods such as cross-validation are applied for determination of a reasonable number of PLS components.

estimation of models with two components

[1] Explanations in this section are based on the PCR and PLSR demonstration from the Matlab statistics toolbox, which can be found at `http://www.mathworks.com/products/statistics/demos.html?file=/products/demos/shipping/stats/plspcrdemo.html`



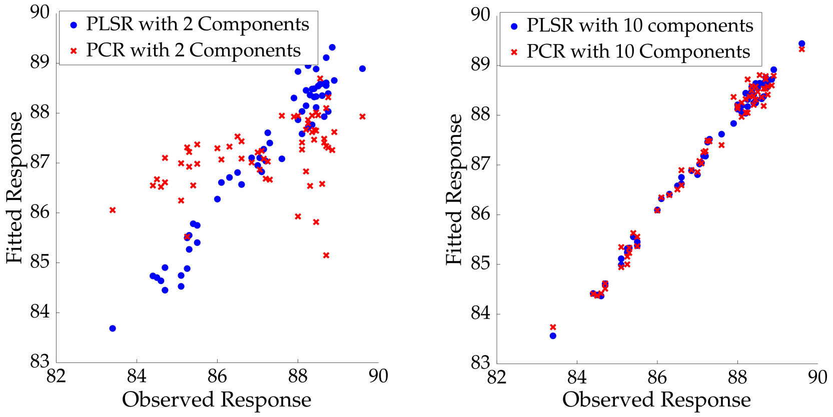
(a) Explained variance in X by PC and PLS regression.

(b) Explained variance in Y by PLS regression.

Figure A.4: Percent of explained variance in X and Y of PC or PLS models under variation of the model components.

Fitting a PCR and PLSR model with two components to the data leads to fitted versus observed response plots as shown in figure A.5a. The fitted PLSR responses do predict the octane ratings very good and better than those achieved by the PCR model. However, the choice of two components is based on observations for the PLSR model and the quality of the fit of the PCR model to the dependent variable is expected to increase by using more components. Figure A.5b demonstrates the low difference in residuals of both methods using ten model components.

Summarizing, the results of PCR and PLSR become quite similar with an increasing number of model components. However, PLSR is able to fit the response variables with a lower number of components due to the incorporation of the response variables in the model estimation.



(a) 2 components

(b) 10 components

Figure A.5: Plot of fitted versus observed octane ratings using two or ten model components.

BIBLIOGRAPHY

[Abdi 03] H. Abdi. "Partial Least Squares (PLS) Regression". In: M. Lewis-Beck, A. Bryman, and T. Futting, Eds., *Encyclopedia of Social Sciences Research Methods*, Thousand Oaks, 2003.

[Aber 09] K. Aberg, E. Alm, and R. J. Torgrip. "The correspondence problem for metabonomics datasets". *Analytical and Bioanalytical Chemistry*, 2009.

[Adam 06] C. P. Adams and V. V. Brantner. "Estimating the cost of new drug development: is it really 802 million dollars?". *Health Affairs*, Vol. 25, No. 2, pp. 420–428, 2006.

[Akse 03] M. Aksela. "Comparison of classifier selection methods for improving committee performance". In: T. Windeatt and F. Roli, Eds., *Multiple Classifier Systems*, pp. 306–316, Springer, 2003.

[Ande 06] N. Anderson and J. Borlak. "Drug-induced phospholipidosis". *FEBS Letters*, Vol. 580, No. 23, pp. 5533–5540, 2006.

[Anth 95] M. L. Anthony, V. S. Rose, J. K. Nicholson, and J. C. Lindon. "Classification of toxin-induced changes in 1H NMR spectra of urine using an artificial neural network". *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 13, No. 3, pp. 205 – 211, 1995.

[Asun 07] A. Asuncion and D. Newman. "UCI Machine Learning Repository". 2007.

[Bald 00] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and H. Nielsen. "Assessing the accuracy of predicition algorithms for classification: an overview". *Bioinformatics*, Vol. 16, No. 5, pp. 412 – 424, 2000.

[Ball 81] D. H. Ballard. "Generalizing the Hough transform to detect arbitrary shapes". *Pattern Recognition*, Vol. 13, No. 2, pp. 111–122, 1981.

[Banf 03] R. E. Banfield, L. O. Hall, K. W. Bowyer, and W. P. Kegelmeyer. "A new ensemble diversity measure applied to thinning ensembles". In: T. Windeatt and F. Roli, Eds., *Multiple Classifier Systems*, pp. 306–316, Springer, 2003.

[Barn 89] R. J. Barnes, M. S. Dhanoa, and S. J. Lister. "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra". *Applied Spectroscopy*, Vol. 43, No. 5, pp. 772–777, 1989.

[Beck 03] O. Beckonert, M. E. Bollard, T. M. Ebbels, H. C. Keun, H. Antti, and E. a. Holmes. "NMR-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches". *Analytica Chimica Acta*, Vol. 490, pp. 3–15, 2003.

[Belo 02a] A. I. Belousov, S. A. Verzakov, and J. von Frese. "Applicational aspects of support vector machines". *Journal of Chemometrics*, Vol. 16, No. 8-10, pp. 482–489, 2002.

[Belo 02b] A. I. Belousov, S. A. Verzakov, and J. von Frese. "A flexible classification approach with optimal generalisation performance: support vector machines". *Chemometrics and Intelligent Laboratory Systems*, Vol. 64, No. 1, pp. 15 – 25, 2002.

[Bezd 98] J. Bezdek, T. Reichherzer, G. Lim, and Y. Attikiouzel. "Multiple-prototype classifier design". *IEEE Transactions on Systems, Man, and Cybernetics, C*, Vol. 28, No. 1, pp. 67–79, 1998.

[Bish 95] C. M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, 1995.

[Brad 07] J. K. Bradley and R. E. Schapire. "FilterBoost: Regression and classification on large datasets". In: J. C. Platt, D. Koller, Y. Singer, and S. T. Roweis, Eds., *Advances in Neural Information Processing Systems*, pp. 185–192, 2007.

[Brat 00] D. C. Brater and W. J. Daly. "Clinical pharmacology in the Middle Ages: Principles that presage the 21st century". *Clinical Pharmacology & Therapeutics*, Vol. 67, No. 5, pp. 447–450, 2000.

[Brei 01] L. Breiman. "Random Forests". *Machine Learning*, Vol. 45, No. 1, pp. 5–32, 2001.

[Brei 84] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen. *Classification and Regression Trees*. Chapman & Hall/CRC, 1984.

[Brei 96a] L. Breiman. "Bagging Predictors". *Machine Learning*, Vol. 24, No. 2, pp. 123–140, 1996.

[Brei 96b] L. Breiman. "Heuristics of instability and stabilization in model selection". *Ann. Statis.*, Vol. 24, No. 6, pp. 2350–2383, 1996.

[Brei 96c] L. Breiman. "Stacked regressions". *Machine Learning*, Vol. 24, No. 1, pp. 49–64, 1996.

[Bril 92] F. Brill, D. Brown, and W. Martin. "Fast genetic selection of features for neural network classifiers". *IEEE Transactions on Neural Networks*, Vol. 3, No. 2, pp. 324–328, 1992.

[Brin 02] J. T. Brindle, H. Antti, E. Holmes, G. Tranter, J. K. Nicholson, H. W. Bethell, S. Clarke, P. M. Schofield, E. McKilligin, D. E. Mosedale, and D. J. Grainger. "Rapid and noninvasive diagnosis of the presence and severity of coronary heart disease using ^{1}H-NMR-based metabonomics". *Nature Medicine*, Vol. 8, pp. 1439–1445, 2002.

[Buhl 03] P. Bühlmann and B. Yu. "Boosting with the L2 loss: regression and classification". *Journal of the American Statistical Association*, Vol. 98, pp. 324–339, 2003.

[Burb 01] R. Burbidge, M. Trotter, S. Holden, and B. Buxton. "Drug design by machine learning: support vector machines for pharmaceutical data analysis". *Comput. Chem*, Vol. 26, pp. 5–14, 2001.

[Bure 05] A. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, and P. V. Eerdewegh. "Identifying SNPs predictive of phenotype using random forests". *Genetic Epidemiology*, Vol. 28, No. 2, pp. 171–182, 2005.

[Byle 06] M. Bylesjö, M. Rantalainen, O. Cloarec, J. K. Nicholson, E. Holmes, and J. Trygg. "OPLS discriminant analysis: combining the strengths of PLS-DA and SIMCA classification". *Journal of Chemometrics*, Vol. 20, No. 8-10, pp. 341–351, 2006.

[Byva 03] E. Byvatov, U. Fechner, J. Sadowski, and G. Schneider. "Comparison of support vector machine and artificial neural network systems for drug/nondrug classification". *Journal of Chemical Information and Computer Sciences*, Vol. 43, No. 6, pp. 1882–1889, 2003.

[Camp 03] N. A. Campbell and J. B. Reece. *Biologie*, Chap. 7, pp. 1134–1142. Spektrum Akademischer Verlag GmbH Heidelberg, Berlin, 2003. (in german).

[Chan 01] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*. 2001.

[Chri 00] N. Christiani and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, 2000.

[Cons 05] M. A. Constantinou, E. Papakonstantinou, M. Spraul, S. Sevastiadou, C. Costalos, M. A. Koupparis, K. Shulpis, A. Tsantili-Kakoulidou, and E. Mikros. "1H NMR-based metabonomics for the diagnosis of inborn errors of metabolism in urine". *Analytica Chimica Acta*, Vol. 542, No. 2, pp. 169–177, 2005.

[Cort 95] C. Cortes and V. Vapnik. "Support Vector Networks". *Machine Learning*, Vol. 20, No. 3, pp. 273–297, 1995.

[Crai 06] A. Craig, O. Cloarec, E. Holmes, J. Nicholson, and J. Lindon. "Scaling and normalization effects in NMR spectroscopic metabonomic data sets". *Analytical Chemistry*, Vol. 78, No. 7, pp. 2262–2267, 2006.

[Csen 07] L. Csenki, E. Alm, R. J. O. Torgrip, K. M. Åberg, L. I. Nord, I. Schuppe-Koistinen, and J. Lindberg. "Proof of principle of a generalized fuzzy Hough transform approach to peak alignment of one-dimensional ^{1}H NMR data". *Analytical and Bioanalytical Chemistry*, Vol. 389, pp. 875–885, 2007.

[Cunn 00] P. Cunningham and J. Carney. "Diversity versus quality in classification ensembles based on feature selection". In: *Machine Learning: ECML 2000*, pp. 109–116, Springer, 2000.

[Deca 97] C. Decaestecker. "Finding prototypes for nearest-neighbor classification by means of gradient descent and deterministic annealing". *Pattern Recognition*, Vol. 30, No. 2, pp. 281–288, 1997.

[Dett 03] M. Dettling and P. Bühlmann. "Boosting for tumor classification with gene expression data.". *Bioinformatics*, Vol. 19, No. 9, pp. 1061–1069, 2003.

[Dett 04] M. Dettling. "BagBoosting for tumor classification with gene expression data". *Bioinformatics*, Vol. 20, No. 18, pp. 3583–3593, 2004.

[Diaz 06] R. Díaz-Uriarte and S. Alvarez de Andrés. "Gene selection and classification of microarray data using random forest.". *BMC Bioinformatics*, Vol. 7, No. 3, 2006.

[Diet 00] T. G. Dietterich. "Ensemble methods in machine learning". In: J. Kittler and F. Roli, Eds., *Multiple Classifier Systems*, pp. 1–15, Springer, 2000.

[Diet 03] C. Dietrich, G. Palm, and F. Schwenker. "Decision templates for the classification of bioacoustic time series". *Information Fusion*, Vol. 4, pp. 101–109, 2003.

[Diet 06] F. Dieterle, A. Ross, G. Schlotterbeck, and H. Senn. "Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in 1H NMR metabonomics". *Analytical Chemistry*, Vol. 78, No. 13, pp. 4281–4290, 2006.

[DiMa 03] J. A. DiMasi, R. W. Hansen, and H. G. Grabowski. "The price of innovation: new estimates of drug development costs". *Journal of Health Economics*, Vol. 22, No. 2, pp. 151–185, 2003.

[DiMa 91] J. A. DiMasi, R. W. Hansen, H. G. Grabowski, and L. Lasagna. "Cost of innovation in the pharmaceutical industry". *Journal of Health Economics*, Vol. 10, No. 2, pp. 107–142, 1991.

[Duda 01] R. O. Duda, P. E. Hart, and D. G.Stork. *Pattern Classification*. John Wiley & Sons, Inc., 2nd Ed., 2001.

[Duda 72] R. O. Duda and P. E. Hart. "Use of the Hough transformation to detect lines and curves in pictures". *Communications of the ACM*, Vol. 15, No. 1, pp. 11–15, 1972.

[Duma 06] M.-E. Dumas, E. C. Maibaum, C. Teague, H. Ueshima, Z. Beifan, J. C. Lindon, J. K. Nicholson, J. Stamler, P. Elliot, Q. Chan, and E. Holmes. "Assessment of analytical reproducibility of ^{1}H NMR spectroscopy based metabonomics for large-scale epidemiological research : The INTERMAP study". *Analytical Chemistry*, Vol. 78, No. 7, pp. 2199–2208, 2006.

[Ebbe 03] T. Ebbels *et al.* "Toxicity classification from metabonomic data using a density superposition approach: CLOUDS". *Analytica Chimica Acta*, Vol. 490, pp. 109–122, 2003.

[Ebbe 07] T. M. D. Ebbels, H. C. Keun, O. P. Beckonert, M. E. Bollard, J. C. Lindon, E. Holmes, and J. K. Nicholson. "Prediction and classification of drug toxicity using probabilistic modeling of temporal metabolic data: the consortium on metabonomic toxicology screening approach". *Journal of Proteome Research*, Vol. 6, No. 11, pp. 4407–4422, 2007.

[Esbe 01] K. H. Esbensen. *Multivariate Data Analysis - in Practice*. Camo Press AS, Norway, 2001.

[Espo 97] F. Esposito, D. Malerba, and G. Semeraro. "A comparative analysis of methods for pruning decision trees". *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 19, No. 5, pp. 476–491, 1997.

[Ferr 99] F. Ferri, J. Albert, and E. Vidal. "Considerations about sample-size sensitivity of a family of edited nearest-neighbor rules".

IEEE Transactions on Systems, Man, and Cybernetics, B, Vol. 29, No. 4, pp. 667–672, 1999.

[Fink 08] G. A. Fink. *Markov Models for Pattern Recognition, from Theory to Applications*. Springer, Heidelberg, 2008.

[Fors 03] J. Forshed, I. Schuppe-Koistinen, and S. P. Jacobsson. "Peak alignment of NMR signals by means of a genetic algorithm". *Analytica Chimica Acta*, Vol. 487, No. 2, pp. 189–199, 2003.

[Fors 05] J. Forshed, R. J. Torgrip, K. M. Aberg, B. Karlberg, J. Lindberg, and S. P. Jacobsson. "A comparison of methods for alignment of NMR peaks in the context of cluster analysis". *Journal of Pharmaceutical and Biomedical Analysis*, Vol. 38, No. 5, pp. 824–832, 2005.

[Free 03] R. Freeman. *Magnetic Resonance in Chemistry and Medicine*. Oxford University Press Inc., New York, 2003.

[Freu 01] Y. Freund. "An adaptive version of the boost by majority algorithm". *Machine Learning*, Vol. 43, No. 3, pp. 293–318, 2001.

[Freu 97] Y. Freund and R. E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting". *Journal of Computer and System Sciences*, Vol. 55, No. 1, pp. 119–139, 1997.

[Frie 00] J. Friedman, T. Hastie, and R. Tibshirani. "Additive logistic regression: a statistical view of boosting". *Annals of Statistics*, Vol. 28, pp. 337–407, 2000.

[Frie 99] H. Friebolin. *Basic One and Two Dimensional Nmr Spectroscopy*. Wiley-VCH Verlag Gmbh, Weinheim, 1999.

[Fung 01] M. Fung, A. Thornton, K. Mybeck, J. H.-H. Wu, K. Hornbuckle, and E. Muniz. "Evaluation of the characteristics of safety withdrawal of prescription drugs from worldwide pharmaceutical markets - 1960 to 1999". *Drug Information Journal*, Vol. 35, pp. 293–317, 2001.

[Gold 89] D. E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., 1989.

[Gray 98] H. F. Gray, R. J. Maxwell, I. Martínez-Pérez, C. Arús, and S. Cerdán. "Genetic programming for classification and feature selection: analysis of ^{1}H nuclear magnetic resonance spectra from human brain tumour biopsies". *NMR Biomed*, Vol. 11, No. 4-5, pp. 217–224, 1998.

[Hall 07] L. O. Hall, R. E. Banfield, K. W. Bowyer, and W. P. Kegelmeyer. "Boosting lite - handling larger datasets and slower base classifiers". In: M. Haindl, J. Kittler, and F. Roli, Eds., *Multiple Classifier Systems*, pp. 161–170, Springer, 2007.

[Hama 97] Y. Hamamoto, S. Uchimura, and S. Tomita. "A bootstrap technique for nearest neighbor classifier design". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 1, pp. 73–79, 1997.

[Han 94] J. H. Han, L. T. Koczy, and T. Poston. "Fuzzy-Hough transform". *Pattern Recognition Letters*, Vol. 15, No. 7, pp. 649–658, 1994.

[Hans 90] L. Hansen and P. Salamon. "Neural network ensembles". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 10, pp. 993–1001, 1990.

[Hart 68] P. Hart. "The condensed nearest neighbor rule". *IEEE Transactions on Information Theory*, Vol. 14, No. 3, pp. 515–516, 1968.

[Ho 98] T. K. Ho. "The random subspace method for constructing decision forests". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 8, pp. 832–844, 1998.

[Holm 01] E. Holmes, J. K. Nicholson, and G. Tranter. "Metabonomic characterization of genetic variations in toxicological and metabolic responses using probabilistic neural networks". *Chemical Research in Toxicology*, Vol. 14, No. 2, pp. 182–191, 2001.

[Holm 94] E. Holmes, P. J. D. Foxall, J. K. Nicholson, G. H. Neild, S. M. Brown, C. R. Beddell, B. C. Sweatman, E. Rahr, J. C. Lindon, and P. Spraul, M. Neidig. "Automatic data reduction and pattern recognition methods for analysis of ^{1}H nuclear magnetic resonance spectra of human urine from normal and pathological states". *Analytical Biochemistry*, Vol. 220, No. 2, pp. 284–296, 1994.

[Holm 98a] E. Holmes, A. W. Nicholls, J. C. Lindon, S. Ramos, M. Spraul, P. Neidig, S. C. Connor, J. Connelly, S. J. Damment, J.Haselden, and J. K. Nicholson. "Development of a model for classification of toxin-induced lesions using ^{1}H NMR spectroscopy of urine combined with pattern recognition.". *NMR Biomed*, Vol. 11, No. 4-5, pp. 235–244, 1998.

[Holm 98b] E. Holmes, J. K. Nicholson, A. W. Nicholls, J. C. Lindon, S. C. Connor, S. Polley, and J. Connelly. "The identification of novel biomarkers of renal toxicity using automatic data reduction techniques and PCA of proton NMR spectra of urine". *Chemometrics and Intelligent Laboratory Systems*, Vol. 44, No. 1-2, pp. 245–255, 1998.

[Hsu 02] C.-W. Hsu and C.-J. Lin. "A comparison of methods for multiclass support vector machines". *IEEE Transactions on Neural Networks*, Vol. 13, No. 2, pp. 415–425, 2002.

[Huan 95] Y. S. Huang and C. Y. Suen. "A method of combining multiple experts for the recognition of unconstrained handwritten numerals". *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol. 17, No. 1, pp. 90–94, 1995.

[Jank 09] A. Jankevics, E. Liepinsha, E. Liepinsha, R. Vilskerstsa, S. Grinbergaa, O. Pugovicsa, and M. Dambrovaa. "Metabolomic studies of experimental diabetic urine samples by ^{1}H NMR spectroscopy and LC/MS method". *Chemometrics and Intelligent Laboratory Systems*, Vol. 97, No. 1, pp. 11–17, 2009.

[Jin 03] R. Jin, Y. Liu, L. Si, J. Carbonell, and A. G. Hauptmann. "A new boosting algorithm using input-dependent regularizer". In: *Proceedings of Twentieth International Conference on Machine Learning (ICML03)*, AAAI Press, 2003.

[Joll 02] I. T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 2nd Ed., 2002.

[Kali 97] J. H. Kalivas. "Two data sets of near infrared spectra". *Chemometrics and Intelligent Laboratory Systems*, Vol. 37, No. 2, pp. 255–259, 1997.

[Kass 98] A. Kassidas, J. F. MacGregor, and P. A. Taylor. "Synchronization of batch trajectories using dynamic time warping". *AIChE Journal*, Vol. 44, No. 4, pp. 864–875, 1998.

[Keun 02] H. C. Keun, T. M. Ebbels, H. Antti, M. E. Bollard, O. Beckonert, G. Schlotterbeck, H. Senn, U. Niederhauser, E. Holmes, J. C. Lindon, and J. K. Nicholson. "Analytical reproducibility in ^{1}H NMR-based metabonomic urinalysis". *Chemical Research in Toxicology*, Vol. 15, No. 11, pp. 1380–1386, 2002.

[Keun 03] H. C. Keun, T. M. D. Ebbels, H. A. ad Mary E. Bollardand Olaf Beckonert, E. Holmes, J. C. Lindon, and J. K. Nicholson. "Improved analysis of multivariate data by variable stability (VAST) scaling:application to NMR spectroscopic metabolic profiling". *Analytica Chimica Acta*, Vol. 490, pp. 265–276, 2003.

[Keun 04] H. Keun, T. Ebbels, M. Bollard, O. Beckonert, H. Antti, E. Holmes, J. Lindon, and J. Nicholson. "Geometric trajectory analysis of metabolic responses to toxicity can define treatment specific profiles". *Chemical Research in Toxicology*, Vol. 17, No. 5, pp. 579–587, 2004.

[Keun 06] H. C. Keun. "Metabonomic modeling of drug toxicity". *Pharmacology & Therapeutics*, Vol. 109, No. 1-2, pp. 92 – 106, 2006.

[Kim 03] H.-C. Kim, S. Pang, H.-M. Je, D. Kim, and S. Y. Bang. "Constructing support vector machine ensemble". *Pattern Recognition*, Vol. 36, No. 12, pp. 2757 – 2767, 2003.

[Kitt 02] J. Kittler, M. Ballette, J. Czyz, F. Roli, and L. Vandendorpe. "Decision level fusion of intramodal personal identity verification experts". In: F. Roli and J. Kittler, Eds., *Multiple Classifier Systems*, pp. 314–324, Springer-Verlag New York, Inc., 2002.

[Kitt 98] J. Kittler, M. Hatef, R. P. Duin, and J. Matas. "On combining classifiers". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 3, pp. 226–239, 1998.

[Koh 08] H.-W. Koh, S. Maddula, J. Lambert, R. Hergenröder, and L. Hildebrand. "Feature selection by lorentzian peak reconstruction for ^{1}NMR post-processing". In: *CBMS '08: Proceedings of the 2008 21st IEEE International Symposium on Computer-Based Medical Systems*, pp. 608–613, IEEE Computer Society, Washington, DC, USA, 2008.

[Koha 95] R. Kohavi. "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *Proceedings of the International Joint Conference on Artifical Intelligence*, pp. 1137–1143, Morgan Kaufmann, 1995.

[Kola 04] I. Kola and J. Landis. "Can the pharmaceutical industry reduce attrition rates?". *Nature Reviews Drug Discovery*, Vol. 3, pp. 711–716, 2004.

[Kunc 01] L. Kuncheva, J. Bezdek, and R. Duin. "Decision templates for multiple classifier fusion: an experimental comparison". *Pattern Recognition*, Vol. 34, pp. 299–314, 2001.

[Kunc 03] L. I. Kuncheva and C. J. Whitaker. "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy". *Machine Learning*, Vol. 51, No. 2, pp. 181–207, 2003.

[Kunc 04] L. I. Kuncheva. *Combining Pattern Classifiers – Methods and Algorithms*. Wiley Interscience, 2004.

[Lam 97] L. Lam and S. Suen. "Application of majority voting to pattern recognition: an analysisof its behavior and performance". *IEEE Transactions on Systems, Man and Cybernetics, Part A*, Vol. 27, No. 5, pp. 553–568, 1997.

[Lati 00] P. Latinne, O. Debeir, and C. Decaestecker. "Different ways of weakening decision trees and their impact on classification accuracy of DT combination". In: J. Kittler and F. Roli, Eds., *Multiple Classifier Systems*, pp. 200–209, Springer, London, UK, 2000.

[Lenz 07] E. Lenz and I. Wilson. "Analytical strategies in metabonomics". *Journal of Proteome Research*, Vol. 6, No. 2, pp. 443–458, 2007.

[Li 05] X. Li, L. Wang, and E. Sung. "A study of AdaBoost with SVM based weak learners". In: *Proceedings of the IEEE International Joint Conference on Neural Networks*, pp. 196–201, 2005.

[Lien 07] K. Lienemann, T. Plötz, and G. A. Fink. "On the application of SVM-Ensembles based on adapted random subspace sampling for automatic classification of nmr data". In: M. Haindl, J. Kittler, and F. Roli, Eds., *Multiple Classifier Systems*, pp. 42–51, Springer, 2007.

[Lien 08a] K. Lienemann, T. Plötz, and G. A. Fink. "SVM ensemble classification of nmr spectra based on different configurations of data processing techniques". In: *Proc. Int. Conf. on Pattern Recognition*, 2008. WeAT8.4.

[Lien 08b] K. Lienemann, T. Plötz, and S. Pestel. "NMR-based urine analysis in rats: Prediction of proximal tubule kidney toxicity and phospholipidosis". *Journal of Pharmacologial and Toxicological Methods*, Vol. 58, No. 1, pp. 41–49, 2008.

[Lien 08c] K. Lienemann, T. Plötz, and G. A. Fink. "Automatic classification of NMR spectra by ensembles of local experts". In: *Structural, Syntactic, and Statistical Pattern Recognition*, pp. 790–800, Springer, 2008.

[Lien 09] K. Lienemann, T. Plötz, and G. A. Fink. "Stacking for ensembles of local experts in metabonomic applications". In: J. A. Benediktsson, J. Kittler, and F. Roli, Eds., *Multiple Classifier Systems*, pp. 498–508, Springer, 2009.

[Lin 03] X. Lin, S. Yacoub, J. Burns, and S. Simske. "Performance analysis of pattern classifier combination by plurality voting". *Pattern Recogn. Lett.*, Vol. 24, No. 12, pp. 1959–1969, 2003.

[Lin 07] H.-T. Lin, C.-J. Lin, and R. C. Weng. "A note on Platt's probabilistic outputs for support vector machines". *Journal of Machine Learning*, Vol. 68, No. 3, pp. 267–276, 2007.

[Lind 03] J. C. Lindon, J. K. Nicholson, E. Holmes, H. Antti, M. E. Bollard, H. Keun, O. Beckonert, T. M. Ebbels, M. D. Reily, D. Robertson, G. J. Stevens, P. Luke, A. P. Breau, G. H. Cantor, R. H. Bible, U. Niederhauser, H. Senn, G. Schlotterbeck, U. G. Sidelmann, S. M. Laursen, A. Tymiak, B. D. Car, L. Lehman-McKeeman, J.-M. Colet, A. Loukaci, and C. Thomas. "Contemporary issues in toxicology the role of metabonomics in toxicology and its evaluation by the COMET project.". *Toxicology and Applied Pharmacology*, Vol. 187, No. 3, pp. 137–146, 2003.

[Lind 05] J. C. Lindon, H. C. Keun, T. M. Ebbels, J. M. Pearce, E. Holmes, and J. K. Nicholson. "The Consortium for Metabonomic Toxicology (COMET): aims, activities and achievements". *Pharmacogenomics*, Vol. 6, No. 7, pp. 691–699, 2005.

[Lisb 98] P. J. G. Lisboa, S. P. J. Kirby, A. Vellido, Y. Y. B. Lee, and W. El-Deredy. "Assessment of statistical and neural networks methods in NMR spectral classification and metabolite selection". *NMR in Biomedicine*, Vol. 11, No. 4–5, pp. 225–234, 1998.

[Loon 97] C. G. Looney. *Pattern Recognition Using Neural Networks. Theory and Algorithms for Engineers and Scientists*. Oxford Univ. Press, 1997.

[Mari 63] T. Marill and D. Green. "On the effectiveness of receptors in recognition systems". *IEEE Transactions on Information Theory*, Vol. 9, No. 1, pp. 11–17, 1963.

[Mart 89] H. Martens and T. Naes. *Multvariate Calibration*. John Wiley and Sons Inc., New York, 1989.

[Maso 07] S. Masoum, C. Malabat, M. Jalali-Heravi, C. Guillou, S. Rezzi, and D. N. Rutledge. "Application of support vector machines to 1H NMR data of fish oils: methodology for the confirmation of wild and farmed salmon and their origins". *Analytical and Bioanalytical Chemistry*, Vol. 387, No. 4, pp. 1499–1510, 2007.

[Meir 03] R. Meir and G. Rätsch. "An introduction to boosting and leveraging". In: S. Mendelson and A. Smola, Eds., *Advanced lectures on machine learning*, pp. 118–183, Springer-Verlag New York, Inc., New York, NY, USA, 2003.

[Mitc 96] M. Mitchell. *An Introduction to Genetic Algorithms*. MIT Press, 1996.

[Nich 01] A. Nicholls, E. Holmes, J. Lindon, J. Shockcor, R. Farrant, J. Haselden, S. Damment, C. Waterfield, and J. Nicholson. "Metabonomic investigations into hydrazine toxicity in the rat". *Chemical Research in Toxicology*, Vol. 14, No. 8, pp. 975–987, 2001.

[Nich 99] J. K. Nicholson, J. C. Lindon, and E. Holmes. "Metabonomics: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data.". *Xenobiotica*, Vol. 29, pp. 1181–1189, 1999.

[Niel 98] N.-P. V. Nielsen, J. M. Carstensen, and J. Smedsgaard. "Aligning of single and multiple wavelength chromatographic profiles for chemometric data analysis using correlation optimised warping". *Journal of Chromatography A*, Vol. 805, No. 1-2, pp. 17–35, 1998.

[Nitz 82] S. Nitzan and J. Paroush. "Optimal decision rules in uncertain dichotomous choice situations". *International Economic Review*, Vol. 23, No. 2, pp. 289–297, 1982.

[Oh 04] I.-S. Oh, J.-S. Lee, and B.-R. Moon. "Hybrid genetic algorithms for feature selection". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 11, pp. 1424–1437, Nov. 2004.

[Opit 99] D. Opitz and R. Maclin. "Popular ensemble methods: an empirical study". *Journal of Artificial Intelligence Research*, Vol. 11, pp. 169–198, 1999.

[Ott 03] K.-H. Ott, N. Araníbar, B. Singh, and G. W. Stockton. "Metabonomics classifies pathways affected by bioactive compounds. Artificial neural network classification of NMR spectra of plant extracts". *Phytochemistry*, Vol. 62, No. 6, pp. 971–985, 2003. Plant Metabolomics.

[Parm 96] B. Parmanto, P. W. Munro, and H. R. Doyle. "Improving committee diagnosis with resampling techniques". In: D. S. Touretzky, M. C. Mozer, and M. E. Hesselmo, Eds., *Advances in Neural Information Processing Systems*, pp. 882–888, MIT Press, Cambridge, MA, 1996.

[Pear 05] M. R. Pears, J. D. Cooper, H. M. Mitchison, R. J. Mortishire-Smith, D. A. Pearce, and J. L. Griffin. "High resolution ^{1}H NMR-based metabolomics indicates a neurotransmitter cycling deficit in cerebral tissue from a mouse model of batten disease". *Journal of Biological Chemistry*, Vol. 280, pp. 42508–42514, 2005.

[Plat 99a] J. C. Platt. *Fast Training of Support Vector Machines using Sequential Minimal Optimization*, pp. 185–208. *Advances in kernel methods: support vector learning*, MIT Press, Cambridge, MA, USA, 1999.

[Plat 99b] J. C. Platt. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". In: *Advances in Large Margin Classifiers*, pp. 61–74, 1999.

[Prav 02] V. Pravdova, B. Walczak, and D. L. Massart. "A comparison of two algorithms for warping of analytical signals". *Analytica Chimica Acta*, Vol. 456, No. 1, pp. 77–92, 2002.

[Pudi 94] P. Pudil, J. Novovičová, and J. Kittler. "Floating search methods in feature selection". *Pattern Recogn. Lett.*, Vol. 15, No. 11, pp. 1119–1125, 1994.

[Quin 86] J. R. Quinlan. "Induction of decision trees". *Machine Learning*, Vol. 1, pp. 81–106, 1986.

[Quin 93] J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann, 1993.

[Rabi 38] I. I. Rabi, J. R. Zacharias, S. Millman, and P. Kusch. "A new method of measuring nuclear magnetic moment". *Physical Review*, Vol. 53, No. 4, pp. 318–318, 1938.

[Rant 07] M. Rantalainen, M. Bylesjö, O. Cloarec, J. K. Nicholson, E. Holmes, and J. Trygg. "Kernel-based orthogonal projections to latent structures (K-OPLS)". *Journal of Chemometrics*, Vol. 21, pp. 376–385, 2007.

[Raym 00] M. Raymer, W. Punch, E. Goodman, L. Kuhn, and A. Jain. "Dimensionality reduction using genetic algorithms". *Evolutionary Computation, IEEE Transactions on*, Vol. 4, No. 2, pp. 164–171, Jul 2000.

[Robe 00] D. G. Robertson, M. D. Reily, R. E. Sigler, D. F. Wells, D. A. Paterson, and T. K. Braden. "Metabonomics: evaluation of nuclear magnetic resonance (NMR) and pattern recognition technology for rapid in vivo screening of liver and kidney toxicants". *Toxicological Sciences*, Vol. 57, pp. 326–337, 2000.

[Rose 62] F. Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, 1962.

[Rose 76] A. Rosenfeld and A. C. Kak. *Digital Picture Processing*. Orlando: Academic Press, 1976.

[Rous 08] R. Rousseau, B. Govaerts, M. Verleysen, and B. Boulanger. "Comparison of some chemometric tools for metabonomics biomarker identification". *Chemometrics and Intelligent Laboratory Systems*, Vol. 91, No. 1, pp. 54–66, 2008. Selected papers presented at the Chemometrics Congress CHIMIOMETRIE 2006 Paris, France, 30 November - 1 December 2006.

[Ruta 03] D. Ruta. *Classifier diversity in combined pattern recognition systems*. PhD thesis, University of Paisland, Scotland, UK, 2003.

[Sako 78] H. Sakoe and S. Chiba. "Dynamic programming algorithm optimization for spoken word recognition". *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol. 26, No. 1, pp. 43–49, 1978.

[Sama 97] A. Samal and J. Edwards. "Generalized Hough transform for natural shapes". *Pattern Recognition Letters*, Vol. 18, No. 5, pp. 473–480, 1997.

[Savi 64] A. Savitzky and M. J. E. Golay. "Smoothing and differentiation of data by simplified least squares procedures". *Analytical Chemistry*, Vol. 36, No. 8, pp. 1627–1639, 1964.

[Scar 98] F. Scarselli and A. C. Tsoi. "Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results". *Neural Networks*, Vol. 11, No. 1, pp. 15–37, 1998.

[Schl 03] R. Schlittgen. *Einführung in die Statistik*. Oldenburg, München, 10 Ed., 2003.

[Scho 00] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett. "New support vector algorithms". *Neural Computation*, Vol. 12, No. 5, pp. 1207–1245, 2000.

[Scho 02] B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

[Schw 00] H. Schwenk and Y. Bengio. "Boosting neural networks". *Neural Comput.*, Vol. 12, No. 8, pp. 1869–1887, 2000.

[Schw 96] G. Schwedt. *Taschenatlas der Analytik*. Thieme Verlag, 1996.

[Shap 84] L. Shapley and B. Grofman. "Optimizing group judgmental accuracy in the presence of interdependencies". *Public Choice*, Vol. 43, No. 3, pp. 329–343, 1984.

[Sied 89] W. Siedlecki and J. Sklansky. "A note on genetic algorithms for large-scale feature selection". *Pattern Recognition Letters*, Vol. 10, No. 5, pp. 335–347, 1989.

[Skal 94] D. B. Skalak. "Prototype and feature selection by sampling and random mutation hill climbing algorithms". In: *Proceedings of the Eleventh International Conference on Machine Learning*, pp. 293–301, Morgan Kaufmann, Los Altos, CA, 1994.

[Skov 06] T. Skov, F. van den Berg, G. Tomasi, and R. Bro. "Automated alignment of chromatographic data". *Journal of Chemometrics*, Vol. 20, No. 11-12, pp. 484–497, 2006.

[Smit 47] C. A. B. Smith. "Some examples of discrimination". *Annals of Eugenics*, Vol. 13, pp. 272–282, 1947.

[Smit 81] T. F. Smith and M. S. Waterman. "Identification of common molecular subsequences". *Journal of Molecular Biology*, Vol. 147, No. 1, pp. 195–197, 1981.

[Spec 90] D. F. Specht. "Probabilistic neural networks". *Neural Networks*, Vol. 3, No. 1, pp. 109–118, 1990.

[Spen 62] W. Spendley, G. R. Hext, and F. R. Himsworth. "Sequential application of simplex designs in optimization and evolutionary operation". *Technometrics*, Vol. 4, pp. 441–461, 1962.

[Spra 94] M. Spraul, P. Neidig, U. Klauck, P. Kessler, E. Holmes, J. K. Nicholson, B. C. Sweatman, S. R. Salman, R. D. Farrant, and E. Rahr. "Automatic reduction of NMR spectroscopic data for statistical and pattern recognition classification of samples". *Journal of Pharmaceutical & Biomedical Analysis*, Vol. 12, pp. 1215–1225, 1994.

[Step 06] N. Stepenosky, D. Green, J. Kounios, C. M. Clark, and R. Polikar. "Majority vote and decision template based ensemble classifiers trained on event related potentials for early diagnosis of alzheimer's disease". In: *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 901 – 904, 2006.

[Stoy 04] R. Stoyanova, A. W. Nicholls, J. K. Nicholson, J. C. Lindon, and T. R. Brown. "Automatic alignment of individual peaks in large high-resolution spectral data sets.". *Journal of Magnetic Resonance*, Vol. 170, No. 2, pp. 329–335, Oct 2004.

[Suet 06] N. Suetake, E. Uchino, and K. Hirata. "Generalized fuzzy hough transform for detecting arbitrary shapes in a vague and noisy image". *Soft Computing*, Vol. 10, No. 12, pp. 1161–1168, 2006.

[Ting 99] K. M. Ting and I. H. Witten. "Issues in stacked generalization". *Journal of Artificial Intelligence Research*, Vol. 10, pp. 271–289, 1999.

[Toma 04] G. Tomasi, F. van den Berg, and C. Andersson. "Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data". *Journal of Chemometrics*, Vol. 18, No. 5, pp. 231–241, 2004.

[Torg 03] R. J. O. Torgrip, M. Åberg, B. Karlberg, and S. P. Jacobsson. "Peak alignment using reduced set mapping". *Journal of Chemometrics*, Vol. 17, pp. 573–582, 2003.

[Torg 06] R. J. O. Torgrip, J. Lindberg, M. Linder, B. Karlberg, S. P. Jacobsson, J. Kolmert, I. Gustafsson, and I. Schuppe-Koistinen. "New methods of data partitioning based on PARS peak alignment for improvedmultivariate biomarker/biopattern detection in ^{1}H NMR spectroscopic metabolic profiling of urine". *Biomedical and Life Sciences*, Vol. 2, No. 1, pp. 1–19, 2006.

[Torg 08] R. J. O. Torgrip, K. M. Åberg, E. Alm, I. Schuppe-Koistinen, and J. Lindberg. "A note on normalization of biofluid 1D ^{1}H-NMR data". *Metabolomics*, Vol. 4, No. 2, pp. 114–121, 2008.

[Tous 74] G. Toussaint. "Bibliography on estimation of misclassification". *IEEE Transactions on Information Theory*, Vol. 20, No. 4, pp. 472–479, 1974.

[Troh 03] U. Tröhler. *James Lind and Scurvy: 1747 to 1795*. The James Lind Library, 2003.

[Tryg 02] J. Trygg and S. Wold. "Orthogonal projections to latent structures (O-PLS)". *Journal of Chemometrics*, Vol. 16, No. 3, pp. 119–128, 2002.

[Tryg 07] J. Trygg, E. Holmes, and T. Lundstedt. "Chemometrics in metabonomics". *Journal of Proteome Research*, Vol. 6, pp. 469–479, 2007.

[Tume 96] K. Tumer and J. Ghosh. "Error correlation and error reduction in ensemble classifiers". *Connection Science*, Vol. 8, No. 3, pp. 385–404, 1996.

[Vapn 95] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 1995.

[Wang 06] Z. Wang and S. B. Kim. "Automatic alignment of high-resolution NMR spectra using a bayesian estimation approach". In: *18th International Conference on Pattern Recognition (ICPR'06)*, pp. 667–670, 2006.

[Wang 09] S. Wang, A. Mathew, Y. Chen, L. Xi, L. Ma, and J. Lee. "Empirical analysis of support vector machine ensemble classifiers". *Expert Systems with Applications*, Vol. 36, No. 3, Part 2, pp. 6466 – 6476, 2009.

[Wate 05] N. Waters, C. Waterfield, R. Farrant, E. Holmes, and J. Nicholson. "Metabonomic deconvolution of embedded toxicity: application to thioacetamide hepato- and nephrotoxicity". *Chemical Research in Toxicology*, Vol. 18, No. 4, pp. 639–654, 2005.

[Whit 71] A. W. Whitney. "A direct method of nonparametric measurement selection". *IEEE Transactions on Computers*, Vol. 20, No. 9, pp. 1100–1103, 1971.

[Wilc 45] F. Wilcoxon. "Individual comparisons by ranking methods". *Biometrics Bulletin*, Vol. 1, No. 6, pp. 80–83, 1945.

[Wils 72] D. L. Wilson. "Asymptotic properties of nearest neighbor rules using edited data". *IEEE Transactions on Systems, Man and Cybernetics*, Vol. 2, No. 3, pp. 408–421, 1972.

[Wold 04] S. Wold, L. Eriksson, J. Trygg, and Nouna. *The PLS method – partial least squares projections to latent structures– and its applications in industrial RPD (research, development, and production)*. 2004.

[Wold 66] H. Wold. *Estimation of Principal Components and Related Models by Iterative Least Squares*, Chap. Estimation and Prediciton, pp. 391–420. New York: Academic Press, 1966.

[Wolp 92] D. H. Wolpert. "Stacked generalization". *Neural Networks*, Vol. 5, pp. 241–259, 1992.

[Wu 06] W. Wu, M. Daszykowski, B. Walczak, B. Sweatman, S. Connor, J. Haselden, D. Crowther, R. Gill, and M. Lutz. "Peak alignment of urine NMR spectra using fuzzy warping". *Journal of Chemical Information and Modeling*, Vol. 46, No. 2, pp. 863–875, 2006.

[Xu 92] L. Xu, A. Krzyzak, and C. Suen. "Methods of combining multiple classifiers and their applications to handwriting recognition". *Systems, Man and Cybernetics, IEEE Transactions on*, Vol. 22, No. 3, pp. 418–435, May/Jun 1992.

[Yang 98] J. Yang and V. Honavar. "Feature subset selection using a genetic algorithm". *Intelligent Systems and their Applications, IEEE*, Vol. 13, No. 2, pp. 44–49, Mar/Apr 1998.

[Yule 00] G. U. Yule. "On the association of attributes in statistics". *Phil. Trans. A*, Vol. 194, pp. 257–319, 1900.

[Zhou 02] Z. Zhou, J. Wu, and W. Tang. "Ensembling neural networks: Many could be better than all". *Artificial Intelligence*, Vol. 137, No. 1-2, pp. 239 – 263, 2002.