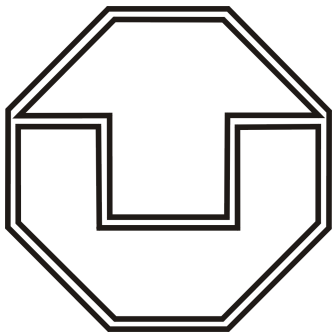


Technische Universität Dresden
Fakultät Mathematik und Naturwissenschaften
Fachrichtung Chemie
Institut für Analytische Chemie

Diplomarbeit

Chemometrische Auswertung von IR-Images und -Maps

vorgelegt von Claudia Beleites,
geb. am 6. Mai 1978 in Bad Nauheim,

eingereicht am 30. Januar 2003

Gutachter: Prof. Dr. rer. nat. habil. R. Salzer
Betreuer: Dr. Ing. G. Steiner

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Übersicht über die verwendeten Formelzeichen und Abkürzungen

Die verwendeten Formelzeichen

$\tilde{\nu}$	Wellenzahl	
n	Anzahl an Spektren	
p	Anzahl an Dimensionen des Merkmalsvektors, Anzahl der Datenpunkte eines Spektrums, Variablenzahl	S. 12
g	Anzahl der Klassen	S. 16
k, $\hat{k}$, l $\in \{(0), 1, \ldots, g\}$	wahre und geschätzte Klassenzugehörigkeit eines Objekts	S. 16
Ω	Grundgesamtheit aller zu klassifizierenden Objekte	S. 16
Ω_i	Gesamtheit der Objekte der Klasse i	S. 16
ω	ein Objekt	S. 16
$\mathbf{x} \in \mathbb{R}^p$	beobachteter Merkmalsvektor eines Objekts, hier: ein Spektrum	S. 16
$\mathbb{S} \subset \mathbb{R}^p$	durch alle $\mathbf{x}$ aufgespannter Stichprobenraum	S. 16
e	Entscheidungsregel	S. 16
$\mathrm{p}(k)$	a priori Wahrscheinlichkeit der Klassenzugehörigkeit, entspricht der relativen Häufigkeit der Klasse k	S. 16
$\mathrm{f}(\mathbf{x}\|k)$	Klassenverteilung von $\mathbf{x}$ in Ω_k, Verteilung der Merkmalsvektoren der Klasse k	S. 16
$\mathrm{f}(\mathbf{x})$	unbedingte Verteilung von $\mathbf{x}$ in Ω, die Verteilung der Merkmalsvektoren aller Klassen	S. 16
$\mathrm{p}(k\|\mathbf{x})$	a posteriori Wahrscheinlichkeit der Klassenzugehörigkeit, Wahrscheinlichkeit, dass ein Objekt mit dem beobachteten Merkmalsvektor $\mathbf{x}$ zur Klasse k gehört	S. 16
F, $\hat{\mathrm{F}}$	wahre und geschätzte Fehlerrate	S. 16
$\mathrm{F}(e\|\mathbf{x})$	Fehlerrate der Entscheidungsregel e für ein Objekt mit dem Merkmalsverktor $\mathbf{x}$	S. 17
$\mathbf{C}$	Kostenmatrix	S. 17
$\mathbf{C}\left(\hat{k}\|\mathbf{x}\right)$	zu erwartende Kosten bei der Zuordnung des Objekts mit dem Merkmalsvektor $\mathbf{x}$ zur Klasse k	S. 17
$\mathbf{S}$	exakte Kovarianzmatrix der Merkmalsvektoren aller Objekte aller Klassen	S. 18
$\mathbf{S}_k$	exakte Kovarianzmatrix der Merkmalsvektoren der Objekte der Klasse k	S. 18
$\boldsymbol{\mu} \in \mathbb{R}^p$	Erwartungswertvektor der multivariaten Normalverteilung	S. 16
$\boldsymbol{\mu}_k \in \mathbb{R}^p$	Erwartungswertvektor der multivariaten Normalverteilung der Merkmalsvektoren der Objekte der Klasse k	S. 16
M	Testgröße nach Box	S. 20
$\chi^2(P = 1 - \alpha; f)$	P-Quantil der χ^2-Verteilung mit f Freiheitsgraden	S. 20
α	Signifikanzniveau, maximale Irrtumswahrscheinlichkeit 1. Art	S. 20
f	Anzahl Freiheitsgrade	S. 20
T, $\hat{T}$, T_0, $\hat{T}_0$	Trefferrate	S. 22
V, $\hat{V}$	Veränderung der Trefferrate	S. 24
$f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$	Zielfunktional der Optimierung	S. 28

$\mathbb{G} \subseteq \mathbb{R}^n$	Suchraum der Optimierung	S. 28
q	Dimensionalität des Optimierungsproblems. Hier gilt: $q = p - 1$	S. 28
$\hat{T}_{Opt}$	Trefferrate mittels Validierung von Optimierung und LDA geschätzt	S. 48
$\hat{T}_{LDA}$	Trefferrate mittels Validierung nur der LDA geschätzt	S. 48
$\hat{T}_R$	Reklassifikations-Trefferrate mittels Validierung von Optimierung und LDA geschätzt	S. 48
E (? - ? cm^{-1})	Mittelwert der Extiktion im angegebenen Spektralbereich	

Folgender Schriftsatz wurde zur Kennzeichnung der Größen verwendet.

x**:** Skalar (nomale Kleinbuchstaben)

x: Vektor (fette Kleinbuchstaben)

X: Matrix (fette Großbuchstaben)

Geschätzte statistische Größen sind durch ein Dach gekennzeichnet: $\mathbf{S}$ bezeichnet die *wahre* Kovarianzmatrix, die *geschätzte* Kovarianzmatrix ist $\hat{\mathbf{S}}$.

Programme, Funktionen und Datenstrukturen sind durch diktengleichen Schriftsatz gekennzeichnet: `Matlab`.

Verwendete Abkürzungen

FPA-Detektor: Focal-Plane-Array-Detektor

LDA: Lineare Diskriminanzanalyse

NMR: Nuclear Magnetic Resonance

MCT-Detektor: Mercury-Cadmium-Telluride-Detektor

Astro II: Astrozytome zweiten Grades

Astro III: Astrozytome dritten Grades

Glio: Glioblastome

HE-Färbung: Hämatoxylin-Eosin-Färbung

WHO: World Health Organization

DNA: Desoxyribonukleinsäure

Teil I.

Einleitung

1. Einleitung

Die Chemometrie wendet statistische Verfahren zur Auswertung sehr komplexer chemischer Problemstellungen an. Diese Methoden kommen daher dort zum Einsatz, wo die klassische Auswertung aufgrund großer Datenvolumina, hochdimensionaler Problemstellungen oder zu geringem Informationsgehalt der Daten an ihre Grenzen stößt.

Krankheiten bedeuten Veränderungen in der molekularen Zusammensetzung der Zellen [1], Tumore gehen mit Veränderungen der betroffenen Gewebe einher. Da Infrarot-Spektren die molekulare Zusammensetzung einer Probe widerspiegeln, ist zu erwarten, dass sich die Krankheit auch in einer Veränderung der Infrarot-Spektren des Gewebes niederschlägt.

Allerdings sind diese Veränderungen in den Spektren mit bloßem Auge nicht in der geforderten Genauigkeit zu erkennen, so dass zur Auswertung der Spektren chemometrische Methoden herangezogen werden müssen. Diese mathematische Vorgehensweise ermöglicht auch die Abschätzung der Genauigkeit der getroffenen Vorhersagen, was einen großen Vorteil reproduzierbarer Methoden gegenüber den subjektiveren Ergebnissen der Begutachtung durch Histologen darstellt. Wie in vielen anderen Gebieten ist es aber auch auf diesem Gebiet schwierig, eine entsprechende jahrelange Berufserfahrung durch mathematisch fassbare Regeln zur Entscheidungsfindung zu ersetzen.

Infrarot-Spektren stellen eine große Menge mathematisch verarbeitbarer Informationen bereit, auch zur Auswertung dieser Informationen und Interpretation der Ergebnisse existieren bewährte Methoden. Der Informationsgehalt der Infrarot-Spektren ist so groß, dass der Versuch, allein aufgrund der spektralen Informationen zu einer hinreichend genauen Klassifikation zu kommen, als legitim betrachtet werden kann.

Ein wichtiger Schritt der Informationsverarbeitung ist die Reduktion der Daten auf den für die vorliegende Fragestellung relevanten Anteil. So unterscheiden sich die Spektren unterschiedlicher Gewebe zwar, weisen aber auch große gemeinsame Anteile auf. Diese gemeinsamen Informationen sind für die Auswertung von untergeordneter Bedeutung und können stark komprimiert oder ganz aus der weiteren Betrachtung ausgeschlossen werden.

Es kann zwischen verschiedenen chemometrischen Methoden mit unterschiedlichen Eigenschaften gewählt werden. Die Anwendung von Verfahren, die wenig Zusatzinformationen in die Auswertung einfließen lassen, kann neue Wege der Interpretation eröffnen. Andererseits macht die Berücksichtigung von Vorwissen die gewünschten Informationen oft schneller und exakter zugänglich. Eine genaue Kenntnis der Methoden und ihrer spezifischen Eigenschaften ist notwendig, um Fehlinterpretationen der Ergebnisse zu vermeiden.

Alle Transformationen der Daten haben das Ziel, den Anteil interpretierbarer Information zu erhöhen. In der Regel kann ein steigender, für Menschen oder weitere Verfahren erkennbarer, Informationsgehalt nur in Verbindung mit einem sinkenden Informationsgehalt im informationstheoretischen Sinn erreicht werden. Dies tritt immer dann auf, wenn Transformationen auf die Daten angewandt werden, die *nicht* zu mathematisch äquivalenten Abbildungen führen, also keine Grundoperationen, sondern Datenmanipulationen darstellen [2].

2. Ziel der Arbeit

Diese Arbeit untersucht Infrarot-Spektren verschiedener Hirntumore mit dem Ziel der Klassifikation nach histologischen Gesichtspunkten.

Inhalt dieser Arbeit ist die Erstellung eines geeigneten Trainingsdatensatzes, die Ermittlung der erforderlichen Parameter zur Auswahl geeigneter Wellenzahlbereiche für die Klassifizierung sowie die Anzahl der Variablen der Klassifizierung selbst. Weiterhin sollen die Einflüsse verschiedener Methoden zur Datenvorbehandlung auf die Ergebnisse der Auswertungsroutinen untersucht werden, um Empfehlungen für eine möglichst allgemeingültige und automatisierbare Datenvorbehandlung aussprechen zu können. Auch die Möglichkeit der Modellbildung mit Daten, die an verschiedenen Geräten gemessen wurden, ist zu untersuchen.

Besondere Aufmerksamkeit ist den Fehlklassifikationen zu widmen. Einerseits im Hinblick auf die Folgen, aber auch unter dem Gesichtspunkt, dass die Trainingsdatensätze histologisch klassifiziert sind, die untersuchten Tumore jedoch ineinander übergehen.

Im Gegensatz zu faktoranalytischen Methoden, die die Daten in ein abstraktes Koordinatensystem transformieren, arbeiten die hier angewandten Verfahren mit realen Teilbereichen der Spektren. Die durch den Algorithmus ausgewählten Wellenzahlbereiche können und sollen damit auch auf ihre biochemische Aussage hin untersucht werden.

Die verwendete Methode zur Klassifikation ist die lineare Diskriminanzanalyse (LDA). Sie wird in der vorliegenden Arbeit mittels des Programms `stackedGen` durchgeführt. Ein weiteres Programm, `ga_ors`, wird zur Auswahl der für die lineare Diskriminanzanalyse genutzten spektralen Regionen eingesetzt.

Auch die Wahl der Parameter dieser beiden Programme ist Gegenstand dieser Arbeit.

3. Stand der Technik

In der Literatur sind spektroskopische Untersuchungen im mittleren Infrarot-Bereich über Tumorerkrankungen verschiedenster Gewebe angeführt, wie zum Beispiel der Lunge [3; 4], des Gebärmutterhalses [5–13], des Dickdarms [14–18], der Leber [19], der Haut [20–22], der Brust [23–30], der Prostata [31], des Mundes [32; 33], von Blutzellen [34; 35] und des Gehirns [36–39]. Diese Arbeiten unterscheiden sich insbesondere in der Anzahl analysierter Spektren und auch in den angewandten Auswertemethoden stark. Generell ist im Laufe der Zeit bei der Spektroskopie von Gewebeproben oder monozellularen Filmen eine Tendenz zu immer feiner ortsaufgelösten Spektren zu verzeichnen. Mit wachsendem Datenvolumen nimmt auch die Anwendung chemometrischer Auswertungen zu.

Grundsätzlich existieren zwei unterschiedliche Wege, die Daten so vorzubereiten, dass die eigentliche Klassifikation erfolgen kann. Zum einen können chemometrische Verfahren diese Vorbehandlung ohne Einbeziehung biochemischen Vorwissens übernehmen. Andererseits kann an die klassische Spektrenauswertung mit der Betrachtung von bestimmten Substanzklassen zugeordneten Banden und deren Parameter wie Intensitätsverhältnisse oder Lage des Maximums angeknüpft werden. Dieser Ansatz wird oftmals auf der Basis des Vorwissens über biochemische Veränderungen im Tumorgewebe gegangen, besonders dort, wo einzelne Spektren jeder Probe aufgenommen wurden. Meist werden Intensitätsverhältnisse einzelner Banden ausgewertet, vereinzelt wird auch die Lage der Banden in die Überlegungen mit einbezogen.

Während jedoch bestimmte Intensitätsveränderungen bei vielen unterschiedlichen Tumorarten auftreten, ist die Verschiebung der Bandenlage nicht immer signifikant [6], beziehungsweise nicht zur prädiktiven Klassifikation geeignet, da die Verteilungen zu stark überlappen [32]. Für die in dieser Arbeit betrachteten Tumore wurde teilweise eine signifikante Verschiebung der Banden beobachtet [38].

Eine Übersicht über die angewandten Auswertungsverfahren gibt Tab. 3.1.

Die gängige Diagnosemethode ist die histologische Untersuchung gefärbter Gewebeschnitte, die auch hier als Referenzmethode dient.

Mit den in der vorliegenden Arbeit genutzten Programmen wurden bereits Untersuchungen an Tumorgeweben durchgeführt [23, LDA[1]][21][14, NMR-Untersuchungen[2]]. Auch im Rahmen des Projekts „molekulare Endospektroskopie“ wurde die lineare Diskriminanzanalyse mittels `ga_ors` und `stackedGen` eingesetzt [36; 39]. Allerdings standen zu jenem Zeitpunkt erst wenige Proben zur Verfügung, so dass die Übertragbarkeit auf neue Proben noch nicht untersucht werden konnte. Allgemeine Vorschläge zur Datenvorbehandlung wurden noch nicht gegeben.

Der Einsatz genetischer Algorithmen zur Lösung von Optimierungsproblemen ist wohluntersucht [42–46]. In der Chemometrie werden genetische Algorithmen erfolgreich zur Variablenselektion angewandt [47–50]. Die genetischen Algorithmen sind in diesem Kontext auch insofern etabliert, als eine Reihe vergleichender Untersuchungen mit anderen

[1] lineare Diskriminanzanalyse
[2] Nuclear Magnetic Resonance

Tabelle 3.1.: In der Literatur auf Tumorproben angewandte Auswertungsverfahren

Verfahren	**Arbeiten**
klassische Auswertung	
Peakintensitätsverhältnisse, evt. Peaklage	[3; 4; 6; 8; 10; 13; 15; 16; 18; 20; 24; 26; 29; 32; 35; 38; 40]
chemometrische Verfahren	
Clusteranalyse	[22; 33]
Korrelationsanalyse	[11; 17]
principal component analysis	[7; 22]
principal component analysis / deskriptive Statistiken	[27; 28]
principal component analysis / soft independent modeling of class analogies	[37]
principal component analysis / logistische Regressionsanalyse	[28; 31]
forward subset selection / LDA	[23]
(NMR-Spektren)	[14]
principal component regression zur Klassifikation	[41]
partial least squares zur Klassifikation	[41]
principal component analysis / Diskriminanzanalyse (MAHALANOBIS-*Distanz*)	[9; 11; 30; 41]
genetic optimal region selection / LDA	[21; 36]
(NMR-Spektren)	[14]
principal component analysis / neuronale Netze	[12; 22]

Verfahren vorliegen (vgl. besonders [51–53]). Besondere Bedeutung kommt dabei den Methoden des *simulated annealing*, ebenfalls stochastischen Algorithmen, zu. Während die Wellenlängenselektion selbst auch für Verfahren, die auf vollständige Spektren angewandt werden können, als wichtig akzeptiert ist [47; 49; 54–60], ist unentschieden, ob genetische Algorithmen oder simulated annealing zur Lösung dieser Optimierungsaufgaben geeigneter sind. Die Leistungsfähigkeit der einzelnen Implementationen scheint stark von der mehr oder minder geschickten Wahl der Parameter abzuhängen [48; 51; 52].

Diskriminanzanalysen werden insbesondere auf der Basis von MAHALANOBIS-Distanzen [61–63] beschrieben. Aber auch das hier benutzte Programm zur linearen Diskriminanzanalyse, `stackedGen` , wurde zur Auswertung in einer Reihe von Veröffentlichungen unterschiedlichster Daten angewandt [14; 21; 36; 64]. Dieses Programmsystem ist allerdings noch in der Entwicklung begriffen, so dass die Beschreibungen bezüglich der Algorithmen nicht immer vollständig mit der derzeitigen Technik übereinstimmen.

4. Medizinische Grundlagen der betrachteten Hirntumore

4.1. Astrozytome und Glioblastome [65–68]

Die in dieser Arbeit betrachteten Hirntumore, Astrozytome ersten und zweiten Grades sowie Glioblastome (WHO-Grad IV), gehören zu den *Gliomen*. Es sind hirneigene Tumore, sie entstehen aus *Astrozyten* (Sternzellen) und können zu steigender Malignität entarten.

4.1.1. Die Makroglia

Das Ursprungsgewebe dieser Tumore ist die *Astro-* oder *Makroglia*, die zwischen den Nervenzellen und Blutgefäßen des Hirns lokalisiert ist und zum Stützgewebe gehört. Die Zellen der Makroglia umschließen die Kapillaren vollständig und haben daher für das Nervengewebe auch Nähr- und Phagozytenfunktion. Allerdings sind sie kein Bestandteil der Blut-Hirn-Schranke, diese Funktion übernimmt das *Endothel* der Blutgefäße. Die Astrocyten regeln den Wasser- und Ionengehalt des umgebenden Hirngewebes, besonders wichtig ist dabei die Aufnahme der bei der Impulsleitung und -übertragung aus den Nervenzellen ausgetretenen Kaliumionen.

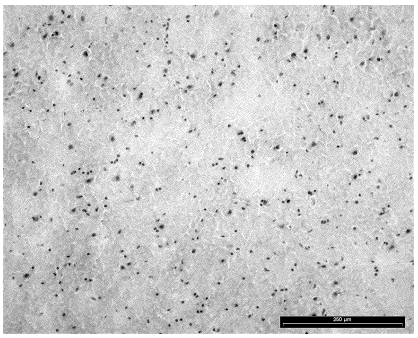


Abbildung 4.1.: Gesundes Gewebe — Der Messbalken kennzeichnet eine Länge von 250 μm

Astrocyten sind *fakultativ postmitotisch*, sie teilen sich normalerweise nicht, können aber unter bestimmten Bedingungen wieder in den Zellzyklus eintreten. Im reifen zentralen Nervensystem teilen sie sich nur unter pathologischen Bedingungen.

Zu den Astrozytomen zweiten Grades (*noch gutartig*, Abb. 4.2) gehören *fibrilläre*, *protoplasmatische* und *gemistozytäre Astrozytome*. Es handelt sich dabei um scharf begrenzte Raumforderungen, die in der Regel langsam wachsen, die anatomischen Strukturen des Hirns berücksichtigen und die Blut-Hirn-Schranke intakt lassen. Allerdings kommt auch *diffuses* und *infiltratives* Wachstum vor. Sowohl die Möglichkeit zur malignen Entartung als auch die Steigerung des Hirndrucks machen bereits bei diesen benignen Tumoren eine Behandlung notwendig.

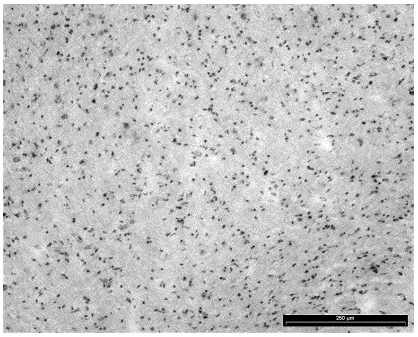


Abbildung 4.2.: Astrozytom zweiten Grades — Der Messbalken kennzeichnet eine Länge von 250 μm. Die Zellkerndichte ist gegenüber dem gesunden Gewebe erhöht.

Anaplastische Astrozytome (Abb. 4.3) entsprechen dem WHO-Grad III und sind also als *bereits bösartig* eingestuft. Charakteristisch sind Ödeme, schnelles Wachstum und Blutungen. *Glioblastome* sind bösartig, ihre Charakteristika sind neben den Merkmalen der anaplastischen Astrozytome auch Nekrosen und die ausgeprägte Polymorphie der Zellkerne (Abb. 4.4).

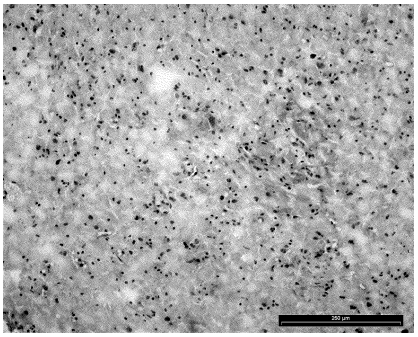


Abbildung 4.3.: Astrozytom dritten Grades — Der Messbalken kennzeichnet eine Länge von 250 μm. Weitere Zunahme des Zellkernanteils und der Größe der Zellkerne.

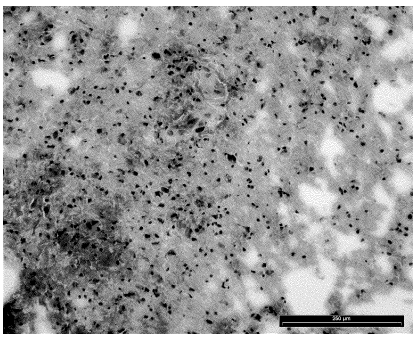


Abbildung 4.4.: Glioblastom — Der Messbalken kennzeichnet eine Länge von 250 μm. Das Gewebe ist sehr inhomogen, die Polymorphie der Zellkerne auffällig.

Die Behandlung dieser Tumore erfolgt in der Regel chirurgisch, oft in Kombination mit Bestrahlungen, eventuell auch mit unterstützender Chemotherapie. Insbesondere diffus wachsende Tumore bereiten Probleme. Im Hirn kann bei der Resektion nicht der sonst übliche Sicherheitsabstand von 3 – 5 cm gesund erscheinendem Gewebes eingehalten werden. Auch für scharf begrenzte Tumore resultiert deshalb eine besondere Notwendigkeit einer verlässlichen *in-vivo*-Diagnostik.

4.2. Überlegungen zur infrarot-spektroskopischen Untersuchung von Tumoren

4.2.1. Unterschiedliche Änderungen der Morphologie und der Spektren

Die histologische Begutachtung von Gewebeproben stützt sich auf die durch verschiedene Färbungen deutlich gemachte Morphologie des Gewebes, wohingegen die Infrarot-Spektroskopie als Schwingungsspektroskopie die chemische Zusammensetzung des Gewebes abbildet. Meist werden Spektren von — außer Trocknung — weitgehend unveränderten Gewebeproben aufgenommen. Allerdings wurden auch von gefärbten Proben Spektren angefertigt und ausgewertet[30].

Die histologische Einstufung einer Probe stützt sich also auf bereits erfolgte *makroskopische* Veränderungen, sowohl der Zellen als auch des Zellverbandes. Diese Veränderungen sollten sich auch in den IR-Spektren abzeichnen, da sie auch unterschiedliche Anteile der Substanzen beziehungsweise Substanzklassen in der betrachteten Probe bedeuten. Die beobachteten Veränderungen in den Spektren müssen jedoch nicht gleichzeitig mit den Veränderungen der Morphologie erfolgen. Im Gegenteil ist davon auszugehen, dass die biochemischen Veränderungen vor den morphologischen Änderungen erfolgen und sich daher auch die Spektren bereits verändern, wenn noch keine — oder nur geringe — morphologischen Veränderungen zu erkennen sind [5; 69]. Dies betrifft in besonderem Maße die hier untersuchten Tumore, da verschiedene Tumorgrade einer Tumorreihe betrachtet werden.

Bei der Klassifikation der Spektren ist also zu beachten, dass die Einstufung der Proben auf morphologischen Veränderungen beruht. Daher existieren eventuell für die spektroskopische Klassifikation geeignetere Grenzen zwischen den Tumoren.

4.2.2. Der Einfluss der örtlichen Auflösung

Die Größe der Astrozyten liegt im Bereich weniger Mikrometer (2 – 5 μm) und damit in der Größenordnung eines Pixels eines unter dem Mikroskop aufgenommenen Images, damit erfasst ein solches Spektrum nur wenige Zellen. Ein Spektrum eines Maps entstammt einer Fläche von 90 $\mu m \times 90\ \mu m$, daher tragen hier hunderte Zellen zu der Beobachtung bei.

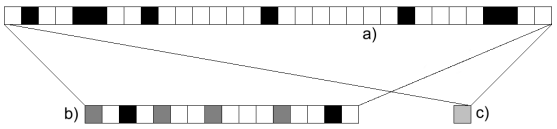


Abbildung 4.5.: Auswirkungen unterschiedlicher Ortsauflösung — a) zugrundeliegende Struktur b) hohe Auflösung c) niedrige Auflösung

Inhomogenitäten werden stark abgebildet, wenn die Größe der homogenen Bereiche innerhalb der Probe der Ortsauflösung ähnlich ist (Abb. 4.5). Eventuell wird dann eine andere Auswertungsstrategie notwendig, die die Häufigkeiten der gefundenen Merkmale in Betracht zieht. Eine niedrige Ortsauflösung bewirkt Mittelwertbildung über große Flächen.

Unter der Annahme, dass es sich um recht homogene Gewebe handelt, also alle Zellen ähnliche Spektren aufweisen, ist eine veränderte Ortsauflösung unproblematisch. Dann sollten die Spektren unabhängig von der Aufnahmetechnik sein. Allerdings sind insbesondere die höhergradigen Tumore durch starke Inhomogenitäten gekennzeichnet.

Deshalb ist zu erwarten, dass die einzelnen Spektren eines Images nicht nur aufgrund des etwas geringeren Signal-Rausch-Verhältnisses [70], sondern auch aufgrund der Inhomogenität der Probe eine deutlich größere Varianz aufweisen als die Spektren der Maps.

Diese vergrößerte Varianz der Spektren kann die Auswertung der Daten deutlich erschweren, so kann es notwendig werden, weitere spektrale Muster zu definieren. Dann ergibt sich natürlich auch die Frage des biochemischen Hintergrunds dieser neuen Klassen und nach der Zuordnung der Spektren zu den Klassen. Einleuchtend ist dieses Vorgehen zum Beispiel für die Glioblastome, die auch morphologisch in nekröse Bereiche und lebende Tumorzellen unterteilt werden können.

Die histologische Einordnung der Proben wird als „Gold-Standard“ zur Modellbildung genutzt. Um die Eignung einer Probe für die Modellbildung bei der Diskriminanzanalyse zu beurteilen, ist jedoch zusätzliches Wissen notwendig. Die Notwendigkeit genauerer Einordnungen der Spektren tritt natürlich weiter in den Vordergrund, wenn weitere Klassen gebildet werden sollen.

Die örtliche Genauigkeit der Erkennung der Grenzen zwischen gesundem und Tumorgewebe hängt natürlich von der Ortsauflösung der Daten ab, insofern ist eine möglichst hohe Ortsauflösung wünschenswert [25]. Verschiedene Faktoren können allerdings bewirken, dass diese Grenze auch bei gesteigerter Ortsauflösung nicht genauer erkannt oder sogar die Erkennung des Tumors erschwert werden kann.

Die Frage nach einer günstigen Ortsauflösung kann nur dann beantwortet werden, wenn die biochemische und örtliche Herkunft der zur Unterscheidung der Gewebearten genutzten spektralen Regionen bekannt ist. Hierbei besteht sowohl die Möglichkeit, dass es sich um Merkmale, die ausschließlich einzelnen Tumorzellen zuzuordnen sind, handelt, als auch die Möglichkeit, dass es sich um Merkmale ganzer Zellverbände handelt. Letzteres umfasst die Möglichkeit, Tumormarker, die auch im umgebenden Gewebe gefunden werden können, oder auch die An- oder Abwesenheit bestimmter Zelltypen im Zellverband (vgl. [32]) zu nutzen.

Auch die zeitlichen Veränderungen in lebendem Gewebe müssen bei dieser Diskussion in Betracht gezogen werden, da die unterschiedlichen Stadien des Zellzyklus starke Veränderungen im biochemischen Sinne bedeuten, die sich insbesondere auch deutlich in den Infrarot-Spektren der Zellen niederschlagen. In der Literatur wird die Hypothese aufgestellt, dass die Desoxyribonukleinsäure (DNA) praktisch nur während der Synthese-Phase sichtbar ist, da sie sonst aufgrund der dichten Packung eine zu hohe optische Dichte aufweist, um in den Spektren beobachtet zu werden. Daher wird in der Literatur geschlossen, dass bei der Betrachtung des DNA-Anteils oft eher die Teilungsaktivität des Zellverbands als der tatsächliche DNA-Gehalt gemessen wird. [20; 34; 40]

Es kann also nicht von vornherein davon ausgegangen werden, dass die Ergebnisse unabhängig von der Ortsauflösung sind, eventuell werden für unterschiedliche Ortsauflösungen auch sehr unterschiedliche Modelle gebildet, weil andere Merkmale zur Klassifikation genutzt werden.

Teil II.

Grundlagen

5. Begriffsbestimmungen zu den Datenstrukturen

5.1. Die Datenstruktur

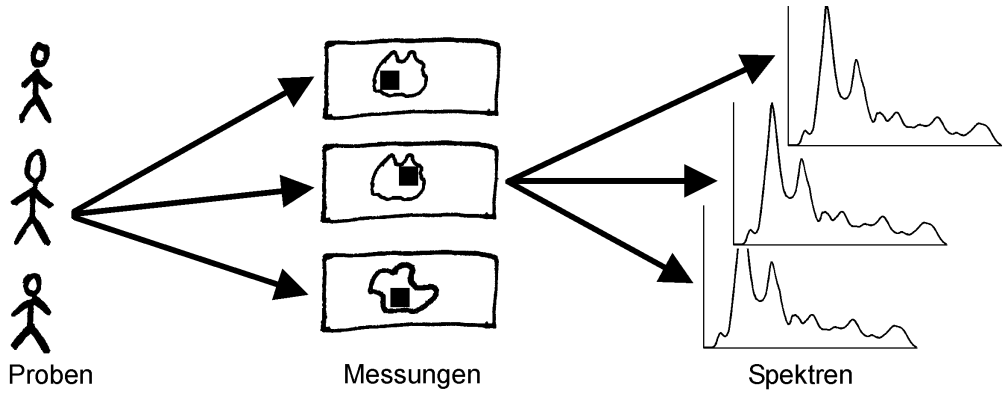


Abbildung 5.1.: Illustration der Datenstruktur — Zu den Proben verschiedener Patienten können mehrere Messungen existieren, die jeweils viele Spektren enthalten.

Die einzelnen Gewebeproben stammen von verschiedenen Patienten. Von diesen Proben werden Mikrotomschnitte angefertigt, die infrarotspektroskopisch gemessen werden. Jede solche Messung besteht aus vielen einzelnen Spektren.

Ein Map ist eine Messung, deren Spektren nacheinander von verschiedenen Positionen der Probe aufgenommen wurden, dagegen werden die einzelnen Spektren eines Images gleichzeitig gemessen.

5.2. Der Ablauf der angewendeten chemometrischen Untersuchungen

Die in dieser Arbeit angewandte Methode zur Klassifikation der Spektren umfasst verschiedene Fachgebiete mit jeweils eigenen Begriffen.

Die Daten bestehen aus den Extinktionswerten verschiedener Wellenzahlbereiche (Abb. 5.2). Die Anzahl dieser Wellenzahlbereiche wird in dieser Arbeit mit p bezeichnet, sie ist eine Schlüsselgröße und wird in der Spektroskopie, der Optimierung und der Klassifikation jeweils anders benannt.

Aus spektroskopischer Sicht werden aus Spektren, bestehend aus den Extinktionswerten verschiedener Wellenzahlen, Mittelwerte einzelner spektraler Regionen gebildet. Dabei bleibt die physikalische Bedeutung erhalten. Die so entstandenen Daten sind also im Prinzip immer noch Spektren.

Die Optimierung versteht p als eine Anzahl an Dimensionen. Vor der Optimierung liegen die Daten in einem hochdimensionalen Suchraum vor, die Spektren werden als Punkte in $\mathbb{R}^p$ aufgefasst. Nach der Optimierung ist die Dimensionalität stark reduziert, da nur die *wichtigen* Dimensionen beibehalten wurden.

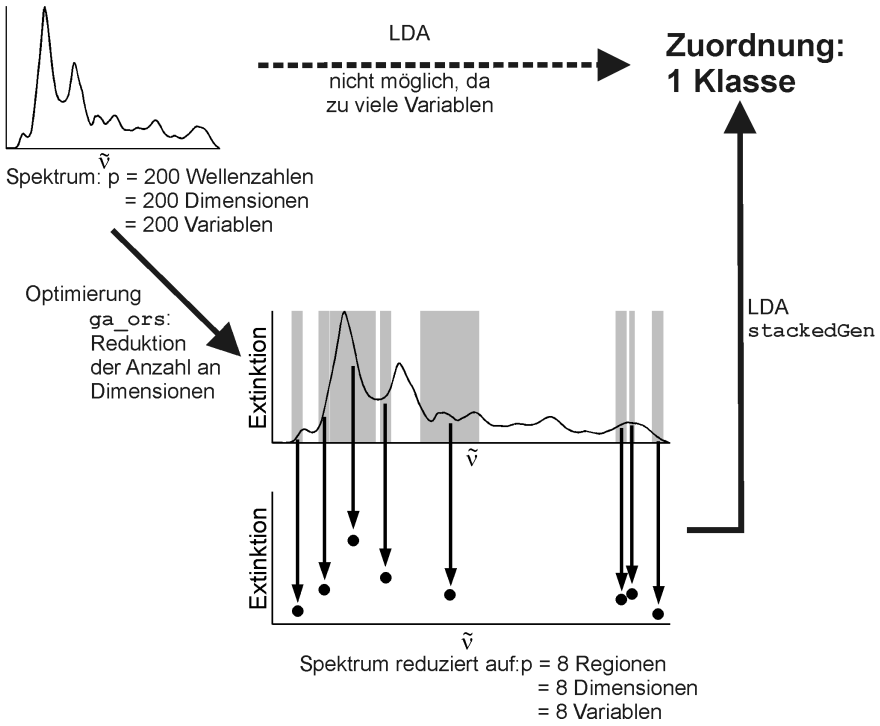


Abbildung 5.2.: Illustration des Ablaufs der Analyse — Die einzelnen Spektren können erst nach einer Dimensionsreduktion klassifiziert werden.

In der Terminologie der Klassifikation stellt p die Anzahl der Variablen dar. Vor der Optimierung ist die Anzahl der Variablen zu groß und ihre Trennfähigkeit zu gering, als dass eine lineare Diskriminanzanalyse erfolgreich durchgeführt werden könnte. Aus Sicht der Diskriminanzanalyse wählt die Optimierung die bedeutsamen Variablen aus einer großen Menge an vorhandenen Variablen aus.

6. Klassifikation[71–74]

Unter Klassifikation versteht man die Zuordnung einzelner Objekte zu Klassen, deren Existenz in der Regel bereits bekannt ist. Daher sind die Klassifikationsverfahren dem *überwachten Lernen* zuzuordnen. Methoden des *unüberwachten Lernens*, die ebenfalls Objekte gruppieren, wie zum Beispiel die Clusteranalyse, werten nie Informationen über bekannte Gruppen aus. Daher können sie im Rahmen von Klassifikationsproblemen zwar zur explorativen Datenanalyse und dadurch eventuell zur Klärung der Ursache bei auftretenden Problemen eingesetzt werden, eine Klassifikation im geforderten Sinn können sie jedoch nicht leisten.

Vereinzelt werden auch Verfahren wie die *principal component regression* und *partial least squares* zur Klassifikation eingesetzt. Dies ist in der Regel allerdings nur dann möglich, wenn zwischen zwei Klassen zu unterscheiden ist.

Bekannte Methoden zur Behandlung von Klassifikationsproblemen sind außer der *Diskriminanzanalyse* die *Methode der k nächsten Nachbarn* und das *soft independent modeling of class analogies (SIMCA)*. Auch bestimmte *Neuronale Netze* und *Heuristiken* wie *Expertensysteme* können auf diese Aufgabenstellungen angewandt werden. Die einzelnen Verfahren unterscheiden sich stark in ihren Voraussetzungen, mathematischen Modellen und Annahmen.

Die lineare Diskriminanzanalyse ist als besonders robustes und schnelles Verfahren bekannt [23; 34; 76; 77]. Ein weiterer wichtiger Vorteil der Diskriminanzanalyse ist, dass außer der Klassenzugehörigkeit noch Vorwissen über die Häufigkeit des Auftretens der Klassen eingebracht und eine Gewichtung der Fehlklassifikationen einfach realisiert werden können [41].

Von den vorliegenden Gewebeproben wurden ortsaufgelöst IR-Spektren aufgenommen, entweder als Maps in Form vieler Einzelmessungen oder simultan als Images. Die Klassifikation kann sich also auf zwei wesentliche Informationsarten stützen: zum einen die einzelnen Spektren, die die molekulare Zusammensetzung des Gewebes und damit auch die Gewebeart widerspiegeln und zum anderen die räumliche Information.

Die hier erfolgte Klassifikation stützt sich allein auf die Information der Spektren, die räumliche Information wird nicht ausgewertet. Bei der Interpretation und Beurteilung der Klassifikationsergebnisse kann jedoch die räumliche Information wieder hinzugezogen werden. Dabei gilt es zu beachten, dass eine Probe gesundes Gewebe und verschiedene Tumorgewebe enthalten kann, da die verschiedenen hier untersuchten Tumorarten auseinander hervorgehen können. Dies sollte sich dann auch in der Zuordnung der einzelnen Spektren widerspiegeln.

Ortsaufgelöste Informationen spielen bei der traditionellen histologischen Begutachtung eine große Rolle. Der Histologe schließt aus *Formen* und *Größen* auf den Gewebetyp. Die Färbung dient eher dem Verstärken des Kontrastes zum *Sichtbar machen* der vorhandenen Strukturen, als dass eine farbliche Differenzierung zwischen verschiedenen Gewebearten erreicht würde.

Grundsätzlich sind zur Klassifikation verschiedene mögliche Anordnungen der einzelnen Objekte im Raum zu berücksichtigen, da sie unterschiedliche Ansätze der Klassifikati-

on beziehungsweise der Datenvorbehandlung vor der eigentlichen Analyse erfordern. So beschreibt [74] folgende wichtige Fälle (Abb. 6.1):

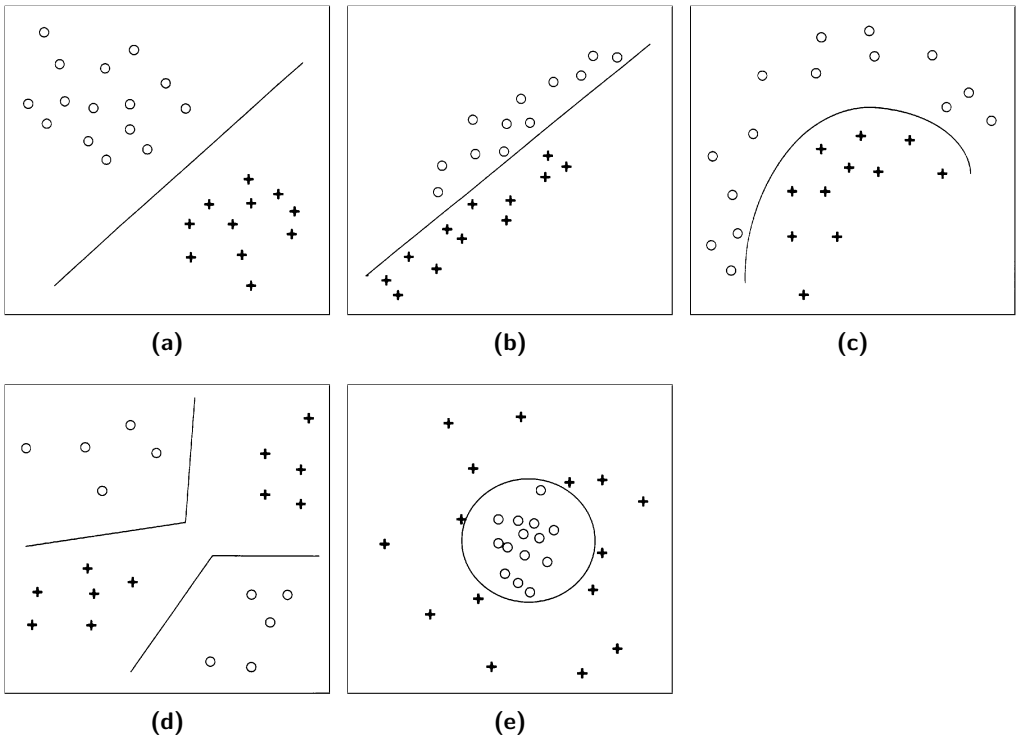


Abbildung 6.1.: Wichtige Fälle der Objektanordnungen [74]

(a) Die Trennung der Gruppen ist bereits mit einer linearen Funktion möglich, der euklidische Abstand vom Gruppenmittelpunkt liefert die richtige Klassifikation.

(b) Die Variablen sind korreliert, so dass die Klassifikation mittels euklidischem Abstand nicht mehr möglich ist, wohl aber mit Hilfe der MAHALANOBIS-*Distanz*. Auch eine lineare Funktion trennt die Gruppen richtig.

(c) Hier reicht auch eine lineare Funktion nicht mehr zur Separierung der Gruppen aus, es muss eine gekrümmte Diskriminanzfunktion verwendet werden. Alternativ können die Daten transformiert werden, so dass wieder eine lineare Funktion zur Trennung ausreicht.

(d) Die einzelnen Gruppen bestehen aus verschiedenen Untergruppen, daher sind mehrere Diskriminanzfunktionen zur korrekten Trennung erforderlich.

(e) Der Asymmetrische Fall ist typisch für Aufgabenstellungen der Qualitätssicherung: eine Gruppe bildet eine Enklave innerhalb einer anderen Gruppe. Es muss mit einer geschlossenen Diskriminanzkurve gearbeitet werden, unter Umständen kann auch der euklidische Abstand vom Gruppenmittelpunkt zur Klassifikation herangezogen werden, nun allerdings in der Form eines Grenzwerts.

Der hier betrachtete Ansatz der *linearen* Diskriminanzanalyse kann in den Situationen entsprechend (a) und (b) angewandt werden, andere Objektanordnungen können jedoch in solche Anordnungen transformiert werden. Die Vielzahl der Messstellen eines IR-Spektrums stellt meist genügend Merkmale mit den geforderten Verteilungen zur Verfügung, deshalb kann auf die Anwendung entsprechender Transformationen verzichtet werden.

6.1. Diskriminanzanalyse [71–74; 78]

Die Diskriminanzanalyse geht von der Annahme aus, dass die *Grundgesamtheit* Ω aus g *disjunkten*, also trennbaren, Teilgesamtheiten $\Omega_1 \dots \Omega_g$, den *Klassen*, besteht.

Ausgehend von den Merkmalsausprägungen $\mathbf{x} \in \mathbb{R}^p$ der einzelnen Objekte $\omega \in \Omega_k$ werden Entscheidungskriterien e zur Zuordnung der Objekte zu den Klassen gesucht. Das heißt, dass die Entscheidungskriterien den aus allen Merkmalsausprägungen $\mathbf{x}$ der Stichprobe gebildeten *Stichprobenraum* $\mathbb{S} \subset \mathbb{R}^p$ auf die Klassenzugehörigkeit k abbilden:

$$\mathrm{e} : \mathbb{S} \to \{1, .., g\} \tag{6.1}$$

$$\mathbf{x} \mapsto \mathrm{e}(\mathbf{x}) = \hat{k} \tag{6.2}$$

Eventuell wird noch $\hat{k} = 0$ verwendet, wenn ein Objekt *keiner* der bekannten Klassen zugeordnet werden kann.

Zu Klassenzugehörigkeit k und Merkmalsausprägung $\mathbf{x}$ eines Objekts existieren folgende wichtige charakteristische Größen:

$\mathrm{p}(k) = \mathrm{p}(\omega \in \Omega_k) > 0$, die *a priori Wahrscheinlichkeit* der Klassenzugehörigkeit.
Die a priori Wahrscheinlichkeit gibt die relative Häufigkeit der einzelnen Klassen an.

$\mathrm{f}(\mathbf{x}|k)$, die Klassenverteilung von $\mathbf{x}$ in Ω_k
$\mathrm{f}(\mathbf{x}|k)$ ist also die Verteilung der Objekte einer Klasse und gibt die Wahrscheinlichkeit an, den Merkmalsvektor $\mathbf{x}$ für ein Objekt der Klasse k zu beobachten.

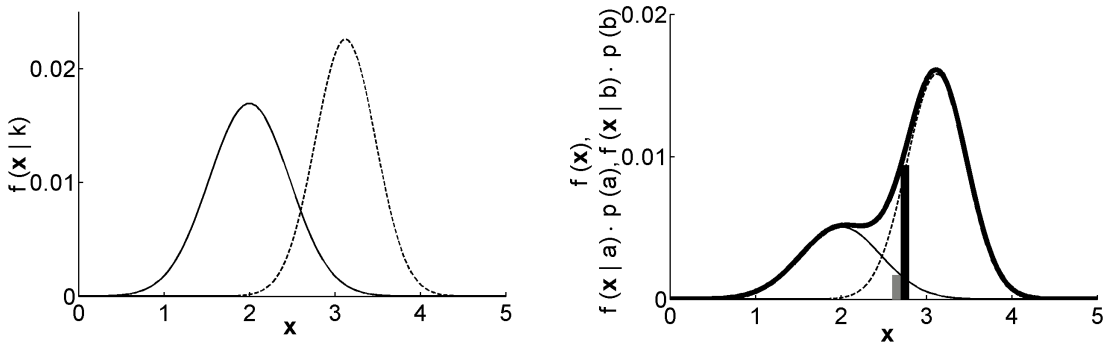
$\mathrm{f}(\mathbf{x}) = \sum_{k=1}^{g} \mathrm{p}(k)\,\mathrm{f}(\mathbf{x}|k)$, die *unbedingte* Verteilung von $\mathbf{x}$ auf Ω.
Das ist die Verteilung aller Objekte aller Klassen, also der Grundgesamtheit.

$\mathrm{p}(k|\mathbf{x})$, die *a posteriori Wahrscheinlichkeit* der Klassenzugehörigkeit,
also die Wahrscheinlichkeit, dass ein Objekt mit dem Merkmalsvektor $\mathbf{x}$ zur Klasse k gehört oder der Anteil aller Beobachtungen $\mathbf{x}$, der von Objekten der Klasse k herrührt.
Nach Bayes gilt

$$\mathrm{p}(k|\mathbf{x}) = \frac{\mathrm{p}(k)\,\mathrm{f}(\mathbf{x}|k)}{\mathrm{f}(\mathbf{x})} \tag{6.3}$$

Abb. 6.2 illustriert diese Größen für folgendes Beispiel:

Abb. 6.2(a) zeigt die Klassenverteilungen $\mathrm{f}(\mathbf{x}|a)$ (durchgezogen) und $\mathrm{f}(\mathbf{x}|b)$ (gestrichelt) zweier Klassen. Sind die a priori Wahrscheinlichkeiten dieser Klassen $\mathrm{p}(a) = 30\%$ und $\mathrm{p}(b) = 70\%$, dann ist $\mathrm{f}(\mathbf{x}) = \mathrm{p}(a) \cdot \mathrm{f}(\mathbf{x}|a) + \mathrm{p}(b) \cdot \mathrm{f}(\mathbf{x}|b)$, die



(a) Klassenverteilungen — Klasse a (durchgezogen), Klasse b (gestrichelt)

(b) Zusammensetzung der unbedingten Verteilung der Merkmalsvektoren $\mathrm{f}(\mathbf{x})$ (fett). Das Verhältnis des grauen zum schwarzen Balken ist die a posteriori Wahrscheinlichkeit

Abbildung 6.2.: Skizze zu den charakteristischen Größen der Verteilungen

in (b) fett gekennzeichnete Verteilung.
Für ein Objekt mit $\mathbf{x} = 2,7$ ist die Wahrscheinlichkeit, dass das Objekt zur Klasse a gehört

$$\mathrm{p}\,(a|\mathbf{x} = 2,7) = \frac{0,3 \cdot \mathrm{f}\,(2,7|a)}{0,3 \cdot \mathrm{f}\,(2,7|a) + 0,7 \cdot \mathrm{f}\,(2,7|b)} \approx 20\ \%.$$

Das Verhältnis des kurzen grauen Balkens $(0,3 \cdot \mathrm{f}\,(2.7|a))$ zum langen schwarzen Balken $(\mathrm{f}\,(2,7))$ in (b) illustriert diese Wahrscheinlichkeit.

Diese Größen sind in der Regel unbekannt und werden daher mit Hilfe einer Lernstichprobe geschätzt.

Weiterhin wird nach dem Ziel der Analyse unterschieden. Die *deskriptive* Diskriminanzanalyse benutzt die gefundenen Entscheidungskriterien, um die Merkmale herauszufinden, in denen sich die Klassen unterscheiden. Dagegen wendet die *prädiktive* Diskriminanzanalyse die aus einer Lernstichprobe geschätzten Entscheidungskriterien auf weitere Stichproben an, um deren Objekte zu den bekannten Klassen zuzuordnen.

Die eingesetzten Verfahren sollten an dieses Ziel angepasst sein [78]. Auch die Kriterien zur Beurteilung der Modellgüte sollten daher das Ziel der Analyse berücksichtigen. Im Folgenden wird gemäß den Zielen dieser Arbeit, der Klassifikation weiterer Infrarotspektren, speziell die prädiktive Diskriminanzanalyse betrachtet.

6.1.1. Entscheidungsregeln

Für verschiedene Aufgabenstellungen existieren unterschiedliche *Entscheidungsregeln*, die der jeweiligen Situation angepasst sind.

Ein unbekanntes Objekt wird nach Bayes derjenigen Klasse zugeordnet, für die die a *posteriori Wahrscheinlichkeit* am größten ist:

$$\mathrm{p}\left(\hat{k}|\mathbf{x}\right) \geq \mathrm{p}\,(l|\mathbf{x}) \qquad \forall\, l \in \{1,\ldots,g\} \tag{6.4}$$

$$\mathrm{p}\left(\hat{k}\right)\mathrm{f}\left(\mathbf{x}|\hat{k}\right) \geq \mathrm{p}\,(l)\,\mathrm{f}\,(\mathbf{x}|l) \qquad \forall\, l \in \{1,\ldots,g\} \tag{6.5}$$

Diese Zuordnung ist nicht zwingend eindeutig. Die BAYES-Regel ist in der Hinsicht optimal, als sie für jedes $\mathbf{x}$ die kleinste bedingte Fehlerrate $\mathrm{F}\,(e|\mathbf{x})$ und damit auch die kleinste *unbedingte* Fehlerrate $\mathrm{F}\,(\mathbf{x})$ erreicht.

Eine Erweiterung sind die *kostenoptimalen* Entscheidungsregeln. Dabei wird eine *Kostenfunktion* angewandt, die es ermöglicht, die verschiedenen Arten der Fehlklassifikation unterschiedlich zu gewichten. Dazu wird eine Matrix $\mathbf{C} = \mathbf{C}\left(k, \hat{k}\right)$, deren Hauptdiagonalelemente null sind, eingeführt. Die einzelnen Elemente $c_{k\hat{k}}$ bedeuten dabei die Kosten der Fehlzuordnung, wenn die wahre Klassenzugehörigkeit k ist, das Objekt aber der Klasse $\hat{k}$ zugeordnet wird. Statt der bedingten Fehlerraten sind nun die zu erwartenden Kosten zu minimieren. Dabei gilt für die *bedingten erwarteten Kosten*

$$\mathrm{C}\left(\hat{k}|\mathbf{x}\right) = \sum_{k=1}^{g} \mathbf{C}\left(k, \hat{k}\right) \mathrm{p}\,(k|\mathbf{x}) \rightarrow \min_{\forall k} \tag{6.6}$$

Die *unbedingten erwarteten Kosten* für jede Klasse erhält man durch Integration über den gesamten Merkmalsraum $\mathbb{S}$:

$$\mathrm{C}\left(\hat{k}\right) = \int_{\mathbb{S}} \mathrm{C}\left(\hat{k}|\mathbf{x}\right) d\mathbf{x}. \tag{6.7}$$

Die BAYES-Regel (Gl. 6.4) sowie die aus der Anwendung der kostenoptimalen Regel folgenden analogen Ungleichungen können durch streng monoton steigende Transformationsfunktionen in besser handhabbare Formen überführt werden. So führt logarithmieren der BAYES-Regel zur äquivalenten Darstellung

$$\ln \mathrm{f}\left(\mathbf{x}|\hat{k}\right) + \ln \mathrm{p}\left(\hat{k}\right) \rightarrow \max_{\forall k} \tag{6.8}$$

Weitere Vereinfachungen dieser Regeln sind möglich, wenn die Verteilung der Objekte bekannt ist. Die wichtigsten Vereinfachungen betreffen die multivariate Normalverteilung.

6.2. Lineare Diskriminanzanalyse

Man spricht von einer *linearen* Diskriminanzanalyse (LDA), wenn die Diskriminanzfunktionen linear bezüglich der Merkmalsvektoren $\mathbf{x}$ sind.

Damit die Regeln nach dem entscheidungstheoretischen Ansatz angewandt werden können, muss die *bekannte* Verteilung der Daten eingesetzt werden. In der Regel geht man hier vom Vorliegen einer *multivariaten Normalverteilung* aus.

Mit der Dichtefunktion

$$\mathrm{f}\,(\mathbf{x}|k) = \frac{1}{\sqrt{(2\pi)^p}\sqrt{|\mathbf{S}_k|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^T \mathbf{S}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)} \tag{6.9}$$

erhält man für die BAYES-Regel (Gl. 6.8)

$$-\frac{1}{2}\ln|\mathbf{S}_k| - \frac{1}{2}\left(\mathbf{x}-\boldsymbol{\mu}_k\right)^T \mathbf{S}_k^{-1}\left(\mathbf{x}-\boldsymbol{\mu}_k\right) + \ln \mathrm{p}\left(\hat{k}\right) \rightarrow \max_{\forall k} \tag{6.10}$$

Dabei wurde ausgenutzt, dass die bezüglich der Gruppenzugehörigkeit k konstanten Terme weggelassen werden können, ohne die zugrundeliegende Ungleichung zu verletzen.

Die so erhaltene Diskriminanzfunktion ist zunächst quadratisch in $\mathbf{x}$, man spricht daher von einer *quadratischen* Diskriminanzanalyse.

Wenn allerdings die Kovarianzmatrizen aller Gruppen gleich sind, also

$$\mathbf{S}_1 = \mathbf{S}_2 = \cdots = \mathbf{S}_g = \mathbf{S} \tag{6.11}$$

gilt, kann weiter vereinfacht werden. Nach Ausmultiplizieren der quadratischen Form (6.10) resultiert eine in $\mathbf{x}$ lineare Diskriminanzfunktion

$$\mathbf{x}^T\mathbf{S}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k^T\mathbf{S}^{-1}\boldsymbol{\mu}_k + \ln \mathrm{p}\left(\hat{k}\right) \rightarrow \max_{\forall k} \tag{6.12}$$

Der analoge Ausdruck bei Anwendung der kostenoptimalen Regel unterscheidet sich nur im Auftreten eines Summanden für die erwarteten Kosten der Fehlklassifikationen.

6.3. Voraussetzungen der linearen Diskriminanzanalyse

Die lineare Diskriminanzanalyse nach dem entscheidungstheoretischen Ansatz beruht also auf zwei wichtigen Annahmen

- Es wird eine multivariate Normalverteilung vorausgesetzt.
- Zusätzlich müssen die Kovarianzmatrizen $\mathbf{S}_k$ aller Klassen gleich sein.
  Ist dies nicht der Fall, so kann eine *quadratische* Diskriminanzanalyse durchgeführt werden.

6.3.1. Testen der Voraussetzungen

Zur Überprüfung dieser Voraussetzungen sind eine Reihe statistischer Tests bekannt, die allerdings in der Regel einen recht großen Strichprobenumfang erfordern. Daher ist es in der Praxis oft schwierig, diese Tests durchzuführen.

Andererseits erweist sich die lineare Diskrimianzanalyse empirisch als ein sehr robustes Verfahren, wobei auch Erfahrungen darüber vorliegen, welche Abweichungen starke Beeinträchtigungen der Leistungsfähigkeit der Diskriminanzanalyse nach sich ziehen und welche Verletzungen der Voraussetzungen als eher unkritisch zu werten sind.

Hinzu kommt, dass oftmals auch kein notwendig besseres Verfahren angegeben werden kann, da die alternativen Methoden in der Regel komplexer sind und damit größere Unsicherheiten in den Schätzungen resultieren.

Die Voraussetzung der multivariaten Normalverteilung

Verschiedene Tests auf Vorliegen einer multivariaten Normalverteilung sind bekannt. Diese Tests haben recht unterschiedliche Eigenschaften und sind zum Teil auf die Erfordernisse der linearen Diskriminanzanalyse abgestimmt [79, Kap. 6.2.2 und 6.2.3].

Diese Tests, die die Aussage, dass keine Abweichung — beziehungsweise keine für die lineare Diskriminanzanalyse problematische Abweichung — von der multivariaten Normalverteilung nachgewiesen werden kann, ermöglichen, sind recht aufwändig. Die Aussage, dass *keine* multivariate Normalverteilung vorliegt, ist dagegen oft recht einfach zu treffen, da einfach zu prüfende notwendige, aber nicht hinreichende, Bedingungen bekannt sind.

univariate Normalverteilung aller Linearkombinationen $\mathbf{a}^T\mathbf{x}$

Aus der Definition der *multivariaten Normalverteilung* folgt, dass alle $\mathbf{a}^T\mathbf{x}$ univariat normalverteilt sein müssen. Das sind auch die $x_1, \ldots, x_p$ selbst. Kann also eine Linearkombination $\mathbf{a}^T\mathbf{x}$ angegeben werden, die *nicht* univariat normalverteilt ist, so kann keine multivariate Normalverteilung vorliegen.

Es ist *nicht hinreichend* für das Vorliegen einer multivariaten Normalverteilung, dass alle $x_1, \ldots, x_p$ univariat normalverteilt sind, aber es handelt sich um eine *notwendige* Bedingung.

Damit kann die Verletzung dieser ersten Voraussetzung der linearen Diskriminanzanalyse eventuell durch mehrere univariate Tests auf Normalverteilung (z. B. χ^2-Test oder Kolmogorow-Smirnow-Test) gezeigt werden.

Homogenität der Kovarianzmatrizen

Um die *Homoskedastizität*, die Homogenität der Kovarianzmatrizen, zu testen, stehen verschiedene Tests auf der Basis von χ^2- und F-Statistiken zur Verfügung [78, S. 69 – 70][71, S. 74 – 75][80], dabei wird

$$H_0: \quad \mathbf{S}_1 = \mathbf{S}_2 = \cdots = \mathbf{S}_k \tag{6.13}$$

$$\text{gegen} \quad H_1: \quad H_0 \text{ falsch} \tag{6.14}$$

getestet.

Als Testgröße wird das Boxsche M verwendet, das unter H_0 asymptotisch χ^2-verteilt ist.

$$M = \sum_{k=1}^{g} n_k \ln \left| \hat{\mathbf{S}}_k^{-1} \hat{\mathbf{S}} \right| \tag{6.15}$$

H_0 ist abzulehnen, wenn

$$M > \chi^2(P = 1 - \alpha; f = \frac{1}{2} p(p+1)(g-1)) \tag{6.16}$$

ist. Es existieren weitere Transformationen, die die χ^2- beziehungsweise die F-Statistik approximieren, insbesondere bessere Näherungen für bestimmte Situationen geben.

Diese Tests sind bereits gegenüber kleinen Abweichungen der Kovarianzmatrizen voneinander recht empfindlich. Daher sollte eine Ausreißererkennung *vor* dem Test auf Homogenität der Kovarianzmatrizen durchgeführt werden [78, S. 64].

Zusätzlich werden diese Statistiken mit wachsendem $\frac{n_k}{p}$ immer schärfer, so dass die Verwendung kleiner Irrtumswahrscheinlichkeiten empfohlen wird [78, S. 64].

Ein weiteres Problem dieser Test-Statistiken ist, dass sie sehr empfindlich gegenüber Abweichungen von der multivariaten Normalverteilung sind. Daher geht Ablehnung der Nullhypothese oft eher auf Abweichungen von der Normalverteilung als auf unterschiedliche Kovarianzmatrizen zurück.

6.3.2. Konsequenzen von Verletzungen der Voraussetzungen

Die lineare Diskriminanzanalyse ist als ein gegenüber Verletzungen der Voraussetzungen als sehr robustes Verfahren bekannt.

Allerdings sind bestimmte Abweichungen von der multivariaten Normalverteilung bekannt dafür, dass sie die Leistung, speziell die erreichbare Trefferrate, der linearen Diskriminanzanalyse stark beeinträchtigen.

Heteroskedastizität hat in vielen Fällen wesentlich weniger gravierende Folgen für die Diskriminanzanalyse. Zum einen, weil dieser Tatsache recht einfach Rechnung getragen werden kann, indem von der linearen auf eine quadratische Diskriminanzanalyse ausgewichen wird. Die erforderlichen Verfahren sind ebenfalls bekannt und stehen in der Regel zur Verfügung. Andererseits wird bei geringen Stichprobenumfängen oft empfohlen, trotz inhomogener Kovarianzmatrizen den linearen Ansatz zu verwenden. Die Abweichungen haben meist geringere Auswirkungen auf die Trefferrate als die gestiegene Komplexität und die damit bei gleicher Probensituation gestiegene Unsicherheit bei der Parameterschätzung bei einer quadratischen Diskriminanzanalyse. Allerdings sind diese Untersuchen bislang auf den Zwei-Gruppen-Fall beschränkt.

Erfahrungen legen für viele Fälle nahe, dass es sich bei der linearen Diskriminanzanalyse um ein robustes Verfahren handelt, das auch bei Verletzungen der Voraussetzungen angewandt werden kann.

Die vorliegende Situation weicht deutlich von den in der allgemeinen Literatur zur Mustererkennung üblichen Fragestellungen ab. Dort handelt es sich meist um Probleme, bei denen eine nicht zu große Anzahl an Variablen zur Analyse zur Verfügung steht, aus denen dann eine geeignete Teilmenge ausgewählt werden muss. Die spektroskopischen Daten stellen jedoch eine sehr große Anzahl an Variablen — bei denen allerdings von starken Korrelationen untereinander ausgegangen werden muss — zur Verfügung. Diese Variablen unterscheiden sich auch insofern von den oben genannten, als sie zwar als lauter einzelne Dimensionen aufgefasst werden können, diese Dimensionen jedoch erst durch die Diskretisierung des zunächst stetigen Spektrums entstehen. Daher können die Extinktionswerte benachbarter Wellenzahlen addiert werden, ohne dass die physikalische Bedeutung verlorengeht.

6.4. Die optimale Variablenwahl

Die lineare Diskriminanzanalyse ist in ihren Ergebnissen stark von den verwendeten Variablen abhängig. Zunächst ist aus theoretischer Sicht festzuhalten, dass weiter hinzukommende *wahre* Größen schlimmstenfalls nicht zur Klassentrennung beitragen, die Trennung der Klassen jedoch nicht negativ beeinflussen.

Anders sieht es aus, wenn — wie es in der Regel der Fall ist — die *wahren* Größen unbekannt sind und daher Schätzungen verwendet werden müssen. Jede hinzukommende Variable bedeutet zusätzliche Komplexität des Modells und die Schätzungen werden daher mit einer größeren Unsicherheit behaftet sein.

Übersteigt diese durch die größere Anzahl an Schätzungen hinzugekommene Unsicherheit die Trennkraft der Variablen, so ist diese für die Modellbildung nicht nur nutzlos, sondern sogar schädlich. Das Einschließen solcher Variablen in das Modell führt zu einem Absinken der Trefferrate (Abb. 6.3, Kreuze).

Daher ist zu erwarten, dass die Trefferrate bei einer bestimmten Auswahl an Variablen ein Maximum durchläuft, während die Fehlerrate entsprechend ein Minimum zeigt, was auch der Grund für die Bezeichnung „Badewanneneffekt“ ist.

Es gibt also eine optimale Auswahl an Variablen für einen gegebenen Sachverhalt. Die Bestimmung dieser besten Untermenge der zur Verfügung stehenden Variablen ist ein

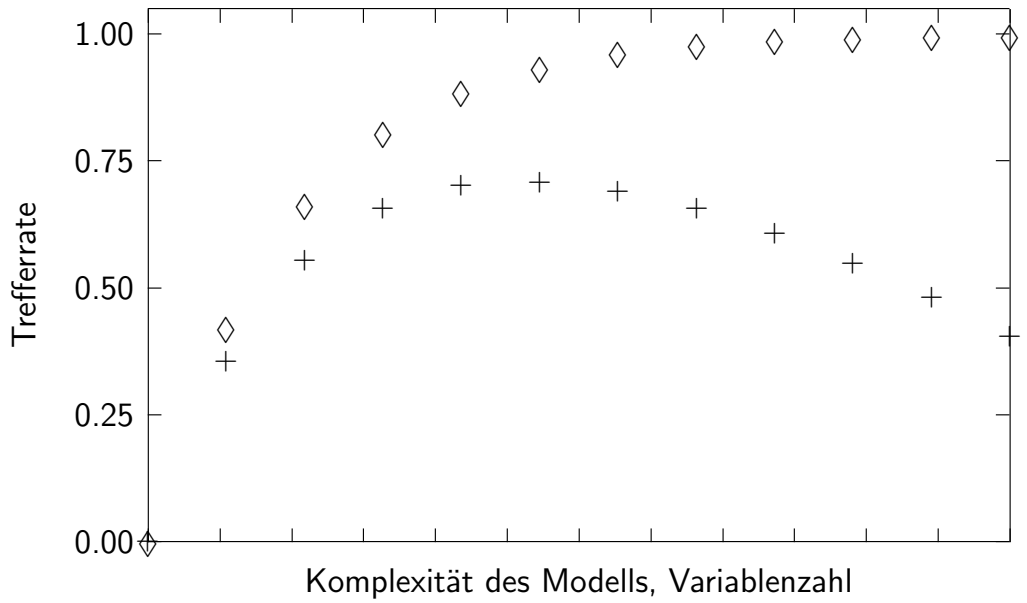


Abbildung 6.3.: Skizze zur Abhängigkeit der Trefferrate von der Variablenzahl — wahrer Verlauf (+) und Reklassifikations-Schätzung (◊)

großes Problem. Bei der *deskriptiven* Diskriminanzanalyse, die die Unterschiede zwischen den Klassen aufzeigen und interpretierbar machen soll, stehen verschiedene Kenngrößen der Gruppentrennung zur Beschreibung der einzelnen Variablen zur Verfügung. Allerdings sollte die Grundlage für die Aufnahme von Variablen in ein *prädiktives* Modell immer die erreichte Modellgüte, zum Beispiel die Trefferrate sein [78, Kap. VIII].

6.5. Beurteilung der Qualität eines Modells

Ein wichtiger Aspekt der Klassifikation sind die Angaben zur Qualität des Modells. Die beiden wichtigsten Größen zur Beurteilung sind die *Trefferrate* T, der Anteil richtiger Zuordnungen, und die *Fehlerrate* F, der Anteil der falschen Zuordnungen an allen erfolgten Zuordnungen. Diese beiden Größen sind also äquivalent. Werden kostenoptimale Regeln angewandt, so sind die zu erwartenden Kosten der Parameter zur Modellbeurteilung.

Die Zuordnungsmatrix stellt die *wahre* Klassenzugehörigkeit (Zeilen) den erfolgten Zuordnungen (Spalten) gegenüber. Damit sind die richtigen Zuordnungen auf der Hauptdiagonale zu finden. Die Trefferrate $\hat{T}$ lässt sich als das Verhältnis der Spur der Zuordnungsmatrix zur Summe über alle Elemente ausrechnen. Diese Darstellung der Validierungsergebnisse ermöglicht eine genauere Analyse der Fehlzuordnungen als es Fehler- oder Trefferrate allein erlauben.

Diese Größen können mit Hilfe geeigneter Testdatensätze geschätzt werden, allerdings ist der Auswahl repräsentativer Datensätze große Aufmerksamkeit zu widmen.

Um eine Schätzung der Trefferrate eines Modells vorzunehmen, wird zusätzlich zur Trainingsstichprobe eine weitere Stichprobe mit bekannter Klassenzugehörigkeit gebraucht. Aus der Zuordnungsmatrix dieser Daten wird die Trefferrate ermittelt. Anders gesagt steht bei einer gegebenen Stichprobe mit bekannter Zuordnung nur noch ein Teil der Daten als

Trainingsstichprobe oder Trainingsset zur Verfügung, beim Teilen muss darauf geachtet werden, dass zwei repräsentative Datensätze gebildet werden, damit sowohl die Schätzung der Modellparameter als auch die Schätzung der Trefferrate durchgeführt werden können.

Da meist, so auch im vorliegenden Fall, der zur Verfügung stehende Stichprobenumfang so gering ist, dass bereits die Schätzung der Modellparameter als mit großen Unsicherheiten behaftet eingestuft werden muss, weicht man auf andere Schätzmethoden, die mit einem geringeren Stichprobenumfang auskommen, aus.

Wichtig ist, dass für eine Schätzung der Trefferraten der Klassifikation unbekannter Proben unabhängige Proben eingesetzt werden müssen, das bedeutet, dass diese Proben in *keiner* Weise an der Modellbildung beteiligt sein dürfen.

6.5.1. Reklassifikation

Werden alle Daten zur Schätzung der Modellparameter verwendet, so wird vermutlich das bestmögliche Modell mit den gegebenen Daten gebildet. Dann kann nur noch die Reklassifikations-Trefferrate bestimmt werden, indem alle Proben mit dem aus ihnen geschätzten Modell klassifiziert werden.

Diese Schätzung der Trefferrate wird einen stark positiven systematischen Fehler aufweisen, ist als Schätzung der Trefferrate also nicht brauchbar. Allerdings kann sie als Obergrenze der Trefferrate interpretiert werden, da nicht zu erwarten ist, dass unbekannte Proben besser klassifiziert werden als die Trainingsproben.

Die durch Reklassifikation geschätzte Trefferrate gibt an, wie gut das gebildete Modell die Trainingsdaten abbildet. Die Rauten in Abb. 6.3 (S. 22) zeigen die Charakteristik einer Reklassifiaktions-Schätzung der Trefferrate gegenüber dem wahren Verlauf (Kreuze) in Abhängigkeit der Komplexität des Modells. Die Reklassifikationsschätzung nähert sich asymptotisch der 1, sie erreicht eine Trefferrate von 100 %, wenn das Modell die Trainingsdaten exakt abbildet. Diese Schätzung gibt also keinen Hinweis auf das Vorliegen einer *Übermodellierung* (*overfitting*), das Maximum der wahren Trefferrate ist nicht zu erkennen. Auch aus diesem Grund ist die Reklassifikations-Trefferrate als Maß der Modellgüte nicht brauchbar.

6.5.2. Kreuz-Validierung

Eine Möglichkeit, die Fähigkeiten des aufgestellten Modells abzuschätzen, ohne eine weitere repräsentative Stichprobe zu benötigen, sind die Verfahren der *Set-* und *Leave-One-Out-Validierung*. Dabei werden die Datensätze zufällig in Gruppen aufgeteilt und der Reihe nach die Daten einer Gruppe aus der Modellbildung ausgeschlossen und diese Spektren dann zugeordnet. Die Leave-One-Out-Validierung ist ein Spezialfall der Set-Validierung: es werden so viele Gruppen gebildet, wie Objekte vorhanden sind, jeweils einzelne Objekte aus dem Modell entfernt und die Vorhersage mit der wahren Klassenzugehörigkeit verglichen.

Bei den vorliegenden Daten sind die Spektren einer Probe untereinander deutlich ähnlicher als Spektren verschiedener Proben. Daher ist hier darauf zu achten, dass die Gruppenaufteilung nach Proben und nicht nach Messungen oder Spektren erfolgt. Im Folgenden soll daher der Begriff der *Leave-One-Out*-Validierung dahingehend verwendet werden, dass einzelne Proben — und nicht etwa einzelne Spektren — ausgeschlossen werden.

Diese Schätzmethoden beurteilen genau genommen nicht die Qualität des aus allen Daten gebildeten Modells, sondern führen einzelne Schätzungen zur Qualität vieler Modelle durch. Daher weisen diese Methoden oft große Varianzen in den Schätzwerten auf.

Die Begründung für dieses Vorgehen ist die Annahme, dass sich die Parameter des Modells nur wenig ändern, wenn einzelne Daten aus der Modellbildung ausgeschlossen werden und daher die erhaltenen Schätzwerte für die vielen gebildeten Modelle nicht zu stark vom wahren Wert für das aus allen Daten geschätzte Modell abweichen. Deshalb sind Abweichungen insbesondere dann zu erwarten, wenn die Schätzproben einen großen Anteil der gesamten Stichprobe umfassen, beziehungsweise wenn insgesamt nur ein geringer Stichprobenumfang realisiert werden kann.

Diese Methoden stellen für den Fall, dass kein eigener Datensatz zur Schätzung der Trefferrate zur Verfügung steht, eine brauchbare Abschätzung der Modellgüte dar.

Allerdings ist über die Eigenschaften solcher Schätzer aus theoretischer Sicht bislang nur wenig bekannt. Das betrifft insbesondere Empfehlungen über die Anteile eines Datensatzes, der für die Validierung reserviert werden soll. Bei der Festlegung der Gruppengröße ist ein Kompromiss zwischen dem hohen Rechenaufwand bei Verwendung vieler Gruppen und der vermutlich besseren Annäherung an das zu beurteilende Modell zu finden. Weiterhin ist zu berücksichtigen, dass auch die Schätzung der Trefferrate mit einer Unsicherheit behaftet sind, die mit steigendem Stichprobenumfang des Testsets abnimmt. Als Faustregel zum Teilen der Daten wird ein Anteil von ungefähr $\frac{1}{4} - \frac{1}{3}$ der gesamten Stichprobe für das Testset empfohlen[81].

6.5.3. Die Veränderung der Klassifikationsergebnisse[78, Kap. VII]

Neben der erreichten Trefferrate ist zur Einschätzung der Bedeutung dieser Größe oft von Interesse, wie viel besser die Klassifikation mit den ermittelten Regeln gegenüber einer anderen Zuordnungsregel ist.

Eine besondere Rolle spielt dabei die Verbesserung gegenüber einer zufälligen Zuordnung und die Verbesserung gegenüber Zuordnungsregeln die alle Objekte einer einzigen Klasse zuordnen.

Die Veränderung der Trefferrate verglichen mit der mit einer anderen Entscheidungsregel beziehungsweise einem anderen Modell erreichbaren ist dann gegeben durch

$$\hat{V} := \frac{\hat{T} - \hat{T}_0}{1 - \hat{T}_0} \tag{6.17}$$

mit $\hat{V}$...Veränderung

$\hat{T}$...mit der betrachteten Zuordnungsregel erreichte Trefferrate

$\hat{T}_0$...Trefferrate der Zuordnungsregel, mit der verglichen wird

6.6. Grundsätzliche Probleme

6.6.1. Falsche Klassifikation in den Referenzdaten

Klassifikationsverfahren zählen zu den Methoden des überwachten Lernens, sie sind auf einen Trainingsdatensatz bekannter Klassenzugehörigkeit angewiesen. Daher sind falsche Klassenzuordnungen der Trainingsdaten durch die Referenzmethode ein grundlegendes Problem beim Schätzen der Modellparameter der Klassifikation.

Dieses Problem ist theoretisch und mit Hilfe von Simulationsrechungen für den Zwei-Klassen-Fall mit dem Ergebnis, dass kleine Anteile an Fehlklassifikationen durch entsprechend größere Stichprobenumfänge ausgeglichen werden können, untersucht worden. Die zu erwartende Fehlerrate wurde nicht zu stark beeinflusst, wenn die Anteile an falsch zugeordneten Trainingsproben für beide Klassen etwa gleich war [82].

Eine empirische Einschätzung dieser Problematik ist besonders dadurch schwierig, dass die Anteile an Fehlzuordnungen in der Regel unbekannt sind.

Im Kontext der medizinischen Diagnostik spielt die Fehlzuordnung in verschiedener Hinsicht eine wichtige Rolle. Zum einen erfolgt die Referenzdiagnose durch eine histologische Begutachtung des gefärbten Referenzschnittes, die Einstufung des Gewebes ist dabei in gewissen Grenzen subjektiv, zumal die hier untersuchten Tumorarten ineinander übergehen. Weiterhin ist die Festlegung der Grenzen des Tumors mit subjektiven Einflüssen behaftet (vgl. z. B. [83, S. 598 – 600]).

Auch entspricht die histologische Einstufung insofern nicht den Anforderungen der Erfordernissen des überwachten Lernens, als sie an die klinischen Anforderungen angepasst ist und daher in der Regel die Diagnose dem höchsten gefundenen Tumorgrad entspricht. Der Tumor kann aber auch weiterhin Zellbereiche niedrigerer Malignität enthalten. Daher muss jede Probe, die für den Trainingsdatensatz verwendet werden soll, eigens untersucht sein und die örtliche Aufteilung in die unterschiedlichen Gewebearten bekannt sein.

Die den Infrarot-Spektren unterschiedlicher Klassen zugrundeliegenden signifikanten Änderungen der molekularen Zusammensetzung des Gewebes müssen nicht gleichzeitig mit dem Übergang zwischen den histologischen Diagnosen der Tumorarten auftreten. Man erwartet sogar, dass die molekularen vor den morphologischen Veränderungen auftreten (Kap. 4.2.1, S. 7). Daher liegt eine weitere mögliche Quelle für Fehlzuordnungen vor.

Ist dies bekannt, so ist eine Verschiebung der Klassifikationsgrenzen zu erwägen. Hinweise auf solche Situationen können Verfahren des unüberwachten Lernens, zum Beispiel Clusteranalysen, geben. Die Bildung neuer Klassen bedeutet allerdings eine wachsende Komplexität des Modells und dadurch eine größere Unsicherheit bei der Parameterschätzung beziehungsweise die Notwendigkeit eines größeren Stichprobenumfangs.

6.6.2. Die zur Verfügung stehende Probenzahl

Aus statistischer Sicht stellt zunächst die Anzahl der Proben (unterschiedlicher Patienten) den Stichprobenumfang dar. Das resultiert aus der Tatsache, dass die Varianz der Spektren zwischen den Proben etwa drei Größenordungen größer ist als die der Spektren innerhalb einer Messung, wobei die Varianz der Spektren zwischen verschiedenen Messungen innerhalb einer Probe vergleichbar mit der zwischen verschiedenen Proben ist. Daher muss man die Spektren einer Messung aus statistischer Sicht eher als Wiederholungsmessungen denn als eigenständige Beobachtungen einordnen.

Die im Rahmen der Wellenlängenselektion neu gebildeten Variablen haben aber unter Umständen andere statistische Eigenschaften als die einzelnen Wellenzahlen der Spektren. Damit ist die für viele statistische Tests wichtige Angabe des Stichprobenumfangs nicht mehr einfach möglich.

Die Anzahl der bei der linearen Diskriminanzanalyse genutzten Variablen soll in der Regel $\frac{1}{5}$ bis $\frac{1}{3}$ des Stichprobenumfangs der Klasse mit der geringsten Probenzahl nicht überschreiten [78; 79; 84].

Allerdings beziehen sich diese Faustregeln auf einfachere Datenstrukturen. Meist wird eine Beobachtung pro Objekt untersucht, eventuell mit einer echten Wiederholungsmessung. In beiden Fällen ist der Stichprobenumfang jedoch eindeutig anzugeben.

Auch die Erfahrungen mit dem Programmsystem `ga_ors/ stackedGen` für NMR-Spektren legen solche Größenordnungen nahe [14].

7. Ermittlung optimaler Wellenzahlbereiche

Das IR-Spektrum einer Substanz spiegelt ihre Zusammensetzung wider, da die Absorption substanzspezifisch und unabhängig erfolgt. Gewebeproben stellen jedoch so komplexe Mischungen der einzelnen Substanzen dar, dass es nicht möglich ist, aus dem Spektrum auf die einzelnen zugrundeliegenden Verbindungen zu schließen. Wohl aber auf einzelne Stoffklassen, vorausgesetzt, sie absorbieren hinreichend stark und es liegen keine zu großen Querempfindlichkeiten vor.

Die zur Verfügung stehenden Daten enthalten zunächst sehr viele Informationen. Eine wichtige Eigenschaft bestimmter chemometrischer Verfahren ist, dass sie auf hochdimensionale Datensätze, unter Umständen sogar ganze Spektren, angewandt werden können. Damit sind sie den Methoden der Untersuchung einzelner diskreter Wellenzahlen überlegen [77]. Allerdings ist dabei zu beachten, dass auch diejenigen Verfahren, die auf komplette Spektren angewandt werden können, in aller Regel bessere Ergebnisse liefern, wenn die Daten vorher auf den bedeutsamen Anteil reduziert wurden [47; 54–58]. Die lineare Diskriminanzanalyse ist empfindlich gegen Variablen mit geringer Trennkraft, daher ist hier der Einsatz von Verfahren zur Variablenselektion beziehungsweise -bildung unumgänglich (vgl. Kap. 6, S. 14).

Die Informationen in den Spektren weisen große Redundanzen auf, die verschiedenen Ursachen zugeordnet werden können. Alle Proben teilen bestimmte Eigenschaften, da es sich immer um menschliches Gewebe handelt. Daher wird eine eingeschränkte Anzahl von Substanzklassen in ähnlichen Anteilen vorkommen. Auch dadurch, dass eine Substanz oder Substanzklasse verschiedene Absorptionsbanden zeigt, kann es unnötig sein, alle diese Wellenlängenbereiche zu betrachten. Wählt man nun charakteristische Regionen aus dem Spektrum aus, so kann davon ausgegangen werden, dass die Daten immer noch redundant sind, da die Spektren — eine hinreichend gute Energieauflösung vorausgesetzt — stetig sind. Eine weitere Datenkompression ist demnach durchführbar. Faktoranalytische Methoden ermöglichen eine exzellente Kompression, allerdings haben die dort neu gebildeten Variablen keine direkte physikalische Bedeutung mehr, wodurch ihre Interpretation deutlich schwieriger wird. Andere Ansätze wie die Mittelwertbildung erlauben auch nach der Datenkompression die direkte spektroskopische Interpretation der erhaltenen Daten.

Die Auswahl der letztlich verwendeten Charakteristika der Spektren kann mit zwei entgegengesetzten Verfahren getroffen werden. Einerseits ermöglicht die Kenntnis der krankheitsbedingten molekularen Veränderungen, Erwartungen über Veränderungen in den Spektren zu formulieren. Andererseits steht auch eine Anzahl an Routinen zur Verfügung, die, zunächst ausschließlich mathematisch, die bedeutsamen Variablen für die Klassifizierung ermitteln.

Das hier verwendete Verfahren sucht eine Untermenge einer gegebenen Variablenmenge, so dass nach wie vor *reale* Variablen, also Variablen mit direkter physikalischer Bedeutung, vorliegen. Daher ist der Versuch, diese ohne die Benutzung biochemischen Wissens ermittelten Variablen im Hinblick auf die molekularen Veränderungen zwischen den einzelnen Klassen zu interpretieren, legitim.

Zu beachten ist aber unbedingt, dass zwar aus der Kenntnis einer Substanz auf das Spektrum geschlossen werden kann, aber die umgekehrte Zuordnung zunächst eine Hypothese bleibt, da unterschiedliche Verbindungen im gleichen Wellenlängenbereich absorbieren können.

Es stehen jedoch statistische Werkzeuge zur Verfügung, um die Wahrscheinlichkeit abzuschätzen, dass eine vermutete Substanz beziehungsweise Substanzklasse tatsächlich zu den im Spektrum beobachteten Absorptionsbanden führt. Eine Prüfung der entsprechenden Hypothese ist jedoch genau genommen nur dann möglich, wenn *alle* in der Mischung vorhandenen Verbindungen bekannt sind.

Insgesamt stellt sich also die Frage einer *optimalen* Auswahl an Wellenzahlen des Spektrums zur weiteren Analyse.

7.1. Optimierung[42–46; 85–87]

Die mathematische Formulierung eines Optimierungsproblems ist

$$f(\mathbf{x}) \to \max_{\mathbf{x}\in\mathbb{G}}, \qquad f : \mathbb{R}^q \to \mathbb{R} \tag{7.1}$$

Grundsätzlich muss zur rechnergestützten Lösung für jedes Optimierungsproblem eine solche Bewertungsfunktion f gefunden werden, die optimiert wird. Mathematisch handelt es sich dabei um ein Funktional, also eine Abbildung aus dem Suchraum $\mathbb{G} \subseteq \mathbb{R}^q$ in die Menge der reellen Zahlen. Viele Optimierungsprobleme können nur dann exakt gelöst werden, wenn der gesamte Suchraum abgesucht wird, weil eine analytische Lösung der Probleme nicht bekannt ist. Eine direkte Umsetzung dieser Idee bezeichnet man als *brute force Lösung*, sie erfordert einen enormen Rechenaufwand, der schon bei zunächst gering erscheinenden Problemen zur praktischen Unlösbarkeit führen kann.

> Als Beispiel sei eine Abschätzung der Rechenzeit für das hier vorliegende Problem gegeben: Ein Spektrum im Bereich von 1000 – 1800 cm^{-1} mit einer Auflösung von 4 cm^{-1} enthält etwa $p = 200$ Messpunkte. Der Suchraum ist $q = (p-1)$-dimensional, wobei jeweils 2 Möglichkeiten pro Dimension bestehen, da zwei Messpunkte zu einer Region verbunden sein können oder nicht (vgl. Kap. 8.1, S. 38). Damit erhält man insgesamt $2^{199} \approx 8 \cdot 10^{59}$ Möglichkeiten. Angenommen, die Bewertung einer Lösung dauere im Schnitt 1 μs, so dauert die Lösung der Optimierungsaufgabe ungefähr $3 \cdot 10^{46}$ Jahre. Da diese Suche problemlos parallelisiert werden kann, könnte man die Rechnungen auf viele Rechner verteilen. Das Programm SETI@home ist ein Beispiel der Verteilung von Rechnungen auf viele Computer, zur Zeit nehmen etwa 4 Millionen Rechner teil (`http://setiathome.ssl.berkeley.edu/`, Meldung vom 2. Oktober 2002). Diese bräuchten also bei der oben angenommenen Rechenleistung zusammen „nur“ ca. $6 \cdot 10^{39}$ Jahre ...
> Zum Vergleich: das Alter des Universums wird auf die Größenordnung von 10^{10} Jahren geschätzt [88].

Allerdings ist dieses Verfahren als einziges sicher in der Lage, das globale Optimum zu ermitteln. Daher kann man entsprechende Rechnungen als Referenz verwenden, wenn zum Beispiel einzelne Lösungen durchaus errechnet werden können, aber insgesamt viele ähnliche Probleme gelöst werden müssen.

In der Regel verwendet man Verfahren, die zwar nicht garantiert das globale Optimum finden, dafür aber mit wesentlich geringerem Rechenaufwand auskommen. Oft wird man sich auch damit zufrieden geben, wahrscheinlich nicht die beste, aber eine hinreichend gute Lösung zu erhalten. Dies soll im Folgenden vereinfachend im Begriff einer optimalen Lösung mit enthalten sein. Eine wichtige Klasse der Optimierungsverfahren sind die *Hill-Climbing*-Verfahren. Diese Verfahren sind streng *deterministisch*, sie nutzen das lokale Verhalten des Zielfunktionals, um eine gute Richtung für die weitere Suche zu bestimmen. Nachteilig wirkt sich diese lokale Auswertung des Zielfunktionals dahingehend aus, dass die Gefahr, nur ein lokales Optimum zu finden, das unter Umständen sehr viel schlechter als große Bereiche des Suchraums ist, besonders groß ist. Weiterhin sind diese Algorithmen auf *stetige* Funktionale angewiesen.

Die *Monte-Carlo*-Verfahren sind dagegen nicht auf stetige Funktionale beschränkt und auch die Gefahr, lokale Optima zu finden, die weit schlechter als das globale Optimum sind, ist sehr viel geringer. Sie werten das Zielfunktional an zufällig bestimmten Punkten aus. Allerdings haben diese Verfahren gravierende Schwierigkeiten bei der exakten Lokalisierung eines Optimums.

Man versucht daher, Verfahren mit zufälligen Anteilen mit deterministischen Verfahren zu kombinieren, um die Vorteile beider Methodenklassen zu nutzen, also lokal schnell das Optimum zu finden, ohne zu früh den gesamten Suchraum aus den Augen zu verlieren. Ein klassischer Ansatz ist, zunächst ein vielversprechendes Teilgebiet des Suchraums einzugrenzen und dann innerhalb dieses Unterraums einen anderen Algorithmus einzusetzen, der sehr effektiv lokal optimiert.

Evolutionäre Algorithmen bieten verschiedene Parameter, mit denen die zufälligen und deterministischen Anteile genau eingestellt werden können.Allerdings hängt die Effektivität dieser Algorithmen wiederum empfindlich von der geeigneten Wahl dieser Parameter ab.

Optimierungsprobleme können in drei Grundklassen aufgeteilt werden. Dies ist bei der Auswahl und Implementierung des Verfahrens zu berücksichtigen.

- Parameter-Optimierung: Die Parameter eines Modells sind zu optimieren, das heißt, es sollen optimale Werte der verschiedenen Parameter gefunden werden.
- Subset-Selection: Aus einer Menge soll nach bestimmten Kriterien eine Untermenge ausgewählt werden.
- Kombinatorische Probleme: Elemente sollen in einer optimalen Reihenfolge angeordnet werden.

Hier handelt es sich um ein Problem der Subset-Selection, es ist eine geeignete Anzahl an Variablen auszuwählen.

7.2. Genetische Algorithmen[42–46; 85–87]

In den 1970er Jahren wurde das Konzept der *evolutionären Algorithmen* zur Lösung von Optimierungsaufgaben entwickelt. Dabei handelt es sich unter anderem um die von John Holland entwickelten *Genetic Algorithms* und die zeitgleich von Ingo Rechenberg und Hans-Paul Schwefel entwickelten *Evolutionsstrategien* als Algorithmen, die zur Optimierung eingesetzt werden können.

Optimierung bedeutet dabei nicht zwingend die Optimierung eines hinreichend gut bekannten Funktionals, evolutionäre Algorithmen werden zum Beispiel auch in der Simulation unterschiedlichster Systeme und in der Forschung zur künstlichen Intelligenz angewandt. Insbesondere bei Methoden des *genetischen Programmierens* ist in der Regel das Problem selbst nicht codiert.

Die Idee dieser Konzepte ist die Nachbildung „natürlicher“ Optimierungsstrategien, wie sie in biologischen Systemen den Lauf der Evolution bestimmen.

Begriffsbestimmungen

Viele Begriffe im Umfeld der genetischen Algorithmen stammen aus der Biologie und unterscheiden sich daher von den bekannten Begriffen der Optimierung. Genetische Algorithmen arbeiten nicht mit einer (möglichen) Lösung des Problems, also einem Punkt im Suchraum, sondern mit einer Menge solcher Punkte im Suchraum. Diese Menge wird *Population* genannt, eventuell enthält eine Population außer den einzelnen Individuen noch weitere Attribute.

Der einzelne Punkt im Suchraum, also jedes Element der Population, ist ein *Individuum* oder *Chromosom*. Hier bedeutet jedes Individuum also eine Möglichkeit, Wellenzahlregionen des Spektrums zur LDA zu benutzen.

Jedes Individuum besteht aus einer Bit- oder Zeichen-Folge, einem *String*, die den Punkt im Suchraum beschreibt, also die Parameter einer Lösung darstellt. Die Parameter werden auch als *Gene* bezeichnet. Meist werden Bit-Strings konstanter Länge verwendet.

Der Bit-String des Lösungsvorschlags, also die Datenstruktur, die im genetischen Algorithmus beeinflusst wird, heißt *Genotyp*. Demgegenüber steht der *Phänotyp*, der tatsächliche Lösungsvorschlag. Genotyp und Phänotyp werden durch die *Decodierungs-* und die *Codierungs-Funktion* ineinander überführt. Bei der Wahl dieser Funktionen sollte darauf geachtet werden, dass geringe Änderungen des Genotyps auch zu geringen Änderungen des Phänotyps führen und umgekehrt. Man spricht in diesem Fall von einem *stark kausalen* Zusammenhang zwischen Geno- und Phänotyp. Das bedeutet, dass Codierungs- und Decodierungsfunktion „möglichst stetig“ sind. Daher kann eine einfache Binärcodierung durch das verwendete Stellenwertsystem bei Parametern mit mehr als zwei möglichen Werten zu Problemen führen.

An dieser Stelle sei erwähnt, dass die Evolutionsstrategien statt des Bit-Strings direkt die Parameter in Form eines Vektors verwenden, also nur mit dem Phänotyp arbeiten. Dieser Unterschied in der Datenstruktur ist auch der Hauptunterschied zwischen den genetischen Algorithmen und den Evolutionsstrategien.

Jedes Individuum wird mit Hilfe der *Fitness-Funktion*, also des bereits erwähnten Zielfunktionals bewertet.

Genetische Algorithmen erzeugen aus der vorliegenden Population eine neue Population mit hoffentlich besseren Eigenschaften. Man spricht auch von der Population der nächsten *Generation* oder der Erzeugung der *Kind-Population* aus der *Eltern-Population*.

Um eine Verbesserung der Fitness der gesamten Population zu erreichen, muss die jeweilige Kind-Population in geeigneter Weise aus der Eltern-Population aufgebaut werden.

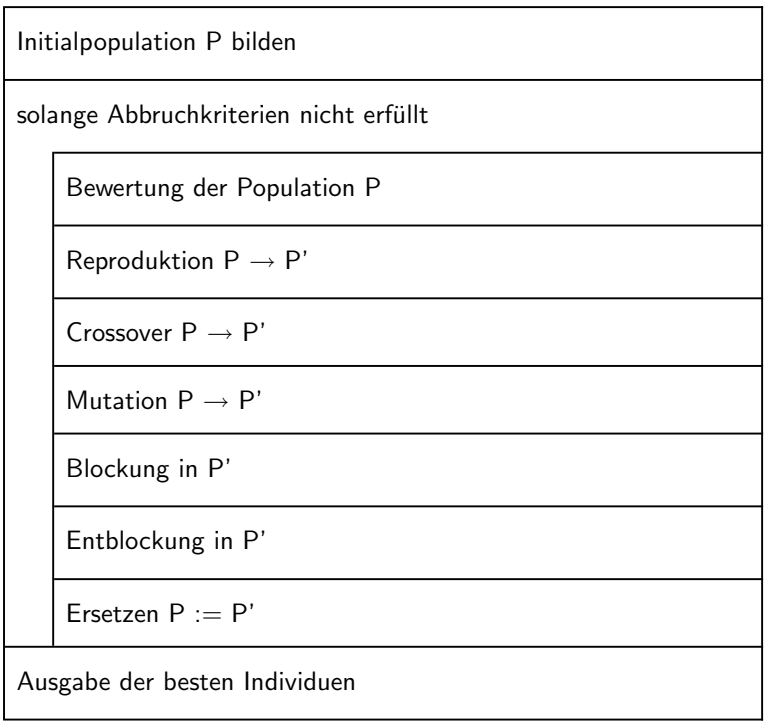


Abbildung 7.1.: Struktogramm eines genetischen Algorithmus

Dazu stehen in der Regel folgende Operationen zur Verfügung:

- Selektion
- Reproduktion
- Mutation
- Crossover
- Inversion
- Blockung und Entblockung

Oft sind allerdings nicht alle diese Operationen zur effektiven Lösung des Problems notwendig, viele Implementationen kommen allein mit Selektion, eventuell Reproduktion, Mutation und Crossover aus.

7.2.1. Implementation eines genetischen Algorithmus

Ein Struktogramm (Nassi-Schneidermann-Diagramm) eines genetischen Algorithmus ist in Abb. 7.1 dargestellt.

Vor den Erläuterungen zu den einzelnen Punkten sei hier noch der Unterschied zweier möglicher Vorgehen herausgestellt. Die neue Population wird entweder direkt aus der Eltern-Population aufgebaut, indem mit jeder entsprechenden implementierten Operation eine gewisse Anzahl an Individuen erzeugt wird. Alternativ dazu wird zunächst eine Intermediär-Population durch Selektion und entsprechende Reproduktion aus der Eltern-Population erzeugt, auf die dann die restlichen Operationen (insbesondere Mutation und

Crossover) der Reihe nach angewandt werden. Natürlich sind auch Mischformen dieser beiden Verfahren möglich.

Abbruchbedingungen

Üblich sind zwei Arten von Abbruchbedingungen: entweder man gibt die Zahl der zu berechnenden Generationen vor oder einen zu erreichenden Grenzwert für die Fitness der Gesamtpopulation. Letzteres entspricht der Angabe, welche Lösung hinreichend gut ist. Auch eine Kombination dieser Abbruchkriterien ist möglich. Dadurch wird sichergestellt, dass der Algorithmus auch dann endet, wenn das vorgegebene Ziel der Fitness nicht erreicht wird.

Populationsgröße

Die Wahl der Populationsgröße läuft auf einen Kompromiss zwischen dem Bestreben, einen möglichst großen Bereich des Suchraums abzudecken und dem entgegengesetzten Bestreben, möglichst wenig Ressourcen zu beanspruchen, hinaus. Allgemeingültige Angaben können nicht gemacht werden, die Entscheidung über die Populationsgröße muss in Abhängigkeit von der konkreten Problemstellung getroffen werden. Es besteht auch die Möglichkeit, mit veränderlichen Populationsgrößen zu arbeiten.

Fitness-Funktion

Die Implementation der Fitness-Funktion ist für den genetischen Algorithmus von enormer Wichtigkeit. Dies ist zum einen in Bezug auf die benötigte Rechenzeit zu sehen, da die Fitness-Funktion in *jeder* Generation für *jedes* Individuum ausgewertet werden muss. Dadurch kommt der Fitness-Funktion in der Regel ein großer Anteil an der gesamten Rechenzeit des Algorithmus zu — sie stellt also einen wichtigen Ansatzpunkt zur Optimierung dar. Meist ist das Ermitteln einer geeigneten Fitness-Funktion eine der größten Schwierigkeiten bei der Formulierung einer konkreten Problemlösung mittels eines genetischen Algorithmus.

Zum anderen stehen in der Regel aus mathematischer Sicht verschiedene mögliche Zielfunktionale zur Verfügung. Weiter ist ein stark kausaler Zusammenhang zwischen dem Genotyp des bewerteten Individuums und dem Wert der Fitness-Funktion wünschenswert. Das bedeutet natürlich, dass auch zwischen Geno- und Phänotyp ein stark kausaler Zusammenhang vorliegen muss, da die Fitnessfunktion immer auf den Phänotyp angewandt wird. Auch hier wird also eine „möglichst stetige“ Funktion gefordert.

Je nach Implementation der Operationen zur Erzeugung der Kind-Population wird die *kanonische* Fitness benötigt, das ist die auf die durchschnittliche Fitness der Population normierte Fitness.

Selektion

Die Selektion ist eine Operation, die zunächst nicht im Beispiel-Algorithmus aufgeführt ist, obwohl sie von grundlegender Bedeutung für die Funktion des Algorithmus ist. Selektion bedeutet, dass Individuen nach ihrer Fitness ausgewählt werden. Praktisch alle Operatoren zum Aufbau der neuen Population benötigen die Selektionsfunktion, da sie die Bewertung

der Fitness der einzelnen Individuen in den Aufbau der neuen Population einfließen läßt. Daher sichert die Selektion eine steigende Fitness im Laufe der Generationszyklen.

Es ist eine Vielzahl verschiedener Selektionsmechanismen bekannt, von denen hier nur zwei grundsätzliche Methoden kurz erwähnt seien:

remainder stochastic sampling: Die Anzahl der Kopien eines Individuums wird von der abgerundeten kanonischen Fitness bestimmt. Die Wahrscheinlichkeit, dass noch eine Kopie des Individuums eingefügt wird, ist vom beim Abrunden gebliebenen Rest abhängig.

roulette-wheel-method: Die Wahrscheinlichkeit, dass ein Individuum kopiert wird, entspricht seiner kanonischen Fitness geteilt durch die Anzahl der Individuen in der Population. Die Summenfunktion dieser Wahrscheinlichkeiten über alle Individuen kann als Grenzwert zur Selektion eines Individuum mittels einer Zufallszahl zwischen null und eins benutzt werden. Auf diese Weise kann mit n Zufallszahlen eine neue Population von n Individuen erzeugt werden.

Die Selektion muss nicht proportional zur Fitness erfolgen. In manchen Fällen ist es günstiger, die Reproduktionsrate oder -wahrscheinichkeit an der *Rangfolge* der Fitness festzumachen. Setzt man die Reproduktionsrate fest, so ergibt sich ein zum remainder stochastic sampling analoges Verhalten, arbeitet man mit Reproduktionswahrscheinlichkeiten, so entspricht das Vorgehen eher der roulette-wheel-method.

Eine Reihe besonderer Selektionsmechanismen soll die Diversität der Population erhalten. Dazu zählt die *Präselektion*, die bei Crossover, Inversion und Mutation angewandt werden kann. Das neu entstandene Individuum wird mit seinem Vorgänger verglichen und ersetzt den Vorgänger nur dann, wenn es eine größere Fitness hat. Ähnlich arbeitet das *Crowding-Schema* nach DE JONG. Hier wird das neue Individuum mit mehreren Individuen der Eltern-Population verglichen und ersetzt das ihm ähnlichste Individuum. Hierfür wird die Implementation eines Ähnlichkeitsmaßes benötigt. Meist kommt die HAMMING-*Distanz* zur Anwendung. Die *beschränkte Kreuzung* läßt für Crossover-Operationen nur Individuen-Paare zu, die sich hinreichend ähnlich sind. Auch hierbei wird oft die HAMMING-*Distanz* eingesetzt.

Eine besondere Form der Selektion ist die Anordnung der Individuen auf einem Gitter. Interaktionen sind nun nur zwischen hinreichend dicht benachbarten Individuen erlaubt. Dadurch können sich mehrere Nischen unabhängig voneinander ausbilden. Oft erreicht man so eine schnelle Konvergenz, ohne die Diversität der Population in Frage zu stellen.

Die Fortführung dieser Idee ist die vollständige Parallelisierung des Algorithmus, indem auf verschiedenen Prozessoren mit kleinen Populationen gerechnet wird.

Erzeugen der Initialpopulation

Die Startpopulation wird in der Regel zufällig erzeugt. Ist das nicht der Fall, so besteht die Gefahr, dass der Algorithmus frühzeitig, das heißt möglicherweise bei einem vom globalen Optimum weit entfernten lokalen Optimum konvergiert.

Reproduktion

Reproduktion bedeutet das direkte Kopieren von Individuen der Eltern-Population in die Kind-Population. Werden die besten Individuen jeweils in die nächste Generation mit übernommen, so sichert dies, dass die maximal erreichte Fitness nicht absinkt.

Mutation

Mutation beschreibt die zufällige Veränderung einzelner Bits eines Individuums und sichert einen Ausweg für den Fall, dass eine Population aus sehr ähnlichen Individuen besteht. Andere Gebiete des Suchraums können in diesem Fall nur noch durch zufällige Änderungen berücksichtigt werden. Allerdings besteht bei Mutation auch die Gefahr, gute Gene zu zerstören. Daher darf die Mutationswahrscheinlichkeit nicht zu groß gewählt werden. Meist sind Werte im Promille- bis unteren Prozentbereich günstig.

Crossover

Beim Crossover werden Informationen zwischen zwei Individuen ausgetauscht. Man unterscheidet

one-point-crossover: Nach der Selektion zweier Individuen wird eine Position zufällig gewählt, ab der der restliche Teilstring zwischen den Individuen ausgetauscht wird.

uniform-crossover: Man entscheidet zufällig anhand einer vorgegebenen Wahrscheinlichkeit für jedes Bit, ob es zwischen den Individuen ausgetauscht wird oder nicht.

Meist erhält man mit dem uniform-crossover bessere Ergebnisse.

Die verschiedenen Implementationen unterscheiden sich noch dahingehend, dass entweder beide neuen Individuen oder nur das bessere in die neue Population übernommen werden.

Inversion

Inversion bedeutet, dass die *Bit-Folge* eines Teilstrings umgekehrt wird, dieses Stück rückwärts wieder eingesetzt wird. Viele genetische Algorithmen implementieren diese Operation nicht.

Blockung und Entblockung

Wenn viele gute Individuen denselben Teilstring besitzen, kann es sinnvoll sein, diesen als Einheit zu schützen, so dass er nicht durch Crossover, Mutationen oder Inversionen zerstört werden darf. Dieses Vorgehen bezeichnet man als Blockung. Die Implementation dieser Operation benötigt eine Funktion zum Test auf gleiche Teil-Strings.

Ein vergleichbarer Effekt kann auftreten, wenn mit variabler Bitzahl gearbeitet wird. Oft sammeln sich dann wirkungslose Bits um die guten Teilstrings an und verringern so die Wahrscheinlichkeit, dass diese zerstört werden. Dieser Effekt ist besonders beim genetischen Programmieren ein bekanntes Phänomen.

Wird ein Blockungs-Operator implementiert, so muss auch ein Entblockungsoperator bereitgestellt werden, der mit einer gegebenen Wahrscheinlichkeit dazu führt, dass ein bestehender Block nicht länger geschützt wird.

7.2.2. Einige ausgewählte Probleme

Bei der Behandlung auftretender Probleme gilt es zu berücksichtigen, dass es sich hier um *stochastische Algorithmen* handelt, also grundsätzlich auch unwahrscheinliche Ergebnisse auftreten können. Daher muss ein Konvergenzproblem nicht zwingend an falschen Parametern oder einer ungeeigneten Fitness-Funktion liegen. Zunächst ist von der Möglichkeit eines Neustarts des Algorithmus Gebrauch zu machen. Normalerweise wird dann auch mit einer veränderten Startpopulation gearbeitet, selbst bei identischen Initialpopulationen sind allerdings die nachfolgenden Generationen aufgrund der erwähnten stochastischen Eigenschaften des Algorithmus unterschiedlich. Wenn das Problem allerdings gehäuft auftritt, sollte an eine entsprechende Veränderung der Parameter, der Fitness-Funktion oder sogar der Operatoren zur Erzeugung der Kind-Population gedacht werden.

Ein Problem besteht darin, dass der Algorithmus fast konvergiert, aber das globale Optimum nicht findet. Insgesamt liegt also eine langsame Konvergenz vor. In diesem Fall kann eine Spreizung der entsprechenden Fitness-Werte helfen, den erforderlichen Gradienten in der Fitness der Population zu erzeugen.

Auch der gegenteilige Fall, eine sehr schnelle Konvergenz des Algorithmus, kann ein Problem darstellen. Das ist dann der Fall, wenn zu befürchten ist, dass die optimale Lösung nicht gefunden wird. Verschiedene Ursachen können diesen Effekt bewirken. Ein überdurchschnittlich gutes Individuum kann sich gegenüber den anderen Individuen durchsetzen, so dass der Suchraum frühzeitig sehr stark eingeschränkt wird. Man vergleicht diesen Fall mit der Nischenbildung in biologischen Systemen. Ganz ähnliche Auswirkungen hat der unwahrscheinliche, aber mögliche, Fall, dass praktisch die gesamte Population einem Genotyp entspricht. Die Entstehung solcher Probleme ist das Resultat eines sehr komplexen Zusammenspiels der Fitness-Funktion mit den Operatoren zur Erzeugung der neuen Population.

Im ersten Fall kann man versuchen, den Wertebereich der Fitness-Funktion zu komprimieren und so der starken Bevorzugung einzelner guter Individuen entgegenzuwirken. Ein anderer Ansatz ist die Modifikation der Parameter, die die Erzeugung der Kind-Population steuern. So kann die Reproduktionsrate nicht proportional zur Fitness der Individuen, sondern abhängig von der Rangfolge der Individuen in der Fitness-Skala gewählt werden. Auch können bekannte, aber unerwünschte, Nischen vermieden werden, indem die Fitness-Funktion entsprechend verändert wird.

Das letztgenannte Problem tritt oft bei sehr kleinen Populationen auf, die Vergrößerung der Population ist ein Ausweg. Ein andere Möglichkeit ist die Vergrößerung der Mutationswahrscheinlichkeit.

Wichtig ist, dass genügende Anzahlen „schlechter“ Individuen in die nächste Population eingehen, da sonst auch eine Nischenbildung eintreten kann.

7.2.3. Wichtige Aspekte aus der Informatik

Heuristiken

Genetische Algorithmen gehören zu den *Heuristiken*, daher können keine im Sinne von mathematischen Beweisen gesicherten Angaben über die Leistungsfähigkeit gemacht werden, die Leistungsfähigkeit solcher Algorithmen muss empirisch abgeschätzt werden.

In der Optimierung haben sie sich aber als Verfahren für Problemstellungen etabliert, bei denen kein *effizienter* Algorithmus bekannt ist.

Determinismus und Determiniertheit

Genetische Algorithmen zählen zu den *stochastischen Algorithmen*, sie sind weder *determinisitisch* noch *determiniert*. Deshalb wird ein genetischer Algorithmus bei mehrfacher Anwendung auf dasselbe Problem nicht immer dieselbe Lösung ermitteln. Es besteht die Wahrscheinlichkeit, dass eine *falsche* Lösung im Sinne nicht optimaler Resultate erhalten werden kann. Grundsätzlich sollte der Algorithmus daher mehrfach auf dieselbe Eingabe angewandt werden, um solche, zwar in der Regel unwahrscheinlichen, aber immer möglichen, falschen Lösungen zu erkennen und zu verwerfen.

Der Nichtdeterminismus der genetischen Algorithmen ist in der Zugehörigkeit zu den stochastischen Algorithmen begründet. Das bedeutet auch, dass Entscheidungen zufällig getroffen werden. Die verwendete Rechentechnik arbeitet jedoch deterministisch, man nutzt also Pseudo-Zufallszahlen. Wichtig ist eine hinreichende Qualität dieser Pseudo-Zufallszahlen, sonst wächst unter Umständen die Gefahr, falsche Lösungen zu erhalten.

Die empirische Abschätzung der Leistungsfähigkeit des Algorithmus muss davon aber nicht beeinträchtigt sein.

7.2.4. Einsatzgebiete und Anforderungen

Jedes parametrisierbare Optimierungsproblem, bei dem ein quantitatives Maß für die Güte eines gegebenen Parametersatzes angegeben werden kann, kann mit genetischen Algorithmen behandelt werden [44]. Allerdings bedeutet das nicht, dass der Einsatz genetischer Algorithmen damit für alle Optimierungsprobleme sinnvoll ist.

Grundsätzlich sollten genetische Algorithmen nur dann eingesetzt werden, wenn kein *effizienter* Algorithmus zur Lösung des Problems bekannt ist [86]. Die Aufgaben der Subset-Selektion gehören zu den *NP-vollständigen* Problemen, zu denen kein effizienter Lösungsweg bekannt ist. Damit ist die Wellenlängen-Selektion ein typisches Aufgabengebiet für genetische Algorithmen.

Genetische Algorithmen eignen sich ideal zur Parallelisierung, weil grundsätzlich mehrere vollkommen unabhängige Programmläufe mit jedem Eingabedatensatz durchgeführt werden sollten, um das Riskio eines nicht optimalen Ergebnisses aufgrund der Konvergenz zu einem lokalen Optimum zu verringern.

Außerdem können die Operatoren Reproduktion, Mutation, Crossover und Inversion entweder der Reihe nach auf eine Intermediärpopulation angewandt werden, die so in die Kind-Generation übergeht. Alternativ kann jeder dieser Operatoren zur Erzeugung eines bestimmten Anteils der Kind-Generation eingesetzt werden. Dann kann auch die Erzeugung der Kind-Generation weitgehend parallelisiert werden, da die einzelnen Operatoren voneinander unabhängig arbeiten.

Teil III.

Die verwendeten Programme

8. Das Programmsystem ga_ors und stackedGen

Die Wellenlängenselektion und Berechnung der zur linearen Diskriminanzanalyse verwendeten Variablen erfolgte in dieser Arbeit mit dem Programmen `ga_ors`, die lineare Diskriminanzanalyse wurde mit dem Programm `stackedGen` durchgeführt.

Diese beiden Programme arbeiten eng zusammen, die Optimierung durch `ga_ors` benutzt als Zielfunktional die durch `stackedGen` ermittelte Qualität des Modells der linearen Diskriminanzanalyse für die angegebenen Variablen. Damit ist ein sehr robustes Verhalten dieser Programmkombination zu erwarten.

8.1. Die Implementation des genetischen Algorithmus in ga_ors [14]

Ziel der Optimierung ist es, aus den ausgewählten Regionen der Spektren eine begrenzte Anzahl an Variablen zu bilden, die eine optimale Klassifikation der Daten ermöglichen. Des Weiteren wird gefordert, dass diese Variablen oder Koordinaten in direkter Beziehung mit dem Spektrum stehen sollen, damit aus den Ergebnissen der Optimierung und Klassifikation auf mögliche Zusammenhänge zwischen Krankheitsbild und molekularer Zusammensetzung des Gewebes geschlossen werden kann.

Von den vorgestellten Operatoren zur Erzeugung der Kind-Population sind hier nur Mutation, Crossover und Reproduktion implementiert.

Zur Codierungs- und Decodierungsfunktion

Der Algorithmus zur Optimierung arbeitet mit Wellenzahl-Regionen des Spektrums, nicht mit einzelnen Wellenzahlen. Diese müssen daher in geeigneter Weise auf eine einzelne Variable abgebildet werden. In erster Linie handelt es sich dabei um die *Mittelwertbildung*. In einzelnen Fällen wurden auch andere Transformationen wie die *Varianzen* oder *Verhältnisse gemittelter Intensitäten* erfolgreich verwendet[14].

Die spektralen Regionen werden in `ga_ors` direkt in einen Bit-String der Länge $p-1$ übersetzt (p bedeutet dabei die Anzahl der Messpunkte pro Spektrum). Damit ist jedes Bit einem Zwischenraum zwischen zwei gemessenen Extinktionen zuzuordnen. Hat das Bit den Wert 1, so sind die beiden Punkte des Spektrums um dieses Bit zu einer ausgewählten Region verbunden (siehe Abb. 8.1).

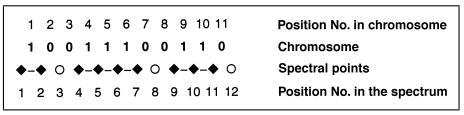


Abbildung 8.1.: Codierung in `ga_ors` [14]

Vorteilhaft gegenüber dem Selektieren der einzelnen Messpunkte (Bit hat Wert 1) eines Bit-Strings der Länge p ist diese Wahl der Codierungsfunktion insbesondere für die anschließende Mittelwertbildung.

Diese Codierungsfunktion erlaubt im Extremfall Individuen, deren Phänotyp $\frac{p}{2}$ Variablen hat. Die Anzahl an tatsächlich gebildeten spektralen Regionen wird allerdings durch den bereits erwähnten Parameter festgelegt.

Die Fitness-Funktion

`ga_ors` verwendet das mittlere Quadrat des Fehlers zwischen der vorhergesagten Klassenzugehörigkeit in Form einer Wahrscheinlichkeit und der für die Trainingsdaten bekannten Klassenzugehörigkeit als Fitness-Funktion:

$$f = \frac{1}{ng} \sum_{\hat{k}=1}^{g} \sum_{j=1}^{n} \left(p_{j\hat{k}} - l_{j\hat{k}} \right)^2 \tag{8.1}$$

mit g .. Anzahl der Klassen

n .. Anzahl der Spektren

$\hat{p}_{j\hat{k}}$ .. *vorhergesagte* a posteriori Wahrscheinlichkeit

k .. wahre (bekannte) Klassenzugehörigkeit

$$l_{j\hat{k}} = \begin{cases} 1 & \text{für } \hat{k} = k \\ 0 & \text{sonst} \end{cases}$$

.. Der *bekannten* Klassenzugehörigkeit des Spektrums j

Zur Bestimmung der a posteriori Wahrscheinlichkeiten $\hat{p}_{j\hat{k}}$ wird, genau wie für die endgültige Klassifikation, eine lineare Diskriminanzanalyse durchgeführt.

Da hier ein Fehler als Fitness-Funktion verwendet wird, *fallen* die Werte der Fitness-Funktion — entgegen den bei genetischen Algorithmen üblichen Konventionen — bei steigender Güte der Lösung.

Das Erzeugen der Initialpopulation und die Populationsgröße

Als sinnvolle Populationsgrößen werden 200 – 600 Individuen genannt [14].

Die Initialpopulation wird zufällig erzeugt, allerdings sind dabei die Wahrscheinlichkeiten für die einzelnen Punkte, in der Initialpopulation in ausgewählten Regionen zu sein, nicht gleich, um eine bessere Initialpopulation und damit hoffentlich eine schnellere Konvergenz des Algorithmus zu erreichen. Diese Wahrscheinlichkeiten werden anhand der *Trennschärfe* der Wellenzahlen bestimmt, das Kriterium ist das Quadrat der statistischen Testgröße t.

Der Reproduktionsoperator

Die besten Individuen der jeweiligen Population werden als *Elite* direkt in die nächste Generation übernommen. Diese Anzahl ist vom Benutzer einstellbar.

Selektion

Die Selektion erfolgt nach der Rangfolge der Individuen. Es werden immer zwei Individuen selektiert, auf die dann der Mutations- und anschließend der Crossover-Operator angewendet werden.

Mutation

Die Implementation des Mutationsoperators unterscheidet sich deutlich von bekannten Lösungen. Hier werden nicht einzelne Bits negiert, sondern ganze Blöcke. Zu Beginn ist die Länge dieser Blöcke mit $\frac{p}{64}$ sehr groß, so dass Mutationen zu großen Sprüngen im Suchraum führt und entsprechend große Bereiche bei der Suche berücksichtigt werden. Im Laufe der Optimierung wird die Blocklänge reduziert. Der Algorithmus endet, wenn sie einen vorgegebenen Wert erreicht.

Kriterien zum Beenden des Algorithmus

Oben genanntes Endkriterium ist mit einem entsprechenden Fortschritt der Optimierung äquivalent, da die Länge des Mutationsblocks aus der Erreichten Modellgüte der linearen Diskriminanzanalyse bestimmt wird.

Das andere Endkriterium ist die vom Benutzer vorzugebende Höchstzahl zu berechnender Generationen. Für die in [14] beschriebenen Probleme wurden 50 – 100 Generationen gewählt. Man kann eine Abschätzung der notwendigen Größe leicht mit Hilfe der Ausgabe des Programms während der Rechnungen vornehmen.

Crossover

Die Crossover-Operation ist als one-point-crossover implementiert, beide neuen Individuen werden in die Kind-Population übernommen.

8.2. Die Beurteilung der Modellgüte durch das Programmsystem

Ein großer Vorteil der im Rahmen dieser Arbeit genutzten Programmkombination besteht darin, dass die Variablenselektion tatsächlich unter Berücksichtigung der erzielten Modellgüte stattfindet.

Die Validierung des gebildeten Modells der linearen Diskriminanzanalyse erfolgt mit einer *Leave-One-Out*-Schätzung, bei der jeweils ein Spektrum nach dem anderen aus der Modellbildung ausgeschlossen und klassifiziert wird. Bei der hier gegebenen Datenstruktur führt dieses Vorgehen allerdings zu einer der *Reklassifikation* sehr ähnlichen Schätzung. Der Grund dafür ist die große Ähnlichkeit der Spektren einer Messung beziehungsweise Probe untereinander. Wenn alle restlichen Spektren einer Probe im Trainingsdatensatz verbleiben, kann das aus der Modellbildung der LDA ausgeschlossende Spektrum nicht als unabhängig bezeichnet werden. Deshalb ist zu erwarten, dass die Qualität der Modelle systematisch überschätzt wird.

Für die hier genutzten Daten sollte die Validierung der gebildeten Modelle als Set-Validierung erfolgen, wobei die Sets so gewählt werden, dass jeweils ganze Proben aus der Modellbildung ausgeschlossen werden.

Weiterhin fließen *alle* Spektren in die Optimierungsrechung zur Bildung der Variablen ein, also war auch das gerade ausgeschlossene Spektrum bereits an der Modellbildung beteiligt. Diese Form der Validierung mit bereits bekannten Daten lässt sich nicht vermeiden, da nicht für jede einzelne Validierung während der Variablenbildung ein eigenes Test-Set zur Verfügung gestellt werden kann. Es besteht jedoch die Hoffnung, dass das letztgenannte Problem von untergeordneter Bedeutung ist und die Ergebnisse der Optimierung nicht beeinträchtigt. Das ist dann der Fall, wenn der systematische Fehler in der Schätzung etwa konstant ist.

Die Implementierung der Set-Validierung wird vermutlich zu einem deutlich erhöhten Rechenaufwand bei der Optimierung führen, da die Invertierung der Kovarianzmatrix für jede Modellbildung dann wahrscheinlich nicht mehr umgangen werden kann (vgl. [14; 79; 89]). Dem gegenüber steht jedoch nicht nur die Möglichkeit, deutlich bessere Modelle zu erreichen, sondern auch ein großes Potential zum Vermeiden von Rechenaufwand. Werden Parameter, die im Rahmen dieser Arbeit „von Hand“ oder überhaupt nicht optimiert wurden und prinzipiell von genetischen Algorithmen optimiert werden können, in die Optimierung mit einbezogen, so kann auf viele einzelne Rechnungen verzichtet werden.

9. Die Wahl der Parameter des Programms ga_ors

9.1. Reproduzierbarkeit der Optimierungsergebnisse

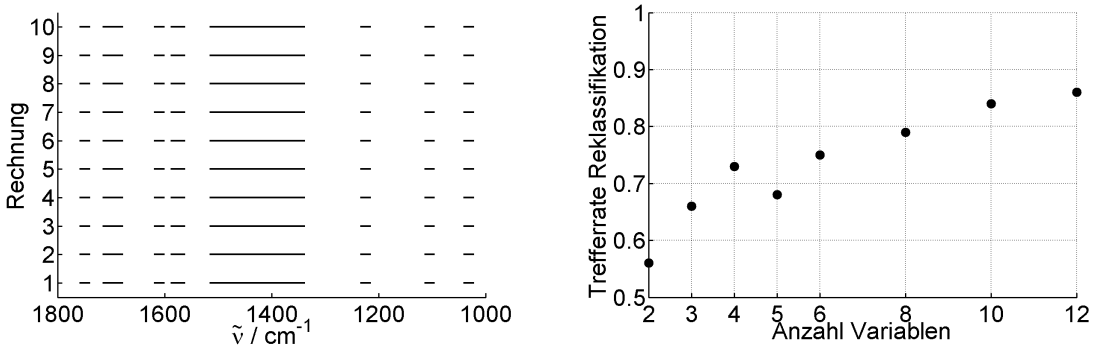
Auf einen Datensatz wurde zehnmal hintereinander der Optimierungsalgorithmus angewendet. Es resultierten identische Modelle.

Allerdings stellte sich bei näherer Betrachtung dieser zunächst sensationell gut erscheinenden Ergebnisse heraus, dass der Grund dieser perfekten Reproduzierbarkeit nicht die Konvergenz des Algorithmus bei dem globalen Maximum des Suchraums ist. Vielmehr erwiesen sich die „zufälligen“ Entscheidungen des Programms, die ja durch Verwendung von Pseudo-Zufallszahlen modelliert werden, als durchaus nicht zufällig. Sie traten mehrfachen Programmstarts exakt in der gleichen Folge auf.

Dieser Verdacht wurde durch die Tatsache untermauert, dass nicht nur die Ergebnisse in ihren Werten ununterscheidbar ähnlich sind, sondern vielmehr bereits die Ausgabedateien von ga_ors sich bei mehrfachen Läufen mit identischen Eingabedaten und Parametern jeweils nur in den angegebenen Dateinamen und Zeitmarken unterschieden. Dies wurde auch nach wenigen Generationen beobachtet, einem Stadium der Rechnung, bei dem identische oder auch nur sehr ähnliche Ergebnisse extrem unwahrscheinlich sein sollten.

Eine diesbezügliche Anfrage an die Entwickler des Programmsystems bewirkte, dass das Problem bearbeitet wird. Leider konnte es nicht rechtzeitig gelöst werden, daher erfolgten alle hier vorgestellten Rechnungen mit der fehlerhaften Programmversion.

Dieser Fehler ist für die hier untersuchten Fragestellungen von entscheidender Bedeutung. In der Folge können keinerlei Abschätzungen über die Qualität der errechneten Modelle bezüglich der mit diesem System für vorgegebene Daten erreichbaren Qualität gemacht werden. Dies bedeutet auch, dass es nicht möglich ist, statistisch gesicherte Aussagen über die Wirkungen der verschiedenen Einflussgrößen, wie zum Beispiel der Datenvorbehandlung, zu machen.



(a) Identische Ergebnisse mehrerer Rechnungen

(b) Probleme beim Auffinden des optimalen Modells mit 5 Variablen

Abbildung 9.1.: Reproduzierbarkeit der Optimierung und Probleme beim Auffinden des optimalen Modells

Weiterhin ist es nicht möglich, zu entscheiden, ob ein Datensatz ungeeignet zur linearen Diskriminanzanalyse ist oder ob die Probleme entstanden sind, weil nur lokale Optima gefunden wurden. Abbildung 9.1(b) zeigt die ermittelten Reklassifikationstrefferraten eines Datensatzes für unterschiedliche Regionenzahlen. Der Verlauf der Kurve legt nahe, dass bei dem Modell mit fünf Regionen nicht das globale Optimum gefunden wurde. Normalerweise könnten und sollten (vgl. Kap. 7.2.3, S. 36) weitere Rechnungen mit dem Datensatz durchgeführt werden, um mit großer Wahrscheinlichkeit ein besseres Modell zu erhalten. Dies ist hier nicht möglich. Auch wenn im Beispiel das Problem sehr deutlich ist, bleibt zu befürchten, dass geringfügig hinter den erwarteten beziehungsweise optimalen Ergebnissen zurückbleibende Modelle nicht einfach als solche zu erkennen sind.

Da auch die Verteilung der erreichten Modellgüten nicht ermittelt werden kann, ist keine Aussage über die Wahrscheinlichkeit ein deutlich schlechteres als das optimale Modell zu erhalten, möglich. Das hat die Folge, dass auch alle Ergebnisse, die auf dem Vergleich der errechneten Modelle beruhen, nicht auf ihre Signifikanz überprüft werden können. In dieser Arbeit betrifft das besonders die Untersuchungen zur Datenvorbehandlung (Kap. 14, S. 61).

Eine weitere Folge ist, dass bezüglich der Parameter des Programms nur sehr bedingt Empfehlungen gegeben werden können. Mutations- und Crossover-Wahrscheinlichkeit sowie die Größe der Elite-Gruppe und der gesamten Population sind Parameter, die Auswirkungen auf die Wahrscheinlichkeit, ungünstige Ergebnisse zu erhalten, haben. Daher bedürfen Empfehlungen über die zu verwendenden Werte immer der Abschätzung dieser Wahrscheinlichkeit, die mittels mehrerer unabhängiger Programmläufe durchgeführt werden müsste. Da diese Einschätzung nicht gegeben werden kann, ist es hier auch nicht möglich, Empfehlungen über eine sinnvolle Anzahl unabhängiger Programmläufe zu geben.

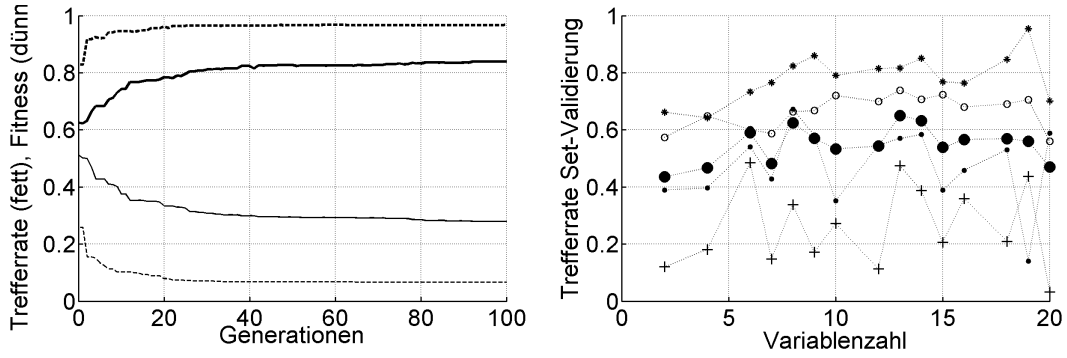
Allerdings hat dieser Fehler keine Auswirkungen auf die absolute Beurteilung des gebildeten Modells, die immer auf der Grundlage der Trefferraten des resultierenden Modells der linearen Diskriminanzanalyse erfolgt.

9.2. Populationsgröße, Größe der Elitegruppe, Crossover- und Mutations-Wahrscheinlichkeit

Wie oben dargelegt, konnte im Rahmen dieser Arbeit aufgrund des Programmfehlers keine Optimierung dieser Parameter stattfinden. Daher wurden alle Rechnungen mit 100 Individuen, davon 10 in der Elite, einer Mutations-Wahrscheinlichkeit von 0,1 % und einer Crossover-Wahrscheinlichkeit von 66 % durchgeführt. Diese Werte basieren auf den in der Literatur gegebenen Empfehlungen [14; 36].

9.3. Anzahl an Generationen

Die notwendige Anzahl an Generationen um eine Konvergenz des Algorithmus beim Optimum zu erlangen, variiert stark mit der Komplexität des Optimierungsproblems. `ga_ors` bietet jedoch mit der Bildschirmausgabe und dem `.log`-File Hinweise auf den Fortschritt der Rechnung, indem für jede gerechnete Generation die erreichte Fitness und die Reklassifikations-Trefferrate (Accuracy) für den Trainingsdatensatz angezeigt werden. Abb. 9.2(a) zeigt beispielhaft den Verlauf zweier Rechnungen, die durchgezogenen Linien geben die Trefferrate, die gestrichelten Linien die Werte der Fitness-Funktion wieder.



(a) typischer Verlauf der Optimierungsrechnungen — Trefferrate ($\hat{T}_R$, fett) und Fitness zweier Rechnungen (durchgezogen bzw. gestrichelt)

(b) Trefferraten in Abhängigkeit von der Variablenzahl — ungerade Variablenzahlen führen in einigen Fällen zu deutlich schlechteren Ergebnissen als gerade Anzahlen.

Abbildung 9.2.: Generationszahl und Einfluss der Variablenanzahl

9.4. Die Anzahl ermittelter Variablen

Die Anzahl der zu bildenden Variablen, also der durch `ga_ors` zu suchenden Regionen, ist ein Schlüsselparameter der linearen Diskriminanzanalyse (Kap. 6.4, S. 21). Die günstigste Wahl hängt dabei stark vom vorliegenden Datensatz ab. Dieser Aspekt der Variablenanzahl und -wahl wird daher in Kapitel 15.3.2 (S. 70) ausführlich für das gebildete Modell dargelegt. Hier werden die Eigenschaften des Programms `ga_ors` bezüglich der Variablenzahl diskutiert.

Die Zahl der Variablen hat zwar großen Einfluss auf die erreichbare Modellgüte, wird jedoch von `ga_ors` praktisch nicht optimiert. Stattdessen wird in aller Regel ein Modell mit der beim Programmaufruf anzugegebenden Variablenzahl gebildet, hin und wieder entstehen auch Modelle mit einer geringeren Variablenzahl. Dies ist vermutlich eine Folge der Fitness-Funktion, die im Wesentlichen der Reklassifikations-Trefferrate entspricht und dementsprechend in Richtung größerer Regionenzahlen wirkt.

Weitere Gründe dafür, dass fast immer die vorgegebene Anzahl an Regionen beibehalten wird, könnten in der Implementierung des genetischen Algorithmus liegen. Werden bei der Erzeugung der Initialpopulation Individuen mit unterschiedlichen Anzahlen an Regionen gebildet, so wird das die Optimierung der Anzahl an Variablen begünstigen. Aber auch die konkrete Realisierung der Crossover- und Mutationsoperatoren kann Auswirkungen auf die Wahrscheinlichkeit, dass Individuen mit unterschiedlichen Regionenzahlen entstehen, haben. Diese programmtechnischen Details spielen zur Zeit nur eine untergeordnete Rolle, da die Optimierung der Variablenzahl außerhalb des Programmsystems stattfindet. Prinzipiell sind genetische Algorithmen jedoch in der Lage, auch die Anzahl an Variablen zu optimieren.

In der Vergangenheit hat es sich als günstig erwiesen, gerade Variablenanzahlen zu verwenden [90]. Oft werden bessere Ergebnisse als bei Verwendung ungerader Regionenzahlen erzielt. Der Grund dieses erstaunlichen Verhaltens ist nicht bekannt, aber auch während der im Rahmen dieser Arbeit durchgeführten Rechnungen konnte dieses Verhalten beobachtet werden (Abb. 9.2(b)).

Insgesamt schwanken die Ergebnisse in Abhängigkeit von der Variablenzahl sehr stark. Dadurch wird auch die Festlegung einer günstigen Anzahl zu suchender Regionen oder Variablen erschwert.

10. Weitere Besonderheiten der genutzten Programme

10.1. Zuordnungswahrscheinlichkeiten

Die eigentliche Klassifikation wird durch `stackedGen` durchgeführt. In der Ergebnisdatei ist eine Tabelle mit Zahlenwerten für jedes Spektrum und jede Klasse, die vermutlich die a posteriori Wahrscheinlichkeiten für die einzelnen Klassen angibt. Daher bezieht sich der Begriff der *a posteriori Wahrscheinlichkeit* im Zusammenhang mit `stackedGen` auf diese Werte.

Abbildung 10.1 zeigt die Verteilung der ermittelten a posteriori Wahrscheinlichkeiten für Trainings- und Testdaten eines Datensatzes, bei dem die Testdaten eine Kopie der Trainingsdaten waren. Ein Datensatz aus identischen Daten im Trainings- und Testset sollte praktisch gleiche Ergebnisse für Test- und Trainingsdaten erreichen, was auch in den Zuordnungsmatrizen für alle Spektren beobachtet wurde.

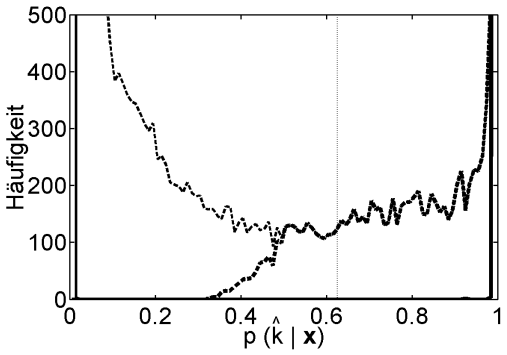


Abbildung 10.1.: Verteilung der a posteriori Wahrscheinlichkeiten für Trainings- (gestrichelt) und Testdaten (durchgezogen). Verteilung a posteriori Wahrscheinlichkeiten für die Klasse, zu der die Spektren zugeordnet wurden (fett).

Größere Differenzen traten in den Zuordnungsmatrizen der „sicher“ klassifizierten Spektren auf. Für den Testdatensatz wurden alle Spektren „sicher“ zugeordnet, im Gegensatz dazu wurden nur etwa 78 % der Trainingsspektren als „sicher“ klassifiziert eingestuft. Der Anteil richtiger Zuordnungen lag mit 89,5 % deutlich über dem Anteil richtiger Zuordnungen bei Betrachtung aller Spektren (83,2 %). Diese Verbesserung spielt bei den Trainingsdaten nur eine geringe Rolle, da ja alle diese Spektren in die Modellbildung eingehen. Dagegen wäre eine solche Unterscheidung bei Testdaten sehr hilfreich.

„Sicher“ und „unsicher“ bezieht sich im Rahmen dieser Auswertung auf die durch `stackedGen` vorgenommene Kennzeichnung der Spektren. Zuordnungen mit einer maximalen a posteriori Wahrscheinlichkeit von weniger als 62,5% sind in der Ergebnisdatei mit einer Tilde (`~`) als „unsicher“ gekennzeichnet. Prinzipiell besteht die Möglichkeit, eine beliebige Grenze zu setzen, da die Ergebnisdatei die a posteriori Wahrscheinlichkeiten für alle Klassen auflistet (vgl. Kap. B.1.3, S. 93).

Die Ergebnisse für die Testdaten können also aufgrund der sehr scharfen Zuordnung nicht hinsichtlich ihrer Zuverlässigkeit beurteilt werden. Das erstaunt insofern, als in der Literatur diese a posteriori Wahrscheinlichkeiten zur Einstufung der Verlässlichkeit der Ergebnisse verwendet wurden [14; 23]. Der Grund dieser stark unterschiedlichen Verteilungen konnte nicht herausgefunden werden.

10.2. Größe der mutierten Blöcke und Abbruchbedingungen

Eine weitere Auffälligkeit betrifft die Abbruchbedingungen des Programms. Wie bereits beschrieben, werden zwei Kriterien genutzt: zum einen die beim Aufruf angegebene Höchstgrenze an Generationen und zum anderen als indirektes Kriterium der Modellgüte die Größe der zu mutierenden Blocks (vgl. Kap. 8.1, S. 38).

Für die in dieser Arbeit verwendeten Spektren mit einer Länge von 208 Datenpunkten beträgt die anfängliche Länge der zu mutierenden Blocks $\frac{p}{64} = 3,25 \approx 4$ (aufgerundet), das entsprechende Abbruchkriterium ist jedoch immer eine Länge von 6 Datenpunkten. Von diesem Kriterium ist gegenwärtig nicht bekannt, wie es verändert werden kann. Das heißt, dass für die hier verwendeten Daten immer nur das Abbruchkriterium der Generationszahl Anwendung findet.

Allerdings bedeutet das feste Abbruchkriterium für Spektren mit mehr Datenpunkten, für die die anfängliche Länge der zu mutierenden Blocks wenig oberhalb von 6 liegt, unter Umständen Probleme bei der Optimierung, weil das Endkriterium der Länge der zu mutierenden Blocks vor der Konvergenz des Algorithmus erreicht werden kann.

10.3. Programmabbrüche

Vereinzelt wurden Programmabbrüche bei der Optimierung beobachtet. Sie traten für einzelne Datensätze in Verbindung mit bestimmten Parametern auf. Mit veränderten Parametern, zum Beispiel einer anderen Regionenzahl, verliefen die Rechnungen jedoch problemlos. Diese Programmabbrüche erfolgten ohne Fehlermeldung und traten auch bei weiteren Starts mit den kritischen Parametern immer an derselben Stelle auf, wobei letzteres aufgrund des nicht-zufälligen Verhaltens des Programms zu erwarten war.

10.4. Schwierigkeiten beim Auffinden des optimalen Modells

Generell lässt sich der Eindruck festhalten, dass verschiedentlich Schwierigkeiten beim Auffinden des globalen Optimums auftraten. Das betrifft die Wahrscheinlichkeit, deutlich schlechtere als die optimalen Modelle zu erhalten, wie es am Beispiel in Kap. 9.1 (S. 41) dargelegt wurde. Modelle, die auffällig hinter den erwarteten Ergebnissen zurückblieben, wurden im Rahmen dieser Arbeit mehrfach beobachtet. Auch die in Abhängigkeit der Variablenzahl sehr stark schwankenden Ergebnisse, sowohl für ganze Modelle als auch für einzelne Klassen, zeigen die Schwierigkeiten bei der Lokalisierug der besten Regionen.

Das legt den Schluss nahe, dass die gewählten Parameter für die Programmläufe von `ga_ors` nicht optimal sind. Sicherlich spielt hierbei auch die in der Literatur beklagte Situation eine Rolle, dass nur wenige und sehr allgemeine Richtlinien zur Wahl der Parameter existieren. Insgesamt besteht, verglichen mit anderen Verfahren, wenig Erfahrung mit dem Verhalten stochastischer Algorithmen in der Anwendung auf chemometrische Probleme (vgl. besonders [48; 51; 52]).

Teil IV.

Chemometrische Untersuchung der Daten

11. Beurteilung der Modellgüte

Im Rahmen dieser Arbeit wurden drei verschiedene Kriterien zur Beurteilung der Qualität der erhaltenen Modelle genutzt, Abb. 11.1 zeigt den Verlauf dieser Schätzmethoden für die in Kap. 15 (S. 67) diskutierten Trainingsdaten.

11.1. Die Reklassifikations-Trefferrate

Zwar ist die Beurteilung der Modellgüte mittels der Reklassifikations-Trefferrate $\hat{T}_R$ problematisch, sie kann jedoch insofern zur Beurteilung herangezogen werden, als sie eine Obergrenze der erreichbaren Trefferrate darstellt. Es ist nicht zu erwarten, dass das Modell mit fremden Daten bessere Trefferraten als mit den Trainingsdaten erzielt.

Die Tendenz, die Modellgüte zu überschätzen, ist bei der absoluten Beurteilung des Modells ein Problem. Vergleiche zwischen unterschiedlichen Modellen können jedoch in gewissem Rahmen erfolgen, wenn ähnlich große systematische Fehler beider Schätzungen vorliegen.

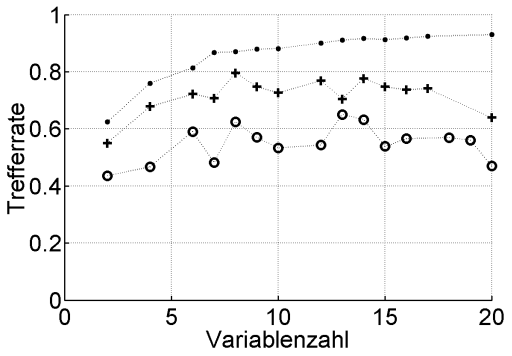


Abbildung 11.1.: Die verschiedenen Schätzmethoden der Trefferrate — Validierung von Optimierung und LDA ($\hat{T}_{Opt}$, o), Validierung der LDA ($\hat{T}_{LDA}$, +) und Reklassifikations-Trefferrate ($\hat{T}_R$, ·)

Für unterschiedliche Anzahlen an Variablen ist der systematische Fehler der Reklassifikationsrate sehr unterschiedlich groß. Bei Verwendung gleicher Datensätze und identischer Regionenzahlen wird jedoch erwartet, dass die Fehler der einzelnen Schätzungen nicht zu unterschiedlich sind.

Da die Reklassifikations-Schätzung der Trefferrate nur sehr geringe Ressourcen beansprucht, wurde sie in dieser Arbeit zur vergleichenden Beurteilung der Modellgüte herangezogen.

Besonders die Rechnungen zur Untersuchung der verschiedenen Datenvorbehandlungsmethoden wurden mit dieser Kenngröße verglichen, die vielversprechendsten Modelle wurden dann genauer untersucht.

11.2. Die Validierung der LDA-Modelle

Als weitere Möglichkeit zur Beurteilung der Modellgüte wurde die Set-Validierung der gebildeten LDA-Modelle auf der Grundlage derselben Variablen, also derselben spektralen Regionen, genutzt. Im Folgenden wird diese Kenngröße als „Trefferrate der LDA“ $\hat{T}_{LDA}$ bezeichnet.

Auch dieser Schätzer weist einen systematischen Fehler in Richtung zu positiver Trefferraten auf, da zur Ermittlung der spektralen Regionen alle Daten benutzt wurden. Allerdings bildet diese Schätzung bereits die Abhängigkeit der Diskriminanzanalyse von der

Anzahl der Variablen ab, das Maximum der Trefferrate gibt einen Hinweis auf geeignete Variablenzahlen.

Diese Methode wurde im Rahmen der vorliegenden Arbeit zur Bestimmung vielversprechender Anzahlen an Variablen genutzt.

Die Berechnung der Trefferrate der LDA bewirkt gegenüber der Schätzung der Reklassifikations-Trefferrate bereits deutlich erhöhten Zeitaufwand und auch der Bedarf an Speicherplatz steigt stark an.

11.3. Set-Validierungen beider Modelle

Diese Beurteilung beider Rechnungen — der Optimierung und der Diskriminanzanalyse — sollte die realistischste Schätzung der Trefferrate, $\hat{T}_{Opt}$, ermöglichen. Allerdings sind die benötigten Ressourcen bei der Berechnung so groß, dass diese Validierung nur für einzelne Modelle durchgeführt werden konnte.

Ein weiteres Problem dieser Validierung ist, dass manchmal die vorgegebene Regionenzahl bei der Optimierung nicht eingehalten wird. Dann können für die vorgegebene Anzahl an Regionen nicht für alle Untermengen des Datensatzes Zuordnungsergebnisse erhalten werden. Das betrifft insbesondere große Regionenzahlen, bei den verwendeten Spektren mit 208 Datenpunkten Länge macht sich das oberhalb von etwa 12 Regionen deutlich bemerkbar. Die Validierungsergebnisse für die vorgegebene Regionenzahl sind dann entsprechend mit einer größeren Unsicherheit behaftet.

Tab. 11.1 gibt einen überblick über den Ressourcenbedarf typischer Rechnungen.

Tabelle 11.1.: Faustregeln für benötigte Ressourcen der Rechnungen

Rechnung	**Ressourcenbedarf**
Rechenzeiten	
Trainingsdaten	
Optimierung (8400 Spektren à 208 Datenpunkte, 10 Regionen)	1 h
pro Wert und Region	ca. 200 μs
Diskriminanzanalyse	<1 min
Trainings- und Testdaten	
Diskriminanzanalyse (8400 Trainings + 2200 Testspektren)	1 – 5 min
Optimierung pro Spektrum, Datenpunkt und Region	190 μs
erzeugen der Eingabedatei (10600 Spektren à 208 Datenpunkte)	ca. 20 min
Speicherplatzbedarf	
Eingabedatei (pro Wert)	ca. 9 Byte
Protokoll-Datei `ga_ors`	10 – 15 kByte
Ergebnis-Datei `stackedGen` (pro Spektrum)	ca. 66 Byte

Wird ein Datensatz mit 10000 Spektren à 200 Wellenzahlen in sechs Sets zur Validierung geteilt, die jeweils für 2, 4, ... 20 Regionen optimiert werden, um die günstigste Variablenzahl für das Trainingsset zu ermitteln, so ist mit einer Rechenzeit von ca. 70 h oder knapp drei Tagen zu rechnen.

12. Herkunft der Daten

12.1. Präparation [37]

Für die Untersuchungen standen Spektren zur Verfügung, die im Rahmen des Projekts „molekulare Endospektroskopie“ aufgenommen wurden [37; 70].

Die Gewebeproben stammen aus Tumor-Resektionen am Universitätsklinikum Carl-Gustav-Carus der TU Dresden, sie wurden in flüssigem Stickstoff schockgefroren und bei -80 °C aufbewahrt. Die Schnitte wurden mit einem Mikrotom bei circa -5 – -15 °C angefertigt, die Schnittdicke wurde dabei zwischen 5 und 25 μm gewählt. Noch während des Schneidens wurden die Schnitte auf CaF_2-Probenträger überführt und dann dort getrocknet. Die Probenträger wurden dazu mit Aceton und Ethanol gereinigt, eventuell verbliebene organische Rückstände durch zehnminütige Behandlung im Plasmacleaner und Absaugen mit Vakuum entfernt. Aus den Absorptionseigenschaften des CaF_2-Probenträgers resultiert die langwellige Grenze der Spektren bei 950 cm^{-1}.

Ein Einbettungsmedium wurde nicht verwendet. Die dadurch verursachten Komplikationen wie Risse, Löcher und unterschiedliche Dicken der Proben wurden in Kauf genommen, um störende Absorptionen des Einbettungsmediums in den Spektren zu vermeiden. Die variierende Dicke innerhalb einer Probe bewirkt auch eine Variation der Extinktionswerte der Spektren und muss daher bei der Datenvorbehandlung ausgeglichen werden.

Zu jedem Schnitt für die spektroskopische Untersuchung wurde ein direkt angrenzender weiterer Schnitt angefertigt, der mit *Hämatoxylin-Eosin* gefärbt wurde und so als Referenz für die histologische Begutachtung zur Verfügung steht.

12.2. Maps

Ein Teil der IR-Maps wurde an einem Nicolet 5PC Spektrometer mit Olympus BH2 IR-Mikroskop (15fache Vergrößerung) aufgenommen, sie umfassen einen Wellenzahlbereich von 950 – 4000 cm^{-1}. Verwendet wurden Spektren, die von einer Probenfläche von 90 μm × 90 μm aufgenommen wurden, diese Begrenzung wurde durch eine Blende realisiert. [37]

Die Spektrendaten dieses Gerätes lagen zunächst im JCAMP-DX-Format vor. Da das Einlesen und Konvertieren einige Zeit benötigt, wurden alle Daten mit weiteren Informationen wie zum Beispiel Diagnose, Herkunft der Daten und Aufnahmedatum als `Matlab`-Datei[1] gespeichert.

Weitere Daten stammen von einem Bruker IFS66/S Spektrometer in Verbindung mit dem Hyperion™-System. Die Maps wurden mit dem Einkanal-MCT-Detektor im Spektralbereich zwischen 950 und 3800 cm^{-1} aufgenommen. Die Messparameter entsprechen denen der Maps des Nicolet-Gerätes [37; 70].

Diese Spektren liegen in der Regel als .SPC-Dateien vor. Obwohl das Konvertieren dieses Dateiformats deutlich schneller geht, als das bei den JCAMP-DX-Dateien der Fall ist, wurden auch diese Daten zuerst in das genutzte `Matlab`-Format umgewandelt.

[1] Matlab, Version 6, ©1984 – 2001 The Mathworks, Inc.

12.3. Images

Die Images wurden ebenfalls an dem Bruker IFS66/S Spektrometer mit Hyperion™-System aufgenommen, sie wurden mit einem 64 × 64 Pixel FPA-Detektor gemessen. Mit der vorliegenden Optik des Hyperion-Systems bildet ein solches Image eine Fläche von etwa 270 μm × 270 μm ab, die Kantenlänge eines Pixels entspricht also ungefähr 4 μm. [70]

Diese Daten liegen im `Matlab`-lesbaren ENVI-Format vor, allerdings besteht hierbei keine Möglichkeit, die $\tilde{\nu}$-Achse zu erhalten. Die Messsoftware OPUS beinhaltet zwar die Option, Daten im JCAMP-DX-Format abzuspeichern, diese Möglichkeit besteht aber nur für einzelne Spektren.

12.3.1. Hintergrundkorrektur der Images

Während die Maps bereits als Hintergrund-korrigierte Extinktionsspektren vorliegen, handelt es sich bei den Images um Einkanal-Absorptionsspektren. Die Hintergrundkorrektur und die Umrechnung der Absorptionswerte in Extinktionswerte bei den Images wurde daher in `Matlab` durchgeführt.

Die Hintergrundkorrektur ist jedoch bei den Images mit größeren Problemen behaftet als bei den Daten der Maps. Es wird empfohlen, am Rand der Meßfläche liegende Spektren des sauberen Probenträgers für die Hintergrundkorrektur zu verwenden [70].

Gegen dieses Vorgehen sprechen die beim FPA starken Unterschiede in der Detektorcharakteristik der einzelnen Pixel sowie die örtlich variierende Ausleuchtung der Probe.Sollen die Spektren mit den hier angewandten Methoden ausgewertet werden, so müssen diese Unterschiede zwischen den einzelnen Punkten kompensiert werden. Daher ist eine Hintergrundkorrektur mit Images leerer Objektträger unumgänglich.

Auch die Beschränkung, nur am Rand der Proben zu messen, ist problematisch. Dort treten besonders oft Falten und Risse auf, so dass die Spektren häufig eine schlechte Qualität haben.

Eine gute Hintergrundkorrektur ist besonders wichtig, wenn die Images gemeinsam mit den bereits vorhandenen Maps ausgewertet werden sollen.

Da zu 40 im Mikroskop gemessenen Images von Proben nur zwei Hintergrund-Images vorhanden waren, wurde versucht, durch Messung weiterer acht Hintergrund-Images und Mittelwertbildung über alle zehn Hintergrund-Images eine Korrektur unter Berücksichtigung der Ortsabhängigkeit zu erzielen. Allerdings kann die Zeitabhängigkeit der Detektorcharakteristik über einen Zeitraum von mehreren Monaten zwischen der Aufnahme der Proben-Images und der Hintergrund-Images nicht von vornherein vernachlässigt werden.

Weder mit den vorhandenen Hintergrund-Images noch mit dem gemittelten Hintergrund konnte eine gute Korrektur erzielt werden. Die meisten Spektren wiesen über weite Bereiche negative Extinktionswerte auf, ein deutliches Zeichen, dass keine gute Hintergrundkorrektur erreicht wurde.

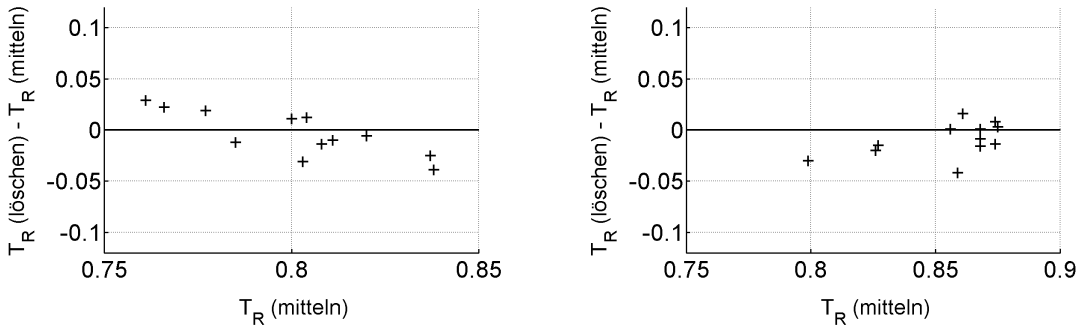
12.4. Modellbildung mit Daten der verschiedenen Geräte

12.4.1. Unterschiedliche Wellenzahl-Achsen

Die Daten des Nicolet-Geräts unterscheiden sich von den Bruker-Maps nicht nur im Dateiformat, sondern auch in der $\tilde{\nu}$-Achse. Die am Nicolet-Gerät gemessenen JCAMP-DX-Dateien bestehen aus 3164 Messstellen im Bereich zwischen 949,8 und 3999,6 cm^{-1}. Die

am Bruker-Gerät gemessenen Spektren wurden dagegen mit 740 Punkten zwischen 948,8 und 3799,1 cm^{-1} abgespeichert. Um eine einheitliche $\tilde{\nu}$-Achse der Spektren zu erhalten, wurde jeweils der erste Messwert der .SPC-Dateien entfernt und die Spektren des Nicolet-Geräts mittels der `matlab`-Funktion `interp1` auf die verkürzte $\tilde{\nu}$-Achse der Bruker-Daten interpoliert. Die so erhaltenen Spektren bestehen aus 739 Punkten im Bereich von 952,7 bis 3799,1 cm^{-1}, auch diese Daten wurden als `Matlab`-Dateien gespeichert, da die Interpolation einen großen Zeitaufwand bedeutet. Die Spektren der Images bestehen aus 739 Punkten zwischen etwa 950 und 3800 cm^{-1}. Da die exakte $\tilde{\nu}$-Achse noch nicht nach `Matlab` exportiert werden konnte, wurde zunächst angenommen, dass sie der $\tilde{\nu}$-Achse der Maps des Bruker-Gerätes entspricht.

Die Spektren des Nicolet-Gerätes umfassen etwa die vierfache Anzahl an Datenpunkten gegenüber den Spektren des Bruker-Gerätes, die durch die benutzte Interpolationsfunktion jedoch nicht alle genutzt werden. Daher wurde untersucht, ob eine Verbesserung der Modelle durch Mittelwertbildung über jeweils vier Punkte vor der Interpolation auf die neue Wellenzahl-Achse erreicht werden kann.



(a) Datensatz „a“ **(b)** Datensatz „l“

Abbildung 12.1.: Wirkungen der Mittelwertbildung vor der Interpolation auf eine gemeinsame $\tilde{\nu}$-Achse

Abb. 12.1 zeigt die Ergebnisse verschiedener Rechnungen mit und ohne Mittelwertbildung für zwei verschiedene Datensätze. Dabei ist die Differenz der erzielten Trefferraten,

$$\hat{\mathrm{T}}_{\mathrm{R}}\,(\text{keine Mittelwertbildung}) - \hat{\mathrm{T}}_{\mathrm{R}}\,(\text{Mittelwertbildung})\,,$$

über den Trefferraten mit Mittelwertbildung aufgetragen.

Die erreichten Reklassifikations-Trefferraten mit Mittelwertbildung waren geringfügig größer. In jeweils sieben von zwölf Fällen waren die mit Mittelwertbildung erhaltenen Modelle besser. Die Unterschiede sind jedoch so gering, dass daraus keine Empfehlung abgeleitet werden kann. Im Rahmen dieser Arbeit wurden die nach Mittelwertbildung erhaltenen Daten genutzt.

12.4.2. Diskriminanzanalyse mit Daten der unterschiedlichen Geräte

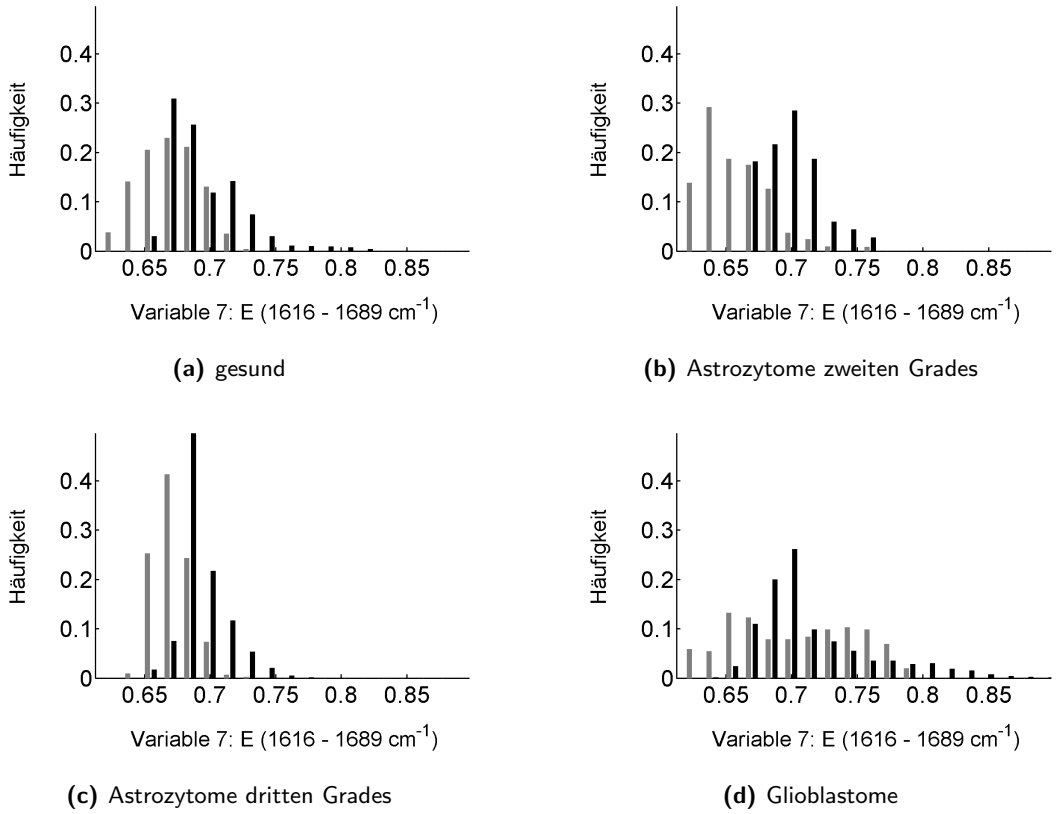
Aus den Maps beider Geräte ließ sich problemlos ein Modell bilden, dagegen traten bei der Integration der Images Probleme auf.

Mittels einer Set-Validierung der LDA mit sechs Sets wurde die Trefferrate bestimmt, sie sank von 79,6 % für den in Kap. 15 (S. 67) beschriebenen Trainingsdatensatz auf

58,9 %, als die zur Verfügung stehenden Images einbezogen wurden. Abb. 12.2 zeigt die Verteilungen der gebildeten Variablen für die einzelnen Klassen, eine genaue Diskussion dieser Auftragung findet in Kap. 13.1.1 (S. 54) statt.

Es fällt auf, dass die Images durchweg niedrigere Werte aufweisen als die Spektren der Maps. Der Unterschied zwischen den Daten der verschiedenen Geräte ist dabei größer als die Unterschiede zwischen den einzelnen Klassen innerhalb der Spektren der Maps.

Gegenwärtig ist es daher nicht möglich, innerhalb eines Modells sowohl Images als auch Maps zu verwenden.



(a) gesund **(b)** Astrozytome zweiten Grades

(c) Astrozytome dritten Grades **(d)** Glioblastome

Abbildung 12.2.: Verteilung der Variablen 7 $E(1616 - 1689cm^{-1})$ der Maps (schwarz) und Images (grau). Die Häufigeiten wurden skaliert.

13. Erstellen eines Trainingsdatensatzes

13.1. Ermittlung zur Modellbildung geeigneter Proben

Im vorliegenden Fall sind zwei weitere Besonderheiten gegenüber der üblichen Problematik, einen geeigneten Trainingsdatensatz zur Schätzung der Modellparameter zu erstellen, zu beachten. Zum einen handelt es sich genau genommen um zwei getrennte Rechenschritte, wenn sie auch unter gegenseitiger Bezugnahme durchgeführt werden. Damit wird gleichzeitig die Güte zweier zusammenwirkender Modelle getestet.

Zum anderen ist die vorliegende Datenstruktur in Ebenen gegliedert: zu jeder Probe können mehrere Schnitte und Messungen existieren, jede dieser Messungen besteht aus vielen Spektren.

Die untersuchten Kriterien zur Erkennung geeigneter Spektren werden im Rahmen der Methoden zur Datenvorbehandlung in Kap. 14.5 (S. 64) vorgestellt, dieses Kapitel behandelt die Erkennung zur Modellbildung geeigneter und ungeeigneter Proben und Messungen.

13.1.1. Ausschluss untypischer Proben und Messungen

Die Erstellung des Trainingsdatensatzes erfolgte, indem zunächst alle zur Verfügung stehenden Proben in den Trainingsdatensatz aufgenommen wurden und dann diejenigen Proben und Messungen ausgeschlossen wurden, die als nicht repräsentativ erkannt wurden.

Die Erkennung ungeeigneter Proben stützt sich dabei auf zwei grundverschiedene Wege. Einerseits ist das Ziel dieser Untersuchungen, möglichst allgemein anwendbare und automatisierbare Wege aufzuzeigen, einen Datensatz zur Modellbildung zusammenzustellen, entsprechend wurde versucht, allein aus den Daten auf die Eignung der Messungen zu schließen. Andererseits beruht der Erfolg der Auswertung darauf, dass zur Modellbildung typische Daten der einzelnen Klassen benutzt werden. Daher wurde auch die Begutachtung der einzelnen Proben unter histologischen Gesichtspunkten berücksichtigt.

Die Erkennung untypischer Objekte setzt immer das Wissen, was „typisch“ ist, voraus. Alle Verfahren, die allein auf den spektroskopischen Daten aufbauen, stützen sich also darauf, dass nur einzelne untypische Spektren zwischen vielen typischen vorliegen. Bei den vorliegenden Daten ist diese Voraussetzung vermutlich nicht erfüllt. Dazu tragen mehrere Gründe bei. Zum einen ist der Stichprobenumfang für gesundes Gewebe und Astrozytome zweiten und dritten Grades sehr gering, so dass jede einzelne Probe stark in die Auswertung eingeht. Zum anderen ist zu erwarten, dass zusätzlich zu den biologischen Unterschieden zwischen den einzelnen Proben die unterschiedliche Behandlung der Proben zu weiteren Varianzen geführt hat.

Die Reklassifikations-Trefferrate

Ein Ansatz zur Entscheidung, ob die Spektren einer bestimmten Probe als Trainingsdaten geeignet sind, ist die Betrachtung der Trefferrate der Reklassifikation. Ist diese gut, so ist zunächst davon auszugehen, dass diese Daten als Trainingsdaten geeignet sind. Allerdings ist dieses Entscheidungskriterium mit einigen Problemen verbunden.

Sind die Reklassifikationsergebnisse schlecht, weil der Datensatz an der Grenze zwischen zwei Klassen liegt, so kann dies ein wichtiger Hinweis sein, ihn als Trainingsprobe zu *verwenden*, weil unter Umständen die Schätzung der Klassengrenzen korrigiert werden muss. Dieses Problem ist insbesondere bei nur geringen Zahlen an geeigneten Trainingsproben gehäuft zu erwarten, da die Schätzung der Klassengrenzen mit entsprechend größeren Unsicherheiten behaftet ist. Ein Hinweis auf das Vorliegen dieser Situation können die ermittelten a posteriori Wahrscheinlichkeiten liefern. Sie sollten in diesem Fall für mindestens zwei Klassen ähnlich groß und damit auch für die Klasse mit der höchsten a posteriori Wahrscheinlichkeit nicht zu groß sein.

Sind die Reklassifikationsergebnisse gebietsweise sehr unterschiedlich, so sollte die Probe nochmals unter dem Mikroskop begutachtet werden, da die histologische Einstufung in der Regel nur den schlimmsten Tumortypen angibt, nicht jedoch, ob eventuell andere Teile der Probe anderen Gewebetypen angehören. Ob dieser Fall der Fehlklassifikationen vorliegen kann, ist leicht anhand einer Darstellung der Reklassifikationsergebnisse im Raster der Spektrenaufnahme zu sehen. Ist das der Fall, so muss die Probe selbstverständlich für Trainingszwecke so aufgeteilt werden, dass nur eindeutig der angegebenen Diagnose zugehörige Spektren verwendet werden.

Besonders die sehr geringen Stichprobenumfänge für gesunde Proben sowie Astrozytome zweiten und dritten Grades lassen keine Beurteilung anhand der Reklassifikations-Trefferrate zu, da jede einzelne Probe einen sehr großen Einfluss auf die Modellbildung ausübt.

In der Gruppe der Astrozytome zweiten Grades wich eine Messung sehr stark von allen anderen Messungen, auch der übrigen Klassen, ab. Das führte dazu, dass die Klasse „Astrozytom zweiten Grades“ um dieses entfernt liegende und daher gut abzutrennende Gebiet modelliert wurde. Daher erzielte diese Probe in der Reklassifikation exzellente Ergebnisse, die restlichen Astrozytome zweiten Grades wurden mit sehr geringen Trefferraten zugeordnet. Eine erneute Begutachtung des *Hämatoxylin-Eosin gefärbten* Schnittes (Abb. 13.1)ergab, dass es sich nicht um eine normale Gewebeprobe handelt. Auch die in den folgenden Abschnitten vorgestellten Wege zur Erkennung untypischer Messungen beziehen sich beispielhaft auf die Messung dieser Probe.

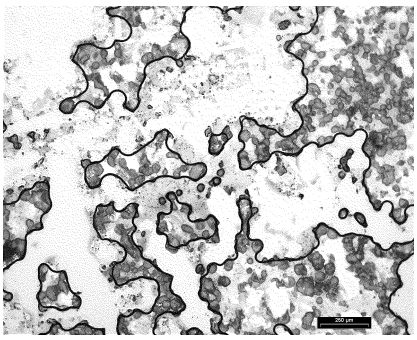


Abbildung 13.1.: HE-Schnitt Probe 149 — der Messbalken entspricht 250 μm

Diese Methode kann also nicht zur Beurteilung der Eignung einzelner Messungen für den Trainingsdatensatz empfohlen werden.

Die Verteilungen der Variablen der linearen Diskriminanzanalyse

Die Verteilungen der bei der Optimierung gebildeten Variablen, die das Koordinatensystem der Diskriminanzanalyse bilden, geben exzellente Hinweise auf das Vorliegen nicht repräsentativer Messungen.

Hilfreich ist dabei, nicht nur die Verteilung aller Daten einer Klasse, sondern auch die Verteilungen für die einzelnen Proben zu begutachten.

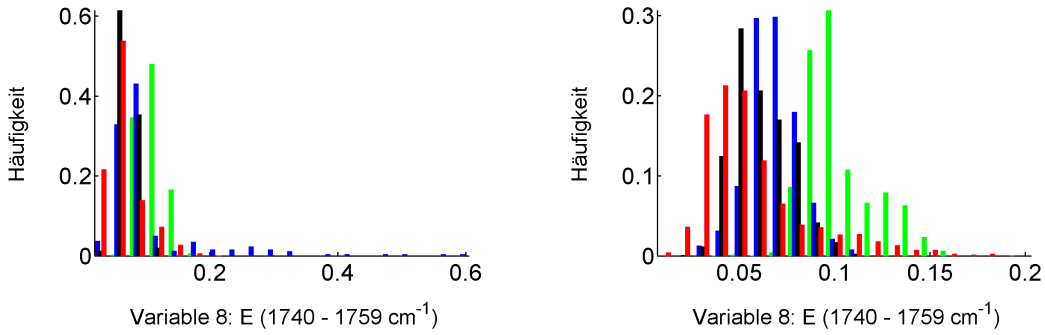
Diese Methode kann dann durchgeführt werden, wenn die spektralen Regionen, die die Variablen der LDA bilden, als Ergebnis einer Optimierungsrechung vorliegen. Gegen die

Begutachtung der Verteilungen an den einzelnen Wellenzahlen der Spektren spricht nicht nur der erhebliche Aufwand, sondern vor allem auch die Tatsache, dass diese Verteilungen in der Regel sehr deutlich nicht-normal sind. Nach dem *zentralen Grenzwertsatz* ist jedoch für genügend breite Regionen eine Annäherung an die Normalverteilung zu erwarten.

Den Darstellungen der Histogramme der einzelnen Proben sind Hinweise auf Proben mit einzelnen problematischen Spektren zu entnehmen, da sich diese durch breit auslaufende Verteilungen äußern. Die genaue Identifikation der problematischen Spektren ist auf diesem Weg zunächst jedoch nicht möglich.

Ein weiterer Vorteil dieser Darstellungen ist, dass sie eine erste Einschätzung der Daten bezüglich der Voraussetzungen der linearen Diskriminanzanalyse ermöglichen. Die Daten müssen innerhalb der Klassen multivariat normalverteilt sein. Die einzelnen Klassen müssen homogene Kovarianzmatrizen und unterschiedliche Mittelwerte aufweisen, damit die lineare Diskriminanzanalyse erfolgreich durchgeführt werden kann.

Grobe Abweichungen von der Normalverteilung und der Homogenität der Kovarianzmatrizen können an den Histogrammen erkannt werden. Dies ist hier besonders wichtig, da die Durchführung statistischer Tests zum Prüfen dieser Voraussetzungen in der Regel an größere Stichprobenumfänge gebunden ist und auch die Angabe des Stichprobenumfangs aufgrund der Datenstruktur problematisch ist.

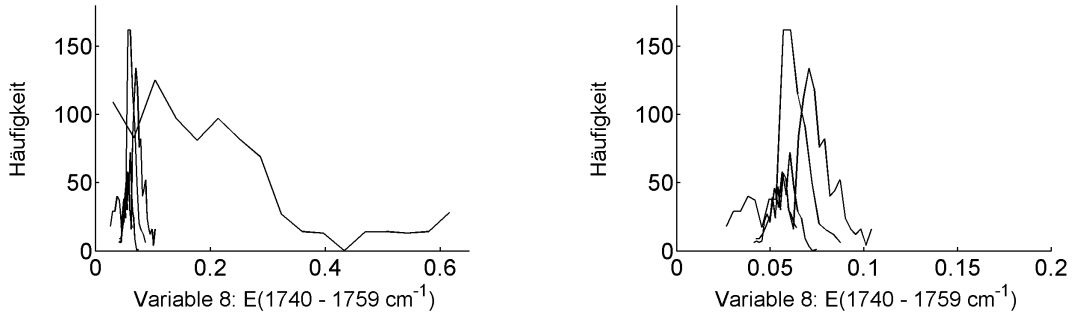


(a) breit auslaufende Verteilung der Astro II) **(b)** nach Ausschluss der untypischen Messung

Abbildung 13.2.: Verteilung einer gebildeten Variablen für die verschiedenen Klassen: Gesund (grün), Astrozytome zweiten Grades (blau), Astrozytome dritten Grades (schwarz), Glioblastome (rot).

Die Abbildungen 13.2 und 13.3 zeigen beispielhaft unterschiedliche Darstellungen dieser Verteilungen einer Variablen.

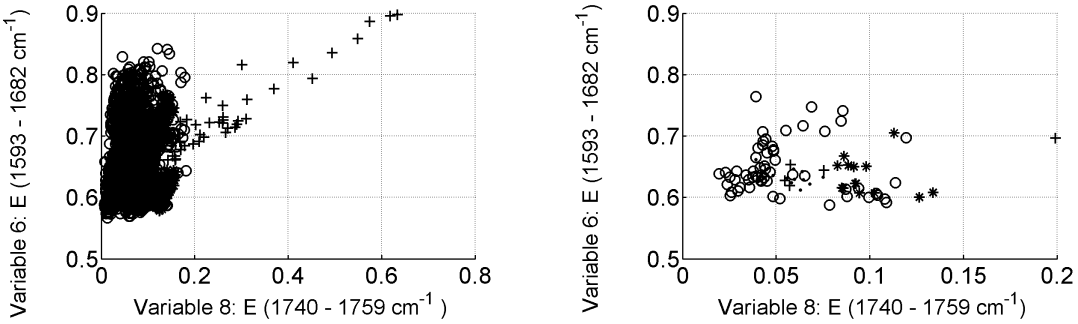
In Abb. 13.2 sind die Verteilungen der Daten der einzelnen Klassen aufgetragen. Bereits mit bloßem Auge ist zu erkennen, dass die Daten weder normalverteilt sind noch homogene Kovarianzmatrizen aufweisen. Die Verteilung der Astrozytome zweiten Grades dominiert den gesamten Wertebereich, sie hat Ausläufer bis zu sehr hohen Werten. Dies ist eine für die lineare Diskriminanzanalyse sehr problematische Verletzung der Voraussetzungen, da die Festlegung der Grenzen zwischen den Klassen erschwert wird. Die hierfür verantwortlichen Daten sollten zunächst aus den Trainingsdaten ausgeschlossen werden. Auch die Verteilung der anderen Klassen ist deutlich nicht-normal, die hier vorliegenden Abweichungen sind jedoch für die lineare Diskriminanzanalyse als nicht so problematisch einzustufen, da die Klassen über einen sehr viel engeren Wertebereich mit verhältnismässig scharfen Grenzen verteilt sind. Die großen Abweichungen von der Normalverteilung verwundern



(a) eine Messung allein ist für die ungewöhnliche Verteilung verantwortlich

(b) nach Ausschluss der untypischen Messung

Abbildung 13.3.: Verteilung einer gebildeten Variablen für die einzelnen Messungen innerhalb der Klasse „Astrozytome zweiten Grades“



(a) Auffällig ist die weit auslaufende Verteilung der Astrozytome II zu hohen Werten beider Variablen hin.

(b) Mittelwertspektren Spektren über zwei der Variablen aufgetragen

Abbildung 13.4.: Darstellung im Koordinatensystem der gebildeten Variablen — Klassen: Gesund (*), Astrozytome zweiten Grades (+), Astrozytome dritten Grades (.), Glioblastome (o)

wegen der geringen Probenzahlen nicht. Gegenüber der Verteilung der Extinktionswerte einzelner Messungen an einer Wellenzahl ist bereits eine deutliche Annäherung an die Normalverteilung zu verzeichnen (vgl Abb. 14.5(a), S. 65).

Abb. 13.3 gibt die Verteilung der einzelnen Messungen — hier aller Messungen der Astrozytome zweiten Grades — wieder. Es wird deutlich, dass eine einzelne Messung für das breite Auslaufen der Verteilung der Astrozytome zweiten Grades zu größeren Werten hin verantwortlich ist. Die Verteilungen der Variablen unter Ausschluß dieser Messung sind in den Abbildungen 13.2(b) und 13.3(b) gezeigt.

Diese Methode der Ausreisser-Erkennung bzgl. einzelner Messungen hat sich im Rahmen dieser Arbeit als sehr wirkungsvoll erwiesen.

Darstellung der Daten im Koordinatensystem der Diskriminanzanalyse

Eine weitere, den diskutierten Histogrammen verwandte, Darstellung ist die Auftragung der Daten über den zur linearen Diskriminanzanalyse verwendeten Variablen. Obwohl

diese Diagramme zunächst anschaulicher als die Histogramme sind, ist ihre Interpretation wesentlich schwieriger, da meist sehr unübersichtliche Punktwolken entstehen Abb. 13.4(a) und die Daten sich gegenseitig verdecken können.

Ein großer Vorteil gegenüber den Histogrammen ist allerdings, dass die Identifikation nicht nur ganzer Proben, sondern auch einzelner Spektren bestimmter Proben erfolgen kann. Hilfreich ist auch die Möglichkeit, nur den Mittelwert jeder Probe eintragen zu lassen (Abb. 13.4(b)). Auch in dieser Darstellung sticht die bereits bekannte Messung als untypisch hervor. Die Erkennung untypischer Daten aus der Auftragung der Mittelwertspektren ist jedoch nicht sicher, da auch Messungen, deren Spektren über einen sehr breiten Wertebereich streuen, ein typisches Mittelwertspektrum aufweisen können.

Diese Methode gibt einen guten Überblick über die Trennung zwischen den Klassen, allerdings kann dies nur für Modelle mit zwei oder drei Variablen direkt eingeschätzt werden, da höherdimensionale Daten nicht bildlich vorstellbar sind. Eine genauere Einschätzung der Optimierungsergebnisse ist mit den beschriebenen Histogrammen möglich.

Hat die Untersuchung der gebildeten Variablen untypische Messungen aufgezeigt, so muss die Optimierungsrechnung unter Ausschluss dieser Messungen wiederholt werden. Danach sind die neu gebildeten Variablen ebenfalls zu prüfen. Für den Fall, dass relativ viele untypische Messungen vorliegen, ist zu erwarten, dass nur einige dieser kritischen Daten zu erkennen sind, da die Optimierungsrechnung ja darauf zielt, möglichst enge Verteilungen der einzelnen Klassen zu finden. Daher können sich die gebildeten Modelle stark unterscheiden.

Histologische Begutachtung

Unerlässlich für die Auswahl der Trainingsdaten ist eine detaillierte histologische Begutachtung der Proben, da diese Klassifizierung die Grundlage für das zu bildende prädiktive Modell ist.

Die Untersuchung der an die spektroskopierten Schnitte angrenzenden gefärbten Schnitte ermöglicht, diesen Bereich der Probe zu charakterisieren. Das ist erforderlich, weil Tumore oft sehr inhomogen sind und der zur Diagnose begutachtete Schnitt eventuell andere Merkmale aufweisen kann. Weiter besteht die Möglichkeit, die Probe dahingehend zu beurteilen, ob sie typisch für die Tumorart ist oder Besonderheiten aufweist, die in der Diagnose nicht angegeben werden. Nicht zuletzt handelt es sich um eine wichtige Kontrolle, um Verwechslungen der Proben vermeiden beziehungsweise erkennen zu können. Auch bei der Präparation der Schnitte entstandene Veränderungen der Gewebe, zum Beispiel ausgetretene Gewebsflüssigkeit, sind durch die *Hämatoxylin-Eosin-Färbung* erkennbar.

Unter Umständen treten präparationsbedingte Veränderungen nur bei einem Schnitt auf, daher kann auch die mikroskopische Untersuchung der ungefärbten Schnitte, die für die IR-Messung benutzt werden, hilfreich sein. Allerdings ist dort aufgrund des geringen Kontrastes keine histologische Untersuchung möglich.

13.2. Die Spektrenanzahl in Trainings- und Testdatensatz

Auch die Festlegung der Spektrenzahlen stellt einen Kompromiss zwischen der möglichst umfassenden Nutzung der vorhandenen Spektren und dem daraus resultierenden größeren Ressourcenbedarf dar. Bei der Festlegung der konkreten Spektrenzahlen des Trainingsdatensatzes sind jedoch einige weitere Randbedingungen einzuhalten.

13.2.1. Trainingsdaten

Bei der Auswertung mit `ga_ors` und `stackedGen` gehen alle Spektren mit gleichem Gewicht in die Rechnung ein, die a priori Wahrscheinlichkeit der Klassenzugehörigkeit wird aus den Daten geschätzt. Das ist für die hier genutzten Daten problematisch, da die a priori Wahrscheinlichkeiten der Klassenzugehörigkeit nicht bekannt sind. Die Probennahme erfolgt nicht repräsentativ im Sinne einer Gesamtstichprobe, die eine solche Schätzung zulässt.

Der überwiegende Teil der vorhandenen Spektren entstammt Glioblastom-Proben. Werden also alle zur Verfügung stehenden Spektren, so werden unbekannte Proben fast immer als Glioblastom eingestuft.

Diese Problematik ist im Kontext der linearen Diskriminanzanalyse bekannt. Es wird empfohlen, ungefähr gleiche Objektanzahlen für alle Klassen zu verwenden [78].

In dieser Arbeit wurden drei Strategien angewandt, um die Randbedingung gleicher Spektrenzahlen in allen Klassen einhalten zu können. Standen für eine Messung mehr Spektren zur Verfügung, als in die Rechnung einfließen sollten, so wurde über örtlich benachbarte Spektren gemittelt und eine zufällige Auswahl an Spektren getroffen, um die benötigte Anzahl an Spektren zu erhalten. Standen weniger Spektren als gefordert zur Verfügung, so wurden Spektren vervielfacht.

Grundsätzlich ist jedoch die Möglichkeit der Vorgabe gleicher a priori Wahrscheinlichkeiten oder besser des Gewichts, mit dem die einzelnen Spektren in die Rechnung einfließen, vorzuziehen. Die Angabe der Gewichtung bietet gegenüber der Festlegung der a priori Wahrscheinlichkeiten die zusätzliche Möglichkeit, innerhalb der einzelnen Klassen die Daten der unterschiedlichen Proben gleich zu gewichten. Bei vorgegebenen a priori Wahrscheinlichkeiten erhalten dagegen innerhalb der Klassen wieder die einzelnen Spektren gleiches Gewicht.

Das Bestreben, Daten möglichst vieler Messungen in die Rechnung einfließen zu lassen, führte zu großen Spektrenanzahlen. Das in Kap. 15 (S. 67) näher diskutierte Modell wurde mit ungefähr 10600 Spektren gerechnet, so dass jede Messung mit mindestens 12 Spektren in die Rechnung einging.

Zufällige Auswahl von Spektren

Die Funktion `sel_rnd` wählt zufällig eine vorgegebene Anzahl an Spektren aus allen Spektren einer Messung aus.

Dies entspricht dem Ziehen einer Stichprobe vorgegebenen Umfangs aus der Menge der vorhandenen Spektren. Diese Auswahl wird nach der Filterung (Kap. 14.5.2, S. 64) getroffen.

Mittelwertbildung örtlich benachbarter Spektren

Eine Möglichkeit, mehr Spektren in der Auswertung zu berücksichtigen, die dann mit entsprechend geringerem Gewicht in die Auswertung eingehen, besteht darin, über mehrere Spektren zu mitteln. Damit die Information über die örtliche Herkunft der Spektren erhalten bleibt — wenn auch mit verringerter Auflösung —, wird die Mittelwertbildung über örtlich benachbarte Spektren durchgeführt.

Die Funktion `dolateral_mean` bildet das Mittelwertspektrum über alle *vorhandenen* Spektren im angegebenen Raster. Ist zum Beispiel nur ein Spektrum in der betracheten Zelle, so wird es direkt übernommen. Dadurch können innerhalb einer Messung Spektren entstehen, die unterschiedlichen Ortsauflösungen entsprechen.

Die Anwendung dieser Methode zur Reduktion der zur Verfügung stehenden Spektrenzahl bewirkte in einer Testrechnung den Anstieg der Trefferrate in der LDA-Validierung mit sechs Sets von 74,3 % auf 79,6 % und ist daher auch für zukünftige Rechnungen zu empfehlen.

Vervielfachen von Spektren

Besteht eine Messung aus weniger als der geforderten Spektrenzahl, so können die Daten mit der Funktion `duplicate_spectra` vervielfältigt werden, um die gesuchte Anzahl an Spektren zu erhalten.

13.2.2. Testdaten

Die Testdaten fließen nicht in die Modellbildung durch `ga_ors` und `stackedGen` ein. Sie werden durch `stackedGen` mit Hilfe des gebildeten Modells zugeordnet. Die Diskriminanzanalyse benötigt gegenüber der Optimierung nur einen Bruchteil an Rechenzeit, daher können problemlos alle geeigneten Spektren einer Messung als Testdaten verwendet werden.

Bei der Validierung der gebildeten Modelle kann es allerdings sinnvoll sein, entweder die Spektrenanzahlen nach den für Trainingsdaten beschriebenen Kriterien festzusetzen oder für die Auswertung der Ergebnisse die einzelnen Spektren unterschiedlich stark zu gewichten. Im Gegensatz zu den Trainingsdaten kann die Gewichtung der einzelnen Spektren der Testdaten jedoch bei der Auswertung der Zuordnungen erfolgen. Bei den Set-Validierungen der Modelle wurden in dieser Arbeit immer die zur Modellbildung genutzten Spektren als Testdaten verwendet.

14. Datenvorbehandlung

Die Vorbehandlung der Daten ist ein wichtiger Schritt der Auswertung, da in der Regel die Rohdaten den Anforderungen der Analysenmethoden nicht genügen. Der für Menschen erkennbare und der für die Auswertealgorithmen nutzbare Informationsgehalt der Daten soll durch diese Behandlung erhöht werden. Allerdings sind viele dieser Methoden keine mathematisch äquivalenten Abbildungen der Daten, sondern stellen Datenmanipulationen dar. Das ist insbesondere bei Methoden der Fall, bei denen subjektiv wählbare Parameter genutzt werden.

Die hier vorgestellten Untersuchungen zum Einfluss von Basislinienkorrektur, Intensitätsnormierung und Filterung sowie das bereits in Kap. 12.4.1 (S. 51) diskutierte Vorgehen bei der Bildung einer einheitlichen Wellenzahl-Achse wurden gemeinsam in Form eines vollständigen Faktorplans für zwei Datensätze gerechnet. Datensatz „l“ stellt eine Untermenge von Datensatz „a“ dar. Die HE-Schnitte der Proben in „l“ waren besonders typisch für die jeweilige Tumorart. Die günstige Wahl des Spektralbereichs und der Auflösung wurden in separaten Rechenserien untersucht.

Die Modellbildung wurde mit maximal acht Regionen durchgeführt. Dieser Wert stellt einen Kompromiss zwischen Modellgenauigkeit und benötigter Rechenzeit dar. In Vorversuchen wurde ein lokales Maximum der Trefferrate bei acht Variablen gefunden. Diese Rechnungen wurden für zwei unterschiedliche Datensätze durchgeführt, eine genauere Beschreibung der Datensätze sowie die Tabellen mit den erzielten Reklassifikations-Trefferraten ist in Anhang C (97) gegeben.

Alle weiteren Untersuchungen sowie die Optimierung der Filterparameter fand in separaten Rechenserien statt.

14.1. Die Auswahl des Spektralbereichs

Hier wurden fast ausschließlich Spektren im Fingerprint-Bereich zwischen 1000 und 1800 cm^{-1} untersucht. Die Anwendung der Programme auf den gesamten zur Verfügung stehenden Spektralbereich zwischen 950 und 3800 cm^{-1} ergab, gemessen an der Reklassifikations-Trefferrate, geringfügig schlechtere Modelle. Das traf auch dann zu, wenn die Daten im Fingerprint-Bereich identisch waren. Außerdem führt die Verwendung des gesamten Spektrums zu einem entsprechend der größeren Anzahl an Datenpunkten erhöhten Rechenaufwand.

Daher ist die Beschränkung der Daten auf den Fingerprint-Bereich zu empfehlen. Dies ist auch aus chemisch-spektroskopischer Sicht insofern zu vertreten, als man davon ausgehen kann, dass sich die hier zu suchenden feinen Veränderungen in den Anteilen der unterschiedlichen Substanzklassen sowie den Substanzen selbst insbesondere im Fingerprint-Bereich bemerkbar machen (vgl. [91, S. 35][92, S. 300]).

14.2. Die spektrale Auflösung

Die Wahl einer guten spektralen Auflösung der Spektren stellt das Finden des Optimums zwischen möglichst guter Genauigkeit einerseits und anderserseits möglichst geringem Aufwand an Ressourcen wie Rechenzeit und Datenvolumen dar.

Generell gilt, dass die Auflösung umso geringer sein kann, je breiter die gefundenen spektralen Regionen sind. Der Aufwand sowohl an Speicherplatz als auch an Rechenzeit wächst hier linear mit der Anzahl der Datenpunkte. Daher sollte die Auflösung nicht höher als erforderlich gewählt werden.

In diesem Zusammenhang muss zwischen zwei unterschiedlichen Bedeutungen der „spektralen Auflösung“ unterschieden werden. Zum einen kann die physikalisch erreichte Auflösung der Rohdaten gemeint sein, zum anderen die Anzahl der Datenpunkte der Spektren wie sie dem Programm `ga_ors` übergeben werden.

An dieser Stelle wird die letztere Bedeutung untersucht. Dabei wurden aus den vorhandenen Daten mit einem Abstand der Datenpunkte von etwa 4 cm^{-1} durch Mitteln über eine vorgegebene Anzahl von Datenpunkten neue Daten mit weniger Messpunkten errechnet. Diese Transformation der Daten beeinflusst bestimmte Eigenschaften, wie die Konvergenz in Richtung normalverteilter Daten der gebildeten Regionen nicht, da letztlich weiterhin alle Datenpunkte in der Region in die Variablenbildung einfließen. Dagegen wird die Genauigkeit, mit der die Grenzen der Regionen ermittelt werden können, mit sinkender Auflösung verringert.

Abb. 14.1 zeigt die erreichten Reklassifikations-Trefferraten für unterschiedliche Abstände der Datenpunkte auf der Wellenzahl-Achse. Bereits eine Mittelwertbildung über zwei Datenpunkte führte zu einem deutlichen Anstieg der Fehlzuordnungen.

Die Rechenserie wurde mit acht Regionen als Vorgabe durchgeführt.

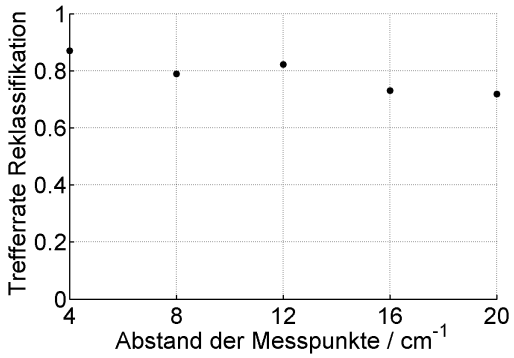


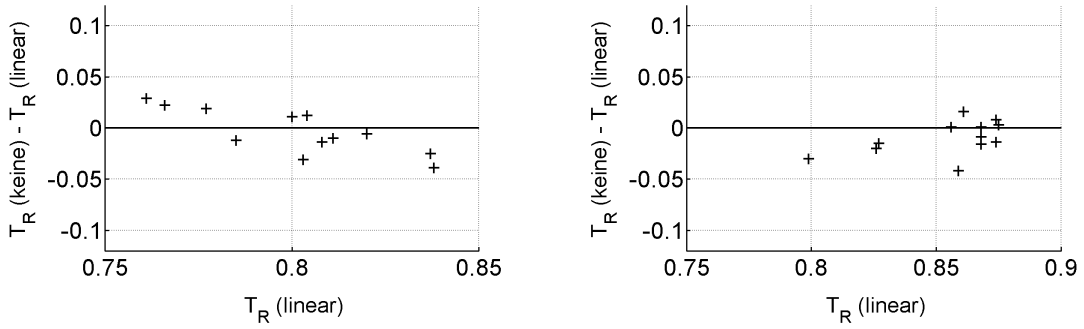
Abbildung 14.1.: Abhängigkeit der Reklassifikations-Trefferrate von der spektralen Auflösung

Für die Zukunft ist mit besseren Modellen mit einer größeren Anzahl an Variablen zu rechnen, da mehr Spektren gemessen und in die Trainingsdatensätze aufgenommen werden. Steigende Regionenzahlen führen aber dazu, dass die einzelnen Regionen schmaler werden. Die begrenzte Auflösung der Spektren wird sich dann also noch stärker bemerkbar machen.

Von einer derartigen Datenreduktion muss daher abgeraten werden.

14.3. Basislinienkorrektur

Die Basislinienkorrektur erfolgte linear zwischen erstem (1000 cm^{-1}) und letztem Datenpunkt (1800 cm^{-1}) der Spektren, eine Suche des genauen Minimums erwies sich nicht als erforderlich. Die erzielten Trefferraten zeigten nur geringe Abhängigkeit von der Basislinienkorrektur (Abb. 14.2), eventuell zugunsten der Basislinien-korrigierten Daten. Jedoch erlaubt die Genauigkeit der Schätzung der Modellgüte mittels der Reklassifikations-Trefferrate keine eindeutige Aussage.

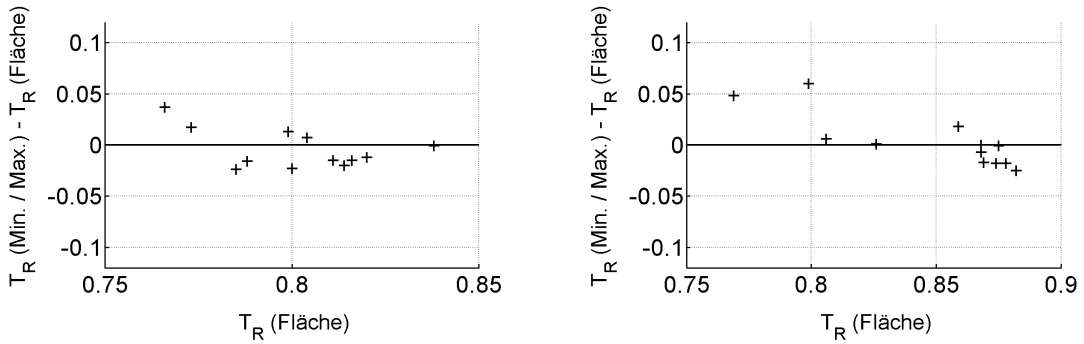


(a) Datensatz „a“ **(b)** Datensatz „I“

Abbildung 14.2.: Wirkung der Basislinienkorrektur — Eventuell liegt ein leichter Trend zugunsten der Basislinienkorrektur vor.

Für die weiteren Rechnungen im Rahmen dieser Arbeit wurde die Basislinienkorrektur angewandt.

14.4. Intensitätsnormierung



(a) Datensatz „a“ **(b)** Datensatz „I“

Abbildung 14.3.: Wirkung unterschiedlicher Intensitätsnormierungen — Es sind keine Vorteile einer Methode erkennbar.

Untersucht wurden die Unterschiede zwischen einer Flächennormierung und einer Minimum-Maximum-Normierung. Vor der Normierung wurde jeweils eine Offset-Korrektur durchgeführt, die minimale Extinktion betrug damit immer 0, aus den sichtbar großen Intensitätsunterschieden der einzelnen Spektren wurde auf die Notwendigkeit einer Intensitätsnormierung geschlossen, diese jedoch nicht gesondert überprüft.

Für die Minimum-Maximum-Normierung wurde das Maximum der Amid-I-Bande zwischen 1600 und 1700 cm^{-1} gesucht und auf 1 skaliert, bei der Flächennormierung wurde durch die aufsummierte Intensität über das Spektrum dividiert.

In Abb. 14.3 sind die erreichten Reklassifikations-Trefferraten für die beiden untersuchten Normierungsmethoden dargestellt, ein Einfluss auf die erreichbaren Trefferraten kann nicht gezeigt werden.

Hier wurde im Weiteren die Minimum-Maximum-Normierung verwendet, da die so behandelten Daten weitestgehend unabhängig vom verwendeten Spektralbereich sind.

14.5. Filterung

Die Eignung der Trainingsdatensätze sollte letztlich immer an der damit erreichten Modellgüte beurteilt werden. Dennoch ist es sinnvoll, verschiedene Kriterien aufzustellen, die bereits *vor* dem Berechnen der Modellparameter eingehalten werden müssen, da so viel Rechenaufwand vermieden werden kann. Dies betrifft vor allem die Erkennung ungeeigneter Spektren.

14.5.1. Kriterien der Spektrenqualität

Damit die Spektren auswertbar sind, ist ein hinreichend hohes Signal-Rausch-Verhältnis nötig, das nur erhalten werden kann, wenn die maximale Extinktion eine Mindestgröße erreicht. Im Rahmen dieser Arbeit wurde die Untergrenze bei 0,1 Extinktionseinheit gewählt.

Für zu hohe Extinktionswerte sind Abweichungen vom Lambert-Beerschen Gesetz zu erwarten, die zu Problemen in der Auswertung führen können. Empfohlen wird eine Obergrenze der Extinktion von 1 [37], die hier auch angewandt wurde.

Auch für die Minima der Spektren wurden eine Unter- und eine Obergrenze festgelegt. Die Spektren sollen positiv sein, negative und zu große Extinktionswerte weisen auf Fehler in der Hintergrund-Korrektur hin. Daher wurden nur Spektren in die Rechnung aufgenommen, deren minimale Extinktion zwischen 0 und 0,1 lag.

Diese Filterkriterien wurden immer angewandt.

14.5.2. Kriterien der Homogenität der Spektren innerhalb einer Messung

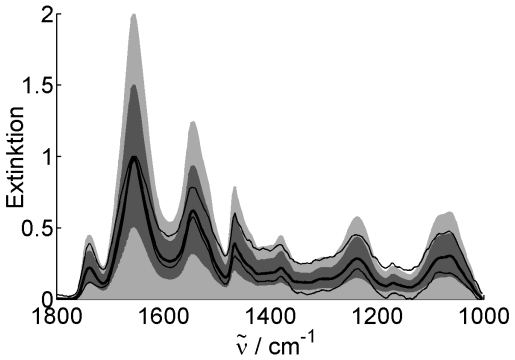
Während bei Testdaten grundsätzlich sehr unterschiedliche, eventuell sogar unterschiedlichen Geweben angehörende, Spektren innerhalb einer Messung auftreten können, müssen die Spektren innerhalb des Trainingsdatensatzes sicher der jeweiligen Klasse angehören. Daher wurden Filterkriterien auf der Basis aller Spektren einer Messung definiert.

Abweichung vom Mittelwertspektrum

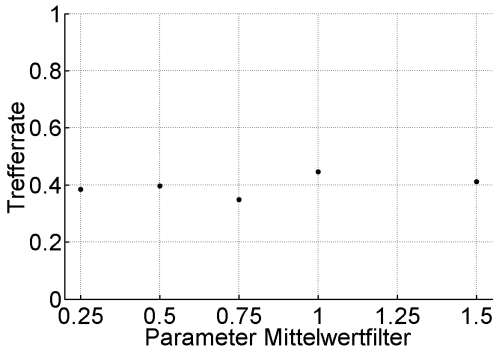
Eine einfache Möglichkeit, Spektren mit stark abweichender Form zu erkennen, besteht darin, nur Spektren in die Modellbildung einzubeziehen, die sich nicht zu stark vom Mittelwertspektrum unterscheiden.

Die erlaubte Abweichung in Vielfachen des Mittelwerts wurde als Parameter angegeben, Abb. 14.4(a) skizziert die Extinktionsbereiche, die dem Filterkriterium entsprechen. Das Mittelwertspektrum sowie die minimalen und maximalen Extinktionswerte der Messung sind eingezeichnet. Abb. 14.4(b) zeigt die erreichten Reklassifikations-Trefferraten in Abhängigkeit dieses Parameters. Als günstige Wahl erwies sich eine Abweichung bis zu $\pm 100\%$ vom Mittelwert, dann erfolgt natürlich nur noch eine Filterung zu hoher Extinktionswerte.

Dieser Filter erwies sich in Kombination mit der Minimum-Maximum-Normierung als sehr scharf. Die Wahl des Filterparameters unterhalb von 1,0 bewirkte die Entfernung fast aller Spektren. Daher ist bei der Verwendung dieses Filters in Kombination mit der Minimum-Maximum-Normierung Vorsicht geboten. Die Optimierung dieses Parameters



(a) Funktions-Skizze — Extinktionsbereiche, die den Filter passieren: Abweichung vom Mittelwertspektrum ± 100 % (hellgrau) bzw. 50 % (dunkelgrau). Mittelwertspektrum (fett) und Spannweite der Spektren vor der Filterung sind eingetragen.



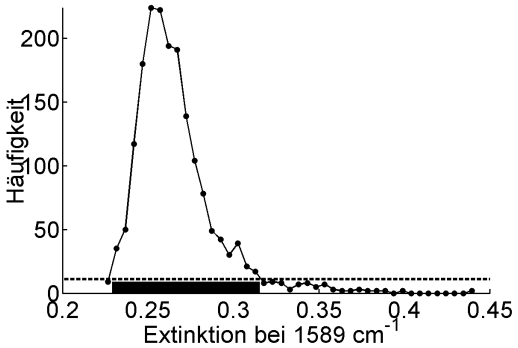
(b) Optimierung des Filterparameters — Trefferrate für einen unabhängigen Datensatz

Abbildung 14.4.: Der Mittelwert-Filter

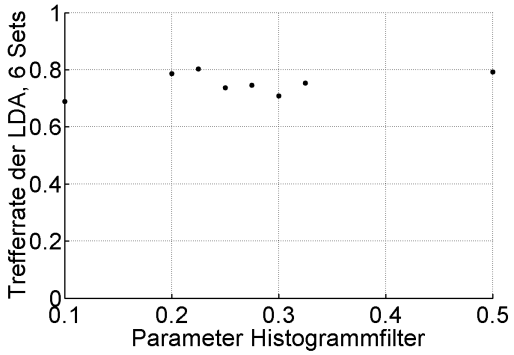
erfolgte, indem der Trainingsdatensatz „l“ gewählt wurde und diejenigen Proben des Datensatzes „a“, die nicht in „l“ enthalten waren, als Testset zugeordnet wurden.

Dieser Filter wurde in der Funktion `sel_filter_01` implementiert.

Histogramm-Filter



(a) Funktions-Skizze — Verteilung der Extinktionswerte und Schwellenwert der Mindestbesetzung (gestrichelt, Filterparameter = 0,25). Der schwarze Balken kennzeichnet den Extinktionsbereich, der den Filter passiert.



(b) Optimierung des Filterparameters

Abbildung 14.5.: Der Histogramm-Filter

Ein weiteres Filter-Kriterim wurde entwickelt, um diese Einseitigkeit zu vermeiden und die oftmals recht *schiefe* Verteilung der Daten zu berücksichtigen. Diese Abweichung von der Normalverteilung kann durch die Datenvorbehandlung zustande kommen.

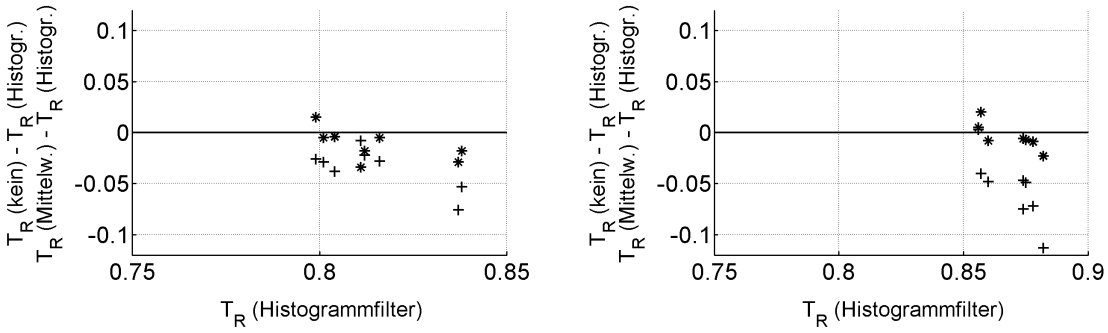
Eine mögliche Ursache ist die Intensitätsnormierung. Diese geschieht unter der Annahme, das Lambert-Beersche Gesetz sei erfüllt und die Extinktion also proportional zu

Schichtdicke und Konzentration der Stoffe. Ist dies nicht der Fall, so resultiert eine schiefe Verteilung der Extinktionswerte.

Grundlage der Filterung ist die Verteilung der Extinktionswerte an der betrachteten Wellenzahl. Der Parameter der Filterfunktion gibt einen Schwellenwert für die Häufigkeit an. Alle Spektren mit Extinktionswerten unterhalb des ersten Erreichens dieser Mindestbesetzung einer Histogramm-Klasse und alle Spektren oberhalb der letzten Histogramm-Klasse mit der vorgegebenen Mindestanzahl an Spektren werden ausgeschlossen. Am Beispiel einer Wellenzahl zeigt Abb. 14.5(a) die Verteilung der verschiedenen Extinktionswerte einer Messung und den Schwellenwert für die Mindestbesetzung der einzelnen Klassen. Dieser Schwellenwert ist das Produkt aus Klassenanzahl und Filterparameter, so dass die durch die stark unterschiedliche Spektrenzahl verschiedener Messungen notwendige Anpassung der Klassenbreite kompensiert wird.

Ein Filterparameter von 0,225 ist eine gute Wahl (Abb. 14.5(b)). Die entsprechende Filter-Funktion ist `sel_filter_02`, der Rechenaufwand ist höher als beim Mittelwertfilter.

Die beiden Filterkriterien müssen dabei für alle Wellenzahlen erfüllt sein, damit ein Spektrum in die Rechnung aufgenommen wird. Daher kann eine eine sinnvolle Filterung nur stattfinden, wenn die Parameter sehr niedrig gewählt werden. Eine Alternative ist die Definition einer Mindestanzahl an Wellenzahlen, an denen das jeweilige Filterkriterium verletzt sein muss, um ein Spektrum auszuschließen, in Verbindung mit verschärften Filterkriterien.



(a) Datensatz „a“ **(b)** Datensatz „l“

Abbildung 14.6.: Wirkung unterschiedlicher Filterkriterien — Differenz zu den aus ungefilterten Daten (+) und den nach Anwendung des Mittelwertfilters erhaltenen Modellen (*). Bei der Anwendung des Histogrammfilters werden durchweg hohe Trefferraten erzielt, dieses Kriterium ist den beiden Alternativen fast immer überlegen.

Die Anwendung der Filter ermöglichte eine deutliche Steigerung der Trefferraten (Abb. 14.6). Zwischen den beiden Filterkriterien war kein so großer Unterschied zu verzeichnen, der Histogramm-Filter war jedoch in den meisten Fällen überlegen.

Daher wurde bei den weiteren Berechnungen der Histogramm-Filter genutzt.

15. Untersuchung des gebildeten Trainingsdatensatzes

15.1. Parameter der Modellerstellung

Tabelle 15.1.: Datenstruktur des Trainingsdatensatzes

Diagnose	Anzahl Proben	Anzahl Messungen	Anzahl Spektren	Spektrenzahl nach Filterung
gesund	6	14	2641	12314
Astro II	3	4	2640	1183
Astro III	6	13	2638	13747
Glio	44	59	2640	48455
gesamt	59	90	10559	75699

Tab. 15.1 gibt eine Übersicht über die Zusammensetzung der zur Modellbildung genutzten Spektren, bei den Daten handelt es sich ausschließlich um Maps.

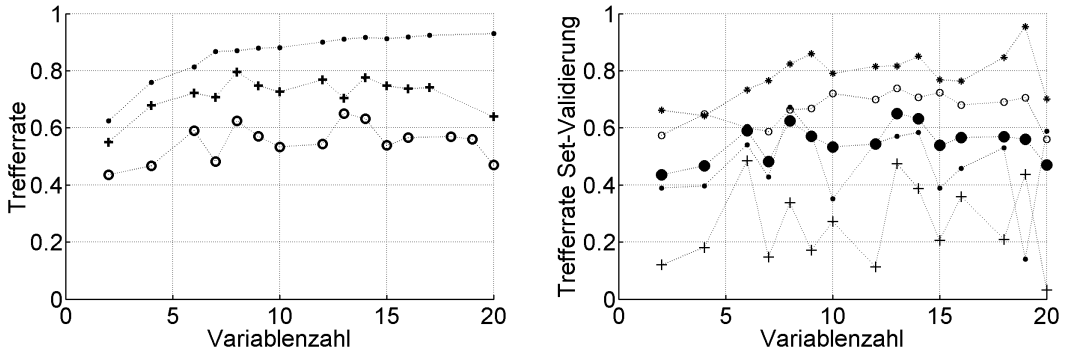
Die $\tilde{\nu}$-Achse der Daten wurde nach Datenreduktion durch Mitteln gebildet, die Spektren überspannen einen Wellenzahlbereich von 1800 – 1000 cm^{-1} und die 208 Datenpunkte liegen jeweils ungefähr 4 cm^{-1} auseinander. Spektren mit minimaler Extinktion unter 0 oder über 0,1 und Spektren mit maximaler Extinktion unter 0,1 oder über 1 wurden verworfen. Lineare Basislinienkorrektur und Minimum-Maximum-Normierung wurden angewandt und danach diejenigen Spektren aus der Rechnung ausgeschlossen, die nicht den Histogrammfilter mit Filterparameter 0,25 passierten.

Eine Spektrenanzahl von mindestens 12 Spektren jeder Messung wurde vorgegeben, diese Spektrenzahlen wurden durch Anwendung der örtlichen Mittelwertbildung und einer zufälligen Spektrenwahl beziehungsweise Vervielfachung der Spektren eingehalten. Insgesamt gingen etwa 75000 Spektren mit unterschiedlichem Gewicht in die Berechnung des Modells ein.

15.2. Die erreichte Modellgüte

Die erreichten Trefferraten der Validierung ausschließlich des LDA-Modells zeigen ein Maximum bei 8 sowie zwei weitere Maxima bei 12 und 14 Variablen. Die Validierung der Optimierung und LDA gleichzeitig weist ein Maximum bei 8 und ein weiteres bei 13 Variablen auf. Da letztere Schätzung der Trefferrate für das Maximum bei 13 Variablen mit 65,0 % Trefferrate keine wesentliche Verbesserung gegenüber 62,4 % Trefferrate mit acht Variablen aufzeigt, wird im Folgenden das Modell aus acht Variablen untersucht.

Auffällig sind die niedrigen Trefferraten der Modelle mit 7 und 15 Variablen, auch der Verlauf der Trefferraten für die Astrozytome zweiten und dritten Grades in (b) zeigt für viele ungerade Variablenzahlen unerwartet schlechte Trefferraten.



(a) Modellgüte in Abhängigkeit der Variablenzahl — $\hat{T}_{Opt}$ (6 Sets, o), $\hat{T}_{LDA}$ (6 Sets, +), $\hat{T}_R$ (.)

(b) Modellgüte in Abhängigkeit der Variablenzahl für die unterschiedlichen Klassen — gesund (*), Astro II (+), Astro III (·), Glio (o), gesamt (•)

Abbildung 15.1.: Modellgüte des Trainingsdatensatzes

Die Astrozytome zweiten Grades erreichen insgesamt nur sehr niedrige Trefferraten, oft liegen sie unter der für zufällige Zuordnungen erwarteten Trefferrate von 25 %. Das ist vermutlich darauf zurückzuführen, dass das Modell insgesamt nur drei Proben dieser Tumore beinhaltet. Die knapp 1200 nach den Filterungen zur Verfügung stehenden Spektren sind weniger als 2 % der in das Modell einfließenden unterschiedlichen Spektren. Daher ist eine deutliche Verbesserung der Trefferrate zu erwarten, wenn mehr Daten von Astrozytomen zweiten Grades in das Modell einfließen.

15.3. Das Modell mit acht Variablen

15.3.1. Voraussetzungen der LDA

Das Programmpaket `SPSS` [1] bietet die Möglichkeit, eine deskriptive lineare Diskriminanzanalyse durchzuführen und einige Kenngrößen dieser Analyse zu berechnen. Die Voraussetzungen der LDA, das Vorliegen multivariat normalverteilter Daten mit homogenen Kovarianzmatrizen zwischen den Klassen und die Trennbarkeit der Gruppen wurden mit Hilfe der in `SPSS` implementierten Funktionen zur linearen Diskriminanzanalyse überprüft.

Die Trennung der Klassen

Auch die Werte der Diskriminanzfunktionen für die einzelnen Spektren können in `SPSS` berechnet werden. Diese Werte sind in der Ergebnisdatei von `stackedGen` nicht tabelliert. Da hier zwischen vier Gruppen unterschieden wird, werden drei Diskriminanzfunktionen gebildet. Abb. 15.2 zeigt die Spektren über den Diskriminanzfunktionen aufgetragen. Die einzelnen Gruppen sind gut zu erkennen, liegen aber sehr dicht beieinander. Einige Spektren der Glioblastome reichen bis in das Gebiet der gesunden Spektren hinein.

[1] SPSS Version 11.0 für Windows, ©1989 – 2001, SPSS Inc.

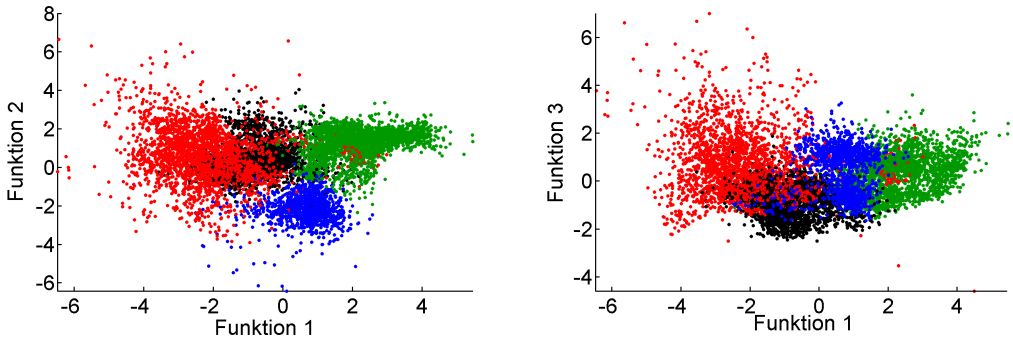


Abbildung 15.2.: Auftragung der Daten über den Werten der Diskriminanzfunktion — gesund (grün), Astro II (blau), Astro III (schwarz) und Gliob (rot). Auffällig ist der Zusammenhang der ersten Diskriminanzfunktion mit dem Grad der Malignität

Test auf Gleichheit der Mittelwerte

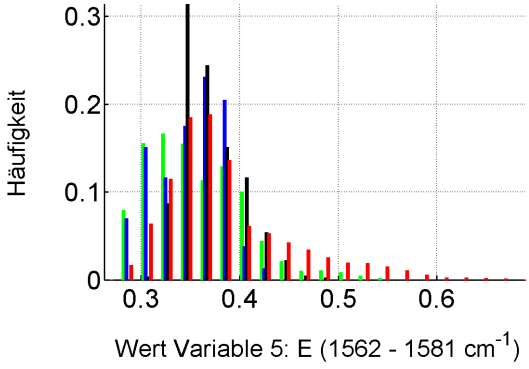
Grundlage für diesen Test ist das Wilkssche Λ, das eine Aussage über die Trennbarkeit der Gruppen macht. Für perfekt separierbare Gruppen geht Λ gegen 0, für untrennbare Gruppen gegen 1 [71].

Die ermittelten Werte für die einzelnen Variablen sind in Tab. 15.2 aufgeführt. Die Nullhypothese gleicher Mittelwerte der Gruppen wird für alle Variablen hochsignifikant abgelehnt.

Tabelle 15.2.: Test auf gleiche Gruppenmittelwerte

Nr.	Region	Λ
1	1014 – 1034 cm^{-1}	0,683
2	1072 – 1080 cm^{-1}	0,424
3	1088 – 1107 cm^{-1}	0,498
4	1385 – 1500 cm^{-1}	0,940
5	1562 – 1581 cm^{-1}	0,923
6	1593 – 1682 cm^{-1}	0,942
7	1686 – 1705 cm^{-1}	0,898
8	1740 – 1759 cm^{-1}	0,509

Abbildung 15.3.: Verteilung der Variablen 5: $E(1562 - 1581 cm^{-1})$ — gesund (grün), Astro II (blau), Astro III (schwarz), Glio (rot)



Test auf Homogenität der Kovarianzmatrizen

Ebenfalls in SPSS wurde ein Box-Test auf Gleichheit der Kovarianzmatrizen der Gruppen durchgeführt. Auch diese Nullhypothese wurde hochsignifikant abgelehnt, mögliche Gründe sind Heteroskedastizität oder das Vorliegen nicht multivariat normalverteilter Daten.

Dieses Ergebnis ist auch an der in Abb. 15.3 gezeigten Verteilung der Variablen 5 abzulesen, man erkennt deutlich, dass die Daten weder normalverteilt sind, noch homogene

Kovarianzmatrizen haben können. Bei der Interpretation des Histogramms ist immer zu beachten, dass nur in dieser Richtung geschlossen werden kann. Univariate Normalverteilungen der einzelnen Variablen sind zwar notwendige, aber keine hinreichenden Bedingungen für das Vorliegen einer multivariaten Normalverteilung.

15.3.2. Die Variablen des Modells

Rangfolge der Variablen für die LDA

Um die Wichtigkeit der Variablen zu ermitteln, wurde jede Variable einmal aus dem LDA-Modell ausgeschlossen und mit einer Leave-One-Out-Validierung die erreichten Trefferraten bestimmt. Die Ergebnisse sind in Tab. 15.3 zusammengefasst, wobei vier Gruppen mit Variablen jeweils ähnlicher Wichtigkeit gebildet wurden. Die Variable, deren Ausschluss den größten Abfall der Trefferrate und die größte Steigerung der Fehlerrate bewirkt, wird als wichtigste Variable für die LDA betrachtet [78, Kap. VIII-3].

Die Steigerung $\frac{\Delta F}{F}$ der Fehlerrate F wird nach

$$\frac{\Delta F}{F} = \frac{\hat{\mathrm{T}}_{\mathrm{LDA}}\,(\text{Modell 8 Variablen}) - \hat{\mathrm{T}}_{\mathrm{LDA}}\,(\text{Modell 7 Variablen})}{1 - \hat{\mathrm{T}}_{\mathrm{LDA}}\,(\text{Modell 8 Variablen})} \tag{15.1}$$

aus den Trefferraten der verglichenen Modelle berechnet. $\frac{\Delta F}{F}$ gibt an, welche Steigerung der Fehlerrate bei Verwendung des Modells mit sieben Variablen gegenüber dem Modell mit acht Variablen zu erwarten sind.

Tabelle 15.3.: Rangfolge der Variablen — Ausschluss einzelner Variablen

Nr.	**ausgeschlossene Variable**	T	$\frac{\Delta F}{F}$
6	1593 – 1682 cm^{-1}	57,9 %	106 %
5	1562 – 1581 cm^{-1}	58,8 %	102 %
7	1686 – 1705 cm^{-1}	63,9 %	77 %
1	1014 – 1034 cm^{-1}	66,6 %	64 %
3	1088 – 1107 cm^{-1}	67,5 %	59 %
2	1072 – 1080 cm^{-1}	68,8 %	53 %
8	1740 – 1759 cm^{-1}	72,4 %	35 %
4	1385 – 1500 cm^{-1}	72,9 %	33 %

Tabelle 15.4.: Schrittweise Analyse durch `SPSS` — wichtigste Variablen zuerst

Nr.	**Variable**	**Rang**
2	1072 – 1080 cm^{-1}	1
1	1014 – 1034 cm^{-1}	2
3	1088 – 1107 cm^{-1}	3
7	1686 – 1705 cm^{-1}	4
5	1562 – 1581 cm^{-1}	5
6	1593 – 1682 cm^{-1}	6
8	1740 – 1759 cm^{-1}	7
4	1385 – 1500 cm^{-1}	8

Auch die Untersuchung des gefundenen Modells mit `SPSS` ermöglicht eine Aussage über die Wichtigkeit und Rangfolge der einzelnen Variablen. `SPSS` nutzt deskriptive Statistiken, um diese Rangfolge zu bestimmen, Tab. 15.4 listet die Ergebnisse auf. Die durch `SPSS` angegebene Rangfolge weicht deutlich von der durch Auslassen einzelner Variablen bestimmten Rangfolge ab.

Die Variablen in ihrer spektroskopischen Bedeutung

Abb. 15.4 zeigt die Mittelwertspektren des Datensatzes zusammen mit den gefundenen Regionen. Die hier angewendeten Bandenzuordnungen sind in Tab. 15.5 zusammengefasst.

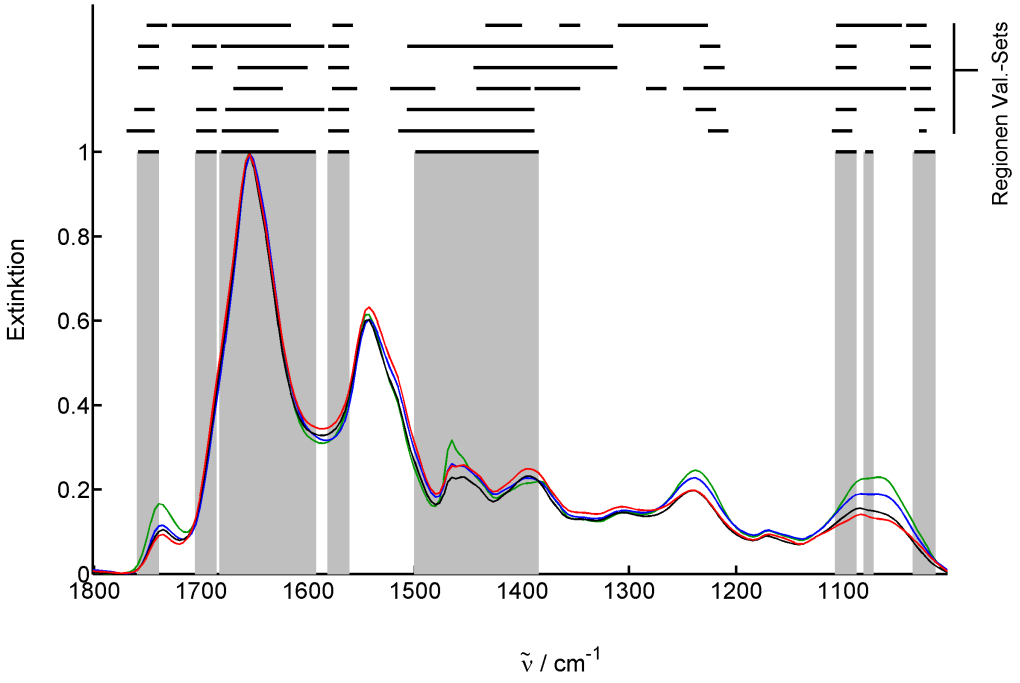


Abbildung 15.4.: Mittelwertspektren der Klassen und ermittelte Regionen — gesund (grün), Astro II (blau), Astro III (schwarz) und Glio (rot), im oberen Teil der Abbildungen sind zusätzlich die für die einzelnen Datensätze der Validierung ermittelten Regionen eingetragen.

Die Region von 1759 – 1740 cm^{-1} liegt an der Flanke der C = O-Streckschwingung der Estergruppe von Lipiden und Cholesterol-Estern. Diese Region wurde mit ähnlichen Grenzen auch bei den Rechnungen zur Validierung des Modells gefunden. Die Mittelwertspektren zeigen mit zunehmendem Tumorgrad geringere Extinktionswerte, das Mittelwertspektrum der gesunden Proben weist mit Abstand die größte Extinktion auf. Dennoch trägt diese Region nur in geringem Maße zur Klassifikation bei, wie beide erfolgten Analysen der Rangfolge ergaben.

Auch die Region zwischen 1705 und 1686 cm^{-1} wird von beiden Verfahren mit recht guter Übereinstimmung als wichtig eingeschätzt. Es handelt sich um den Beginn der Amid-I-Bande. Die zweite gefundene Region im Bereich der Amid-I-Bande reicht von 1682 bis 1593 cm^{-1}. Sie umfasst damit den oberen Teil der steigenden und die fallende Flanke der Bande. Diese Region wird durch `SPSS` auf einem der hinteren Ränge eingeordnet, dagegen ist sie nach dem Ausschlussverfahren als sehr wichtig einzuschätzen. Die Amid-I-Bande ist auch in den Validierungsrechnungen meist ähnlich aufgeteilt.

Zwischen 1500 und 1385 cm^{-1} überspannt eine weitere, sehr breite Region mehrere Banden. Dabei handelt es sich insbesondere um Deformationsschwingungen verschiedener CH_2- und CH_3-Gruppen in Lipiden und Proteinen und die symmetrische Streckschwingung von Carboxylat-Gruppen der Fettsäuren und in Aminosäure-Seitenketten. Da sich die Mittelwertspektren der verschiedenen Klassen mehrfach schneiden, erstaunt es nicht, dass diese Region nach der Rangfolge der Variablen den geringsten Beitrag zur Klassifikation liefert.

Tabelle 15.5.: Zuordnung der IR-Banden nach [93; 94]

Lage des Maximums / cm^{-1}	**Schwingung**	**Substanzklasse**
1733 – 1739	ν(C = O Ester)	Lipide, Cholesterol-Ester
1737	ν(C = O Ester)	
1655	Amid I	Proteine (α-Helix)
1636	Amid I	Proteine (β-Faltblatt)
1545	Amid II	Proteine (α-Helix)
1468 / 1467	$\delta_s(CH_2)$	Lipide, Proteine
1455	versch.$\delta(CH_2, CH_3)$	Proteine
1400 / 1397	$\nu_s(COO^-)$	Aminosäure-Seitenketten, Fettsäuren
1380 / 1379	$\delta_s(CH_3)$	Lipide, Proteine
1240 / 1237	$\nu_{as}(PO_2^-)$	Phospholipide, Nukleinsäuren
1083 / 1080	$\nu_s(PO_2^-)$	Phospholipide, Nukleinsäuren
1063 / 1060	$\nu_s(CO - O - C)$	Phospholipide, Cholesterol-Ester
1050	ν(CO)	Kohlenhydrate, Muzin
1030	ν(CO)	Glycogen

Die Region von 1107 – 1088 cm^{-1} liegt an der Flanke der symmetrischen Streckschwingung der PO_2^--Gruppen von Phospholipiden und Nukleinsäuren. Die Mittelwertspektren der einzelnen Klassen unterscheiden sich in dieser Region stark. Auch `SPSS` hat diese Variable als besonders wichtig zur Klassifikation eingestuft. Dagegen ergab die Ranganalyse durch Ausschluss eine niedrige Bedeutung der Variablen. Diese Region — zum Teil allerdings deutlich verbreitert —, auch in allen Validierungsrechnungen gewählt.

Der Spektralbereich zwischen 1080 und 1072 cm^{-1} wurde in den Validierungsrechnungen nie separat ausgewählt. Er beinhaltet die Absorption der Phospholipide und Cholesterol-Ester bei 1060 cm^{-1}. Die Region von 1034 – 1014 cm^{-1} liegt an der Flanke der Absorption durch die C − O-Streckschwingung von Kohlenhydraten. Die entsprechende Absorption des Glycogens liegt mit 1030 cm^{-1} in dieser Region, allerdings lässt der Verlauf der Spektren dort kein Maximum erkennen und das Vorkommen von Glycogen im Hirn ist nicht zu erwarten. Die letzten drei Regionen wurden durch `SPSS` als wichtigste Variablen zur Klassentrennung angegeben, dagegen lagen sie bei der Ranganalyse durch Ausschluss auf den mittleren bis hinteren Plätzen.

Auffällig ist, dass in allen Validierungsrechnungen Teile der Bande bei 1240 cm^{-1}, die der asymmetrischen Streckschwingung der PO_2^--Gruppen von Phospholipiden und Nukleinsäuren zugeordnet wird, selektiert waren.

15.4. Analyse der erfolgten Zuordnungen und Interpretation im Hinblick auf die klinischen Erfordernisse

Die Zuordnungsmatrix der Validierung des Modells mit acht Variablen ist in Tab. 15.6 gezeigt. Das gebildete Modell weist mit einer geschätzten Trefferrte von etwa 62 % noch nicht die für klinische Anforderungen notwendige Genauigkeit auf.

Der hohe Anteil falsch zugeordneter Astrozytome zweiten Grades wurde bereits mit der äußerst geringen Probenzahl dieser Tumore in Verbindung gebracht. Diese eingeschränkte Datenmenge bewirkt größe Unsicherheiten in der Schätzung der Klassengrenzen. Daher ist noch keine genauere Analyse der Fehlzuordnungen möglich.

Tabelle 15.6.: Die Zuordnungsmatrix des Modells mit acht Variablen für die Validierung von Optimierung und linearer Diskriminanzanalyse

wahre Klasse	**zugeordnete Klasse $\hat{k}$**					
	Anzahlen					
k	gesund	Astro II	Astro III	Glio	**richtig**	Gesamt
gesund	*2175*	71	376	19	2175	2641
Astro II	709	*892*	778	261	892	2640
Astro III	232	137	*1775*	494	1775	2638
Glio	155	106	629	*1750*	1750	2640
Gesamt	3271	1206	3558	2524	*6592*	10559
	Anteile					
	gesund	Astro II	Astro III	Glio	**richtig**	
gesund	*82,4* %	2,7 %	14,2 %	0,7 %	82,4 %	
Astro II	26,9 %	*33,8* %	29,5 %	9,9 %	33,8 %	
Astro III	8,8 %	5,2 %	*67,3* %	18,7 %	67,3 %	
Glio	5,9 %	4,0 %	23,8 %	*66,3* %	66,3 %	
Gesamt	31,0 %	11,4 %	33,7 %	23,9 %	*62,4* %	

Die Trennung zwischen gesundem und Tumorgewebe sehr viel genauer möglich, als es die Gesamt-Trefferrate erwarten lässt. Die Trefferrate für Spektren gesunder Proben wird auf 82,4 % geschätzt. Die als gesund zugeordneten Spektren enthalten etwa 33,5 % falsch negative Zuordnungen. Mehr als die Hälfte davon geht allerdings auf falsch zugeordnete Astrozytome zweiten Grades zurück, so dass auch hier mit einem starken Absinken der Fehlerrate bei Einbeziehung weiterer Proben von Astrozytomen zweiten Grades gerechnet werden kann.

Falsch negative Zuordnungen maligner Tumore sind sehr bedenklich. Allerdings muss hier zunächst überprüft werden, ob es sich tatsächlich um falsch negative Befunde handelt. Die Wahrscheinlichkeit, dass Proben maligner Tumore Anteile gesunden Gewebes enthalten ist sehr viel größer als die, dass gesunde Proben Tumorgewebe enthalten. Solche gesunden Gewebe können zum Beispiel Blutgefäße sein. Die Probe kann auch dem Rand des Tumors entstammen, so dass etwas gesundes Gewebe enthalten sein kann. Wichtig ist hier die genaue histologische Untersuchung der Proben in Verbindung mit der Lokalisation dieser als gesund klassifizierten Spektren.

Ein auffallend großer Anteil der falsch zugeordneten Spektren der gesunden Proben wurde als Astrozytom dritten Grades klassifiziert. Eine genauere Betrachtung der Zuordnungen ergab, dass es sich dabei nicht um eine einzelne falsch klassifizierte Messung oder Probe handelt, vielmehr weisen nur zwei der fünf gesunden Proben keine Spektren auf, die malignen Tumoren zugeordnet wurde. Zwischen 7 und 45 % der Spektren der restlichen drei Proben wurden den Astrozytomen dritten Grades, bis zu 3,2 % der Spektren sogar den Glioblastomen zugeordnet. Die Ursachen dieser Fehlzuordnungen können einerseits in untypischen Spektren der gesunden Proben oder in untypischen Spektren der Astrozytome dritten Grades liegen, so dass die Klassengrenze zugunsten der Astrozytome verschoben wird. Zu ersterem Grund kann beitragen, dass die Proben zwar aus im Sinne von *nicht tu-*

morös gesundem Gewebe bestehen, aber im Rahmen einer Autopsie entnommen wurden. Das bedeutet, dass unter Umständen bereits biochemische Veränderungen in den Zellen stattgefunden haben, die sich in den Spektren zeigen, aber noch nicht zu morphologischen Auffälligkeiten führten. Solche Veränderungen können sehr schnell eintreten [5]. Eine andere Ursache könnte darin liegen, dass diese Spektren bereits sehr lange vorliegen, also eventuell zu einem Zeitpunkt gemessen wurden, als die optimalen Geräteparameter noch nicht bekannt waren. In diesem Fall sollten die Proben erneut gemessen werden.

Weitere große Anteile an Fehlzuordnungen sind zwischen den Astrozytomen dritten Grades und den Glioblastomen zu beobachten. Fasst man jedoch diese beiden Klassen für die Interpretation der erfolgten Zuordnungen als maligne Tumore zusammen, so wird für diese Gruppe eine exzellente Trefferrate von über 88 % erwartet. Glioblastome weisen nekröse Bereiche auf, für die aufgrund der Zersetzungsvorgänge in den abgestorbenen Zellen eine deutlich andere biochemische Zusammensetzung als für lebende Zellen zu erwarten ist. Daher sind hier auch Veränderungen in den Spektren wahrscheinlich. Das Modell der Diskriminanzanalyse wurde gebildet, ohne dieser Möglichkeit einer weiteren Klasse von Spektren Rechnung zu tragen. Bereits eine von der histologischen Grenzziehung abweichende Unterscheidung der spektroskopischen Charakteristika der Tumore ist eine plausible Erklärung für diese Fehlzuordnungen. Aber auch die histologischen Unterschiede zwischen diesen beiden malignen Tumorgraden sind bezüglich der lebenden Teile des Gewebes gering. Aus diesen Gründen erstaunen die wechselseitigen Schwierigkeiten bei der Zuordnung nicht.

Diese Verwechslungen zwischen Astrozytomen dritten Grades und Glioblastomen sind zwar im gebildeten Modell recht häufig, aber gegenüber den Fehlzuordnungen der gesunden Proben klinisch von untergeordneter Bedeutung.

15.4.1. Die Berücksichtigung der klinischen Erfordernisse

Diese Überlegungen machen deutlich, dass für die Beurteilung des Modells unter Aspekten der klinischen Anwendung ist die Diskussion der erfolgten Zuordnungen allein nicht ausreichend ist. Klinisch bedeutsam sind nicht die Fehlzuordnungen an sich, sondern ihre Folgen. Die Festlegung einer Gewichtung der einzelnen Fehlzuordnungsmöglichkeiten eröffnet die Möglichkeit, die Leistungsfähigkeit des Modells anhand der zu erwartenden Folgen zu beurteilen. Dieses Vorgehen entspricht der Definition einer Kostenmatrix, die auch zur Modellbildung genutzt werden kann und sollte.

Die lineare Diskriminanzanalyse würde dann unter Anwendung kostenoptimaler Regeln (Kap. 6.1.1, S. 17) durchgeführt, so dass bereits die Optimierung und Bildung des Modells der linearen Diskriminanzanalyse die Erfordernisse der klinischen Anwendung berücksichtigt.

Dieses Modell wurde unter der Annahme gebildet, dass alle vier Klassen gleich häufig auftreten, also die a priori Wahrscheinlichkeiten gleich sind. Dieser Fall ist jedoch sehr unwahrscheinlich. Die Angabe der a priori Wahrscheinlichkeiten kann daher eine bedeutende Verbesserung der Modelle bewirken. Diese Wahrscheinlichkeiten sind allerdings sehr schwer zu ermitteln, sie hängen vermutlich stark von der Definition der Grundgesamtheit ab.

Teil V.

Folgerungen und Zusammenfassung

16. Zusammenfassung und Ausblick

Die in dieser Arbeit geschilderten Verfahren, sowohl zur Erstellung eines Trainingsdatensatzes als auch die Untersuchungen über die Auswirkungen der verschiedenen Datenvorbehandlungsmethoden beeinflussen sich in komplexer Weise gegenseitig. Daher kann eine solche Untersuchung nicht als abgeschlossen betrachtet werden, die in der als günstig erkannten Situation eventuell veränderten Wirkungen der restlichen Parameter bedürfen Untersuchungen. Auf diese Weise können schrittweise Verbesserungen erreicht werden.

16.1. Trainingsdaten

Im Rahmen dieser Arbeit konnte ein Trainingsdatensatz zusammengestellt werden, für den die Trefferrate für neue Spektren auf etwa 62 % geschätzt wurde. Dabei wirkt sich die geringe Probenzahl besonders der Astrozytome zweiten Grades vermutlich stark negativ auf die erreichte Modellgüte aus. Diese Situation kann sich deutlich entspannen, wenn die Integration der Images in den — zur Zeit ausschließlich aus Maps bestehenden — Trainingsdatensatz gelingt und weitere Proben vermessen werden. Auch die erneute Messung bereits zu Beginn des Projekts untersuchter Proben kann vorteilhaft sein, da inzwischen günstige Parameter und Vorgehen sowohl für die Präparation der Schnitte als auch die Messung der Spektren bekannt sind.

Weitere Gründe sprechen jedoch dafür, dass auch mit den vorliegenden Daten noch Verbesserungen erzielt werden können. Die genutzte Auswertestrategie ist in verschiedener Hinsicht noch nicht exakt an die Erfordernisse der vorliegenden Problemstellung angepasst.

16.2. Die Klinische Anwendbarkeit der Methode

Das diskutierte Modell genügt den Anforderungen an eine Diagnosemethode für die klinische Klassifikation in die betrachteten vier Gewebearte noch nicht. Allerdings ist die Genauigkeit der Zuordnungen für gesundes Gewebe sehr viel gößer. Die Trefferrate wurde mittels einer Set-Validierung auf 82 % geschätzt. Werden die malignen Tumore, also die Astrozytome dritten Grades und die Glioblastome, zur Auswertung zusammengefasst, so resultiert eine Trefferrate von 88 %.

Aber auch die Beurteilung der Leistungsfähigkeit eines Modells ist noch nicht an die klinischen Erfordernisse angepasst. Es bestehen jedoch verschiedene vielversprechende Wege, um die Leistungsfähigkeit der gebildeten Modelle zu erhöhen. Die lineare Diskriminanzanalyse erlaubt eine genaue Abstimmung der Auswertung an die konkrete Anwendung unter Einbeziehung unterschiedlichen Vorwissens. Solches medizinisches Vorwissen sind die unterschiedlich schweren Folgen der verschiedenen Fehlzuordnungen und die Häufigkeiten des Auftretens der unterschiedlichen Tumore.

Die Anwendung dieser Möglichkeiten ist dabei nicht auf das Erstellen des Datensatzes und die Interpretation der Ergebnisse beschränkt, auch die Modellbildung durch `ga_ors` und `stackedGen` könnte — nach einigen Anpassungen der Programme — dieses Vorwissen

berücksichtigen. Dann sind im Hinblick auf die medizinische Anwendung deutlich verbesserte Modelle zu erwarten.

16.3. Anpassung von `ga_ors`

16.3.1. Zufallszahlen

Wie ausführlich dargelegt wurde, hat das beobachtete Fehlverhalten des Optimierungs-Algorithmus umfangreiche Auswirkungen. Da für identische Eingaben identische Programmläufe resultieren, ist es weder möglich, das Risiko, schlechte Ergebnisse zu erhalten, zu bestimmen, noch, die Ergebnisse durch erneute Rechnungen zu verbessern.

Eine Überprüfung oder Optimierung der Parameter zur Steuerung des Programms außer der Generationenzahl konnte daher nicht vorgenommen werden.

Insgesamt müssen die Optimierungsergebnisse daher zur Zeit als unzuverlässig eingeordnet werden.

16.3.2. Gewichtung der Daten

Zur Zeit müssen für alle Klassen möglichst ähnliche Spektrenzahlen in die Modellbildung eingehen, damit die resultierenden Zuordnungen nicht zugunsten einzelner Klassen ausfallen. Könnten zu den einzelnen Spektren Gewichte, mit denen sie in die Modellbildung einfließen sollen, angegeben werden, so wäre die zur Zeit notwendige Auswahl nur eines Teils der zur Verfügung stehenden Daten einerseits und das Duplizieren von Spektren andererseits vermeidbar.

Auch die Möglichkeit, die a priori Wahrscheinlichkeiten vorzugeben, wäre bereits hilfreich, wenn auch nicht so weitreichend wie der erstgenannte Vorschlag.

16.3.3. Angabe der a posteriori Wahrscheinlichkeiten

Die Auswertung der für Trainingsdaten angegebenen a posteriori Wahrscheinlichkeiten der Klassenzugehörigkeit eines Spektrums ermöglicht dort eine weitere Steigerung der Trefferrate. Für Spektren, die mit hohen a posteriori Wahrscheinlichkeiten zugeordnet wurden, traten sehr viel weniger Fehlzuordnungen auf als für Spektren, die nur mit einer geringen Wahrscheinlichkeit der Klasse angehörten, der sie zugeordnet wurden.

Diese Möglichkeit, die Zuverlässigkeit der einzelnen Zuordnung abzuschätzen, sollte auch für Testdaten bestehen.

16.3.4. Set-Validierung

Da die Spektren einer Probe untereinander sehr ähnlich sind, müssen für eine zuverlässige Schätzung der Modellgüte bei der Validierung alle Spektren einer Probe gleichzeitig aus der Modellbildung ausgeschlossen werden. Das entspricht der Möglichkeit, eine Set-Validierung durchzuführen, wobei die zugehörigkeit der Spektren zu den Proben bei der Bestimmung der Grenzen der einzelnen Sets berücksichtigt werden.

Die Implementierung einer solchen Set-Validierung ist außerordentlich wichtig, da die Schätzung der Modellgüte als Zielfunktional der Optimierung dient und daher die Optimierungsergebnisse empfindlich von der Qualität dieser Schätzung abhängen.

16.3.5. Optimierung der Variablenanzahl

Ist oben genanntes Zielfunktional vorhanden, so sollte neben der Bestimmung der für die lineare Diskriminanzanalyse verwendeten spektralen Regionen auch gleichzeitig die Optimierung der Anzahl dieser Regionen möglich sein.

Allerdings ist zu überprüfen, ob die Veränderung allein der Fitness-Funktion ausreicht, um auch die Regionenanzahl zu optimieren. Unter Umständen müssen auch die Operatoren zur Erzeugung der Kind-Population des genetischen Algorithmus angepasst werden.

16.3.6. Kostenoptimale Zuordnungen

Eine kostenoptimale Zuordnung kann helfen, die gebildeten Modelle besser an die klinischen Anforderungen anzupassen, da den einzelnen Möglichkeiten der Fehlzuordnungen unterschiedliches Gewicht begemessen wird. Dies würde eine weitere Änderung des Zielfunktionals der Optimierung bedeuten.

16.4. Empfohlenes Vorgehen bei der Analyse mit `ga_ors` und `stackedGen`

Die hier gegebenen Empfehlungen sind mit Vorsicht anzuwenden, da sie unter Benutzung eines fehlerhaften und auch weiterhin in der Entwicklung befindlichen Programmsystems entstanden sind. Daher ist in der Zukunft mit einem veränderten Verhalten des Programms zu rechnen, dessen Folgen im Hinblick auf gute oder ungünstige Datenvorbehandlung oder auch Programmparameter nicht abgeschätzt werden können.

16.4.1. Datenvorbereitung

Ein wichtiges Hilfsmittel zur Beurteilung der Datensätze ist auch die Untersuchung der gefärbten und ungefärbten Schnitte. Die allein auf den Daten beruhenden Verfahren zur Beurteilung der Eignung als Trainingsdaten sind auf eine bereits erfolgte Optimierung durch `ga_ors` angewiesen. Die Untersuchung der Verteilung der durch `ga_ors` gebildeten Variablen hat sich als leistungsfähig erwiesen. Dies kann mittels der entsprechenden Histogramme oder in der Auftragung der Spektren im Koordinatensystem der Variablen erfolgen. Auch die Darstellung der Spektren über den Diskriminanzfunktionen ist geeignet. Diese Werte werden jedoch nicht durch `ga_ors` und `stackedGen` zur Verfügung gestellt. Sie können mit Hilfe der linearen Diskriminanzanalyse in `SPSS` ermittelt werden, der Export dieser Daten ist jedoch sehr aufwändig. Weiterhin ist nicht sichergestellt, dass `SPSS` und `stackedGen` exakt gleiche Ergebnisse ermitteln. Von der Anwendung der Reklassifikations-Treffer- oder Fehlerrate zur Beurteilung der Daten muss abgeraten werden.

Die Spektrenanzahlen sollten für die einzelnen Klassen annähernd gleich groß gewählt werden. Auch innerhalb der Klassen ist zu bedenken, dass die Wahl der Spektrenzahlen eine Gewichtung der einzelnen Messungen und Proben bewirkt. Die Mittelwertbildung über örtlich benachbarte Spektren zur Reduktion der Spektrenanzahl ist zu empfehlen. Die genaue Einhaltung der angegebenen Anzahlen wurde im Rahmen dieser Arbeit zusätzlich durch eine zufällige Auswahl beziehungsweise durch duplizieren der vorhandenen Spektren erreicht.

Die Auflösung der zur Verfügung stehenden Spektren mit einem Abstand von etwa 4 cm^{-1} sollte nicht weiter verringert werden.

Die Images konnten nicht in den aus den Maps gebildeten Trainingsdatensatz integriert werden.

Die Untersuchung ausschließlich des Fingerprint-Bereichs zwischen 1000 und 1800 cm^{-1} bewirkte gegenüber der Untersuchung des Spektrums zwischen 950 und 3800 cm^{-1} geringfügig verbesserte Modelle. Unter Berücksichtigung des deutlich verringerten Ressourcenbedarfs ist eine Beschränkung auf den Fingerprint-Bereich zu bevorzugen.

Die Datenreduktion durch Mittelwertbildung vor der Interpolation auf eine einheitliche Wellenzahl-Achse erbrachte keine größen Vorteile. Auch die Anwendung einer linearen Basislinienkorrektur bewirkte keine wesentliche Änderung in der Modellgüte.

Zwischen Flächen- und Minimum-Maximum-Normierung traten bezüglich der erreichten Modellgüte keine Unterschiede auf. Eine Intensitätsnormierung ist jedoch notwendig. Die schiefe Verteilung der Extinktionswerte könnte eine Folge der Routinen zur Intensitätsnormierung sein, daher ist hier ein weiterer Ansatzpunkt für Verbesserungen.

Zur Anwendung des beschriebenen Histogramm-Filters kann geraten werden, da eine deutliche Steigerung der Modellgüte erreicht wurde. Dies ist insbesondere der Fall, solange die Verteilungen der Extinktionswerte schief sind.

16.4.2. Programmparameter für die Optimierung

Die gewählte Iterationszahl von 100 Generationen bei der Optimierung erwies sich als ausreichend. Die weiteren Parameter konnten aufgrund des Programmfehlers in `ga_ors` nicht überprüft werden.

16.4.3. Validierungsverfahren

Die Reklassifikations-Trefferrate wurde zur vergleichenden Beurteilung ähnlicher Modelle benutzt. Dabei wurden die Variablenzahl der LDA und die Größe des Datensatzes konstant gehalten.

Hinweise auf die optimale Anzahl an Variablen können aus der Validierung der LDA gewonnen werden, da diese Schätzung die Übermodellierung der Trainingsdaten widergibt.

Die absolute Einschätzung der Modellgüte benötigt eine Validierung durch Spektrendaten, die weder an der Suche geeigneter spektraler Regionen durch `ga_ors` noch an der Bildung des Modells der linearen Diskriminanzanalyse durch `stackedGen` beteiligt sind.

Bei der Anwendung der beiden letztgenannten Verfahren ist darauf zu achten, dass der Ausschluss kompletter Proben aus der Modellbildung erfolgen muss, um diese Daten zur Validierung nutzen zu können.

17. Dank

Mein Dank gilt Herrn Prof. R. Salzer, der mir dieses interessante Thema zur selbstständigen Bearbeitung überlassen hat.

Auch Herrn Dr. G. Steiner danke ich für die Betreuung der Arbeit mit vielen fachlichen Anregungen und Diskussionen.

Viele Informationen, Hinweise und Ideen stammen aus den zahlreichen Diskussionen mit Kollegen hier am Institut und am Thema interessierten Freunden. An dieser Stelle möchte ich mich auch bei ihnen ganz herzlich bedanken.

18. Ehrenwörtliche Erklärung

Ich erkläre, dass ich die vorliegende, unter Betreuung von Herrn Dr. Ing. G. Steiner angefertigte Arbeit selbstständig verfasst habe. Andere als die angegebenen Hilfsmittel wurden von mir nicht genutzt, alle angeführten Zitate wurden kenntlich gemacht.

Dresden, 30. Januar 2003

Teil VI.

Anhang

A. Beschreibung ausgewählter Funktionen, Scripte und Datenstrukturen

Die im Rahmen dieser Arbeit entstandenen Datenstrukturen und Routinen wurden sorgfältig und im Hinblick auf ihre weitere Verwendbarkeit entwickelt. Da die Arbeit insbesondere sehr vielfältige Möglichkeiten der Datenbehandlung untersucht, handelt es sich jedoch nicht um statische Gebilde, die Funktionen und auch die Datenstrukturen wurden entsprechend des jeweilgen Untersuchungsziels verändert. Auch weiterhin sind größere Veränderungen in den Strukturen und Routinen zu erwarten.

Daher sind die hier gezeigten Datenstrukturen und Funktionen als derzeitiger Stand sowie als Anregung zur Weiterentwicklung zu verstehen. An dieser Stelle kann jedoch nur eine stichwortartige Kurzbeschreibung der wichtigsten Funktionen erfolgen, ein Abdruck der kompletten Dokumentation würde den Rahmen dieser Arbeit sprengen.

A.1. Datenstrukturen

Die Basis der entwickelten Datenstruktur ist die *Messung*. Messung steht hier für ein Image oder Map. Die so entstandene Datenstruktur ermöglicht eine große Flexibilität in der Darstellung und Gruppierung der Daten, ohne dabei zu viele hierarchische Ebenen in der Datenstruktur zu benötigen. `Matlab` stellt verschiedene Datenstrukturen zur Verfügung, die die Gruppierung der Daten bequem ermöglichen. Zum einen können `cell`-Matrizen die gruppierten `mstruct`-Strukturen mit den Daten aufnehmen, zum anderen besteht auch die Möglichkeit, die Indizes der `mstruct`-Struktur nach verschiedenen Gesichtspunkten gruppiert oder sortiert abzulegen.

A.1.1. Die Struktur `mstruct`

Diese Struktur kann die Daten einer Messung aufnehmen. Eine leere `mstruct`-Struktur kann mittels der Funktion `get_mstruct` erhalten werden.

Die Struktur enthält folgende Felder:

A.1.2. Die Struktur `minf`

Die Struktur `minf` enthält einen Teil der Felder der `mstruct`-Struktur. Dabei handelt es sich um die notwendigen Informationen zum Laden der Daten aus den `Matlab`-Dateien. Alle `Matlab`-Funktionen, die `minf`-Strukturen als Eingabe verarbeiten, können auch `mstruct`-Strukturen bearbeiten, da die vorhandenen Felder identisch sind.

Die Felder der `minf` sind: `mname`, `ldaclass`, `ldatest`, `diagnose`, `maxz`, `maxs`, sowie `orgformat`.

Tabelle A.1.: Die Felder der Struktur `mstruct`

Name	Dimension	Typ	Beschreibung
`x`	1 × `npoints`	double	$\tilde{\nu}$-Achse der Spektren
	2 × `npoints`	double	Anfang und Ende der Regionen in Wellenzahlen
`y`	`nspc` × `npoints`	double	die Spektrendaten
`maxz`	1 × 1	double	Zeilenzahl des Aufnahmerasters der Spektren
`maxs`	1 × 1	double	Spaltenzahl des Aufnahmerasters der Spektren
`nspc`	1 × 1	double	Anzahl an Spektren
`nsel`	1 × 1	double	Anzahl ausgewählter Spektren
`pic`			Bild der Messung (Lichtmikroskop)
`picname`			Pfad und Dateiname des Bildes
`index`	`maxz` × `maxs`	double	Spaltenindex für `y` des Spektrums an der gegebenen Rasterposition
`flags`	`maxz` × `maxs`	double	enthält Information, ob Spektrum an der gegebenen Rasterposition vorhanden (`SP_EXISTS`) oder ausgewählt (`SP_SELECT`) ist
`dir`	1 × ?	char	Verzeichnis bzw. Datei mit den Rohdaten
`npoints`	1 × 1	double	Anzahl an Wellenlängen der Spektren
`data_processing`	1 × ?	char	erfolgte Datenbehandlung
`comments`	1 × ?	char	Kommentare
`date`	1 × ?	char	Aufnahmedatum der Spektren
`orgformat`	1 × ?	char	ursprüngliches Datenformat der Spektren
`ldaclass`	1 × 1	double	Klasse bei der Diskriminanzanalyse, kann in der Bedeutung gegenüber `diagnose` variieren, da die Eingabedatei für `ga_ors` fortlaufende Werte für die Klassen benötigt.
`mname`	1 × ?	char	Benennung der Messung, wird gleichzeitig als Dateiname zum Ablegen der `mstruct` benutzt.
`ldatest`	1 × 1	double	wie `ldaclass`, aber einige zusätzliche Werte: ausgeschlossen (0) und Testdaten (-1)
`ldaerg`	1 × 1 ×	double	Matrix für die Zuordnungen der LDA
`diagnose`	1 × 1	double	Diagnose zur Probe, verwendet werden die Konstanten `GESUND`,`ASTRO2`, `ASTRO3`, `GLIO`

A.2. Funktionen nach Kategorien geordnet

A.2.1. Funktionen zur Anzeige der Daten

`plot_flags` Zeigt Existenz und Auswahl der Spektren an.

`plot_hellfeld` Zeigt die Summe der Absorptionswerte der Spektren an.

`plot_ldaerg` Ergebnisse der LDA als farbcodierter Plot

`print_hitrate` Tabelle mit Hitratematrix in Anzahlen und Anteilen

`print_hitrate_m` Tabelle mit Hitratematrix in Anzahlen und Anteilen für jede Messung

`print_minf` gibt die Informationen in `minf` aus

A.2.2. Funktionen zur Auswertung der Ergebnisse der Programme `ga_ors` und `stackedGen`

`do_read_erg` Liest .erg-Datei (Ergebnisse der LDA) und trägt Ergebnisse in `messg. ldaerg` ein.

`do_read_log` Extrahiert aus dem LOG-File von GA-ORS die gefundenen Regionen.

`do_regions` Transformiert die Spektren der Proben in das durch die gefundenen Regionen definierte Koordinatensystem

`get_zuordngn_m` Berechnet die Zuordnungen der Spektren der einzelnen Messungen aus `messgn. ldaerg`

`get_zuordngn_s` Berechnet die Zuordnungsmatrix aus `messgn. ldaerg`

`plot_ldaerg` Ergebnisse der LDA als farbcodierter Plot

`print_hitrate` Tabelle mit Hitratematrix in Anzahlen und Anteilen

`print_hitrate_m` Tabelle mit Hitratematrix in Anzahlen und Anteilen für jede Messung

`read_ergs` Einlesen von .erg-Dateien. Schnittstelle zu do_read_erg.

`read_logs` Liest die ermittelten Regionen aus .log-Datei.

`write_lda_set` Erzeugt Eingabe für Set-Validierung nur der LDA.

A.2.3. Schnittstelle zu `ga_ors`

`do_read_erg` Liest .erg-Datei (Ergebnisse der LDA) und trägt Ergebnisse in `messg. ldaerg` ein.

`do_read_log` Extrahiert aus dem LOG-File von GA-ORS die gefundenen Regionen.

`read_ergs` Einlesen von .erg-Dateien. Schnittstelle zu do_read_erg.

`read_logs` Liest die ermittelten Regionen aus .log-Datei.

`write_ga_in` Erzeugt eine Eingabe-Datei für GA-ORS.

`write_lda` Erzeugt Eingabe für Set-Validierung nur der LDA.

`write_lda_set` Erzeugt Eingabe für Set-Validierung nur der LDA.

`write_set` Erzeugt Dateien für Set-Validierung.

A.2.4. Konstanten

`ASTRO2` Konstante für die Diagnose Astrozytom zweiten Grades für `mstruct.diagnose`

`ASTRO3` Konstante für die Diagnose Astrozytom dritten Grades für `mstruct.diagnose`

`GESUND` Konstante für `mstruct. diagnose`

`GLIO` Konstante für `mstruct. diagnose`

`SP_EXISTS` Definition des Flags für existierende Spektren

`SP_SELECT` Definition des Flags für ausgewählte Spektren

A.2.5. Funktionen zur Konvertierung der Spektrendateien

`do_const_x` Interpoliert die Spektrendaten auf eine neue, kürzere $\tilde{\nu}$-Achse.

`read_messg_jdx_qd` Schnittstelle zu read_jdxs_qd, lädt .JDX-Dateien.

`read_messg_spc` Schnittstelle zu spc_load, lädt .SPC-Dateien.

A.2.6. Funktionen zur Arbeit mit den Datenstrukturen

`get_mstruct` Erzeugt eine leere Struktur für die Probendaten.

`print_minf` Gibt die Informationen in `minf` aus.

`split_messgn` Gruppiert die einzelnen Messungen nach den Proben.

`split_minf_ldatest` Teilt `minf` nach gleichen `. ldatest` auf.

`split_mname` teilt `minf. ldaname` in Probennummer und einen alphanumerischen Teil.

A.2.7. Funktionen zur Datenvorbehandlung

`calc_nspc_train` Berechnet die Spektrenzahlen für die Messungen des Trainings-Sets, so dass gleiche Spektrenzahlen für alle Klassen resultieren.

`do_blcorr_01` Lineare Basislinienkorrektur über gesamtes Spektrum

`do_intnorm_a` Flächennormierung

`do_intnorm_minmax` Minimum-Maximum-Normierung

`do_lateral_mean` Mittelwertbildung über örtlich benachbarte Spektren

`do_ofscorr` Offsetkorrektur

`do_xmean` Verkürzt die $\tilde{\nu}$-Ausdehnung der Probe durch Mittelwertbildung um den Faktor `teiler`

`do_xrange` Schneidet die Spektrenbereiche zwischen `xstart` und `xend` aus.

`duplicate_spectra` Vervielfältigt die Spektren der einzelnen Messungen, so dass die geforterte Anzahl eingehalten wird.

`get_sel` Liefert Probe mit nur noch den ausgewählten Spektren zurück.

`sel_filter_01` Mittelwert-Filter

`sel_filter_02` Histogramm-Filter

`sel_max_in` Selektiert alle Spektren, deren Maximum zwischen `min` und `max` liegt.

`sel_min_in` Selektiert alle Spektren, deren Minimum zwischen `minimum` und `maximum` liegt.

`sel_rnd` Wählt zufällig `nsel` Spektren der Probe aus.

A.3. Alphabetische Liste der Funktionen — Syntax und Kurzbeschreibung

Funktion `r = ASTRO2`
Konstante für die Diagnose Astrozytom zweiten Grades für `mstruct.diagnose`

Funktion `r = ASTRO3`
Konstante für die Diagnose Astrozytom dritten Grades für `mstruct.diagnose`

Funktion `[minf, nspc] = calc_nspc_train (minf, nmin)`
Berechnet die Spektrenzahlen für die Messungen des Trainings-Sets, so dass gleiche Spektrenzahlen für alle Klassen resultieren.

Weiterhin werden die einzelnen Proben innerhalb jeder Klasse und in den Proben die Messungen gleich gewichtet.

Funktion `messg = do_blcorr_01 (messg)`
Lineare Basislinienkorrektur über gesamtes Spektrum.

do_blcorr_01 führt eine lineare Basislinienkorrektur durch, indem die Basislinie durch den ersten und letzten Messpunkt jedes Spektrums gelegt wird.

Funktion `do_const_x (minf, xnew)`
Interpoliert die Spektrendaten auf eine neue, kürzere $\tilde{\nu}$-Achse.

Die Proben werden aus .mat-Dateien im Verzeichnis
rohdaten geladen und abhängig vom Orignal-Datenformat weiterbehandelt: bei Spektren des Bruker-Geräts (Orignal-Format SPC) wird die erste Spalte von `messg. x` und `messg. y` gelöscht, die Spektren des Nicolet-Geräts werden mit `interp1` auf die neue $\tilde{\nu}$-Achse interpoliert. Die bearbeitete Probe wird als .mat-Datei im Verzeichnis
xconst gespeichert.

Funktion `messg = do_intnorm_a (messg)`
Flächennormierung

Die Gesamt-Intensität der einzelnen Spektren wird auf 1 normiert, indem jeder Wert durch die Gesamt-Intensität des Spektrums geteilt wird.

Funktion `messg = do_intnorm_minmax (messg)`
Minimum-Maximum-Normierung

Vor der Normierung wird eine Offset-Korrektur mittels `do_ofscorr` durchgeführt, so dass das Minimum bei 0 liegt. Das Maximum der Amid-I-Bande wird zwischen 1600 und 1700 cm^{-1} gesucht und das Spektrum durch diese Extinktion geteilt.

Funktion `new = do_lateral_mean (messg, groesse)`
Mittelwertbildung über örtlich benachbarte Spektren

Über die Spektren der Messung wird ein quadratisches Raster mit der Kantenlänge `groesse` gelegt. Dabei werden eventuell vorhandene leere Rasterzeilen und -Spalten beachtet. Die Mittelwertbildung erfolgt jeweils über alle vorhandenen Spektren innerhalb einer solchen Rasterzelle.

Funktion `messg` = `do_ofscorr` (`messg`)
Offsetkorrektur

Die minimale Extinktion jedes Spektrums wird vom gesamten Spektrum abgezogen.

Funktion `do_read_erg` (`filename, messg, train, index`)
Liest .erg-Datei (Ergebnisse der LDA) und trägt Ergebnisse in `messg. ldaerg` ein.

Die Datei wird geöffnet und das in `pattern` abgelegte Muster gesucht. Einige Zeilen weiter beginnt die Zuordnungstabelle.
Aus dem Namen des Datensatzes werden `i` und `n` ermittelt und die zugeordnete Klasse in `assclass` abgelegt und in `messg (i). ldaerg (n, index, 1)` eingetragen. Ist die Zuordnung nicht als unsicher gekennzeichnet (kein `~`), so wird auch `messg (i). ldaerg (n, index, 2)` mit `assclass` belegt.

Funktion `do_read_log` (`filename, regs`)
Extrahiert aus dem LOG-File von GA-ORS die gefundenen Regionen.

Die Regionen werden in die bereits angelegte Matrix `regs` eingetragen, dabei enthält `regs (:, 1)` den Beginn und `regs (:, 2)` das Ende der Region, jeweils als Index für die entsprechende Spalte in `messg. x` und `messg. y` . Wurden weniger Regionen gefunden, als beim Aufruf von GA-ORS angegeben (also auch weniger als aus dem Dateinamen ablesbar sind), so bleiben die restlichen Elemente von `regs` unverändert.

Funktion `messg` = `do_regions` (`messg, regions`)
Transformiert die Spektren der Proben in das durch die gefundenen Regionen definierte Koordinatensystem.

Die Werte in `messg. y` (Extinktionswerte der Spektren) werden durch die Mittelwerte in den einzelnen Regionen ersetzt, die neue x-Achse enthält in der ersten Zeile die Start-Wellenzahlen und in der zweiten Zeile die End-Wellenzahlen der Regionen.

Funktion `messg` = `do_xmean` (`messg, teiler`)
Verkürzt die x-($\tilde{\nu}$)-Ausdehnung der Probe durch Mittelwertbildung um den Faktor `teiler`.

Berechnet aus `messg. x` und `messg. y` jeweils den Mittelwert über `teiler` Punkte in x-Richtung.

Funktion `messg` = `do_xrange` (`messg, xstart, xend`)
Schneidet die Spektrenbereiche zwischen `xstart` und `xend` ab.

Funktion `messgn` = `duplicate_spectra` (`messgn, nspc`)
Vervielfältigt die Spektren der einzelnen Messungen, so dass die geforterte Anzahl eingehalten wird.

Zunächst werden an das Ende von `.y` so viele Kopien von `.y` angehängt, wie ganzzahlige Vielfache gefordert sind. Die restlichen notwendigen Spektren werden durch Kopie einer zufälligen Auswahl an Spektren erhalten.

Funktion `r` = `GESUND`
Konstante für `mstruct. diagnose`

Funktion `mstruct` = `get_mstruct`
Erzeugt eine leere Struktur für die Probendaten

Funktion `messgnew` = `get_sel` (`messg`)
Liefert Probe mit nur noch den ausgewählten Spektren zurück.

Die mit `messg. flags ==` SP_SELECT gewählten Spektren werden behalten, die restlichen Spektren gelöscht.

Funktion `hitrates` = `get_zuordngn_m` (`messgn, index, nonfuzzy, nkl`)
Berechnet die Zuordnungen der Spektren der einzelnen Messungen aus `messgn. ldaerg`.

Diese Zuordnungsmatrix stellt für die einzelnen Messungen (Zeile) die zugeordneten Klassen (Spalte) dar. Die Anzahl der richtig zugeordneten Spektren steht jeweils in der `messg. ldaclass`-ten Spalte.

Funktion `hitrates` = `get_zuordngn_s` (`messgn, index, nonfuzzy, nkl`)
Berechnet die Zuordnungsmatrix aus `messgn. ldaerg`.

Die Zuordnungsmatrix stellt den wahren Klassen (Zeile) die zugeordneten Klassen (Spalte) gegenüber. Die Hauptdiagonale enthält die Anzahl der richtigen Zuordungen (pro Klasse), alle Nebendiagonalelemente geben falsche Zuordnungen wider.

Funktion `r = GLIO`
Konstante für `mstruct. diagnose`

Funktion `plot_flags (messg)`
Zeigt Existenz und Auswahl der Spektren an.

Eine neue `figure` wird erzeugt, in der `messg. flags` mit der Colormap `cmap` angezeigt wird.
Legende:
schwarz: Spektrum ist nicht vorhanden
blau: Spektrum ist vorhanden
hellgrün: Spektrum ist vorhanden und ausgewählt
rot: Spektrum ist ausgewählt, obwohl es nicht existiert.

Funktion `plot_hellfeld (messg)`
Zeigt die Summe der Absorptionswerte der Spektren an.

Entsprechend des Probenrasters wird die Summe der Absorptionswerte des jeweiligen Spektrums als Graustufe dargestellt.

Funktion `plot_ldaerg (messg, nonfuzzy)`
Ergebnisse der LDA als farbcodierter Plot

Trägt die Ergebnisse im Probenraster ein, daher kann das Bild mit plot_flags und plot_hellfeld verglichen werden.

Funktion `print_hitrate (messgn, hitratematrix, fid, desc)`
Tabelle mit Hitratematrix in Anzahlen und Anteilen

Funktion `print_hitrate_m (messgn, hitratematrix_m, fid, desc)`
Tabelle mit Hitratematrix in Anzahlen und Anteilen für jede Messung

Funktion `print_minf (minf, fid, anzeige)`
Gibt die Informationen in `minf` aus.

Funktion `messgn = read_ergs (filter, messgn)`
Einlesen von .erg-Dateien. Schnittstelle zu do_read_erg.

Liest alle Dateien ein, die `filter` entsprechen. Die Ergebnisse werden dem Dateinamen entsprechend in `messg. ldaerg abgelegt` .

Funktion `read_logs`
Liest die ermittelten Regionen aus .log-Datei. Schnittstelle zu do_read_log.

Funktion `messg = read_messg_jdx_qd (path, maxs, maxz)`
Schnittstelle zu read_jdxs_qd, lädt .JDX-Dateien.

Alle gefundenen Dateien werden geladen und die entsprechenden weiteren Informationen ebenfalls in eine `mstruct` eingetragen.

Funktion `messg = read_messg_spc (path, maxs, maxz)`
Schnittstelle zu spc_load, lädt .SPC-Dateien.

Alle gefundenen Dateien werden geladen und die entsprechenden weiteren Informationen ebenfalls in eine `mstruct` eingetragen.

Funktion `messg = sel_filter_01 (messg, thresh)`
Mittelwert-Filter

Selektiert alle Spektren, die für alle Wellenzahlen maximal um das in `thresh` angegebene Vielfache vom Mittelwert der Extinktionswerte aller Spektren der Messung abweichen. Ausgewählt sind Spektren, deren zugehöriges Flag `messg. flags (z, s)` auf SP_SELECT gesetzt ist.

Funktion `messg = sel_filter_02 (messg, level)`
Histogramm-Filter

Die Verteilung der Extinktionswerte der Spektren wird gebildet. Spektren, deren Extinktionswerte für alle Wellenzahlen zwischen dem Mittelwert der ersten Klasse und dem Mittelwert der letzten Klasse mit der geforderten Mindestbesetzung liegen, werden ausgewählt. Die Mindestbesetzung wird durch `level`

festgelegt, sie berechnet sich zu `level` · sqrt (`messg. nspc`). Dadurch wird die aufgrund der unterschiedlichen Spektrenzahlen notwendige variierende Klassenbreite ausgeglichen. Ausgewählt sind Spektren, deren zugehöriges Flag `messg. flags (z, s)` auf SP_SELECT gesetzt ist.

Funktion `messg` = `sel_max_in` (`messg, min, max`)
Selektiert alle Spektren, deren Maximum zwischen `min` und `max` liegt.

Ausgewählt sind Spektren, deren zugehöriges Flag `messg. flags (z, s)` auf SP_SELECT gesetzt ist.

Funktion `messg` = `sel_min_in` (`messg, mininmum, maximum`)
Selektiert alle Spektren, deren Minimum zwischen `minimum` und `maximum` liegt.

Ausgewählt sind Spektren, deren zugehöriges Flag `messg. flags (z, s)` auf SP_SELECT gesetzt ist.

Funktion `messg` = `sel_rnd` (`messg, nsel`)
Wählt zufällig `nsel` Spektren der Probe aus.

Funktion `SP_EXISTS` = `SP_EXISTS`
Definition des Flags für existierende Spektren

Funktion `proben` = `split_messgn` (`messgn`)
Gruppiert die einzelnen Messungen nach den Proben.

Funktion `minfs` = `split_minf_ldatest` (`minf`)
Teilt `minf` nach gleichen `. ldatest` auf.

Funktion [`hlp, hlpstr`] = `split_mname` (`minf`)
teilt `minf. ldaname` in Probennummer und einen alphanumerischen Teil.

Funktion `SP_SELECT` = `SP_SELECT`
Definition des Flags für ausgewählte Spektren

Funktion `write_ga_in` (`messg`)
Erzeugt eine Eingabe-Datei für GA-ORS.

Schnittstelle zu write_ga_in_file. Speichert außerdem `messgn` als .mat-Datei, damit für die Auswertung die richtigen Klassenzuordnungen zur Verfügung stehen.

Funktion `write_lda` (`messgn, regs`)
Erzeugt Eingabe für Set-Validierung nur der LDA.

Die Spektren in `messgn` werden in die gefundenen Regionen umgerechnet, dann mit write_ga_in die Eingabe-Datei für `stackedGen` erzeugt.

Funktion `write_lda_set` (`messgn, regs, nsets`)
Erzeugt Eingabe für Set-Validierung nur der LDA.

Die Spektren in `messgn` werden in die gefundenen Regionen umgerechnet, dann mit write_set die Test-Sets erzeugt.

Funktion `write_set` (`messg, nsets`)
Erzeugt die Dateien für eine Set-Validierung.

Die Messungen werden probenweise nach Diagnose gruppiert und zufällige Test-Sets gebildet, wobei die Zahl an Proben mit einer bestimmten Diagnose in allen Sets ungefähr gleich ist. Für die Test-Proben werden jeweils alle Messungen als Test-Datensatz geschrieben, für die Trainingsmessungen nur die zum Training gekennzeichneten Messungen.

B. Dateiformate der Programme `ga_ors` und `stackedGen`

B.1. Dateiformate

Die Dateien dürfen beliebige Namen haben, es wurde das in Tab. B.1 gezeigte Schema benutzt.

Tabelle B.1.: Übersicht über die Dateiformate

Dateiendung	Bedeutung	Inhalt
`.in`	Eingabedatei für `ga_ors`	Spektren
`.log`	Protokolldatei von `ga_ors`	Verlauf der Optimierung, gewählte Regionen
`.out`	Ausgabedatei von `ga_ors` Eingabedatei für `stackedGen`	Spektren in Form der gebildeten Variablen
`.erg`	Ergebnisse von `stackedGen`	Ergebnisse der LDA, Zuordnungstabellen mit a posteriori Wahrscheinlichkeiten, Zuordnungsmatrix für Trainingsdaten

B.1.1. Das Format der Eingabedateien

Das Format der Eingabedateien für `ga_ors` und `stackedGen` ist identisch. Die Eingabedatei für `stackedGen` kann dabei eine Ausgabe von `ga_ors` sein oder extern erzeugt werden (`write_ga_in` (A.3, S. 90)).

Eine solche Datei hat folgendes Aussehen:

```
1   Classification problem with  10559 spectra
    10559 8379  Sample set size and train set size
    8 4  Number of attributes and number of classes
    0001#0001 0.797510 1.78110 1.97987 2.61339 3.90027 6.40445 3.03538 0.623198 1.000000
5   0001#0002 0.758712 1.73815 1.91448 2.57910 3.87500 6.37995 2.95800 0.592495 1.000000

    ⋮

    0009#0001 0.853483 1.49998 1.56955 2.00328 2.53125 5.71569 1.95963 0.575358 -1.000000
8384  0009#0002 0.790160 1.54956 1.48790 1.63881 2.17290 5.47613 1.78791 0.524821 -1.000000
```

Im Kopf der Datei sind einige Angaben über die Größe und Struktur des Datensatzes zu finden: In der ersten Zeile wird die Gesamtzahl aller Spektren in der Datei angegeben, in der zweiten Zeile stehen nochmals die Gesamtzahl der Spektren sowie die Anzahl der Trainingsspektren. Die dritte Zeile beschreibt die Anzahl p der Variablen und die Anzahl der Klassen g.

Danach beginnen die eigentlichen Daten, jede weitere Zeile beschreibt ein Spektrum. Zunächst besteht die Möglichkeit, einen Namen für das jeweilige Spektrum anzugeben, der maximal 10 Zeichen umfassen darf. Danach folgen die p Messwerte des Spektrums und schließlich wird die Klassenzugehörigkeit angegeben. Für diese letzte Zahl in jeder Zeile gilt

zu beachten, dass in einer Datei alle Werte von 1 bis g vorhanden sein müssen. Wird die Datei mittels `write_ga_in` erzeugt, so ist diese Randbedingung automatisch eingehalten, da `write_ga_in` eine neue Nummerierung der Klassen vornimmt, falls das erforderlich ist. Der Wert -1 kennzeichnet Testdaten.

B.1.2. Das Format der Protokolldatei von `ga_ors`

`ga_ors` erzeugt während des Programmlaufs Meldungen über den Fortschritt der Optimierung, die auch in der Protokolldatei gespeichert werden.

Diese Protokolldatei liefert Informationen über den Fortschritt der Optimierung und gibt die ermittelten Regionen an.

```
1   PROGRM Logfile: ga_ors.log                    Jan 17,2003   18:56:40


    File "ga.in" has been loaded
5   10559 samples, 208 attributes each
    10559 samples in training set
    4-class classification problem

    Attribute's structure (1 continuous regions)
10
    1 1 - 208 mean value

    Genetic algorithm for feature extraction

15  Parameters are:

    Total population size - 100
    Size of population kept intact - 10
    Number of generations - 100
20  Probability of crossover - 0.660000
    Probability of point mutation - 0.001000
    Initial size of point mutation - 4
    Minimal size of point mutation - 6
    Maximal number of regions to look for - 8
25
    Command to optimize: disc l 0 0


              Fitness            Q1          Accuracy
30  Gen N   Train   Test    Train   Test    Train   Test

        0  0.4148  1.0000  0.4713  0.0000  0.7049  0.0000
        1  0.4017  1.0000  0.4998  0.0000  0.7270  0.0000


 ⋮


       99  0.2046  1.0000  0.7688  0.0000  0.8707  0.0000
      100  0.4537  1.0000  0.7693  0.0000  0.8700  0.0000

              Fitness            Q1          Accuracy
135 Member  Train   Test    Train   Test    Train   Test


        1  0.4537  1.0000  0.7693  0.0000  0.8700  0.0000
    8 regions
140
    4-9 mean value
    19-21 mean value
    23-28 mean value
```

```
⋮

Profile data are saved in ga.in.profile

Results sorted by Q1 for test set:

           Fitness            Q1          Accuracy
Member  Train   Test    Train   Test    Train   Test

    1  0.6536  1.0000  0.7645  0.0000  0.8675  0.0000
8 regions

4-9 mean value
16-21 mean value

⋮

Profile data are saved in ga.in.profile
Saving set file: ga.8.out
```

Zunächst werden die Kenngrößen des Datensatzes angegeben (Zeile 4 – 11), danach folgen die Parameter des genetischen Algorithmus (Z. 15 – 26). In Zeile 26 ist angegeben, welches Zielfunktional verwendet wird, „`disc 1`“ steht für die lineare Diskriminanzanalyse.

Die Zeilen 29 – 132 zeigen den Status der Optimierung nach der angegebenen Generationszahl. Danach folgt die Ausgabe der Ergebnisse, das sind die ermittelten Regionen für die besten Individuen der letzten Population. Die erreichte Fitness und Reklassifikations-Trefferrate (Spalte 2 bzw. 6) werden von der Anzahl der Regionen (Z. 139) gefolgt. Daran schließt sich die Liste dieser Regionen an (Z. 141 – 148), weitere Individuen folgen. Die Zahlen sind die Spaltenindizes der Datenmatrix in der Eingabedatei für Start- und Endwert.

Eine weitere, nach einem anderen Kriterium sortierte, Liste erscheint im Anschluss, und am Dateiende ist aufgeführt, in welcher Datei die umgewandelten Daten gespeichert wurden (Z. 387).

B.1.3. Das Format der Ergebnisse von `stackedGen`

Eine weitere Datei enthält die Ergebnisse der Zuordnungen durch `stackedGen` .

```
Pattern file is 'ga_3.8.out'
No configuration file supplied!  Using single LDA classifier.
Leave-one-out method used for training.

*******************************************************************************
Classifier 1 (LDA):
*******************************************************************************

CLASSIFIER 1 FULL TRAINING DATA CLASSIFICATION TABLE:
-------------------------------------------------------------------------------
Desired          Assigned Class
 Class      1    2    3    4  %Correct  of   SP(%)  PPV(%) NPV(%) Lift

   1     1984    0  203   14    90.1  2201    98.0   94.2   96.5  3.58
   2        0 1756    4    0    99.8  1760    99.2   96.9   99.9  4.61
   3       26   28 1946  198    88.5  2198    89.2   74.5   95.6  2.84
   4       97   28  458 1637    73.7  2220    96.6   88.5   91.1  3.34

Totals   2107 1812 2611 1849          8379
```

```
Overall accuracy on training data =  87.4%

Agreement measure:  95% confidence interval
    = 0.831582 { almost perfect } +/- 0.009516
    = ( 0.822066 { almost perfect }, 0.841097 { almost perfect } ).

CLASSIFIER 1 TRAINING DATA NON-FUZZY CLASSIFICATION TABLE:
-------------------------------------------------------------------------------
Desired          Assigned Class
 Class      1    2    3    4  %Correct  of   SP(%)  PPV(%) NPV(%) Lift %Crisp

   1     1883    0  128    6    93.4  2017    98.2   95.0   97.6  3.58   91.6
   2        0 1744    0    0   100.0  1744    99.4   98.0  100.0  4.28   99.1
   3       12   16 1776  124    92.1  1928    91.8   79.2   97.2  3.13   87.7
   4       88   20  338 1479    76.8  1925    97.7   91.9   92.6  3.64   86.7

Totals   1983 1780 2242 1609          7614

Overall accuracy on training data (excl. fuzzy classifications) =  90.4%

Agreement measure:  95% confidence interval
    = 0.871725 { almost perfect } +/- 0.008826
    = ( 0.862898 { almost perfect }, 0.880551 { almost perfect } ).

Percentage of classifications non-fuzzy =  90.9%

Q1 =  83.1%,  Q2 =  90.3%

CLASSIFIER 1 PATTERN-BY-PATTERN PERFORMANCE ON TRAINING DATA:
-------------------------------------------------------------------------------
                  Assigned      Class Memberships
      Pattern       Class          1       2       3       4
-------------------------------------------------------------------------------
          0001#0001   1          1.000   0.000   0.000   0.000
          0001#0002   1          1.000   0.000   0.000   0.000

    ⋮

          0002#0071   3 * ~ ( 1) 0.479   0.004   0.497   0.021

    ⋮

CLASSIFIER 1 FULL TEST DATA CLASSIFICATION TABLE:
-------------------------------------------------------------------------------
Desired          Assigned Class
 Class      1    2    3    4  %Correct  of   SP(%)  PPV(%) NPV(%) Lift

   1        0    0    0    0     N/A     0    N/A    N/A     0.0   N/A
   2        0    0    0    0     N/A     0    N/A    N/A     0.0   N/A
   3        0    0    0    0     N/A     0    N/A    N/A     0.0   N/A
   4        0    0    0    0     N/A     0    N/A    N/A     0.0   N/A

Totals      0    0    0    0             0

Overall accuracy on test data =   N/A

Agreement measure:  95% confidence interval
    = N/A +/- N/A
    = ( N/A, N/A ).

CLASSIFIER 1 TEST DATA NON-FUZZY CLASSIFICATION TABLE:
-------------------------------------------------------------------------------
Desired          Assigned Class
 Class      1    2    3    4  %Correct  of   SP(%)  PPV(%) NPV(%) Lift %Crisp
```

```
        1       0    0    0    0     N/A     0    N/A    N/A     0.0   N/A    0.0
        2       0    0    0    0     N/A     0    N/A    N/A     0.0   N/A    0.0
        3       0    0    0    0     N/A     0    N/A    N/A     0.0   N/A    0.0
8460    4       0    0    0    0     N/A     0    N/A    N/A     0.0   N/A    0.0

      Totals    0    0    0    0             0

      Overall accuracy on test data (excl. fuzzy classifications) =   N/A
8465
      Agreement measure:  95% confidence interval
          = N/A +/- N/A
          = ( N/A, N/A ).

8470  CLASSIFIER 1 PATTERN-BY-PATTERN PERFORMANCE ON TEST DATA:
      --------------------------------------------------------------------------------
                         Assigned      Class Memberships
            Pattern       Class          1       2       3       4
      --------------------------------------------------------------------------------
8475            0006#0001   1 ?        1.000   0.000   0.000   0.000
                0006#0002   1 ?        1.000   0.000   0.000   0.000
```

⋮

Die Datei beginnt mit der Angabe der verwendeten Daten und Methode, es folgen die Zuordnungsmatrizen aller und nur der sicheren Zuordnungen für die Trainingsdaten (Z. 9 – 45). Daran schließt sich die Zuordnungstabelle der Trainingsdaten an.

Die einzelnen Zeilen beginnen mit dem Namen des Spektrums, wie er in der Eingabedatei festgelegt ist, gefolgt von der bestimmten Klasse. Falsche Zuordnungen sind mit einem Stern und „unsichere“ Zuordnungen mit einer Tilde gekennzeichnet. Bei falschen Zuordnungen ist die wahre Klasse in Klammern angegeben (vgl. Z. 271). Die weiteren Spalten geben die Wahrscheinlichkeit der Zugehörigkeit zu den einzelnen Klassen an.

Diese Angaben wiederholen sich für die Testdaten (ab Z. 8434) mit dem Unterschied, dass die Zuordnungsmatrizen leer sind und die Spektren durch Fragezeichen als Testdaten gekennzeichnet werden.

B.2. Aufruf der Programme

Die Optimierung wird durch `ga_ors` vorgenommen. Das Programm kann dabei wahlweise mit einer Autopilot-Datei gestartet oder über eine Kommandozeile bedient werden. Im Folgenden wird ein beispielhaftes Script zur Erzeugung einer solchen Autopilot-Datei, der Durchführung der Optimierungsrechnung sowie anschließend der Durchführung der linearen Diskriminanzanalyse durch `stackedGen` vorgestellt.

```
1   #!/bin/ksh
     echo "Verzeichnis: " `pwd`

     chmod -u+rw *.in
5    fnames=`cat inh`
     for fil  in $fnames ;
     do
        echo "Datei: $fil" ;
        #echo "auto ----------------------------------"
10      echo "load $fil.in" > auto ;
        echo "GeneticAlg $1 100 10 100 .001 .66 disc l 0 0" >> auto ;
        echo "save" >> auto ;
        echo "data" >> auto;
        echo "$fil.$1.out" >> auto;
15      echo "quit" >> auto;
        #echo "---------------------------------------"
```

```
        time ~nikulin/bin/ga_ors -a auto ;
        cp ga_ors.log $fil.$1.log
        time ~dolenko/bin/stackedGen -p $fil.$1.out > -f ../stack.cfg > $fil.$1.erg ;
20  done
    #rm ga_in
    cd ~
```

In Zeile 5 wird eine Hilfs-Datei eingelesen, die die Namen (ohne Endung) der zu verarbeitenden Eingabedateien enthält. Die Schleife (Z. 6 – 20) arbeitet alle diese Dateien mit identischen Parametern ab.

In den Zeilen 10 – 16 wird die Autopilot-Datei erzeugt. Sie enthält genau die Befehle, die auch beim Kommandozeilen-gesteuerten Programmablauf eingegeben würden.

`GeneticAlg` (Z. 11) startet die eigentliche Optimierung. Der erste Parameter gibt die Maximalzahl zu suchender Regionen an. Es folgt die Anzahl an Generationen, die gerechnet werden soll. Die nächsten beiden Werte sind die Größe der Elitegruppe und die Populationsgröße. Schließlich folgen die Mutations- und die Crossover-Wahrscheinlichkeit. `disc l` wählt die lineare Diskriminanzanalyse als Fitness-Funktion. Die Bedeutung der beiden auf null gesetzten Parameter am Ende der Zeile ist nicht bekannt.

In Zeile 17 wird die Optimierung durchgeführt, dann wird die Protokoll-Datei kopiert, um ein Überschreiben beim nächsten Schleifendurchlauf zu vermeiden. Schließlich wird die lineare Diskriminanzanalyse mittels `stackedGen` ausgeführt (Z. 19).

C. Details zu den durchgeführten Rechnungen zur Datenreduktion, Basislinienkorrektur, Intensitätsnormierung und Filterung

Die Zusammensetzug der verwendeten Datensätze ist in den Tabellen C.1 und C.2 angegeben.

Für den Datensatz „a“ wurde eine Mindestanzahl von 5 Spektren pro Messung gefordert, damit enthält er 2550 Spektren jeder Klasse und umfasst insgesamt 10200 Spektren.

Datensatz „l“ besteht aus 10000 Spektren, 2500 in jeder Klasse. Das sind mindestens 50 Spektren je Messung. Dieser Datensatz ist ein Teil des Datensatzes „a“, dessen Proben histologisch als besonders typisch für die einzelnen Tumorklassen gewertet wurden.

Die Daten aus „a“, die *nicht* auch zu „l“ gehören, können als Testdaten zum Trainingsdatensatz „l“ verwendet werden.

Tabelle C.1.: Zusammensetzung Datensatz „a“

Diagnose	**Anzahl Proben**	**Anzahl Messungen**
gesund	6	14
Astro II	4	6
Astro III	7	14
Glio	51	79
gesamt	68	113

Tabelle C.2.: Zusammensetzung Datensatz „l“

Diagnose	**Anzahl Proben**	**Anzahl Messungen**
gesund	6	14
Astro II	3	4
Astro III	5	10
Glio	25	30
gesamt	39	58

Die Tabellen C.3 und C.4 geben die Ergebnisse in Abhängigkeit von der erfolgten Datenvorbehandlung wider.

Tabelle C.3.: Trefferraten in Abhängigkeit der Datenvorbehandlung — Datensatz „a“

Rechnung Nr.	Daten-reduktion	Filter	Intensitäts-normierung	Basislinien-korrektur	Reklass.-Trefferrate
24	mitteln	Histogramm	Min. / Max.	linear	83.8%
18	mitteln	Histogramm	Fläche	linear	83.7%
22	mitteln	Mittelwert	Min. / Max.	linear	82.0%
11	löschen	Histogramm	Min. / Max.		81.6%
21	löschen	Mittelwert	Min. / Max.	linear	81.4%
17	löschen	Histogramm	Fläche	linear	81.2%
6	mitteln	Histogramm	Fläche		81.1%
9	löschen	Mittelwert	Min. / Max.		81.1%
16	mitteln	Mittelwert	Fläche	linear	80.8%
12	mitteln	Histogramm	Min. / Max.		80.4%
2	mitteln		Fläche		80.3%
5	löschen	Histogramm	Fläche		80.1%
10	mitteln	Mittelwert	Min. / Max.		80.0%
23	löschen	Histogramm	Min. / Max.	linear	79.9%
3	löschen	Mittelwert	Fläche		79.6%
15	löschen	Mittelwert	Fläche	linear	79.4%
13	löschen		Fläche	linear	79.0%
7	löschen		Min. / Max.		78.8%
20	mitteln		Min. / Max.	linear	78.5%
4	mitteln	Mittelwert	Fläche		77.7%
19	löschen		Min. / Max.	linear	77.3%
1	löschen		Fläche	keine	77.2%
8	mitteln		Min. / Max.		76.6%
14	mitteln		Fläche	linear	76.1%

Tabelle C.4.: Trefferraten in Abhängigkeit der Datenvorbehandlung — Datensatz „l“

Rechnung Nr.	Daten-reduktion	Filter	Intensitäts-normierung	Basislinien-korrektur	Reklass.-Trefferrate
35	löschen	Histogramm	Min. / Max.		88.2%
47	löschen	Histogramm	Min. / Max.	linear	87.8%
27	löschen	Mittelwert	Fläche		87.7%
48	mitteln	Histogramm	Min. / Max.	linear	87.5%
36	mitteln	Histogramm	Min. / Max.		87.4%
42	mitteln	Histogramm	Fläche	linear	87.4%
45	löschen	abw. MW	Min. / Max.	linear	86.9%
34	mitteln	Mittelwert	Min. / Max.		86.8%
40	mitteln	Mittelwert	Fläche	linear	86.8%
46	mitteln	abw. MW	Min. / Max.	linear	86.8%
28	mitteln	Mittelwert	Fläche		86.1%
41	löschen	Histogramm	Fläche	linear	86.0%
26	mitteln		Fläche		85.9%
33	löschen	Mittelwert	Min. / Max.		85.9%
29	löschen	Histogramm	Fläche		85.7%
30	mitteln	Histogramm	Fläche		85.6%
39	löschen	Mittelwert	Fläche	linear	85.2%
38	mitteln		Fläche	linear	82.7%
44	mitteln		Min. / Max.	linear	82.6%
25	löschen		Fläche		81.7%
37	löschen		Fläche	linear	81.2%
43	löschen		Min. / Max.	linear	80.6%
32	mitteln		Min. / Max.		79.9%
31	löschen		Min. / Max.		76.9%

D. Glossar

benigner Tumor: gutartiger Tumor [65]

determiniert: Ein Algorithmus heißt determiniert, wenn er jede Eingabe auf genau eine Ausgabe abbildet.

Nichtdeterminierte Algorithmen können bei gleicher Eingabe und Startbedingungen unterschiedliche, möglicherweise auch falsche, Ergebnisse liefern. Man nutzt nichtdeterminierte Algorithmen, wenn sie mit wesentlich geringeren Ressourcen eine Problemlösung liefern und *nicht zu oft* falsche Ergebnisse auftreten. Beispiele sind die *stochastischen Algorithmen*.

Auch bestimmte Fragestellungen, die auf eine spezielle Eigenschaft einer Lösung und nicht auf die Lösung selbst zielen, können mit nichtdeterminierten Algorithmen bearbeitet werden, z. B. die Frage *ob* es eine Lösung gibt (wenn egal ist, *welche*). In diesem Beispiel ist allerdings der gesamte Algorithmus (der die Frage beantwortet, ob es eine Lösung gibt) determiniert, wenn auch nicht *deterministisch*.

[42]

deterministisch: *deterministisch* Ein Algorithmus heißt deterministisch, wenn zu jedem Zeitpunkt der Folgeschritt eindeutig bestimmt ist. Deterministische Algorithmen sind immer auch *determiniert*. In der Regel arbeiten Rechner und auch die verwendeten Programmiersprachen deterministisch. Man kann also streng genommen *nichtdeterministische* Algorithmen nicht implementieren, jedoch nähert man sich diesem Verhalten z. B. durch die Verwendung von Pseudo-Zufallszahlen an. Grund für diese Bemühungen ist, dass eine Reihe von Problemen mit nichtdeterministischen Algorithmen knapper und klarer als mit deterministischen Algorithmen zu lösen sind.

vgl. *NP-Vollständigkeit*

[42]

disjunkt: svw. getrennt, unvereinbar, sich gegenseitig ausschließend. Zwei disjunkte Ereignisse können nicht gleichzeitig eintreten. Z. B. kann ein Objekt nicht gleichzeitig mehreren Klassen angehören.[80]

Effizienz: Ein Algorithmus heißt effizient, wenn er mit möglichst geringem Rechenaufwand (Laufzeit, Speicherbedarf, ...) ein gegebenes Problem löst.

Ein Algorithmus mit *polynomialer Laufzeit* gilt als effektiv, ein Algorithmus mit exponentieller Laufzeit nicht [42; 95]. Es existiert eine Reihe von Problemen, für die keine effizienten Algorithmen bekannt sind, insbesondere ist bei den NP-vollständigen Problemen unbekannt, ob sie überhaupt effizient bearbeitbar sind.

Endothel: Gewebe, das die Blut- und Lymphgefäße sowie die Herzräume auskleidet.[65]

Extinktion: Die Extinktion E ist definiert als

$$E := -\log T = -\log\left(\frac{I}{I_0}\right),$$

wobei T die *Transmission* ist.

Siehe auch *Lambert-Beersches Gesetz*

Grundgesamtheit: Die Menge *aller* Ojekte, die zur betreffenden Fragestellung untersucht werden könnte.
Bsp: für eine Wahlprognose ist die Grundgesamtheit die Menge aller Wahlberechtigten.
Bei analytischen Fragestellungen ist die Grundgesamtheit oft sogar unendlich groß. In aller Regel können jedoch bereits endliche Grundgesamtheiten nicht untersucht werden, daher wird mit *Stichproben* gearbeitet. [96][74]

Hamming-Distanz: Maß für die Distanz zwischen zwei Bit-Strings. Die HAMMING-Distanz ist die Summe aller unterschiedlichen Bits. [43]

Hämatoxylin-Eosin-Färbung: gebräuchliche histologische Färbemethode, Zellkerne und Knorpel werden blau, die restlichen Zellbestandteile rot gefärbt.[66]

Heuristik: *(Informatik)* Strategie zur Problemlösung, die auf plausiblen Annahmen, Vermutungen und Erfahrungen beruht. Die Leistungsfähigkeit solcher Strategien ist nicht beweisbar, aber durch Experimente an typischen Problemen abschätzbar. Viele Heuristiken liefern für sehr komplexe Probleme im Allgemeinen schnell gute Ergebnisse.

Die Nachbildung menschlicher Vorgehensweisen zur Problemlösung, das schrittweise Vorantasten an die Lösung und die Anwendung von Faustregeln sind typische Grundmuster für Heuristiken. Ein wichtiges Anwendungsgebiet der Heuristiken sind die NP-vollständigen Probleme, typische Implementationen sind evolutionäre Algorithmen und viele Verfahren zum Durchmustern von Entscheidungsbäumen.

[42]

Hyperplasie: Größenzunahme eines Organs durch Vermehrung der Zellen und Gewebebestandteile [66]

Exzess: auch *Überhöhung* oder *Kurtosis* einer Verteilung.

$$\hat{\epsilon} := \frac{\sum n_i (x_i - \bar{x})^4}{n\hat{s}^4} - 3 \tag{D.1}$$

mit
- $n_i \cdots$ Zahl der Messwerte in der i-ten (Histogramm-)Klasse
- $n \cdots$ Zahl aller Messwerte
- $\bar{x} \cdots$ Mittelwert der Messwerte
- $\hat{s} \cdots$ Standardabweichung

Für eine Normalverteilung erhält man eine Kurtosis von 0, ist das Maximum der vorliegenden Verteilung höher als das der Normalverteilung, so wird der Exzess positiv. [97]

Lambert-Beersches Gesetz: Für verdünnte Lösungen gilt:

$$E = \epsilon \cdot c \cdot d, \text{bzw. für Gemische}$$
$$E = \sum_{\forall i} \epsilon_i \cdot c_i \cdot d,$$

dabei ist E die Extinktion, $\epsilon_i = \epsilon_i(\lambda)$ der molare dekadische Extinktionskoeffizient der Substanz i, c_i deren Konzentration und d die durchstrahlte Schichtdicke.

Laufzeit: Die Laufzeit eines Algorithmus ist die Anzahl der durchgeführten Rechenschritte. Die Laufzeit ist ein wichtiger Teil der Betrachtungen zur Komplexität eines Algorithmus, man gibt die *worst-case*(maximale) und die *average-case* (mittlere) Laufzeit an.
In der Regel gibt man die *Ordnung* bezüglich der Eingabe an: $\mathrm{t}(n) = O(n)$.
[42]

Mahalanobis-Distanz: ein Distanzmaß, das die Kovarianz-Struktur der Daten berücksichtigt. Dadurch wird Unabhängigkeit gegenüber linearen Transformationen erreicht.

$$d_M(\mathbf{x}_a, \mathbf{x}_b) = \sqrt{(\mathbf{x}_a - \mathbf{x}_b)^T \mathbf{S}^{-1} (\mathbf{x}_a - \mathbf{x}_b)}$$

Die Mahalanobis-Distanz steht in engem Zusammenhang mit der Datenbeschreibung bei der linearen Diskriminanzanalyse (Kap. 6.2 (S. 18)). [74][72][71]

maligner Tumor: bösartiger Tumor [65]

multivariate Normalverteilung: Eine Zufallsvariable $\mathbf{Y} \in \mathbb{R}^m$ heißt *multivariat normalverteilt*, wenn alle Linearkombinationen $\mathbf{Z} = \mathbf{a}^T\mathbf{Y}$ mit $\mathbf{Z} \in \mathbb{R}$ normalverteilt sind.
Man schreibt:

$$\mathbf{Y} \sim N_m(\boldsymbol{\mu}; \mathbf{S}) \tag{D.2}$$

Dann exisistieren auch

$$\begin{aligned} \mathrm{E}(\mathbf{Y}) &= \boldsymbol{\mu} \text{ und} \\ \mathbf{S} &= \mathbf{COV}(\mathbf{Y}) \end{aligned} \tag{D.3}$$

Ist $\mathbf{S}$ *positiv definit*, das heißt $\operatorname{rang}(\mathbf{S}) = m$, so exisitiert die Dichtefunktion

$$\mathrm{f}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^p}\sqrt{|\mathbf{S}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \mathbf{S}^{-1}(\mathbf{x}-\boldsymbol{\mu})}. \tag{D.4}$$

Ist $\mathbf{S}$ *positiv semidefinit*, also $\operatorname{rang}(\mathbf{S}) = p < m$, so spricht man von einer *singulären* oder *degenerierten* Normalverteilung.
Gilt $\boldsymbol{\mu} = \mathbf{0}$ und $\mathbf{S} = \mathbf{I}_p$, so liegt eine *multivariate Standardnormalverteilung* vor.

$$\mathbf{U} \sim N_p(\mathbf{0}; \mathbf{I}_p) \tag{D.5}$$

Eine wichtige Eigenschaft der multivariaten Normalverteilung ist, dass alle linearen Transformationen wieder zu multivariaten Normalverteilungen führen. Damit ist es möglich, eine beliebige multivariat normalverteilte Variable $\mathbf{Y} \sim N_m(\boldsymbol{\mu}; \mathbf{S})$ in eine multivariat standardnormalverteile Variable $\mathbf{U} \sim N_p(\mathbf{0}; \mathbf{I})$ zu überführen.

Auch die Transformation einer multivariaten Standardnormalverteilung in eine beliebige multivariate Normalverteilung ist möglich.

[72]

Nekrose: örtlich begrenztes Absterben von Zellen, Geweben oder Organen während des Lebens des Organismus [66]

Neoplasie: Neubildung (von Gewebe)[66]

NP-vollständig: Die Menge NP umfasst alle Probleme, die mit *nichtdeterministischen* Algorithmen in *polynomialer Laufzeit* gelöst werden können, die Menge P alle Probleme, die mit *deterministischen* Algorithmen in *polynomialer Laufzeit* lösbar sind. Es gilt $P \subseteq NP$.

Eine wichtige Frage der Informatik ist, ob P eine echte Untermenge von NP ist oder $P = NP$ gilt. Die *NP-vollständigen* Probleme zeichnen sich dadurch aus, dass, wenn ein für *ein NP-vollständiges* Problem ein Lösungsweg in P gefunden wird, gleichzeitig gezeigt ist, dass $NP = P$ gilt.

Dies ist bislang nicht gelungen. Für die NP-vollständigen Probleme sind also keine *effizienten* Lösungswege bekannt. [42; 95]

overfitting (Übermodellierung): Mit wachsender Komplexität kann ein Modell immer besser an einen gegebenen Datensatz angepaßt werden. Allerdings soll meist keine perfekte Abbildung der *Trainingsdaten* erfolgen, sondern nur der *verallgemeinerbare Anteil* an Informationen dieser Daten soll in das Modell einfließen.

Ist das Modell nicht komplex genug, so werden weniger Informationen aus den Daten zur Modellbildung genutzt als nutzbar sind, die Ergebnisse sowohl bei der Anwendung auf die Trainingsdaten als auch bei Anwendung auf *Testdaten* erreichen nicht die mögliche Genauigkeit.

Wird ein Modell über diesen Punkt hinaus an die vorhandenen Daten angepasst, so spricht man von Übermodellierung oder overfitting. Dann sind die Ergebnisse für die Trainingsdaten sehr gut, aber für Testdaten werden sie mit steigender Komplexität des Modells immer schlechter.

Phagozytose: Aktive Aufnahme an der Zelloberfläche angelagerter Teilchen in die Zelle. Wichtiger Abwehrmechanismus. [66]

positiv definit: Eine symmetrische (genauer: hermitesche) Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ heißt *positiv definit*, wenn für alle $\mathbf{x} \in \mathbb{R}^n \backslash \{0\}$ gilt:

$$\mathbf{x}^T \mathbf{A} \mathbf{x} > 0$$

Weitere notwendige und hinreichende Bedinungen für das Vorliegen einer positiv definiten Matrix sind:

- alle Eigenwerte sind positiv (größer 0)
- es existiert eine Matrix $\mathbf{W}$, so dass $\mathbf{W}^T\mathbf{W} = \mathbf{A}$
- alle linken, oberen Untermatrizen haben positive Eigenwerte

Positiv definite Matrizen sind *nicht-singulär* und daher *invertierbar*.

[98; 99]

positiv semidefinit: Eine symmetrische Matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ heißt *positiv semidefinit*, wenn für alle $\mathbf{x} \in \mathbb{R}^n \backslash \{0\}$ gilt:

$$\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$$

Eine äquivalente Forderung ist, dass die Eigenwerte nicht negativ sein dürfen.
siehe auch *positiv definit*
[98; 99]

Proliferation: Wucherung, erhebliche Zell- und Kapillarvermehrung [66]

Schiefe:

$$\hat{\varrho} := \frac{\sum n_i (x_i - \bar{x})^3}{n\hat{s}^3} \tag{D.6}$$

mit $n_i \cdots$ Zahl der Messwerte in der i-ten (Histogramm-)Klasse
$n \cdots$ Zahl aller Messwerte
$\bar{x} \cdots$ Mittelwert der Messwerte
$\hat{s} \cdots$ Standardabweichung

Für symmetrische Verteilungen ist die Schiefe 0, für rechtsseitig-asymmetrische Vrteilungen negativ. [97]

stochastischer Algorithmus: Algorithmus, dessen Ausgabe und / oder Reihenfolge der Bearbeitung der Anweisungen von *zufälligen* Ereignissen abhängt. Die Wahrscheinlichkeit, dass eine bestimmte Entscheidung getroffen wird, kann gegeben sein. Stochastische Algorithmen sind im Allgemeinen weder deterministisch noch determiniert, sie können auch falsche Lösungen ausgeben, jedoch ist das Verhalten der Algorithmen bzgl. solcher Fehler sehr unterschiedlich.

Die Implementation solcher Algorithmen benötigt Zufallszahlen, in der Regel nutzt man aber Pseudo-Zufallszahlen, die im Rechner erzeugt werden können. Diese sind meist gleichverteilt, können aber in beliebig verteilte Pseudo-Zufallszahlen überführt werden.

[42; 100]

Testdaten: Daten, auf die ein gebildetes Modell angewandt wird, die jedoch an der Modellerstellung völlig unbeteiligt waren. Siehe auch *Trainingsdaten*.

Trainingsdaten: Daten, die zur Erstellung eines Modells benutzt werden. Siehe auch *Testdaten*.

Transmission: Die Transmission T ist definiert als

$$T := \frac{I}{I_0},$$

wobei I die Intensität der durch die Probe hindurchgelangenden Strahlung und I_0 die Intensität der eingestrahlten elektromagnetischen Welle ist. Die Transmission gibt den Anteil der Strahlung an, der die Probe durchdringt.

Wellenzahl: Die Wellenzahl $\tilde{\nu}$ ist als der Kehrwert der Wellenlänge definiert und damit der Energie der Welle proportional.

$$\tilde{\nu} := \frac{1}{\lambda} = \frac{\nu}{c} = \frac{E}{hc}$$

In der IR-Spektroskopie wird meist die Einheit $\frac{1}{\mathrm{cm}} = \mathrm{cm}^{-1}$ verwendet.

zentraler Grenzwertsatz: Sei X_k eine Folge unabhängiger Zufallsgrößen, charakterisiert durch die Erwartungswerte $\mathrm{E}(X_k) = \mu_k$ und Varianzen σ_k^2 und Z_n die daraus wie folgt gebildete neue Zufallsgröße

$$Z_n = \frac{1}{b_n}\left(\sum_{k=1}^{n} X_k - \sum_{k=1}^{n} \mu_k\right) \qquad \text{mit} \qquad b_n = \sqrt{\sum_{k=1}^{n} \sigma_k^2}$$

mit der Verteilungsfunktion $\mathrm{F}_{\mathrm{Z_n}}$.

Ist weiterhin die Lindenberg sche Bedingung

$$\lim_{n\to\infty} \frac{1}{b_n^2} \sum_{k=1}^{n} \int\limits_{|x-\mu_k|>\epsilon b_n} (x-\mu_k)^2 d\mathrm{F}_{\mathrm{X_k}}(x) = 0$$

für beliebige $\epsilon > 0$ erfüllt, so gilt:

$$\lim_{n\to\infty} F_{Z_n}(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int\limits_{-\infty}^{x} \mathrm{e}^{-\frac{1}{2}t^2} dt \qquad \text{und} \qquad \lim_{n\to\infty} b_n \to \infty$$

die neue Zufallsvariable ist also für $n \to \infty$ normalverteilt mit Erwartungswert null und Standardabweichung 1. Die Lindenberg sche Bedingung ist dabei *hinreichend* und *notwendig*.

Sie bedeutet im Wesentlichen, dass die Verteilung einer neuen Zufallsvariablen, die die Summe vieler Zufallsvariablen ist, gegen die Normalverteilung konvergiert, wenn die einzelnen Summanden jeweils nur einen kleinen Anteil an der Summe haben, oder, noch anders gesagt, keine der einzelnen Zufallsvariablen die neue Variable dominiert.

[73; 80; 101]

Literaturverzeichnis

Die Wissenschaft, sie ist und bleibt,
was einer ab vom andren schreibt.

(Eugen Roth)

[1] Mantsch, Henry ; Jackson, Michael: Molecular Spectroscopy in Biodiagnostics (From hippocrates to Herschel and beyond). In: *J Mol Struct* 347 (1995), S. 187 – 206

[2] Thiele, Dr. S.: *Chemometrie*. WS 2001/02. – Vorlesung

[3] Yano, Kazuyuki ; Ohoshima, Susumu ; Gotou, Yoshiya ; Kumaido, Kuniyoshi ; Moriguchi, Takeshi ; Katayama, Hiroo: Direct Measurement of Human Lung Cancerous and Noncancerous Tissues by Fourier Transform Infrared Microscopy: Can an Infrared Microscope Be Used as a Clinical Tool? In: *Analytical Biochemistry* 287 (2000), Nr. 2, S. 218 – 225

[4] Benedeti, Enzo ; Teodori, Laura ; Trinca, Maria L. ; Vergamini, Piergiorgio ; Salvati, Franco ; Mauro, Francesco ; Spremolla, Giuliano: A New Approach to the Study of Human Solid Tumor Cells by Means of FT-IR Microspectroscopy. In: *Applied Spectroscopy* 44 (1990), Nr. 8, S. 1276 – 1280

[5] Wong, Patrick T. T. ; Senterman, Mary. K. ; Jackli, Pascale ; Wong, Rita K. ; Salib, Sylvia ; Campbell, Craig E. ; Feigel, Roman ; Faught, Wylam ; Fung Kee Fung, Micheal: Detailed Account of Confounding Factors in Interpretation of FTIR Spectra of Exfoliated Cervical Cells. In: *Biopolymers — Biospectroscopy Section* 67 (2002), Nr. 6, S. 376 – 386

[6] A, Cohenford M. ; B, Rigas: Cytologically normal cells from neoplastic cervical samples display extensive structural abnormalities on IR spectroscopy: implications for tumor biology. In: *Proc Natl Acad Sci U S A* 95 (1998), Nr. 26, S. 15327 – 15332

[7] Romeo, Melissa J. ; Quinn, Michael A. ; Burden, Frank R. ; McNaughton, Don: Influence of Benign Cellular Changes in Diagnosis of Cervical Cancer Using IR Microspectroscopy. In: *Biopolymers — Biospectroscopy Section* 67 (2002), Nr. 4 + 5, S. 362 – 366

[8] Wong, Patrick T. T. ; Wong, Rita K. ; Fung, Michael Fung K.: Pressure-Tuning FT-IR Study of Human Cervical Cells. In: *Applied Spectroscopy* 47 (1993), S. 1058 – 1063

[9] Ge, Zhengfang ; Brown, Chris W. ; Kisner, Harold J.: Screening Pap Smears with Near-Infrared Spectroscopy. In: *Applied Spectroscopy* 49 (1995), Nr. 4, S. 432 – 436

[10] Chiriboga, Luis ; Xie, Ping ; Yee, Herman ; Zarou, Donald ; Zakim, DAvid ; Diem, Max: Infrared spectroscopy of human tissue. IV. Detection of dysplastic and

neoplastic changes of human cervical tissue via infrared microscopy. In: *Cellular and Molecular Biology* 44 (1998), Nr. 1, S. 219 – 229

[11] Lowry, Stephen R.: The analysis of exfoliated cervical cells by infrared microscopy. In: *Cellular and Molecular Biology* 44 (1998), Nr. 1, S. 169 – 177

[12] Romeo, Melissa ; Burden, Frank ; Quinn, Micheal ; Wood, Bayden ; McNaughton, Don: Infrared microspectroscopy and artificial neural networks in the diagnosis of cervical cancer. In: *Cellular and Molecular Biology* 44 (1998), S. 179 – 187

[13] Wong, Patrick T. T. ; Wong, Rita K. ; Caputo, Thomas A. ; Godwin, Thomas A. ; Rigas, Basil: Infrared spectroscopy of exfoliated human cervical cells: Evidence of extensive structural changes during carcinogenesis. In: *Proc Natl Acad Sci USA* 88 (1991), S. 10988 – 10992

[14] Nikulin, Alexander E. ; Dolenko, Brion ; Bezabeh, Tedros ; Somorjai, Ray L.: Near-optimal region selection for feature space reduction: novel preprocessing methods for classifying MR spectra. In: *NMR in Biomedicine* 11 (1998), S. 209 – 216

[15] Wong, P. T. T. ; Papavassiliou, E. D. ; Rigas, B.: Phosphodiester Stretching Bands in the Infrared Spectra of Human Tissues and Cultured Cells. In: *Applied Spectroscopy* 45 (1991), Nr. 9, S. 1563 – 1567

[16] Wong, Patrick T. T. ; Lacelle, Suzanne ; Yazdi, Hossein M.: Normal and Malignant Human Colonic Tissues Investigated by Pressure-Tuning FT-IR Spectroscopy. In: *Applied Spectroscopy* 47 (1993), S. 1830 – 1836

[17] Bindig, Uwe ; Winter, Harald ; Wäsche, Wolfgang ; Zelianeos, Konstantinos ; Müller, Gerhardt: Fiber-optical and microscopic detection of malignant tissue by use of infrared spectrometry. In: *Journal of Biomedical Optics* 7 (2002), Nr. 1, S. 100 – 108

[18] Rigas, Basil ; Morgello, Susan ; Goldman, Ira S. ; Wong, Patrick T. T.: Human colorectal cancers display abnormal Fourier-transform infrared spectra. In: *Proc Natl Acad Sci USA* 87 (1990), S. 8140 – 8144

[19] Chiriboga, Luis ; Yu, Herman ; Diem, Max: Infrared Spectroscopy of Human Cells and Tissue Part IV: A Comparative Study of Histopathology and Infrared Microspectroscopy of Normal, Cirrhoic, and Cancerous Liver Tissue. In: *Applied Spectroscopy* 54 (2000), Nr. 1, S. 1 – 8

[20] Lasch, Peter ; Pacifico, Anthony ; Diem, Max: Spatially Resolved IR Microspectroscopy of Single Cells. In: *Biopolymers* 67 (2002), Nr. 4 – 5, S. 335 – 338

[21] Mansfield, James R. ; McIntosh, Laura M. ; Crowson, A. N. ; Mantsch, Henry H. ; Jackson, Michael: LDA-Guided Search Engine for the Nonsubjective Analysis of Infrared Microscopic Maps. In: *Applied Spectroscopy* 53 (1999), Nr. 11, S. 1323 – 1330

[22] Lasch, P. ; Naumann, D.: FT-IR microspectroscopic imaging of human carcinoma thin sections based on pattern recognition techniques. In: *Cellular and Molecular Biology* 44 (1998), Nr. 1, S. 189 – 202

[23] Jackson, Michael ; Mansfield, James R. ; Dolenko, Brion ; Somorjai, Rajmund L. ; Mantsch, Henry H. ; Watson, Peter H.: Classification of Breast Tumors by Grade and Steroid Receptor Status Using Pattern Recognition Analysis of Infrared Spectra. In: *Cancer Detection and Prevention* 23 (1999), Nr. 3, S. 245 – 253

[24] Ci, Yun L. ; Gao, Ti Y. ; Feng, Jun ; Guo, Zhen Q.: Fourier Transform Infrared Spectroscopic Characterization of Human Breast Tissue: Implications for Breast Cancer Diagnosis. In: *Applied Spectroscopy* 53 (1999), 3, S. 312 – 315

[25] Fabian, H. ; Lasch, P. ; Boese, M. ; Haensch, W.: Mid-IR Microspectroscopic Imaging of Breast Tumor Tissue Sections. In: *Biopolymers — Biospectroscopy Section* 67 (2002), Nr. 4 + 5, S. 354 – 357

[26] Eckel, Rainer ; Huo, Hong ; Guan, Hong-Wei ; Hu, Xiang ; Che, Xun ; Huang, Wei-Dong: Characteristic infrared spectroscopic patterns in the protein bands of human breast cancer tissue. In: *Vibrational Spectroscopy* 27 (2001), S. 165 – 173

[27] Malins, Donald C. ; Polissar, Nayak L. ; Gunselman, Sandra J.: Progression of human breast cancers to the metastatic state is linked to hydroxyl radical-induced DNA damage. In: *Proc Natl Acad Sci U S A* 93 (1996), S. 2557 – 2563

[28] Malins, Donald C. ; Polissar, Nayak L. ; Gunselman, Sandra J.: Tumor progression to the metastatic state involves structural modifications in DNA markedly different forom those associated with primary tumor formation. In: *Proc Natl Acad Sci U S A* 94 (1997), S. 259 – 264

[29] T, Gao ; J, Feng ; Y, Ci: Human breast carcinomal tissues display distinctive FTIR spectra: implication for the histological characterization of carcinomas. In: *Anal Cell Pathol* 18 (1999), Nr. 2, S. 87 – 93

[30] Dukor, Rina K. ; Liebman, Michael N. ; Johnson, Beh L.: A new, non-destructive method for analysis of clinical samples with FT-IR microspectroscopy. Breast cancer tissue as an example. In: *Cellular and Molecular Biology* 44 (1998), Nr. 1, S. 211 – 217

[31] Malins, Donald C. ; Polissar, Nayak L. ; Gunselman, Sandra J.: Models of DNA structure achieve alsmost perfect discrimination between normal prostate, benign prostatic hyperplasia (BPH) and adenocarcinoma and habe a high potential for predicting BPH and prostate cancer. In: *Proc Natl Acad Sci U S A* 94 (1997), S. 259 – 264

[32] Wu, Jin-Guang ; Xu, Yi-Zhuang ; Sun, Cuan-Wen ; Soloway, Roger D. ; Xu, Duan-Fu ; Wu, Qi-Guang ; Sun, Kai-Hua ; Weng, Shi-Fu ; Xu, Guang-Xian: Distinguishing Malignant from Normal Oral Tissues Using FTIR Fiber-Optic Techniques. In: *Biopolymers* 62 (2001), Nr. 4, S. 185 –192

[33] Schultz, Christian P. ; Mantsch, Henry H.: Biochemical imaging and 2D classification of keratin pearl structures in oral squamous cell carcinoma. In: *Cellular and Molecular Biology* 44 (1998), Nr. 1, S. 203 – 210

[34] Diem, Max ; Boydston-White, Susie ; Chiriboga, Luis: Infrared Spectroscopy of Cells and Tissues: Shining Light onto a Novel Subject. In: *Applied Spectroscopy* 53 (1999), Nr. 4, S. 148A – 161A

[35] Benedetti, Enzo ; Palatresi, Maria P. ; Vergamini, Piergiorgio ; Papineschi, Federico ; Andreucci, Maria C. ; Spremolla, Giuliano: Infrared Charakterization of Nuclei Isolated from Normal and Leucemic (B-CLL) Lymphocytes: Part III. In: *Applied Spectroscopy* 40 (1986), S. 39 – 43

[36] Steiner, G. ; Shaw, R. A. ; Choo-Smith, L.-P. ; Steller, W. ; Shapoval, L. ; Schackert, G. ; Sobottka, S. ; Salzer, R. ; Mantsch, H. H.: Detection and grading of human gliomas by FTIR spectroscopy and a genetic classification algorithm. In: *proc.SPIE-Int.Soc.Opt.Eng.* 4614 (2002), S. 127 – 133. – Biomedical Vibrational Spectroscopy

[37] Steller, Wolfram: *Differenzierung von gesundem und tumorösem Gewebe mittels optospektroskopischer Methoden*, Technische Universität Dresden, Diplomarbeit, 2000

[38] Lorenz, Anja: *Infrarot Spektroskopische Klassifizierung von Hirntumoren*, Technische Universität Dresden, Diplomarbeit, 2002

[39] Steiner, G. ; Shaw, R. A. ; ChooSmith, L.-P. ; Schackert, G. ; Steller, W. ; Abuid, H. M. ; Salzer, R. ; Mantsch, H. H.: *Distinguishing and grading human gliomas by infrared spectroscopy.* – submitted

[40] Diem, Max ; Chiriboga, Luis ; Lasch, Peter ; Pacifico, Anthony: IR Spectra and IR Spectral Maps of Individual Normal and Cancerous Cells. In: *Bioplymers — Biospectroscopy Section* 67 (2002), April, Nr. 4 + 5, S. 349 – 353

[41] Haaland, David M. ; Jones, Howland D. T. ; Thomas, Edward V.: Multivariate Classification of the Infrared Spectra of Cell and Tissue Samples. In: *Applied Spectroscopy* 51 (1997), Nr. 3, S. 340 – 345

[42] Engesser, Hermann (Hrsg.): *Duden „Informatik“: ein Sachlexikon für Studium und Praxis.* 2., vollst. überarb. und erw. Aufl. Dudenverlag, Mannheim; Leipzig; Wien; Zürich, 1993

[43] Claassen, M.: *Einführung in die Genetischen Algorithmen.* 2002 (Proseminar Evolutionäre Algorithmen, A. Zell, Wilhelm-Schickard-Institut für Informatik, Uni Tübingen). – Ausarbeitung zum Vortrag, `http://www-ra.informatik.uni-tuebingen.de/lehre/WS01/pro_ea_ausarbeitung/claassen_ws01.pdf`

[44] Schatten, Alexander: *Genetische Algorithmen - Einführung in die Angewandte Optimierung.* 1997. – `http://www.ifs.tuwien.ac.at/~aschatt/info/ga_lecture_notes/ga.pdf`

[45] Stark, D.: *Anwendungen der EA in der Parameteroptimierung.* 2002 (Proseminar Evolutionäre Algorithmen, A. Zell, Wilhelm-Schickard-Institut für Informatik, Uni Tübingen). – Ausarbeitung zum Vortrag, `http://www-ra.informatik.uni-tuebingen.de/lehre/WS01/pro_ea_ausarbeitung/stark_ws01.pdf`

[46] Ritzinger, B.: *Evolutionäre Algorithmen im Operations Research.* 2002 (Proseminar Evolutionäre Algorithmen, A. Zell, Wilhelm-Schickard-Institut für Informatik, Uni Tübingen). – Ausarbeitung zum Vortrag, `http://www-ra.informatik.uni-tuebingen.de/lehre/WS01/pro_ea_ausarbeitung/ritzinger_ws01.pdf`

[47] Leardi, Riccardo: Genetic Algorithms in Feature Selection. In: *NAmICS Newsletter* (2002), Jan, Nr. 22, S. 7 – 10

[48] Schaffer, R E. ; Small, G W.: Learning Optimization from Nature. In: *Analytical Chemistry* 69 (1997), Nr. 7, S. 236 A – 242 A

[49] Lavine, Barry K.: Chemometrics. In: *Analytical Chemistry* 70 (1998), Nr. 12, S. 209R – 228R

[50] Arcos, M. J. ; Oritz, M. C. ; Villahoz, Belén ; Sarabia, Luis A.: Genetic-algorithm-based wavelength selection in multicomponent spectrometric determinations by PLS: application on indomethacin and acemethacin mixture. In: *Analytica Chimica Acta* 339 (1997), S. 63 – 77

[51] Lucasius, C. B. ; Beckers, M. L. M. ; Kateman, G.: Genetic algorithms in wavelength selection: a comparative study. In: *Analytica Chimica Acta* 286 (1994), S. 135 – 153

[52] Hörchner, Uwe ; Kalivas, John H.: Further investigation on a comparative study of simulated annealing and genetic algorithm for wavelength selection. In: *Analytica Chimica Acta* 311 (1995), S. 1 – 13

[53] Jouan-Rimbaud, Delphine ; Massart, Désirée-Luc ; Leardi, Riccardo ; de Noord, Onno E.: Genetic Algorithms as a Tool for Wavelength Selection in Multivariate Calibration. In: *Analytical Chemistry* 67 (1995), S. 4295 – 4301

[54] Jouan-Rimbaud, D. ; Walczak, B. ; Massart, D. L. ; Last, I. R. ; Prebble, K. A.: Comparison of multivariate metods based on latent vectors and metods based on wavelength selection for the analysis of near-infrared spectroscopic data. In: *Analytica Chimica Acta* 304 (1995), S. 285 – 295

[55] Wu, W. ; Walczak, B. ; Massart, D. L. ; Prebble, K. A. ; Last, I. R.: Spectral transformation and wavelength selection in near-infrared spectra classification. In: *Analytica Chimica Acta* 315 (1995), S. 243 – 255

[56] Jouan-Rimbaud, D. ; Khots, M. S. ; Massart, D. L. ; Last, I. R. ; Prebble, K. A.: Calibration line adjustment to faciliate the use of synthetic calibration samples in near-infrared spectrometric analysis of pharmaceutical production samples. In: *Analytica Chimica Acta* 315 (1995), S. 257 – 266

[57] Leardi, R. ; Seasholtz, M. B. ; Pell, R. J.: Variable selection for multivariate calibration using a genetic algorithm: prediction of additive concentrations in polymer films from Fourier transform-infrared spectral data. In: *Analytica Chimica Acta* 461 (2002), S. 189 – 200

[58] Hayden, Charlene M. ; Morris, Michael D.: Effects of Sampling Parameters on Principal Components Analysis of Raman Line Images. In: *Applied Spectroscopy* 50 (1996), Nr. 6, S. 708 – 714

[59] Broadhurst, David ; Goodacre, Royston ; Jones, Alun ; Rowland, Jem J. ; Kell, Douglas B.: Genetic algorithms as a method for variable selection in multiple linea regression and partial least squares regression, with applications to pyrolysis mass spectrometry. In: *Analytica Chimica Acta* 348 (1997), S. 71 – 86

[60] Brenchley, Jason M. ; Hörchner, Uwe ; Kalivas, Hohn M.: Wavelength Selection Characterization for NIR Spectra. In: *Applied Spectroscopy* 51 (1997), Nr. 5, S. 689 – 699

[61] Mark, Howard L. ; Tunnell, David: Qualitative Near-Infrared Reflectance Analysis Using Mahalanobis Distance. In: *Analytical Chemistry* 57 (1985), S. 1449 – 1456

[62] Mark, H. L.: Normalized Distances for Qualitative Near-Infrared Reflectance Analysis. In: *Analytical Chemistry* 58 (1986), S. 379 – 384

[63] Mark, Howard: Use of Mahalanobis Distances To Evaluate Spamle Preparation Methods for Near-Infrared Reflectance Analysis. In: *Analytical Chemistry* 59 (1987), S. 790 – 795

[64] Mansfield, James R. ; Sowa, Michael G. ; Mazjels, Claudine ; Collins, Cathy ; Cloutis, Edward ; Mantsch, Henry H.: Near infrared spectroscopic reflectance imaging: supervise vs. unsupervised analysis using an art conservation application. In: *Vibrational Spectroscopy* 19 (1999), S. 33 – 45

[65] *Roche Lexikon Medizin*. 1999. – im Internet: `http://www.gesundheit.de/roche/`

[66] Zetkin-Schaldach ; David, H. (Hrsg.): *Wörterbuch der Medizin, Zahnheilkunde und Grenzgebiete*. 7., völlig neubearb. u. erw. Aufl. Thieme, Stuttgart; Deutscher Taschenbuch Verlag, München, 1985

[67] Riede, Ursus-Nikolaus (Hrsg.) ; Schaefer, Hans-Eckhart (Hrsg.) ; Wehner, Herbert (Hrsg.): *Allgemeine und spezielle Pathologie*. 2., neu bearb. Auflage. Thieme Verlag, Stuttgart, 1989

[68] *Medicine Worldwide: Krebs*. 2002. – `http://www.m-ww.de/krankheiten/krebs`

[69] Kap. 11. Biomedical infrared spectroscopy In: Mantsch, Henry H. (Hrsg.) ; Chapman, Dennis (Hrsg.): *Infrared Spectroscopy of Biomolecules*. Wiley-Liss, New York, 1996

[70] Gorlier, Philippe: *Infrarot Spektroskopie an Hirngewebeschnitten. Vergleich und Optimierung von Transmissions-, Reflexions- und ATR-Messungen*, Technische Universität Dresden, Diplomarbeit, 2002

[71] Fahrmeir, Ludwig (Hrsg.) ; Hamerle, Alfred (Hrsg.): *Multivariate statistische Verfahren*. de Gruyter, Berlin, 1984

[72] Böker, F.: *Multivariate Verfahren*. SS 2001. – Script zur Vorlesung

[73] Otto, M.: *Chemometrie: Statistik und Computereinsatz in der Analytik*. VCH Weinheim, 1997

[74] Danzer, K. ; Hobert, H. ; Fischbacher, C. ; Jagemann, K.-U.: *Chemometrik: Grundlagen und Anwendungen*. Springer-Verlag, Berlin, Heidelberg, New York, 2001

[75] Galactic, Thermo: *Internet-Seiten von Thermo Galactic: Algorithmen*. 2002. – `http://www.galactic.com/algorithms`

[76] van der Broek, W. H. A. M. ; Wienke, D. ; Melssen, W. J. ; Feldhoff, R. ; Huth-Fehre, T. ; Kantimm, T. ; Buydens, L. M. C.: Application of a Spectroscopic Infrared Focal Plane Array Sensor for On-Line Identification of Plastic Waste. In: *Applied Spectroscopy* 51 (1997), Nr. 8, S. 1144 – 1153

[77] Salzer, R. ; Steiner, G. ; Mantsch, H. H. ; Mansfield, J. ; Lewis, E. N.: Infrared and Raman Imaging of biological and biomimetic samples. In: *Fresenius Journal of Analytical Chemistry* 366 (2000), S. 712 – 726

[78] Huberty, Carl J.: *Applied discriminant analysis*. John Wiley & Sons, Inc., New York, 1994 (Wiley series in probabilitiy and mathematical statistics)

[79] McLachlan, Geoffrey J.: *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons, Inc., New York, 1992 (Wiley series in probabilitiy and mathematical statistics)

[80] Bosch, Karl: *Lexikon der Statistik — Nachschlagewerk für Anwender*. 2. Oldenbourg, München, Wien, 1997

[81] Picard, Richard R. ; Berk, Kenneth N.: In: *The American Statistican* 44 (1990), Nr. 2, S. 140 – 147

[82] Lachenbruch, Peter A.: Discriminant Analysis When the Initial Samples Are Misclassified. In: *Technometrics* 8 (1968), Nr. 4, S. 657 – 662

[83] Kap. 10 In: Bowyer, Kevin W.: *Handbook of Medical Imaging*. Bd. 2. Medical Image Processing and Analysis: *Validation of Medical Image Analysis Techniques*. SPIE Press, Bellingham

[84] Lachenbruch, Peter A.: *Discriminant Analysis*. Hafner Press A Division of Macmillan Publishing Co., Inc., New York, 1975

[85] Pönisch, Dr. G.: *Mathematik VI*. WS 2000/01. – Vorlesung

[86] Heitkötter, Jörg ; Beasley, David: *Hitchhiker's Guide to Evolutionary Computation*. 12. April 2001. – http://surf.de.uu.net/encore/www/

[87] Bilger, Uwe E.: *Classifier Systems*. 2002 (Proseminar Evolutionäre Algorithmen, A. Zell, Wilhelm-Schickard-Institut für Informatik, Uni Tübingen). – Ausarbeitung zum Vortrag, http://www-ra.informatik.uni-tuebingen.de/lehre/WS01/pro_ea_ausarbeitung/claassen_ws01.pdf

[88] Keller, Hans-Ulrich (Hrsg.): *Kosmos Himmelsjahr 2002*. Franckh-Kosmos Verlags-GmbH & Co., Stuttgart, 2001

[89] Hawkins, Douglas M.: A New Test for Multivariate Normality and Homoskedsticity. In: *Technometrics* 23 (1981), Nr. 1, S. 105 – 110

[90] Steiner, G. *private Mitteilungen*

[91] Hesse, Manfred ; Meier, Herbert ; Zeeh, Bernd: *Spektroskopische Methoden in der organischen Chemie*. 4., überarb. Aufl. Georg Thieme Verlag, Stuttgart, New York, 1991

[92] Skoog, D. A. ; Leary, J. J.: *Instrumentelle Analytik: Grundlagen, Geräte, Anwendungen*. Springer, Berlin, Heidelberg, New York, 1996. – Übers. D. Brendel, S. Hoffstetter-Kuhn

[93] Jackson, Micheal ; Ramjiawan, Bram ; Hewko, Mark ; Mantsch, Henry H.: Infrared microscopic functional group mapping and spectral clustering analysis of hypercholesterolemic rabbit liver. In: *Cellular and Molecular Biology* 44 (1998), Nr. 1, S. 89 – 98

[94] Lasch, Peter ; Boese, Matthias ; Pacifico, Anthony ; Diem, Max: FT-IR spectroscopic investigations of single cells on the subcellular level. In: *Vibrational Spectroscopy* 28 (2002), S. 147 – 157

[95] Sedgewick, Robert: *Algorithmen*. Addison Wesley, Bonn, München, Reading, 1992

[96] Gottwald, W.: *Statistik für Anwender*. Wiley VCH, Weinheim, Berlin, New York, 2000 (Die Praxis der instrumentellen Analytik)

[97] Doerffel, Klaus: *Statistik in der analytischen Chemie*. 5., erw. u. überarb. Aufl. Deutscher Verlag für Grundstoffindustrie, Leipzig, 1990

[98] Scheffczyk, Jan: *Numerik*. 1999. – Scriptum zur Vorlesung „Numerik“ von Prof. Dr. N. Jacob im FT 1999 — http://www.rz.unibw-muenchen.de/~j8sj0499/documents/numerik/ps.d/numerik.pdf

[99] Weisstein, Eric W.: *Eric Weisstein's World of Mathematics (MathWorld™)*. – http://mathworld.wolfram.com/

[100] Press, William H. ; Teukolsky, Saul A. ; Vetterling, William T. ; Flannery, Brian P.: *Numerical Recipes in C: The Art of Scientific Computing*. 2nd edition. Cambridge University Press., 1988 – 1992

[101] Müller, P. H. (Hrsg.): *Wahrscheinlichkeitsrechnung und mathematische Statistik: Lexikon der Stochastik*. 1. Akademie-Verlag, Berlin, 1991