UNIVERSITÄT DORTMUND

Entwicklung eines Verfahrens zur Klassifikation von Ionenmobilitätsspektrometerdaten

Bertram Bödeker

Algorithm Engineering Report
TR07-2-003
April 2007
ISSN 1864-4503

Universität Dortmund
Fachbereich Informatik
Algorithm Engineering (LS 11)
44221 Dortmund / Germany
http://ls11-www.cs.uni-dortmund.de/

UNIVERSITÄT DORTMUND
FACHBEREICH INFORMATIK

Diplomarbeit

Entwicklung eines Verfahrens zur Klassifikation von Ionenmobilitätsspektrometerdaten

Bertram Bödeker
16. Februar 2007

INTERNE BERICHTE
INTERNAL REPORTS

Diplomarbeit am
Fachbereich Informatik
der Universität Dortmund

Betreuer:
Prof. Dr. G. Rudolph
PD Dr. J. I. Baumbach

Inhaltsverzeichnis

Kapitel 1

Einführung

Die Ionenmobilitätsspektrometrie ist ein Verfahren zur Detektion flüchtiger organischer Verbindungen in Gasen. Als chemisches Analyseverfahren hat es sich bereits in einigen Anwendungsbereichen durchgesetzt. Dazu zählen die Sprengstoff- und Drogendetektion an Flughäfen und die Emissionsüberwachung industrieller Anlagen. Doch vermutlich ist das Potenzial dieser relativ jungen Methode weitaus größer. Durch die Erkennungsmöglichkeit einer Vielzahl von Analyten bei Umgebungsdruck und Raumtemperatur besteht die Hoffnung auf ein universelles Gerät zur Ermittlung der Bestandteile komplexer Gasgemische. Doch bis dahin ist es noch ein weiter Weg.

Ein Ionenmobilitätsspektrometer liefert Messdaten in Form von Spektren. In diesen Spektren zeigen Signalausschläge das Vorhandensein von Analyten im untersuchten Gasgemisch an. Jedoch unterliegen diese Messdaten während der Messung einer Reihe von Einflussfaktoren, wobei deren Auswirkungen auf die Daten oftmals nicht bekannt sind. Es existiert kein allgemein gültiger Standard, in den die Daten transformiert werden könnten. Demnach ist eine universelle Tabelle, in der alle möglichen Analytensignale verzeichnet sind, derzeit noch Utopie.

Das Ziel dieser Diplomarbeit ist es, eine Methode zur Analyse und Klassifikation von IMS-Daten bereitzustellen. Dazu werden die Grundzüge eines Verfahrens entwickelt, welches eine Zuordnung von Gasproben anhand der gemessenen Ionenmobilitätsspek-

trometerdaten zu vorher festgelegten Klassen ermöglicht. Die Messungen erfolgen dabei unter konstanten, klar definierten Bedingungen. Eine Menge vorklassifizierter Beispielmessdaten stellt dabei die Trainingsgrundlage zur Herausarbeitung derjenigen Merkmale dar, die für eine Unterscheidung der verschiedenen Klassen von Bedeutung sind.

Der Schlüssel für ein erfolgreiches Verfahren liegt vor allem in der Vorverarbeitung der Messdaten. Zunächst soll eine Antwort auf die Frage gefunden werden, wie eine Vergleichbarkeit der Daten aus unterschiedlichen Messungen sichergestellt werden kann. Danach wird eine Strategie zur Identifizierung von Signalausschlägen in den Spektren benötigt, die es weiter ermöglicht Signalausschläge, die auf den gleichen Analyten zurückzuführen sind, in verschiedenen Messdaten wiederzufinden. Am Ende sollte ein Messdatum durch die Angabe einer Liste der identifizierten Merkmale repräsentiert werden. Mit Hilfe der erstellten Merkmalslisten der Trainingsdaten sollen im nächsten Schritt Merkmale, die eine Trennung der vorgegebenen Klassen ermöglichen, aufgedeckt und ein Klassifikationsmodell erstellt werden. Weiter wird eine Überprüfung dieses Modells anhand zusätzlicher Daten gefordert.

Das folgende Kapitel erläutert zunächst die Grundlagen der Klassifikation und Mustererkennung, stellt das Verfahren der Ionenmobilitätsspektrometrie vor und beschreibt die verwendeten Methoden, Geräte und die Form der Messdaten. Das dritte Kapitel beschreibt die Entwicklung eines neuartigen Verfahrens zur Vorverarbeitung der Messdaten inklusive der sich anschließenden Merkmalsextraktion. Im vierten Kapitel wird eine Methode zur Klassifikation der gewonnenen Merkmale vorgestellt. Das fünfte Kapitel schildert die Anwendung des gesamten entwickelten Verfahrens auf Basis realer Messdaten von Ausatemluftmessungen im klinischen Umfeld. Schließlich folgen im sechsten Kapitel die Zusammenfassung und der Ausblick. Im Rahmen dieser Diplomarbeit wurde eine prototypische Software zur Darstellung und Analyse der IMS-Daten realisiert, die eine Implementierung der vorgestellten Algorithmen enthält. Eine Beschreibung dazu befindet sich im Anhang.

Kapitel 2

Grundlagen

2.1 Klassifikation und Mustererkennung

2.1.1 Überblick

Klassifikation bezeichnet im Allgemeinen die geordnete Darstellung von Objekten bzgl. eines vorgegebenen Ordnungsprinzips. Klassifizierung bezeichnet den Vorgang der Zuordnung bestimmter Objekte zu sogenannten Klassen. Dabei entscheidet die Ausprägung definierter Merkmale, welcher Klasse ein Objekt zugehörig ist. Die jeweiligen Klassen können fest vorgegeben sein oder sich aus Gemeinsamkeiten und Unterschieden von Beispielobjekten ergeben.

Häufig kommt die Klassifikation als Teilaufgabe der Mustererkennung vor. Unter Mustererkennung versteht man das Auffinden von Strukturen, Regelmäßigkeiten und Invarianten in Datenmengen. Dabei wird versucht Erkenntnisse über den (unbekannten) statistischen Prozess abzuleiten, der diese Daten generiert. Ein Modell bildet diese Erkenntnisse ab, die Eignung wird durch Beispieldaten bestätigt oder widerlegt. Ist ein „gutes“ Modell gefunden, kann es zur Erkennung bzw. Klassifikation von Datensätzen verwendet werden.

2.1.2 Arbeitsweise von Mustererkennungssystemen

Der Prozess der Mustererkennung umfasst prinzipiell die in Abbildung 2.1 dargestellten Teilaufgaben. Alle folgenden Grundlagen der Mustererkennung sind nach [Duda01] beschrieben.

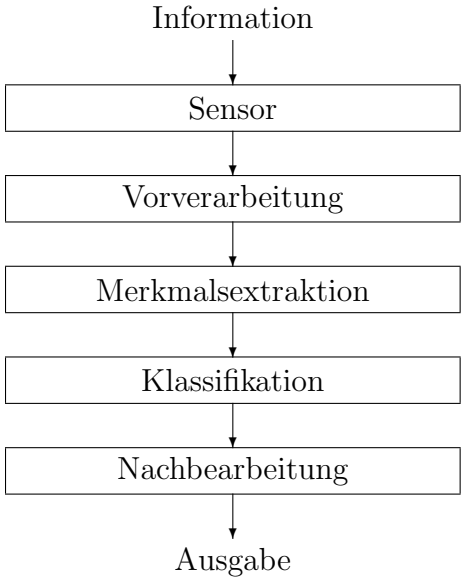


Abbildung 2.1: Arbeitsweise von Mustererkennungssystemen

Die Erfassung der Information geschieht durch Sensoren, Menschen beispielsweise nehmen die Umwelt durch ihre Sinnesorgane wahr. In der Vorverarbeitung werden die Informationen gefiltert, z.B. können visuelle Informationen in Vordergrund und Hintergrund getrennt werden. Die Merkmalsextraktion reduziert die Information auf seine wesentlichen Merkmale. Sie ist ein entscheidender und zugleich schwieriger Schritt in der Mustererkennung, der ausschließlich anwendungsbezogen durchgeführt werden kann. Denkbar sind vielfältige Merkmale, für ein Auto könnten z.B. Form, Farbe, Preis und weitere technische Kennzahlen wie Beschleunigung oder Benzinverbrauch die beschreibenden Merkmale sein. Diese werden in einem sogenannten Merkmalsvektor (auch Featurevektor) zusammengefasst, der als Beschreibung des Objektes die Eingabe des nachfolgenden Schrittes der Klassifikation darstellt. Die Klassifikation beschreibt jede Entscheidung oder Einteilung, die aufgrund des gewählten Klassifikationsverfahrens und den im Featurevektor beschriebenen

Ausprägungen der Merkmale getroffen wird. Dabei gibt das Klassifikationsverfahren die Anleitung vor, nach der die Informationen zur Einteilung in Klassen verwendet werden. In der Nachbearbeitung wird das Ergebnis der Klassifikation benutzt, um eine Entscheidung zu treffen oder eine Handlung durchzuführen. Dabei kommt es durchaus vor, dass dafür die Ergebnisse mehrerer Klassifizierer geeignet miteinander verknüpft werden. Das Ergebnis der Nachbearbeitung stellt schließlich die Ausgabe dar.

2.1.3 Entwicklung eines Systems zur Mustererkennung

Dieser Abschnitt beschreibt die Grundzüge eines allgemeinen Entwurfsmusters für ein System zur Mustererkennung. Dafür müssen die Teilaufgaben Datensammlung, Merkmalsauswahl, Modellauswahl, Training und Evaluierung gelöst werden.

- **Datensammlung**

  Diese Teilaufgabe umfasst die Sammlung der Daten, die im späteren Prozess sowohl zum Training, als auch zur Evaluierung des Klassifizierers verwendet werden. Dabei sollte die Auswahl typische Daten im Hinblick auf das spätere Einsatzgebiet enthalten. Wieviele Daten die Sammlung enthalten soll, ist stark vom jeweiligen Problembereich und der Komplexität der Klassifikationsaufgabe abhängig. Zusätzlich ist die Variabilität der Daten selbst zu beachten.

- **Merkmalsauswahl**

  Die anwendungsbezogene Auswahl der wesentlichen Merkmale benötigt Hintergrundwissen aus dem Problembereich. Mit Hilfe dieses Expertenwissens sollten Merkmale gewählt werden, die hilfreich bei der Einteilung der Muster in Klassen scheinen, dabei möglichst unanfällig gegenüber Störungen (Rauschen innerhalb der Messgröße) und zudem leicht zu extrahieren sind.

- **Modellauswahl**

  Es wird ein Modell erstellt, welches versucht den statistischen Prozess, der die Messdaten generiert, zu beschreiben und dabei möglichst viel Hintergrundwissen aus dem Problembereich ausnutzt. Unbekannte Modellparameter werden in der anschließenden Trainingsphase „erlernt“.

- **Training**

  Im Training werden die unbekannten Parameter des aufgestellten Modells anhand einer Menge von Trainingsdaten erlernt. Lernen bedeutet dabei, eine Einstellung der Parameter zu finden, die den Vorhersagefehler des Modells bzgl. der Trainingsdaten reduziert. Man unterscheidet zwischen

  – überwachtem Lernen,

  – unüberwachtem Lernen und

  – verstärkendem Lernen.

  Beim überwachten Lernen werden für die Trainingsbeispiele gleichzeitig deren Klassen vorgegeben, sodass sich der Klassifizierer in Richtung der Sollausgabe anpassen kann. Beim unüberwachten Lernen hingegen werden Trainingsbeispiele zu Ähnlichkeitsklassen (sogenannten Clustern) zusammengefasst. Häufig wird die Anzahl (maximaler) Klassen vorgegeben. Welche Beispiele in ein Cluster fallen, hängt stark von der Repräsentationsform der Merkmalsvektoren und dem verwendeten Algorithmus ab. Beim verstärkenden Lernen bekommt der Klassifizierer während des Trainings jeweils nur die Rückmeldung, ob die von ihm prognostizierte Klasse für ein Trainingsbeispiel richtig oder falsch war. Diese Information kann er zur weiteren Verbesserung nutzen.

- **Evaluierung**

  Die Evaluierung eines Klassifizierers dient der Überprüfung der Güte bzw. Performanz des Modells. Durch die Betrachtung unbekannter Beispieldaten wird die Fehlerrate des Modells ermittelt. So werden eventuelle Schwachstellen aufgedeckt, für welche dann Verbesserungen abgeleitet werden können. Unter Umständen wird das ganze Modell verworfen und der Entwurf beginnt von vorne. Weiterhin können verschiedene Modelle bzgl. ihrer Performanz verglichen werden, um das beste Modell zum Einsatz zu bringen.

2.1.4 Klassifikationsverfahren

Klassifikationsverfahren arbeiten prinzipiell in zwei Phasen: einer Trainigsphase, in der die Parameter des gewählten Verfahrens mit Hilfe von Trainingsdaten eingestellt werden und einer Einsatz- oder Vorhersagephase, in der das aufgestellte Klassifikationsmodell zur Vorhersage der Klasse unbekannter Instanzen genutzt wird. Jedoch ist es vor dem realen Einsatz sinnvoll die Güte des aufgestellten Modells zu bestimmen.

Es wird zunächst die im Folgenden von mir verwendete Notation beschrieben:

- n : Anzahl der Attribute im Merkmalsvektor
- $\mathbf{x} = (x_1, \ldots, x_n)$: Merkmalsvektor (Belegung der Attribute)
- $\mathbf{X}$: Menge der Merkmalsvektoren
- m : Anzahl der Klassen
- $C = \{c_1, \ldots, c_m\}$: endliche Menge der Klassen
- $D = (\mathbf{x}, c_i)$: vorklassifizierte Instanz mit Merkmalsvektor $\mathbf{x}$ und Klasse c_i
- l : Anzahl der Trainingsbeispiele
- $\mathcal{D} = \{D_1, \ldots, D_l\}$: Menge der Trainingsinstanzen
- $\mathcal{D}_i \subseteq \mathcal{D}$, $i \in \{1, \ldots, m\}$: Menge der Trainingsinstanzen der Klasse c_i
- θ : Belegung der Parameter eines Klassifikationsverfahrens
- Θ : Menge möglicher Parameterbelegungen eines Klassifikationsverfahrens

Ein Klassifikationsverfahren kann damit durch eine Funktion

$$class : \mathbf{X} \times \Theta \rightarrow C \cup \{\emptyset\} \tag{2.1}$$

beschrieben werden die angibt, welcher Klasse $c_i \in C$ ein Merkmalsvektor $\mathbf{x} \in \mathbf{X}$ bei gegebenen Parametern $\theta \in \Theta$ zugeordnet wird:

$$class(\mathbf{x}, \theta) = \begin{cases} \emptyset & \text{falls eindeutige Klassenzuordnung nicht möglich} \\ c_i & \text{sonst} \end{cases} . \tag{2.2}$$

Naive Bayes Methode

Die Naive Bayes Methode gehört zu den Bayes'schen Klassifikationsverfahren, die auf dem *Satz von Bayes* beruhen. Dieser gibt eine Rechenregel für bedingte Wahrscheinlichkeiten an (vgl. [Witten01]):

Hat man eine Hypothese H und ein zu dieser Hypothese passendes Ereignis E, dann gilt

$$Pr(H|E) = \frac{Pr(E|H) \cdot Pr(H)}{Pr(E)} \ . \tag{2.3}$$

Dabei gibt $Pr(A)$ die Wahrscheinlichkeit des Ereignisses A an und $Pr(A|B)$ die bedingte Wahrscheinlichkeit, dass A gilt falls B sicher gilt.

Die Wahrscheinlichkeit für die Gültigkeit der Hypothese H unter dem Ereignis E kann demnach aus der (oft einfacher zu bestimmenden) Wahrscheinlichkeit für das Ereignis E unter Gültigkeit der Hypothese H und den Wahrscheinlichkeiten von H und E berechnet werden.

Bezogen auf die Klassifikationsaufgabe ist bei einer durch den Merkmalsvektor $\mathbf{x}$ gegebenen Instanz diejenige Klasse $c_i \in C$ gesucht, die die größte Wahrscheinlichkeit der Zugehörigkeit aufweist:

$$class(\mathbf{x}, \theta) = \begin{cases} c_i & \text{falls } \exists c_i \in C : \forall c_j \in C \setminus \{c_i\} : Pr(c_i|\mathbf{x}) > Pr(c_j|\mathbf{x}) \\ \emptyset & \text{sonst} \end{cases} \tag{2.4}$$

Dabei gibt $Pr(c_i|\mathbf{x})$ die Wahrscheinlichkeit der Zugehörigkeit einer Instanz mit Merkmalsvektor $\mathbf{x}$ zur Klasse c_i an.

Im Folgenden wird die Bestimmung der Wahrscheinlichkeiten $Pr(c_i|\mathbf{x})$ mit Hilfe des Satzes von Bayes hergeleitet:

$$Pr(c_i|\mathbf{x}) = \frac{Pr(\mathbf{x}|c_i) \cdot Pr(c_i)}{Pr(\mathbf{x})} \tag{2.5}$$

Da die Menge der Klassen C endlich ist und jede Instanz in eine der Klassen fällt, gilt:

$$\sum_{i=1}^{m} Pr(c_i|\mathbf{x}) = 1 \tag{2.6}$$

Daraus folgt mit (2.5):

$$\sum_{i=1}^{m} \frac{Pr(\mathbf{x}|c_i) \cdot Pr(c_i)}{Pr(\mathbf{x})} = 1 \tag{2.7}$$

$$\Leftrightarrow \quad \frac{1}{Pr(\mathbf{x})} \sum_{i=1}^{m} (Pr(\mathbf{x}|c_i) \cdot Pr(c_i)) = 1 \tag{2.8}$$

$$\Leftrightarrow \quad Pr(\mathbf{x}) = \sum_{i=1}^{m} (Pr(\mathbf{x}|c_i) \cdot Pr(c_i)) \tag{2.9}$$

Weiter geht die Naive Bayes Methode naiv von der Unabhängigkeit der Attribute im Merkmalsvektor aus. Damit kann die Wahrscheinlichkeit, dass eine Instanz mit Merkmalsvektor $\mathbf{x}$ der Klasse c_i angehört, als Produkt der Wahrscheinlichkeiten, dass die Attributbelegungen x_j mit $j \in \{1, \ldots, n\}$ der Klasse c_i angehören, ausgedrückt werden:

$$Pr(\mathbf{x}|c_i) = \prod_{j=1}^{n} Pr(x_j|c_i) \ . \tag{2.10}$$

Das Problem der Bestimmung von $Pr(\mathbf{x}|c_i)$ ist nun auf die Aufgabe reduziert, die Wahrscheinlichkeiten $Pr(x_j|c_i)$ und $Pr(c_i)$ zu bestimmen. Es wird hier der Fall betrachtet, dass die Attributbelegungen rein numerische Werte sind, die einer Normalverteilung unterliegen.

Die vorklassifizierten Trainingsbeispiele der Menge $\mathcal{D}$ werden nun zur Abschätzung der Wahrscheinlichkeiten verwendet. Für die Wahrscheinlichkeit der Klassenzugehörigkeit wird die relative Häufigkeit der jeweiligen Klasse in der Trainingsmenge genutzt:

$$Pr(c_i) = \frac{|\mathcal{D}_i|}{|\mathcal{D}|} \ . \tag{2.11}$$

Zur Bestimmung der Wahrscheinlichkeiten $Pr(x_j|c_i)$ wird die Dichtefunktion $f(x; \mu, \sigma)$ verwendet, die die Wahrscheinlichkeitsdichte der Normalverteilung bei gegebenen Parametern μ und σ beschreibt:

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \ . \tag{2.12}$$

Dazu werden für jedes Attribut j und jede Klasse c_i zunächst die Mittelwerte $\mu_{j,i}$ und die Standardabweichungen $\sigma_{j,i}$ des j-ten Attributes über alle Beispiele aus $\mathcal{D}_i$ berechnet. Daraus ergibt sich die Wahrscheinlichkeit $Pr(x_j|c_i)$ wie folgt:

$$Pr(x_j|c_i) = \epsilon \cdot f(x_j; \mu_{j,i}, \sigma_{j,i}) \ . \tag{2.13}$$

Wahrscheinlichkeit und Wahrscheinlichkeitsdichte sind nicht dasselbe. Die Dichtefunktion $f(x)$ beschreibt die Wahrscheinlichkeit, dass die betrachtete Größe innerhalb des Intervalls $x - \epsilon/2$ und $x + \epsilon/2$ liegt mit $\epsilon \cdot f(x)$.

Als Beispiel sei für die Klasse c_1 und das dritte Attribut $\mu_{3,1} = 73$ und $\sigma_{3,1} = 6,2$. Die Wahrscheinlichkeit für eine Attributbelegung $x_3 = 66$ berechnet sich dann wie folgt:

$$Pr(x_3 = 66|c_1) = \epsilon \cdot f(x_3; \mu_{3,1}, \sigma_{3,1}) = \epsilon \cdot \frac{1}{\sqrt{2\pi} \cdot 6,2} e^{-\frac{(66-73)^2}{2 \cdot 6,2^2}} = \epsilon \cdot 0,0340 \ . \qquad (2.14)$$

Die Wahrscheinlichkeit dieses Ereignisses in einer ϵ-Umgebung von 1 beträgt demnach $3,4\%$.

Damit sind alle nötigen Berechnungen beschrieben, um $Pr(c_i|\mathbf{x})$ für alle Klassen $c_i \in C$, bei einen gegebenen Merkmalsvektor $\mathbf{x} \in \mathbf{X}$ und einer gegebenen Menge vorklassifizierter Trainingsinstanzen $\mathcal{D}$, zu bestimmen.

Beim Einsatz der Naiven Bayes Methode muss folgendes beachtet werden: Die Attribute werden so behandelt, als ob sie unabhängig voneinander wären. Redundante Attribute verfälschen den Lernprozess dahingehend, dass eine Vervielfältigung eines Attributes zu einer Vervielfältigung des Einflusses dieses Attributes führt. [Langley94] beschreibt eine selektive Bayes Methode die, durch eine geeignet gewählte Untermenge der Attribute, dieses Problem behebt.

Für die Naive Bayes Methode ist die Annahme der Normalverteilung für numerische Werte nicht immer haltbar. Ist eine andere Verteilung bekannt, kann eine Standardschätzung für diese verwendet werden. Weiter könnte die Prozedur einer Kerndichteschätzung, die eine Schätzung der Wahrscheinlichkeitsverteilung aus den Trainingsdaten ermöglicht, verwendet werden. Außerdem wäre es möglich die Daten im Vorhinein zu diskretisieren und die Naive Bayes Methode für diskrete Attribute, die hier nicht beschrieben ist, anzuwenden.

2.1.5 Evaluierung mittels Kreuz-Validierung

Die n-fache Kreuz-Validierung ist eine Methode zur Überprüfung der Güte eines Klassifikationsverfahrens. Dazu wird die Menge der vorklassifizierten Daten in n (soweit möglich) gleich große Teilmengen partitioniert, von denen $n-1$ Mengen zum Training und die übrige Menge zum Test eines Klassifikationsmodells verwendet werden. Dieser Vorgang wird n-fach wiederholt, wobei jede der n Teilmengen einmal die Trainingsmenge wird. Dabei kann n zwischen zwei und der Anzahl der Daten der Ausgangsmenge liegen. Üblicherweise wird in der Praxis $n = 10$ gewählt. Bei einem geringen Umfang von vorklassifizierten Daten kann n auch der Anzahl dieser Daten entsprechen. Dann heißt die Methode „leave-one-out“ Kreuz-Validierung, da beim Training des Klassifizierers je genau ein Datum zum anschließenden Test zurückgehalten wird.

Die Güte eines Klassifikationsverfahrens wird durch die Anzahl der richtig klassifizierten Testdaten bestimmt. Da bei der Kreuzvalidierung jedes Datum der Gesamtmenge von einem Klassifikationsmodell klassifiziert wird, in dessen Trainingsmenge dieses nicht vorkam, gibt der prozentuale Anteil der richtig klassifizierten Daten an der Gesamtzahl der Daten einen Hinweis auf die Klassifikationsperformanz des Verfahrens bei Daten dieser Art. Je repräsentativer die Ausgangsdatenmenge ist, umso verlässlicher ist die ermittelte Performanz. Es wird allerdings keine Aussage zu einem bestimmten erstellten Modell getroffen.

2.2 Ionenmobilitätsspektrometrie

2.2.1 Überblick

Die Ionenmobilitätsspektrometrie ist ein analytisches Verfahren zur Identifizierung unterschiedlicher chemischer Substanzen in Gasgemischen. Das Ionenmobilitätsspektrometer (IMS) zeichnet sich durch eine hohe Trennleistung, kurze Analysezeiten und niedrige Nachweisgrenzen im ppm_v bis ppb_v Bereich aus. Analysen können bei Umgebungsdruck und Raumtemperatur durchgeführt werden, wobei die Raumluft als Trägergas dienen

kann. Besonders interessant ist die Kopplung des Gerätes mit einer geeigneten gaschromatographischen Säule, die zur Vortrennung des Gasgemisches benutzt wird und damit die Trennleistung nochmals erhöht. Zusätzlich haben die Geräte eine portable Größe erreicht, was eine Analyse vor Ort mit tragbaren Geräten ermöglicht.

Die Ionenmobilitätsspektrometrie hat sich in den letzten Jahren als chemisches Analyseverfahren in vielen Einsatzgebieten durchgesetzt. Dazu gehören unter anderem die Sprengstoff- und Drogendetektion an Flughäfen, die Emissionsüberwachung industrieller Anlagen oder die Überwachung der Luftqualität an Bord der Internationalen Raumstation ISS[1]. Ein weiteres interessantes Einsatzgebiet stellt die Identifizierung von Produkten der biochemischen Umwandlung in Organismen dar. Die Detektion und Charakterisierung dieser sogenannten Metabolite, z.B. in der menschlichen Ausatemluft, könnte der verbesserten Frühdiagnose und therapeutischen Kontrolle von Krankheiten und Krankheitsverläufen dienen (vgl. [Baumbach05a],[Baumbach05b]).

2.2.2 Physikalische Grundlagen

Das Prinzip dieser Analysemethode beruht auf der Ionisierung der gasförmigen Analyten und anschließender Messung der Durchlaufzeit (Driftzeit t_d) der gebildeten positiven oder negativen Ionenschwärme in einem elektrischen Feld durch eine bestimmte Wegstrecke (Driftstrecke l_d). Dabei gibt die Driftzeit Aufschluss über die Art des Analyten. Analyten bestehen aus Molekülen, und die aus unterschiedlichen Molekülen gebildeten Ionen brauchen verschieden lange, um die Driftstrecke zurückzulegen.

Im Folgenden wird eine stoffspezifische Größe hergeleitet, anhand derer verschiedene Moleküle unterscheidbar sind (vgl. [Eiceman05]).

Ein Ionenschwarm hat eine mittlere Driftgeschwindigkeit v_d:

$$v_d = \frac{l_d}{t_d} \tag{2.15}$$

[1]Environmental Monitoring of the International Space Station, NASA, http://exploration.nasa.gov/programs/station/Environmental-Monitoring.html

Diese Geschwindigkeit v_d ist proportional zur elektrischen Feldstärke E mit dem Faktor K, das elektrische Feld wird dabei als homogen angenommen.

$$v_d = K \cdot E \tag{2.16}$$

E ergibt sich aus der Driftspannung U_d und der Länge der Driftstrecke l_d:

$$E = \frac{U_d}{l_d} \tag{2.17}$$

Insgesamt ergibt sich folgende Gleichung die nach K aufgelöst wird:

$$\frac{l_d}{t_d} = K \cdot \frac{U_d}{l_d} \Leftrightarrow K = \frac{l_d^2}{U_d \cdot t_d} \tag{2.18}$$

K heißt Mobilitätskoeffizient und beschreibt die Beweglichkeit der Ionen eines Analyten. Unterschiedliche Analyten haben einen unterschiedlichen Mobilitätskoeffizienten. Sie können nach Formel 2.18 allein über ihre gemessene Driftzeit unterschieden werden, falls alle weiteren Parameter konstant sind.

Die Beweglichkeit von Ionen wird unter anderem beeinflusst von der Umgebungstemperatur und dem Umgebungsdruck. Eine charakteristische temperatur- und druckunabhängige Eigenschaft eines Stoffes ist die reduzierte Mobilität K_0, bei der die Mobilität K auf eine Standardtemperatur $T_0 = 273,15°K$ und einen Standarddruck $p_0 = 1013,25$ hPa normalisiert ist. Dabei verringert eine Erhöhung der Temperatur die Gasdichte und erhöht so die Mobilität. Anders beim Druck: Eine Erhöhung erhöht auch die Gasdichte und verringert damit die Mobilität. Mit gegebenen (bzw. zur Analysezeit gemessenen) Größen für Temperatur T und Druck p ergibt sich die reduzierte Mobilität K_0.

$$K_0 = K \cdot \frac{T_0}{T} \cdot \frac{p}{p_0} \tag{2.19}$$

Jeder Stoff kann damit durch seine spezifische reduzierte Mobilität K_0 charakterisiert werden. Sie wird meist in der Einheit $\frac{cm^2}{Vs}$ angegeben.

2.2.3 Funktionsweise

Dieser Abschnitt beschreibt die Funktionsweise des verwendeten IMS.

Ionenmobilitäts-Spektrometer

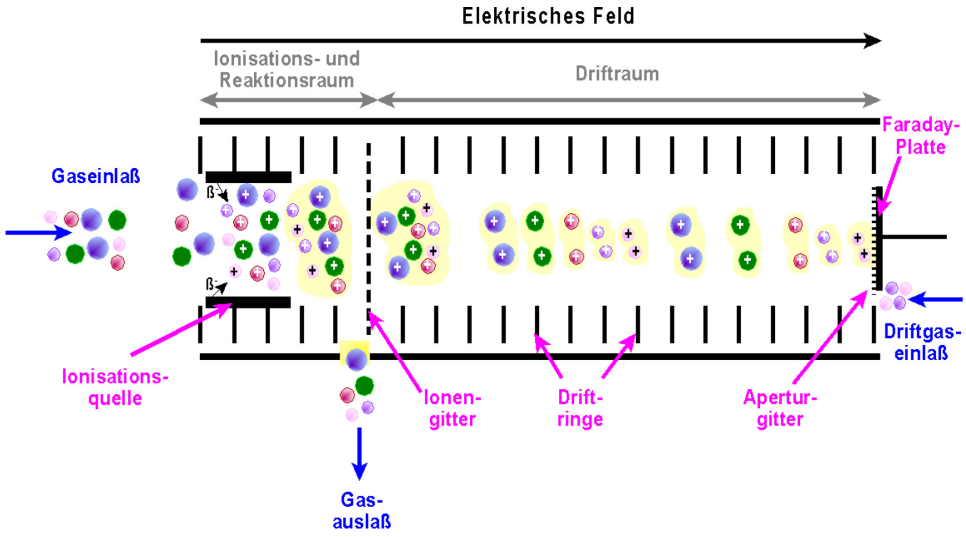


Abbildung 2.2: Funktionsweise eines Ionenmobilitäts-Spektrometers (Quelle: ISAS)

Eine Probe wird durch den Gaseinlass in den Ionisations- und Reaktionsraum eingeleitet, wobei z.B. Luft als Trägergas verwendet wird. Mittels einer Ionisationsquelle werden die Moleküle des Gases ionisiert. Als Ionisationsquelle kommt in herkömmlichen IMS radioaktives Nickel (^{63}Ni) zum Einsatz. Diese Quelle emittiert Elektronen mit einer Energie von bis zu 67 keV. Durch Kollisionen der Moleküle des Trägergases mit den energiereichen Elektronen entstehen sogenannte Reaktionsionen. Wird Luft (Hauptbestandteile Stickstoff N_2, Sauerstoff O_2 und Wasserdampf H_2O) als Trägergas verwendet, entstehen hauptsächlich $H^+(H_2O)_n$ Ionen positiver Polarität und $O_2^-(H_2O)_n$ Ionen negativer Polarität als Reaktionsionen, mit $n \in \{1, \ldots, 7\}$. Im Spektrum zeichnen sie sich durch einen charakterischen Signalausschlag - den Reaktionsionen-Peak (RIP)- ab. Die Gesamtladung aller Reaktionsionen hängt von der Stärke der Ionisationsquelle ab und bildet eine obere Grenze für die Anzahl der Moleküle die ionisiert werden können.

Durch weitere Reaktionen der Moleküle M des zu untersuchenden Gasgemisches

mit den Reaktionsionen entstehen Produkt-Ionen, die im Spektrum in Form weiterer Signalausschläge (Peaks) erkennbar sind. Produkt-Ionen positiver Polarität entstehen z.B. dann, wenn die Moleküle eine höhere Protonenaffinität als die Reaktionsionen haben, Produkt-Ionen negativer Polarität entstehen z.B. durch eine höhere Elektronenaffinität.

Die folgenden Reaktionsgleichungen mit $n \in \{1, \ldots, 7\}$ und $x \in \{0, 1\}$ verdeutlichen die Entstehung (vgl. [Eiceman05]).

Produkt-Ionen positiver Polarität als Monomer:

$$M + H^+(H_2O)_n \longrightarrow MH^+(H_2O)_n \longrightarrow MH^+(H_2O)_{n-x} + xH_2O \tag{2.20}$$

Produkt-Ionen positiver Polarität als Dimer:

$$MH^+(H_2O)_n + M \longrightarrow M_2H^+(H_2O)_{n-x} + xH_2O \tag{2.21}$$

Produkt-Ionen negativer Polarität:

$$M + O_2^-(H_2O)_n \longrightarrow MO_2^-(H_2O)_{n-x} + xH_2O \tag{2.22}$$

Ist die Konzentration von M hoch, reagieren die Produkt-Ionen (Monomere nach 2.20) untereinander. Es entstehen protonengebundene Dimere (nach 2.21), die zu einem zusätzlichen Peak im Spektrum führen.

Durch das periodisch öffnende Einlassgitter gelangen die Ionen in den Driftraum, in dem sie durch das elektrische Feld in Richtung des Detektors (Faraday-Platte) beschleunigt werden. Die Anordnung der Driftringe ist dabei ausschlaggebend für die Homogenität des elektrischen Feldes. Den Ionen strömt ein Driftgas entgegen, mit dessen neutralen Molekülen die Ionen zusammenstoßen und so ihre charakteristische Driftgeschwindigkeit bekommen. Das Aperturgitter führt zu einer kapazitiven Entkopplung der Ionen in der Driftstrecke und der Faraday-Platte. Treffen die Ionen auf den Detektor, erzeugen sie ein elektrisches Signal. Das Auftragen der Signalintensität gegen die Driftzeit ergibt das sogenannte Ionenmobilitäts-Spektrum.

Die Messung kann auf zwei Arten erfolgen:

- Messung von Ionen positiver Polarität
- Messung von Ionen negativer Polarität

Bei allen in dieser Arbeit verwendeten Daten wurden die Ionen positiver Polarität gemessen, es sollte jedoch keinen Unterschied zur Messung der Ionen negativer Polarität geben.

Vortrennung mittels Multi-Kapillarsäule

Bei komplexen Gasgemischen führt die gleichzeitige Ionisierung aller Analyten zu einer starken Überlagerung der Signalausschläge, wodurch eine qualitative Identifizierung schwierig und eine quantitative Identifizierung quasi unmöglich wird. Daher wird dem IMS eine Multi-Kapillarsäule (Multi Capillary Column, MCC) zur Vortrennung des Gasgemisches vorgeschaltet. Diese Säule besteht aus einer Vielzahl von gebündelten Einzelkapillaren, die unterschiedliche Analyten verschieden lange zurückhalten. Somit bekommen die Messdaten eine weitere Dimension: die Retentionszeit, die die jeweilige Verzögerung der Gasbestandteile durch die Säule beschreibt.

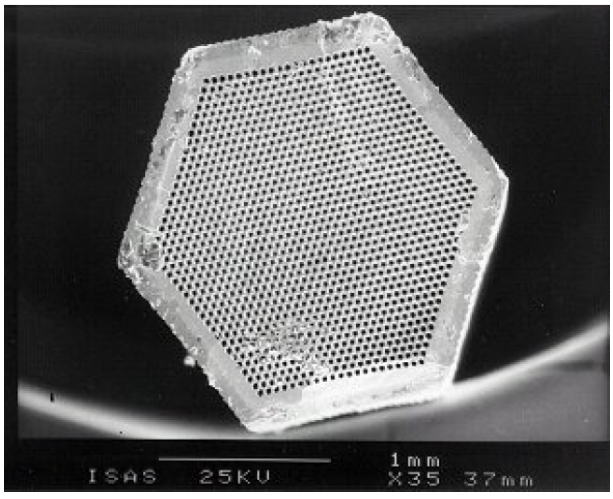


Abbildung 2.3: Querschnitt einer Multi-Kapillarsäule (Quelle: ISAS)

Eine hohe relative Feuchte des verwendeten Trägergases führt zu der Bildung von Ionenclustern mit den Wassermolekülen. Diese erzeugen viele zusätzliche, störende

Peaks. Der Einsatz einer MCC führt zu einer Abtrennung der Feuchte zu Anfang der Analyse. Dadurch wird die Anzahl der störenden Peaks, insbesondere bei niedrigen Retentionszeiten, verringert. Besonders bei der Analyse von Ausatemluft, die ein komplexes Gasgemisch mit 100 % relativer Feuchte darstellt, eignet sich die Multi-Kapillarsäule daher zur Vortrennung (vgl. [Ruzsányi05]).

Geräteaufbau MCC / ^{63}Ni - IMS

In diesem Abschnitt wird der Aufbau eines MCC / ^{63}Ni - IMS Gerätes des ISAS - Institute for Analytical Sciences, Dortmund, anhand von zwei Abbildungen gezeigt.

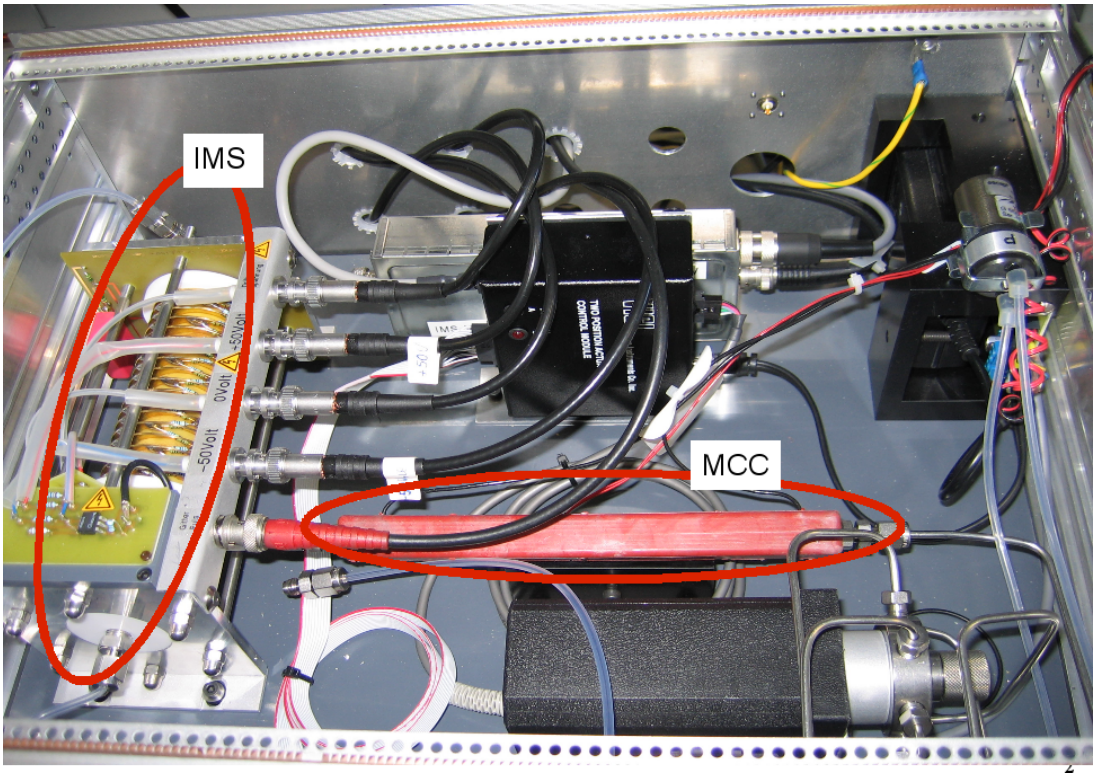


Abbildung 2.4: Geräteaufbau MCC / ^{63}Ni - IMS von innen (Quelle: ISAS)

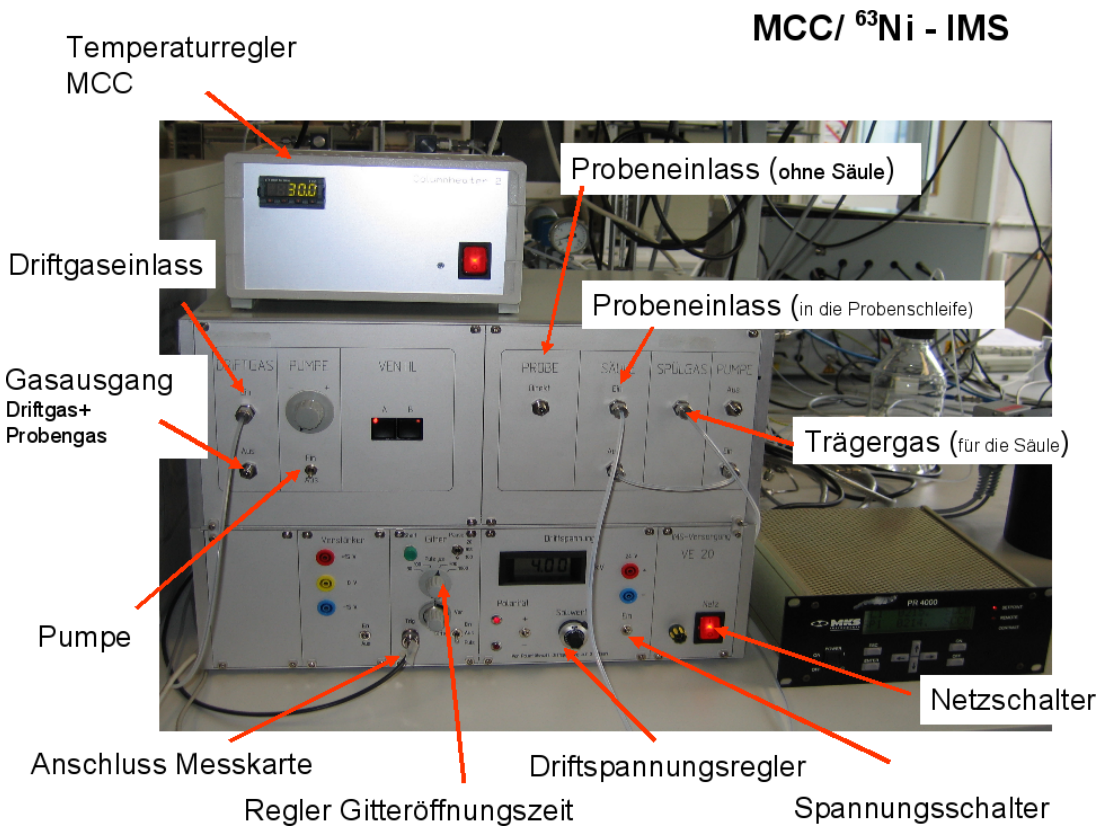


Abbildung 2.5: Geräteaufbau MCC / ^{63}Ni - IMS Außenansicht mit Einstellungsmöglichkeiten (Quelle: ISAS)

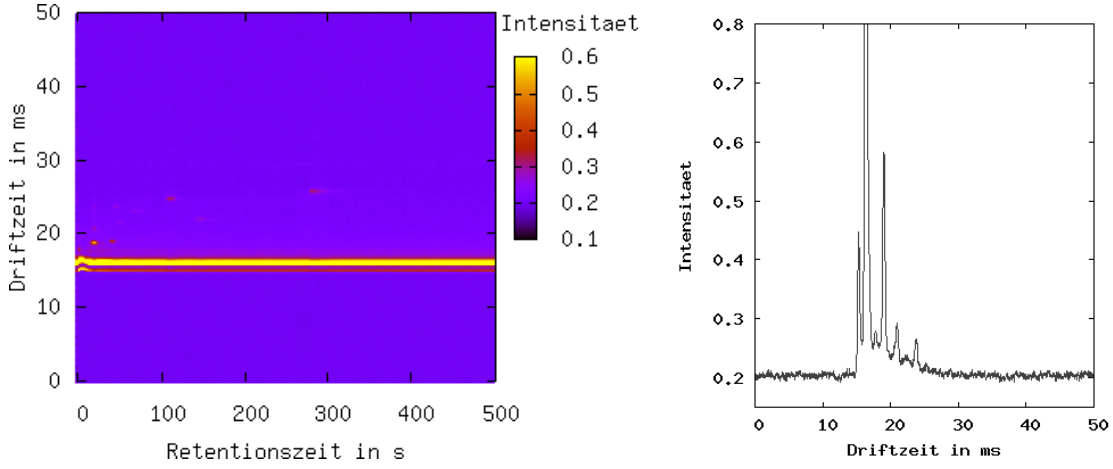
2.2.4 Messdaten

Ein Spektrum wird durch einen Vektor $\mathbf{S} = (z_0, \cdots, z_{1999})$ der Länge 2000 beschrieben, wobei jedes z_i die gemessene Intensität zur Driftzeit $dt_i = i \cdot 0,0025$ ms angibt. Ein vollständiges Spektrum deckt somit einen Driftzeitraum von $dt_0 = 0$ ms bis $dt_{1999} = 49,975$ ms ab.

Ein Messdatum $\mathbf{M}$ besteht aus einer Serie von bis zu 501 Spektren $\mathbf{M} = (\mathbf{S}_0, \ldots, \mathbf{S}_{500})$, die zu unterschiedlichen Retentionszeiten $rt_0, \ldots, rt_{500}$ mit $rt_i < rt_j$ für $i < j$, $i, j \in \{0, \ldots, 500\}$ aufgenommen werden.

Die Abbildung 2.6(a) zeigt, wie die Daten einer Messung als Heatmap mit den Achsen Drift- und Retentionszeit dargestellt werden können. Die Signalintensität wird

durch die Farbe beschrieben, Peaks sind als ovale Flecken erkennbar. Als Beispiel wurden die Daten einer Atemluftmessung verwendet. Die Abbildung 2.6(b) zeigt ein einzelnes Spektrum dieser Atemluftmessdaten. Der hohe Signalausschlag ist der Reaktionsionen-Peak, der Ausschlag links daneben ist durch die Feuchte der Atemluft bedingt. Zusätzlich sind einige weitere Peaks erkennbar.



(a) Darstellung als Heatmap

(b) Einzelnes Spektrum

Abbildung 2.6: Messdaten einer Ausatemluftmessung

Wird der Bereich der Driftzeit auf 15 bis 25 ms und der Bereich der Retentionszeit auf die ersten 100 Sekunden begrenzt, so ergibt sich die Abbildung 2.7.

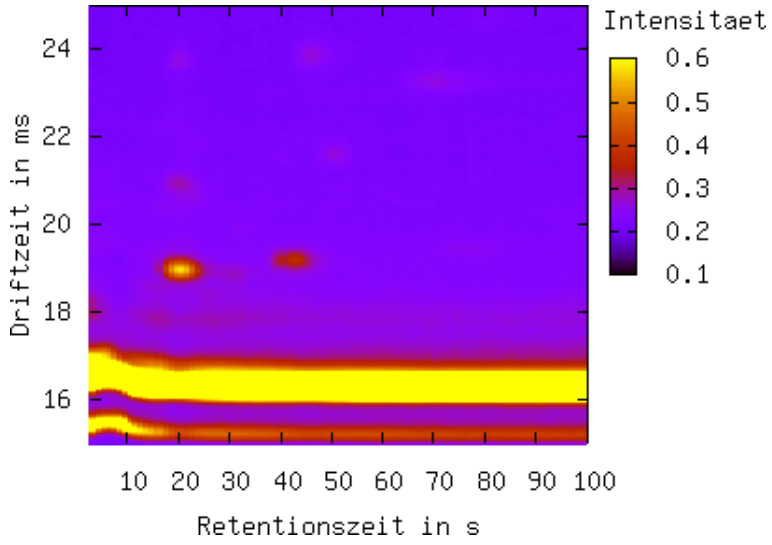


Abbildung 2.7: Teilbereich der Daten einer Ausatemluftmessung

Es sind einzelne rote und rot-gelbe, ovale Flecken zu erkennen, die verschiedene Peaks anzeigen. Der RIP zeichnet sich durch den breiten, durchgehenden, gelben Streifen ab. Der durchgehende Streifen unterhalb des RIP resultiert aus der Feuchte des Trägergases.

Einflüsse bei der Messung

Ionenmobilitätsspektrometer sind sensible Messinstrumente, die empfindlich auf viele unterschiedliche Einflüsse reagieren. Bisher sind keine Prozeduren gefunden, um Messdaten, die unter verschiedenen Messparametern aufgezeichnet werden, in einen einheitlichen Standard zu transformieren. Um die Daten vergleichbar zu halten, wird hier die Messung unter möglichst konstanten Bedingungen gefordert.

Die folgende Aufzählung enthält die wichtigsten Einflussgrößen und beschreibt, welche Annahmen für alle Messdaten, die in dieser Diplomarbeit untersucht werden, gelten.

- **Temperatur und Umgebungsdruck:** nicht konstant
- **Ionisationsquelle:** radioaktive ^{63}Ni Quelle
- **Polarität:** positiv
- **Driftspannung:** $U_d = 4,0$ kV
- **Länge der Driftstrecke:** beträgt $l_d = 12$ cm
- **Elektrisches Feld:** $E = 333,3$ V/cm (nach Formel 2.17)
- **Gitteröffnungszeit:** 300 μs.
- **Gitterpause:** 100 ms.
- **Multi-Kapillarsäule:** MCC, OV-5 Sibertech, LTD, Novosibirsk, Russland
- **Temperatur der MCC:** $30°C$
- **Driftgas:** N_2 (Stickstoff)

- **Driftgasfluss:** 100 mL/min
- **Probengasflussrate MCC:** 150 mL/min
- **Vorverstärker:** 10^{10} V/A -Strom-Spannungswandler, ISAS Dortmund

2.3 Optimierung mittels Evolutionsstrategie

Unter Optimierung wird hier die Suche nach Parametern verstanden, die eine Zielfunktion entweder minimieren oder maximieren. Die Menge der möglichen Parameter sowie die Zielfunktion selbst sind dabei vorgegeben.

Die Evolutionsstrategie (ES) ist ein heuristisches Optimierungsverfahren für reellwertige Parametereinstellungen, dem der biologische Prozess der Evolution zu Grunde liegt. Die Idee ist es, jede Belegung der einzustellenden Parameter als Individuum anzusehen. Aus einer vorgegebenen Menge von sogenannten Eltern-Individuen wird durch zufällige Mutationen eine Generation von Nachkommen generiert. Für jedes Individuum wird dabei durch ein vorgegebenes Gütekriterium die Güte (Fitness) bestimmt. Die fittesten Nachkommen bilden die neue Elterngeneration, aus der wiederum neue Nachkommen erzeugt werden. Das Verfahren endet, sobald ein Abbruchkriterium erfüllt ist, z.B. sobald in mehreren Generationen keine Verbesserung der Güte der Nachkommen mehr eintritt. Im Sinne der Optimierung entspricht die Güte eines Individuums $\mathbf{X}$ bei einer zu maximierenden Zielfunktion dem Funktionswert der Zielfunktion an der Stelle $\mathbf{X}$, bei einer zu minimierenden Zielfunktion dem reziproken Wert der Zielfunktion an dieser Stelle.

Es gelte die folgende Notation:

- $n \in \mathbb{N}^+$: Anzahl der zu optimierenden Parameter
- $\mathbf{X} = (x_1, \ldots, x_n) \in \mathbb{R}^n$: Individuum bzw. mögliche Belegung der zu optimierenden Parameter

- $\Sigma = (\sigma_1, \ldots, \sigma_n) \in \mathbb{R}_+$: Parameter der Schrittweite von Mutationen, dabei gibt σ_i die Mutationsschrittweite für x_i an.
- $f(\mathbf{X})$: Gütekriterium, welches die Fitness des Vektors $\mathbf{X}$ beschreibt
- λ : Anzahl der Nachkommen einer Generation
- $\mathcal{N}(0,1)$: standardnormalverteilt-zufällige Zahl

Das Ziel der Evolutionsstrategie ist es, ein möglichst optimales Individuum $\mathbf{X}$ bzgl. des Gütekriteriums $f(\mathbf{X})$ zu finden. Es wird das Verfahren der $(1,\lambda)$-ES mit adaptiver Schrittweitenanpassung vorgestellt. Eine adaptive Schrittweitenanpassung bedeutet, dass bei der Generierung der Nachkommen auch die Mutationsschrittweiten mutiert werden. Dies entspricht der Erstellung eines Nachkommens $(\mathbf{X}, \Sigma)$. Der Algorithmus durchläuft folgende wesentliche Schritte:

(1) **Initialisierung:** Es wird ein Eltern-Individuum $\mathbf{X}^{(0)}$ mit zufällig gewählten Werten bestimmt und dazu die initiale Mutationsschrittweite $\Sigma^{(0)}$ festgelegt.

(2) **Generierung von λ Nachkommen:** ein Nachkomme $(\mathbf{X}^{(t+1)}, \Sigma^{(t+1)})$ wird durch

$$\sigma_i^{(t+1)} = \sigma_i^{(t)} \cdot e^{\frac{1}{\sqrt{n}} \cdot \mathcal{N}(0,1)} \tag{2.23}$$

$$x_i^{(t+1)} = x_i^{(t)} + \sigma^{(t+1)} \cdot \mathcal{N}(0,1) \tag{2.24}$$

für alle $i \in \{1, \ldots, n\}$ generiert.

(3) **Berechnung der Fitness:** durch Berechnung von $f(\mathbf{X}^{(t+1)})$ für alle λ Nachkommen

(4) **Neuer Elter:** wird der fitteste Nachkomme

(5) **Abbruchkriterium erfüllt?** Falls ja, stellt der aktuelle Elter das Ergebnis der Evolutionsstrategie dar, falls nein, weiter bei (2).

Kapitel 3

Datenvorverarbeitung

Ein wichtiger Schritt zur Klassifikation ist die geeignete Vorverarbeitung der vorliegenden Daten. Häufig stellt diese Teilaufgabe den größten Aufwand bei der Erstellung eines Mustererkennungssystems dar. Sie umfasst zum einen die Auswahl und die Normalisierung der Daten, zum anderen die Extraktion von Merkmalen zur Erstellung des Featurevektors.

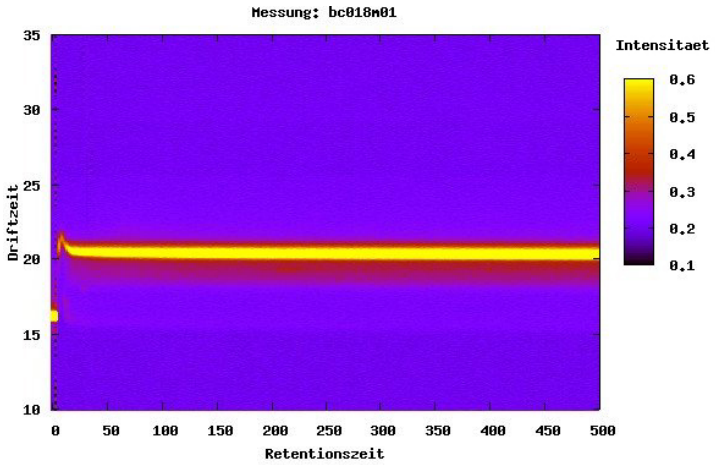
Dieses Kapitel beschreibt die Vorverarbeitungsschritte für MCC / ^{63}Ni - IMS Messdaten. Diese beziehen sich einmal auf die initiale Erstellung einer Instanzmenge als Trainings- und Testgrundlage des Klassifikationsverfahrens, und weiter auf die Verarbeitung aller neuen Messdaten, die anhand des fertigen Klassifikationsmodells klassifiziert werden sollen.

Ausgangspunkt dieser Datenvorverarbeitung sind die rohen Messdaten, wie sie unter Abschnitt 2.2.4 beschrieben sind.

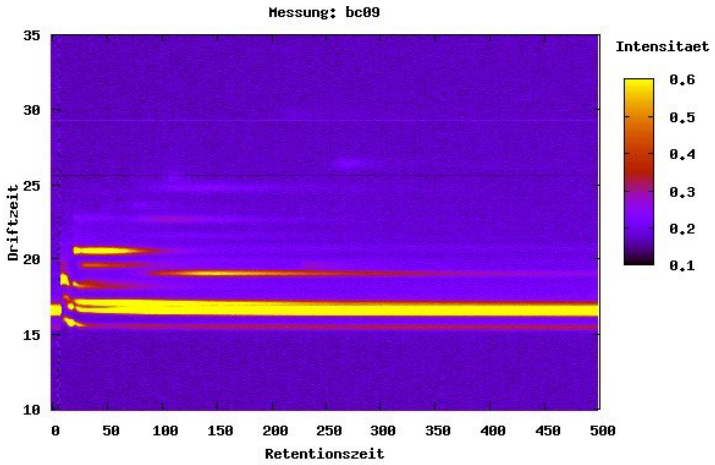
3.1 Datenauswahl

Das Ziel der Datenauswahl ist die Zusammenstellung geeigneter Beispiele für jede mögliche Klasse, in die die Messdaten fallen können und die Bereinigung dieser Datenmenge um fehlerhafte Datensätze. Dabei ist darauf zu achten, dass möglichst repräsentative Bei-

spiele jeder Klasse ausgewählt werden. Diese Aufgabe wird am besten von Experten der Problemdomäne durchgeführt. Die Analyse von Atemluftmessungen bzgl. Krankheiten erfordert z.B. die Mitarbeit des behandelnden Arztes, der das Fachwissen um die Krankheit, das Stadium der Krankheit, Medikamenteneinflüsse usw. mitbringt. So gehen bestimmte Krankheiten beispielsweise oft mit einer Bakterieninfektion einher, durch die ein Klassifikationsverfahren unter Umständen fehlgeleitet wird. Weiter ist das Wissen eines Experten für IMS-Messungen nötig, der am Chromatogramm offensichtliche Fehler identifizieren kann (z.B. wenn ein Proband nicht die ausreichende Menge an Atemluft zur Verfügung stellen konnte oder eine Messung durch elektromagnetische Störungen unbrauchbar ist). Die Abbildung 3.1 zeigt Beispiele für offensichtlich gestörte Messungen.



(a) RIP verschoben und unförmig



(b) Peaks verwischt

Abbildung 3.1: Offensichtlich gestörte Beispiele sind von der Trainingsmenge auszuschließen.

Am Ende der Datenauswahl steht eine bereinigte Menge vorklassifizierter Beispielmessdaten zur weiteren Verarbeitung bereit.

3.2 Datennormalisierung

Der Schritt der Datennormalisierung soll die Einflussgrößen auf IMS-Messungen so weit wie möglich ausgleichen, sodass verschiedene Messdaten vergleichbar werden. Dazu werden Invarianten innerhalb der Daten gesucht, mit deren Hilfe die Daten in eine normierte Form gebracht werden.

3.2.1 Normierung der Peaklagen durch Achsentransformation

Die Lage eines Signalausschlages im Chromatogramm wird durch ein Tupel, bestehend aus Drift- und Retentionszeit (dt_i, rt_j) mit $i \in \{0, \cdots, 1999\}$, $j \in \{0, \cdots, 500\}$ beschrieben (siehe Abschnitt 2.2.4).

Die Retentionszeit rt_j eines Analyten ist abhängig von der verwendeten MCC, der Temperatur innerhalb der Säule und der Durchflussgeschwindigkeit des zu analysierenden Gases durch die Säule. Für die verwendeten Messdaten sollte gelten, dass sie mit der gleichen Gerätekonfiguration aufgenommen wurden (siehe Abschnitt 2.2.4). Folglich ist der Einfluss der Säule und die Durchflussgeschwindigkeit, die als Parameter beim Geräteaufbau bestimmt wird, idealerweise bei jeder Messung gleich. Da es sich um eine beheizte Säule handelt, wird auch deren Temperatur als konstant angenommen. Damit kann insgesamt die Retentionszeit als ein für alle Messungen vergleichbarer Wert angenommen werden. Ein Vergleich von Messungen mit verschiedenen vorgeschalteten Säulen erfordert eine Normalisierung bzgl. der Retentionszeit. Dies könnte z.B. durch eine Kalibrierung mittels Markeranalyten geschehen.

Die Driftzeit dt_i eines Analyten ist, wie unter Abschnitt 2.2.2 beschrieben, keine stoffspezifische Eigenschaft. Sie hängt von weiteren Parametern wie Temperatur und

Umgebungsdruck ab, die meist nicht durch den Geräteaufbau konstant gehalten werden. Erst die reduzierte Mobilität ermöglicht einen Vergleich zweier Peaks im Spektrum, da sie verschiedene Umwelteinflüsse der gemessenen Driftzeit ausgleicht. Der erste Ansatz zur Normalisierung der Driftzeitachse besteht also darin, die reduzierten Mobilitäten zu jedem Driftzeitpunkt zu berechnen. Dazu müssen diese Parameter zusätzlich bei der Analytenmessung erhoben werden. Hierbei spielt die Genauigkeit dieser einzelnen Parametermessungen eine entscheidende Rolle. Beispielsweise führt ein Messfehler der Temperatur von 1° K zu einer Abweichung um ca. $0,4$ %, ein Messfehler des Drucks von 10 hPa zu einer Abweichung um ca. 1 % bezüglich der Driftzeit.

Um die Probleme der Erhebung und der Güte zu vermeiden, wird eine weitere Methode zur Normierung der Driftzeitachse vorgeschlagen. Dazu werden die Positionen der Signalmessungen relativ zur Position des RIPs angegeben. Die Position des RIPs dt_{RIP} wird für das gesamtes Chromatogramm als diejenige Stelle gewählt, an der über alle Retentionszeiten am häufigsten das Maximum des Spektrums liegt (Modus des Spektrenmaximums). Die normierte Driftzeit $normdt_i$ ergibt sich daraus wie folgt:

$$normdt_i = \frac{dt_i}{dt_{RIP}} \ . \tag{3.1}$$

Damit hat der RIP jedes Chromatogramms eine normierte Driftzeit von $normdt_{RIP} = 1,0$.

Zur Einschätzung, welche Verbesserung die beiden Normalisierungsansätze der Driftzeit bringen, werden hier 44 Beispielmessdaten untersucht. Diese Messdaten zeigen alle genau einen einzelnen Peak in einem Driftzeitbereich von 26 ms bis 28 ms und einem Retentionszeitbereich von 185 s bis 310 s. Die Stelle maximaler Signalintensität in diesem Bereich kennzeichnet für jedes dieser Beispiele die Position dieses Peaks. Für die drei Darstellungsformen *„gemessene Driftzeit“*, *„reduzierte Mobilität“* und *„am RIP normierte Driftzeit“*, wird jeweils die mittlere Position über alle Beispiele bestimmt und für jedes Beispiel die Abweichung von diesem Mittel berechnet. Die Ergebnisse sind in Abbildung 3.2 dargestellt.

Ohne eine Normierung ergibt sich eine mittlere Abweichung von $1,53$ %, die reduzierte Mobilität verbessert die mittlere Abweichung immerhin auf $1,25$ %. Deutlich

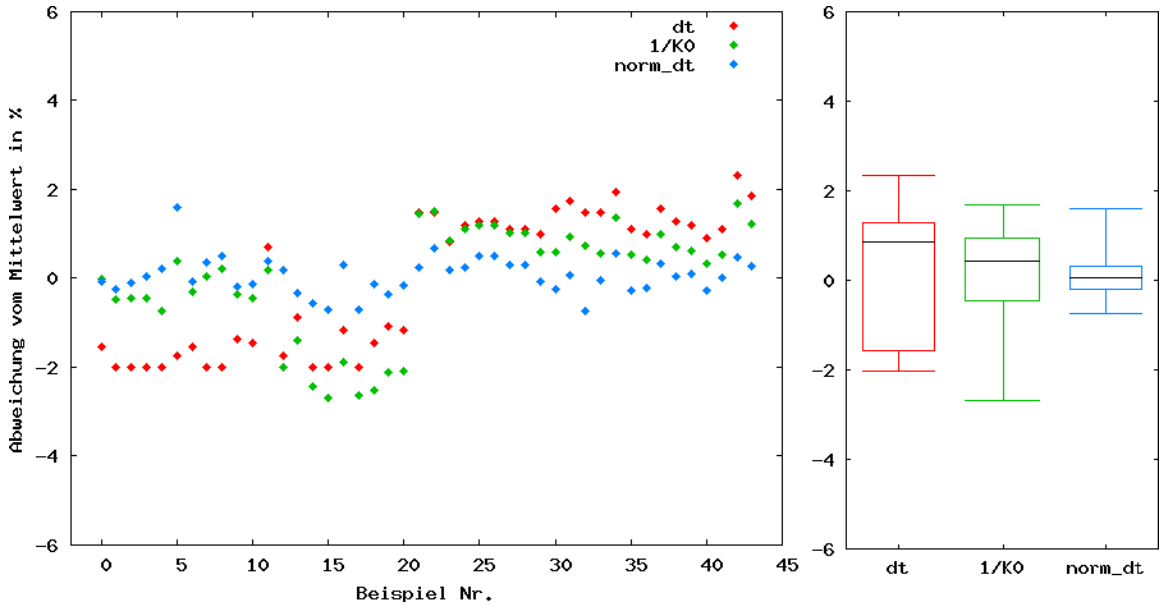


Abbildung 3.2: Vergleich der Schwankung der Peaklage eines festen Peaks für die Darstellungsformen Driftzeit, reduzierte Mobilität und am RIP normierte Driftzeit.

besser wird die Vergleichbarkeit der Messungen mit der am RIP normierten Driftzeit bei einer mittleren Abweichung von $0,53$ %.

Die Ergebnisse dieser Stichprobe und die Probleme bei der Erhebung der Einflussparameter sind die Gründe dafür, dass für die in dieser Diplomarbeit betrachteten Messdaten eine Normalisierung der Driftzeitachse am RIP durchgeführt wird.

In der Literatur sind hingegen oft die reduzierten Mobilitäten von Analyten angegeben. Es besteht aber eine einfache Möglichkeit, die am RIP normierte Driftzeit in diese international vergleichbare Größe umzurechnen. Dafür werden zunächst alle Einflussparameter nach Formel 3.4 in einem Faktor k zusammengefasst (vgl. Formeln 2.18 und 2.19).

$$K_0 = \frac{l_d^2}{U_d \cdot dt_i} \cdot \frac{T_0}{T} \cdot \frac{p}{p_0} \tag{3.2}$$

$$dt_i = \frac{normdt_i}{dt_{RIP}} \tag{3.3}$$

$$K_0 = \frac{l_d^2 \cdot dt_{RIP}}{U_d \cdot normdt_i} \cdot \frac{T_0}{T} \cdot \frac{p}{p_0} = \frac{1}{normdt_i} \cdot k \tag{3.4}$$

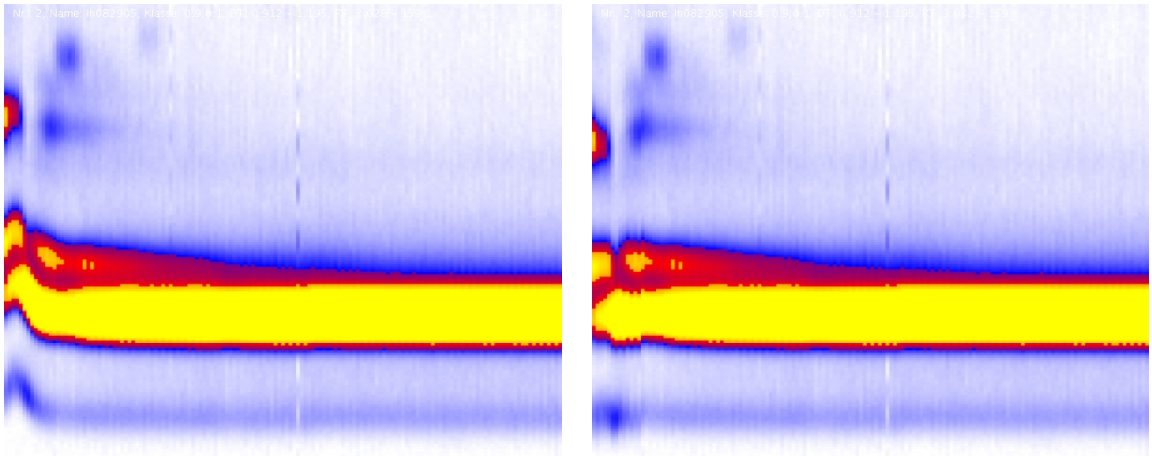
Dieser Faktor kann durch die Angabe eines Vergleichpunktes, für den der K_0-Wert $K_0^{(Vergleich)}$ und der normierte Driftzeitwert $normdt^{(Vergleich)}$ bekannt ist, berechnet werden. Für Aceton mit einem K_0 Wert von $1,82$ und einer normierten Driftzeit von $1,104$ ergibt sich:

$$k = K_0^{(Vergleich} \cdot normdt^{(Vergleich)} = 2,00928 \ . \tag{3.5}$$

Somit existiert die Möglichkeit der Umrechnung zwischen reduzierten Mobilitäten und normierten Driftzeitwerten.

3.2.2 Ausgleich des Feuchteeinflusses

Im Bereich niedriger Retentionszeiten fällt eine Verschiebung des RIPs und des Feuchtepeaks in Richtung höherer Driftzeiten auf. Diese Verschiebung sollte ausgeglichen werden, damit Peaks, über die Retentionszeit gesehen, die gleiche Position auf der Driftzeitachse haben. Dadurch wird ein späteres Auseinanderhalten überlagerter Peaks erleichtert. Die Abbildung 3.3 verdeutlicht den Einfluss der Feuchte.



(a) ohne Ausgleich (b) mit Ausgleich

Abbildung 3.3: Die feuchtebedingte Verschiebung der Spektren bei niedrigen Retentionszeiten in Richtung höherer Driftzeiten sollte ausgeglichen werden.

Zum Ausgleich werden die Signale jedes Spektrums derart verschoben, dass $normdt_{RIP}$

die neue Driftzeitposition des Maximums aus einem engen Bereich um die $normdt_{RIP}$-Position ist. Wichtig dabei ist es, den Bereich nicht zu groß zu wählen, da der RIP nicht zwingend die höchste Position des gesamten Spektrums darstellt.

3.2.3 Hervorhebung der Peaks

Um die Peaks der Chromatogramme vom Rauschbereich abzusetzen, wird für alle Spektren eine Basislinienkorrektur als Transformation der Signalintensität durchgeführt. Dies entspricht der Anpassung der Matrix $\mathbf{M}$ (siehe Abschnitt 2.2.4).

Eine einfache Möglichkeit dafür ist es, die Signalintensität jedes Messpunktes $z_{i,j}$ gleichmäßig um einen absoluten Wert c zu senken, so dass die Signale im Rauschbereich um Null schwanken.

$$\mathbf{M}' = \mathbf{M} - c \cdot \mathbf{1} \cdot \mathbf{1}^T \tag{3.6}$$

Um c zu bestimmen wird die Tatsache ausgenutzt, dass vor dem Feuchtepeak an Position $dt_{i_{Feuchte}}$ kein Peak im Spektrum auftritt. Der Wert c ergibt sich als Mittelwert eines ausgewählten Bereiches (Driftzeit $[dt_{i_1}; dt_{i_2}]$, Retentionszeit $[rt_{j_1}; rt_{j_2}]$ mit $0 \leq i_1 < i_2 < i_{Feuchte} < 1999$, $0 \leq j_1 < j_2 \leq 500$).

$$c = \sum_{i=i_1}^{i_2} \sum_{j=j_1}^{j_2} \frac{z_{i,j}}{(i_2 - i_1 + 1) \cdot (j_2 - j_1 + 1)} \tag{3.7}$$

Im Ausläuferbereich des RIPs sind Peaks durch das flache Abfallen der Spektrenkurven (das sogenannte „RIP-Tailing“) schwer zu identifizieren. Eine verbesserte Hervorhebung dieser Peaks wird durch eine Basislinienkorrektur auf Spektrenebene erreicht. Dafür wird von jedem Spektrum $\mathbf{S}_j$ ein für die Messung konstantes Korrekturspektrum $\mathbf{S}^{Korrektur}$ subtrahiert.

$$\mathbf{S}'_j = \mathbf{S}_j - \mathbf{S}^{Korrektur} \text{ , für } i \in \{0,\dots,500\} \tag{3.8}$$

Dabei ergibt sich der i-te Wert des Korrekturspektrums $\mathbf{S}_i^{Korrektur}$, $i \in \{0,\dots,1999\}$ als Median der Intensitäten $z_{i,j}$ zu Driftzeit dt_i über die Retentionszeit.

$$\mathbf{S}_i^{Korrektur} = Median\{z_{i,0},\dots,z_{i,500}\} \tag{3.9}$$

Anstatt des Medians den Mittelwert über die Retentionszeit zu berechnen führt zu keinem guten Ergebnis, da die Mittelwertbildung zu anfällig auf Peaks oder Ausreißer im Signal reagiert.

Die folgenden Abbildungen zeigen die Daten einer beispielhaften Atemluftmessung vor der Basislinienkorrektur (Abb. 3.4), nach der Korrektur um eine Konstante (Abb. 3.5) und nach der Korrektur um ein Korrekturspektrum (Abb. 3.6). Die erzielte Verbesserung der Darstellung ist eindeutig zu erkennen.

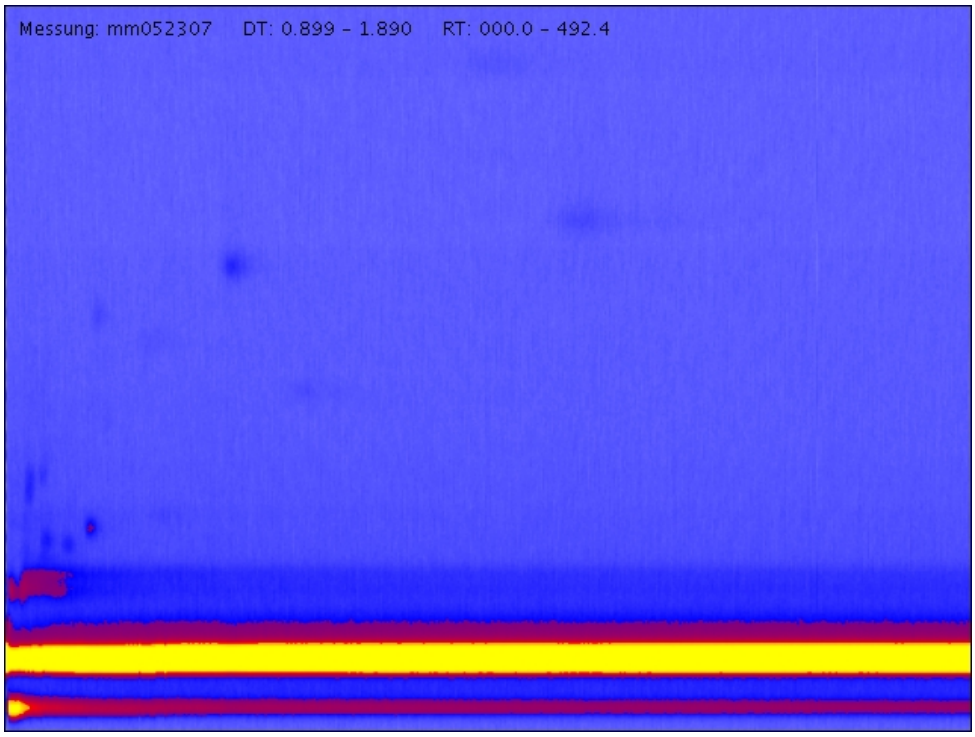


Abbildung 3.4: Die Messdaten ohne Basislinienkorrektur

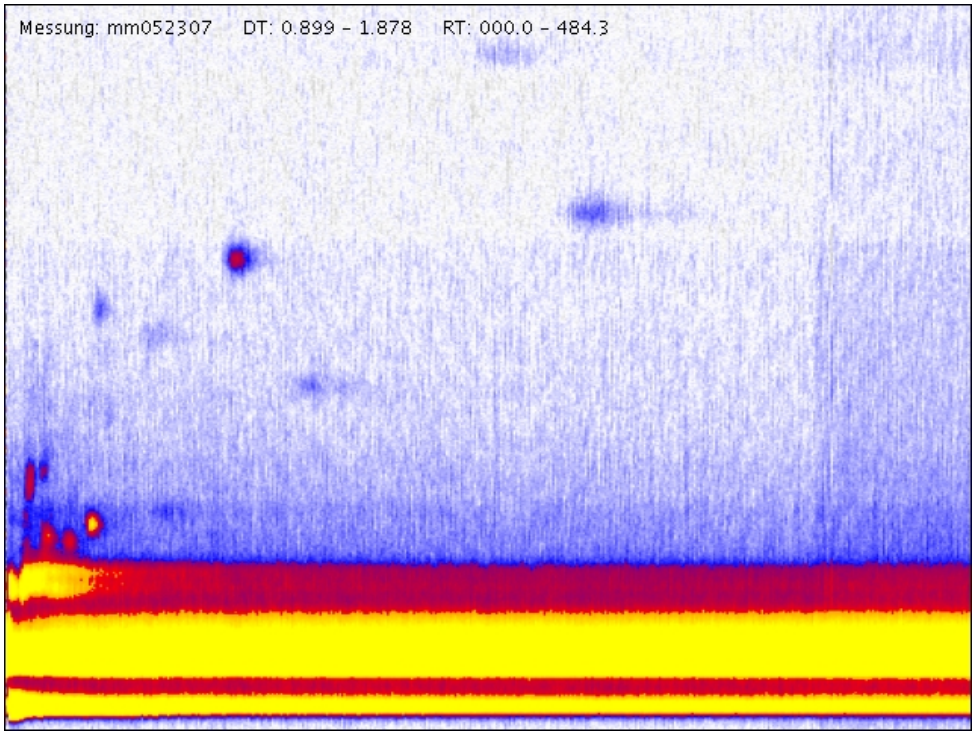


Abbildung 3.5: Die Messdaten nach der Korrektur um eine Konstante, die Signalintensität ist um den Faktor 7 verstärkt

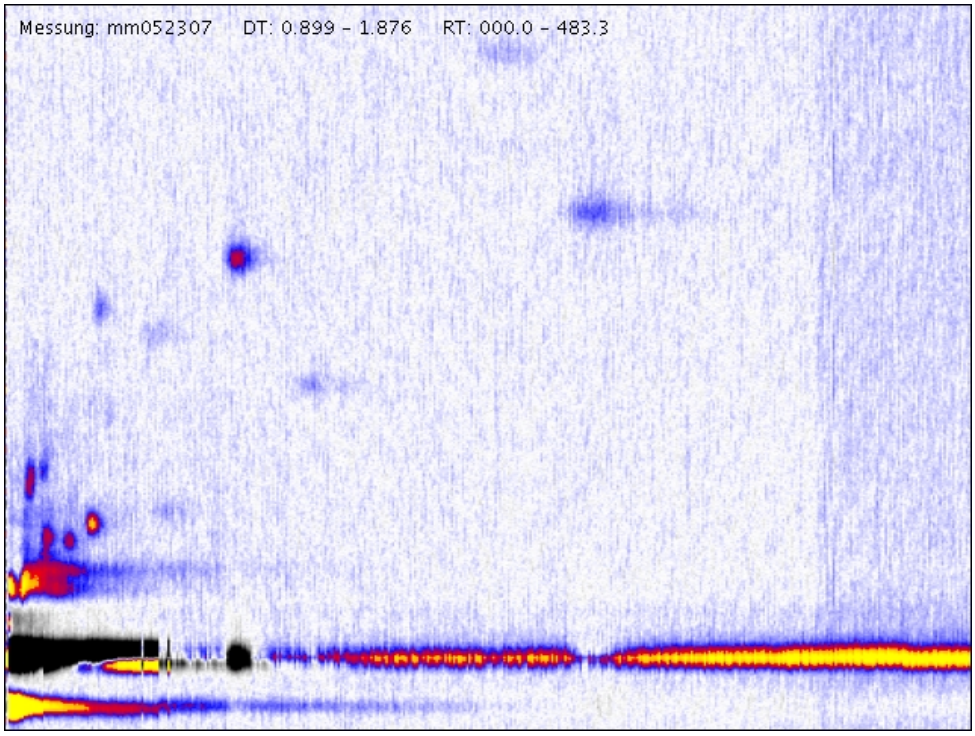


Abbildung 3.6: Die Messdaten nach der Korrektur um ein Korrekturspektrum, die Signalintensität ist um den Faktor 7 verstärkt

3.3 Extraktion von Merkmalen

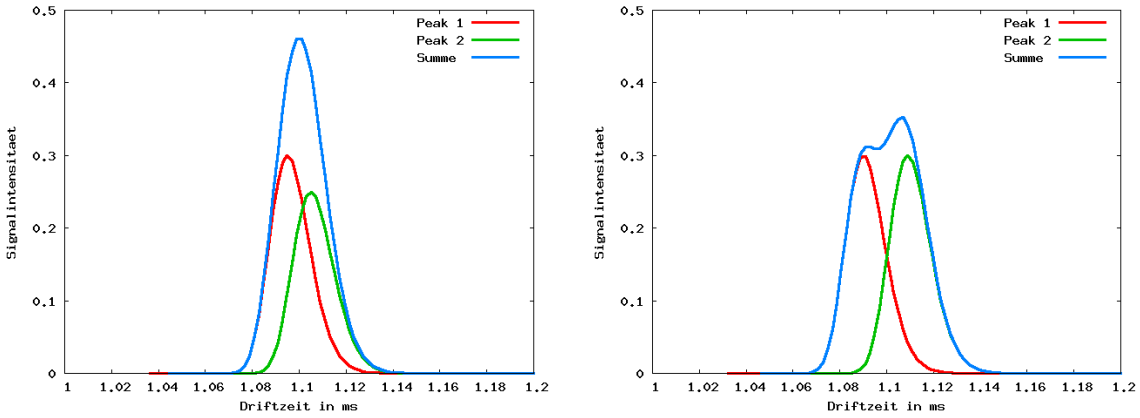
Nachdem die Vergleichbarkeit der Chromatogramme durch die Datennormalisierung sichergestellt ist, folgt die Extraktion von Merkmalen aus den Chromatogrammen.

Bei der manuellen Interpretation der Chromatogramme wird die Datenmatrix häufig in Form einer Heatmap dargestellt und auf auffällige Signalausschläge untersucht. Diese Signalausschläge oder Peaks treten als farbige, meist ovale Flecken in der Heatmap auf. Als Lagebeschreibung wird die Position des Maximums eines Peaks angegeben, als Intensitätbeschreibung die maximale Höhe oder das Raumvolumen. Eine Liste der vorkommenden Peaks stellt eine gute Merkmalsbeschreibung eines Chromatogrammes dar.

Ziel der automatischen Extraktion von Merkmalen ist es, für ein gegebenes Chromatogramm zu entscheiden, welche Peaks in welcher Intensität vorliegen. Prinzipiell gibt es dazu zwei mögliche Vorgehensweisen: Die erste ist eine unabhängige Untersuchung verschiedener Chromatogramme auf auftretende Peaks. Anschließend müssen die Peaklisten der einzelnen Chromatogramme verglichen werden, um Peaks des gleichen Analyten zu identifizieren. Die andere Möglichkeit besteht darin, für jeden bestimmten Peak das Vorkommen in den verschiedenen Chromatogrammen zu prüfen.

Für diese Diplomarbeit wird die zweite Methode gewählt. Der Hauptgrund ist eine Schwierigkeit bei der Entdeckung von Peaks. Peaks mit ähnlichen Drift- und Retentionszeiten überlagern sich gegenseitig. Abbildung 3.7 verdeutlicht dies an künstlich erzeugten Peaks. Liegen zwei oder mehr Peakmaxima eng beieinander, sieht das Gesamtsignal aus wie ein größerer Peak Abb. 3.7(a), oder die lokalen Maxima des Gesamtsignals beschreiben nicht die eigentlichen Peaklagen Abb. 3.7(b).

Eine Idee zur Lösung dieses Problems ist der Aufbau einer Peakdatenbank. Dafür werden zunächst für jeden Peak, der in die Datenbank aufgenommen werden soll, Beispiele gesammelt, in denen der Peak ohne Überlagerungen in verschiedenen Intensitäten



(a) Das Gesamtsignal besteht aus zwei Peaks.

(b) Die lokalen Maxima des Gesamtsignals entsprechen nicht den wirklichen Peaklagen.

Abbildung 3.7: Die Überlagerung von Peaks erschwert deren Identifizierung bzgl. Anzahl, Intensität und Lage.

vorkommt. Aus diesen Beispielen wird eine Funktion erlernt, die das Aussehen der Peaks in Abhängigkeit von der Lage (Drift- und Retentionszeit) und der Intensität beschreibt. Weiter wird aus den Beispielen ein Toleranzfenster für die Lage jedes Peaks errechnet. Die Extraktion der Peaks eines neuen Chromatogramms beschränkt sich dann auf die Aufgabe, eine Kombination der Peaks der Datenbank mit richtig gewählten Intensitäten zu erstellen, die das Chromatogramm am besten „erklärt“. Daraus ergibt sich auch direkt der Merkmalsvektor eines Chromatogramms $\mathbf{x} = (x_1, \ldots, x_n)$, wobei x_i die Intensität des i-ten Peaks der Datenbank beschreibt. Die folgenden Abschnitte beschreiben die notwendigen Schritte zum Aufbau der Peakdatenbank detaillierter.

3.3.1 Aufbau einer Peakdatenbank mit Beispielen

Der erste Schritt beim Aufbau der Peakdatenbank besteht in der Identifizierung von Peaks und der Sammlung von Beispielen zu diesen Peaks. Dies geschieht durch ein anwendergeführtes, halbautomatisiertes Verfahren, welches in die Software integriert ist, die im Rahmen dieser Diplomarbeit entwickelt wurde. Zunächst stellt ein Experte eine Sammlung von möglichst fehlerfreien Chromatogrammen (die Menge B) zusammen. Diese werden nach den Methoden aus Kapitel 3.2 vorverarbeitet.

Aus einem dieser normierten Chromatogramme wird ein Peak P_p gewählt, der neu in die Datenbank aufgenommen werden soll. Dazu wird durch je zwei Werte für die Drift- und Retentionszeit $normdt_{p_{i1}}$, $normdt_{p_{i2}}$ und $rt_{p_{j1}}$, $rt_{p_{j2}}$ ein rechteckiger Bereich angegeben in dem sich der Peak P_p befindet, und ein Grenzwert festgelegt, welche Intensität ein Beispielpeak mindestens haben sollte. Die hier entwickelte Software durchsucht die Sammlung B der Chromatogramme im angegeben Fensterbereich und zeigt diejenigen Fenster an, deren maximale Intensität mindestens dem angegebenen Grenzwert entspricht. Die Abbildung 3.8 verdeutlicht dies an einem Beispiel.

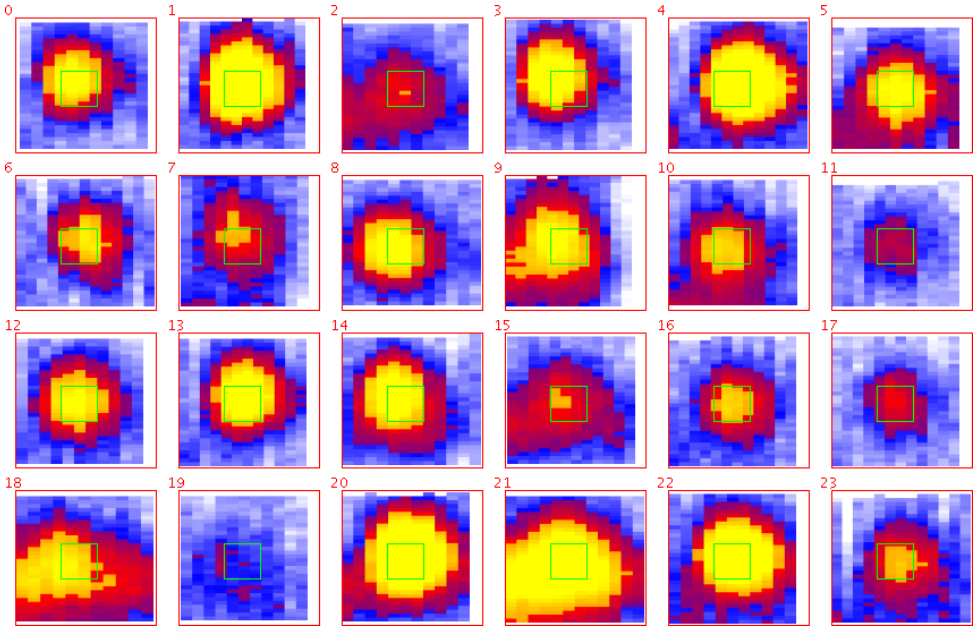


Abbildung 3.8: Für einen Driftzeitbereich $[1,15; 1,20]$ und einen Retentionszeitbereich $[37; 49]$ wird in 24 von 30 Beispielen einer Menge B der Peak mit einer maximalen Intensität $\geq 0,03$ gefunden. Das grüne Fenster zeigt den Toleranzbereich der Peaklage. Einige Beispiele weisen eine Störung (z.B. Nummer 18) oder Überlagerung mit Randbereichen anderer Peaks (z.B Nummer 2, 15 und 21) auf.

Der Anwender wählt daraus diejenigen Beispiele aus, die er aus seiner Erfahrung für nicht überlagert und relativ ungestört erachtet. Diese so verbleibenden Beispiele $B^{(p)} \subseteq B$ dienen als Trainingsmenge zur Parameterfindung der Funktion, die diesen Peak beschreibt (siehe Abbildung 3.9).

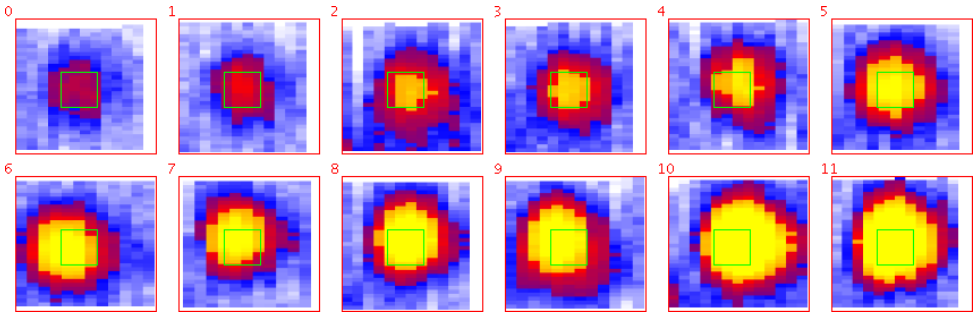


Abbildung 3.9: Von den Beispielen aus Abb. 3.8 werden diese 12 vom Anwender als die besten Repräsentanten des Peaks gewählt. Die Anzeige erfolgt sortiert bzgl. der maximalen Intensität im Fensterbereich.

Für jedes Beispiel $B_k^{(p)} \in B^{(p)}$ eines Peaks P_p wird die Lage des Peaks als Tupel $(normdt_{max}^{B_k^{(p)}}, rt_{max}^{B_k^{(p)}})$ und die Intensität durch $h_{max}^{B_k^{(p)}}$ beschrieben. Dabei entspricht

- $normdt_{max}^{B_k^{(p)}}$ der normierten Driftzeit der Position maximaler Intensität,
- $rt_{max}^{B_k^{(p)}}$ der Retentionszeit der Position maximaler Intensität und
- $h_{max}^{B_k^{(p)}}$ dem Wert der maximalen Intensität.

Das Toleranzfenster $\mathcal{T}_p = (\tau_p^{(dt_{min})}, \tau_p^{(dt_{max})}, \tau_p^{(rt_{min})}, \tau_p^{(rt_{max})})$ der Peaklage des Peaks P_p berechnet sich aus den Extremwerten der Lagen aller Beispielpeaks zu P_p, erweitert um die Toleranzwerte τ_{dt} bzw. τ_{rt}. Der Toleranzbereich der normierten Driftzeit des p-ten Peaks reicht von

$$\tau_p^{(dt_{min})} = \min_{k \in \{1,\ldots,|B^{(p)}|\}} \left\{ normdt_{max}^{B_k^{(p)}} \right\} - \tau_{dt} \quad \text{bis} \quad \tau_p^{(dt_{max})} = \max_{k \in \{1,\ldots,|B^{(p)}|\}} \left\{ normdt_{max}^{B_k^{(p)}} \right\} + \tau_{dt} \ , \tag{3.10}$$

der Toleranzbereich der Retentionszeit von

$$\tau_p^{(rt_{min})} = \min_{k \in \{1,\ldots,|B^{(p)}|\}} \left\{ rt_{max}^{B_k^{(p)}} \right\} - \tau_{rt} \quad \text{bis} \quad \tau_p^{(rt_{max})} = \max_{k \in \{1,\ldots,|B^{(p)}|\}} \left\{ rt_{max}^{B_k^{(p)}} \right\} + \tau_{rt} \ . \tag{3.11}$$

Der nächste Schritt besteht in der Angabe einer Funktion, die die Form der Peaks für eine gegebene Lage und Intensität beschreibt. Diese wird im folgenden Abschnitt entwickelt.

3.3.2 Entwurf einer Funktion zur Beschreibung von Peaks

Dieser Abschnitt beschreibt den Entwurf einer analytischen Funktion zur Beschreibung der dreidimensionalen Kurvenform eines Peaks innerhalb eines Chromatogrammes. Dabei reicht die Angabe der Lage (Driftzeit, Retentionszeit) und der maximalen Höhe eines Peaks aus, um seine Form zu beschreiben.

Vorüberlegungen zeigen, dass sich eine erweiterte Form der Dichtefunktion der logarithmischen Normalverteilung als Grundfunktion zur Beschreibung eignet. Im Anschluss werden die fixen Parameter dieser Funktion mittels eines Optimierungsverfahrens und einer Menge von Beispielen des Peaks bestimmt.

Vorüberlegung

Bei der Betrachtung von IMS-Spektren fällt auf, dass tendenziell die Peaks zu niedrigen Driftzeiten schmaler und höher erscheinen als die Peaks zu höheren Driftzeiten. Außerdem führen bei einer IMS-Messung höhere Spannungen zu einem früheren Auftreten der Peaks. Das führt zu einem Versuch den Reaktionsionen-Peak unter verschiedenen Driftspannungen zu messen, wobei alle weiteren Messparameter konstant gehalten werden. Die Abbildung 3.10 zeigt die Spektren der Messdaten bei unterschiedlichen Driftspannungen.

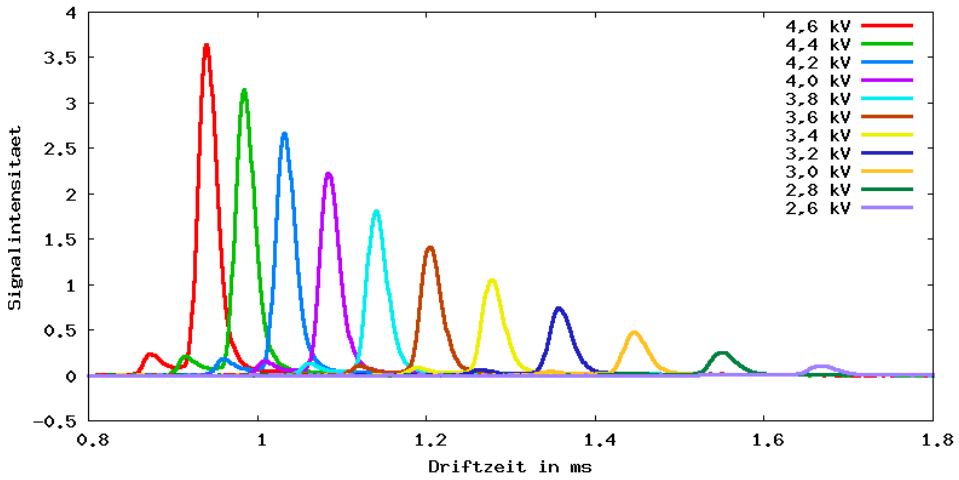


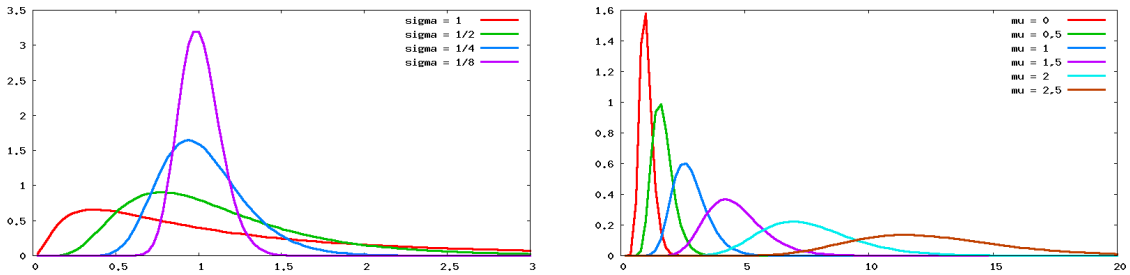
Abbildung 3.10: RIP bei verschiedenen Driftspannungen

Die einzelnen Peaks weisen eine Ähnlichkeit zu der Form der Dichtefunktion der logarithmischen Normalverteilung $lgn(x;\mu,\sigma)$ auf. Sie berechnet sich wie folgt:

$$lgn(x;\mu,\sigma) = \frac{1}{x \cdot \sigma\sqrt{2\pi}} \cdot e^{\frac{-(\log x-\mu)^2}{2\sigma^2}} \tag{3.12}$$

für $\sigma > 0$, $-\infty < \mu < \infty$ und $x > 0$.

In Abbildung 3.11(a) ist $lgn(x;\mu,\sigma)$ für $\mu = 0$ und verschiedene σ-Werte dargestellt, in Abbildung 3.11(b) variiert hingegen μ bei festem $\sigma = 0.25$.



(a) $\mu = 0$, verschiedene σ-Werte (b) verschiedene μ-Werte, $\sigma = 0.25$

Abbildung 3.11: Dichtefunktion der Log-Normalverteilung

Die Idee ist es, die Dichtefunktion der logarithmischen Normalverteilung derart zu erweitern, dass sie durch geeignete Parameterwahl die Form der Peaks möglichst gut beschreiben kann.

Anpassung der Log-Normalverteilung

Das erste Ziel ist es, eine Funktion $peak_{dt}(dt; dt_{max}, h_{max})$ zu erhalten, die unter Angabe der Driftzeit dt_{max} und der maximalen Höhe h_{max} eines Peaks den Kurvenverlauf des Peaks über die Driftzeit beschreibt.

Die Position des Maximums (Modus) x_{max} von $lgn(x;\mu,\sigma)$ liegt bei

$$x_{max} = e^{\mu-\sigma^2} \ . \tag{3.13}$$

Zur Anpassung an einen Peak ist es einfacher, das Maximum des Peaks x_{max} anzugeben und daraus μ zu berechnen. Dies geschieht durch eine Umrechnung der Gleichung (3.13), wodurch sich

$$\mu = \log(x_{max}) + \sigma^2 \tag{3.14}$$

ergibt.

Der nächste Schritt ist die Einführung eines Skalierungsfaktors a, mit dem die Höhe der Funktion angepasst wird. Hier ist es wünschenswert, a aus der maximalen Höhe h_{max} zu berechnen. Dafür wird die Eigenschaft genutzt, dass das Maximum lgn_{max} von $lgn(x; \mu, \sigma)$ folgendes beträgt:

$$lgn_{max} = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\sigma^2/2-\mu} \ . \tag{3.15}$$

Mit $h_{max} = a \cdot lgn_{max}$ ergibt sich für a, mit μ nach (3.14):

$$a = \frac{h_{max} \cdot \sigma\sqrt{2\pi}}{e^{\sigma^2/2-\mu}} \ . \tag{3.16}$$

Als Zwischenergebnis kann ein Peak über die Funktion

$$peak_\sigma(x; x_{max}, h_{max}) = a \cdot lgn(x; \mu, \sigma) \tag{3.17}$$

mit konstantem σ, μ nach (3.14) und a nach (3.16) beschrieben werden.

Der Versuch, diese Funktion den RIP-Messdaten anzupassen, zeigt, dass der Parameter σ alleine keine ausreichende Anpassung erlaubt. Die Abbildung 3.12 verdeutlicht dies. Mittels Evolutionsstrategie (vgl. Abschnitt 3.3.2) wurde hier $\sigma = 0,011$ als derjenige Wert bestimmt, bei dem die Summe des quadrierten Fehlers zwischen den Peakfunktionen und den jeweiligen Spektrenwerten des RIP am geringsten ist. Dabei ist zu beachten, dass die Anpassung jeweils bzgl. des größeren (rechten) Peaks jedes Spektrums durchgeführt wurde.

Es werden zwei zusätzliche Parameter b und c eingeführt, die eine weitere Anpassung ermöglichen. Dabei gibt

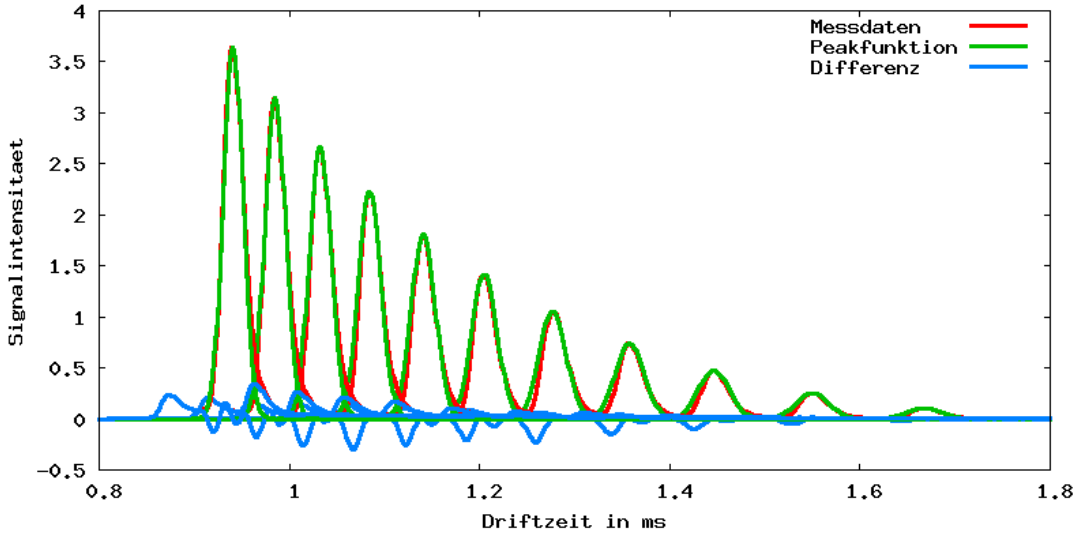


Abbildung 3.12: Anpassung der Funktion $peak_\sigma$ an den Kurvenverlauf des RIP

- b die Verschiebung auf der X-Achse und
- c den Skalierungsfaktor der X-Achse

an. Wird x durch $z(x)$ substituiert, mit

$$z(x) = (x - x_{max}) \cdot c + x_{max} - b \tag{3.18}$$

mit b und c fest gewählt, so ergibt sich die Funktion:

$$peak_{\sigma,b,c}(x; x_{max}, h_{max}) = a \cdot lgn(z(x); \mu, \sigma) \tag{3.19}$$

für feste Parameter σ, b, c, $\mu = \log(z(x_{max})) + \sigma^2$ (vgl. (3.14)) und a nach (3.16). Es bleibt die Eigenschaft erhalten, dass x_{max} das Maximum angibt. Dafür ist zu zeigen, dass

$$z(x_{max}) = e^{\mu - \sigma^2} \tag{3.20}$$

gilt.

Beweis:

$$z(x_{max}) = (x_{max} - x_{max}) \cdot c + x_{max} - b = x_{max} - b \tag{3.21}$$

$$e^{\mu-\sigma^2} = e^{(\log(x_{max}-b)+\sigma^2)-\sigma^2} = x_{max} - b \quad \blacksquare \tag{3.22}$$

Eine Anpassung der Funktion $peak_{\sigma,b,c}$ an den Kurvenverlauf des RIP, bei der die Parameter mittels einer Evolutionsstrategie (vgl. Abschnitt 3.3.2) zu $\sigma = 0,13$, $b = -1,45$ und $c = 27,65$ bestimmt werden, zeigt die Abbildung 3.13. Es ist eine wesentlich verbesserte Anpassung im Vergleich zu Abbildung 3.12 zu sehen.

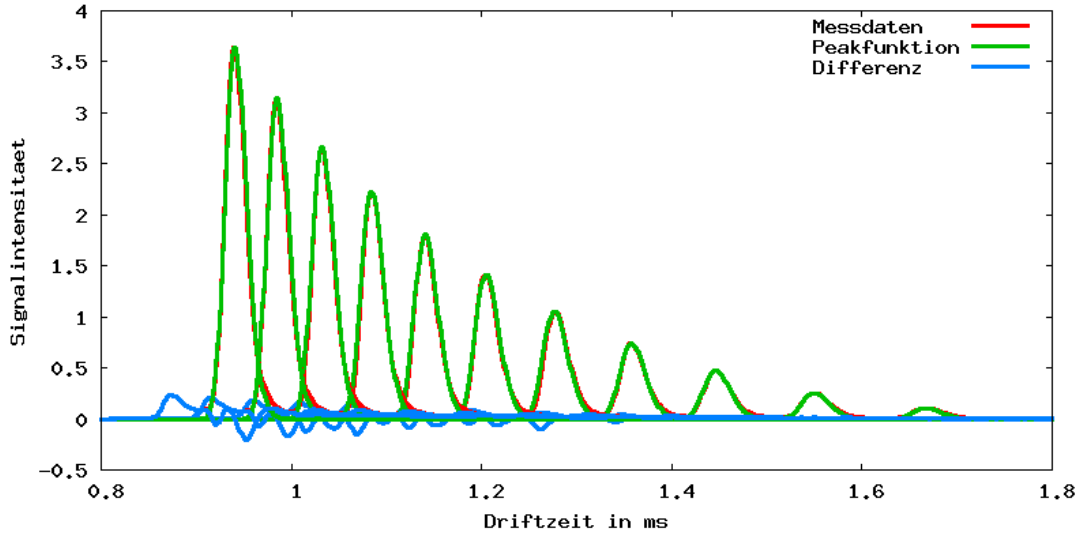


Abbildung 3.13: Anpassung der Funktion $peak_{\sigma,b,c}$ an den Kurvenverlauf des RIP

Daraus folgend wird die Funktion $peak_{dt}(dt, dt_{max}, h_{max}, \sigma_{dt}, b_{dt}, c_{dt})$ definiert, die den Kurvenverlauf eines Peaks unter der Angabe der Position des Maximums dt_{max} und der Höhe des Maximums h_{max} beschreibt:

$$peak_{dt}(dt; dt_{max}, h_{max}, \sigma_{dt}, b_{dt}, c_{dt}) = a_{dt} \cdot lgn(z_{dt}(x); \mu_{dt}, \sigma_{dt}) \tag{3.23}$$

mit

$$z_{dt}(x) = (x - dt_{max}) \cdot c_{dt} + dt_{max} - b_{dt} \tag{3.24}$$

$$\mu_{dt} = \log(z(dt_{max})) + \sigma_{dt}^2 \tag{3.25}$$

$$a_{dt} = \frac{h_{max} \cdot \sigma_{dt}\sqrt{2\pi}}{e^{\sigma_{dt}^2/2-\mu_{dt}}} \; . \tag{3.26}$$

Analog wird die Funktion $peak_{rt}(rt; rt_{max}, h_{max})$ aufgestellt, die unter Angabe der Retentionszeit rt_{max} und der maximalen Höhe h_{max} eines Peaks den Kurvenverlauf des Peaks über die Retentionszeit beschreibt:

$$peak_{rt}(rt; rt_{max}, h_{max}, \sigma_{rt}, b_{rt}, c_{rt}) = a_{rt} \cdot lgn(z_{rt}(x); \mu_{rt}, \sigma_{rt}) \tag{3.27}$$

mit

$$\begin{aligned} z_{rt}(x) &= (x - rt_{max}) \cdot c_{rt} + rt_{max} - b_{rt} && (3.28)\\ \mu_{rt} &= \log(z(rt_{max})) + \sigma_{rt}^2 && (3.29)\\ a_{rt} &= \frac{h_{max} \cdot \sigma_{rt}\sqrt{2\pi}}{e^{\sigma_{rt}^2/2 - \mu_{rt}}} \ . && (3.30) \end{aligned}$$

Insgesamt ergibt sich daraus die Funktion

$$\begin{aligned} &peak(dt, rt; dt_{max}, rt_{max}, h_{max}, \sigma_{dt}, b_{dt}, c_{dt}, \sigma_{rt}, b_{rt}, c_{rt})\\ &= peak_{dt}(dt; dt_{max}, h_{max}, \sigma_{dt}, b_{dt}, c_{dt}) \cdot peak_{rt}(rt; rt_{max}, h_{max}, \sigma_{rt}, b_{rt}, c_{rt})/h_{max} \end{aligned} \tag{3.31}$$

die unter Angabe der Driftzeit dt_{max}, der Retentionszeit rt_{max} und der maximalen Höhe h_{max} eines Peaks und der konstanten Vorgabe von σ_{dt}, b_{dt}, c_{dt}, σ_{rt}, b_{rt} und c_{rt} den dreidimensionalen Kurvenverlauf des Peaks beschreibt.

Parameterfindung

Dieser Abschnitt beschreibt die Suche nach den vorzugebenden Parametern σ_{dt}, b_{dt}, c_{dt}, σ_{rt}, b_{rt} und c_{rt} für einen bestimmten Peak der Peakdatenbank. Diese sollen mit Hilfe einer Evolutionsstrategie (siehe Abschnitt 2.3) und den vom Experten identifizierten Beispielen für diesen Peak ermittelt werden.

Als Einheit der Driftzeit wird im Folgenden die am RIP normierte Driftzeit gewählt, die Retentionszeit wird in Sekunden angegeben.

Ein Individuum $\mathbf{X}$[1] besteht aus einer möglichen Belegung der 6 Parameter:

$$\mathbf{X} = (x_1, \ldots, x_6) \tag{3.32}$$

mit

- x_1 entspricht der Belegung des Parameters σ_{dt}
- x_2 entspricht der Belegung des Parameters b_{dt}
- x_3 entspricht der Belegung des Parameters c_{dt}
- x_4 entspricht der Belegung des Parameters σ_{rt}
- x_5 entspricht der Belegung des Parameters b_{rt}
- x_6 entspricht der Belegung des Parameters c_{rt}

Als Elter werden diese Werte initial mit einer Kombination belegt, die in Tests für viele Peaks bereits eine gute Näherung liefert. Der initiale Wert des Parameters $\Sigma = (\sigma_1, \ldots, \sigma_6)$ wird von mir auf $(1, 1, 2, 1, 1, 2)$, die Anzahl der Nachkommen λ auf 100 festgelegt. Die Wahl des einzigen neuen Elters einer Generation folgt der „Komma-Strategie“. Das Abbruchkriterium wird von mir als das Erreichen der 100. Generation definiert ($(1, \lambda)$-ES) mit 100 Generationen). Diese Parameterwerte erwiesen sich in Tests als geeignet, die Aufgabe der Funktionsparameterbestimmung zu lösen.

Die Güte eines Individuums ist umgekehrt proportional zum mittleren quadratischen Fehler zwischen Beispielpeak und den Funktionswerten der Peakfunktion mit der aktuellen Belegung der Parametervariablen, gemittelt über alle Beispiele des Peaks. Seien die Bezeichnungen wie unter 3.3.1 gewählt, $B^{(p)}_{k_{i,j}}$ bezeichne den Intensitätswert des k-ten Beispiels für Peak P_p an Stelle ($normdt_i$, rt_j) und sei weiter (vgl. Formel 3.31)

$$\begin{aligned} &peak(dt, rt; dt_{max}, rt_{max}, h_{max}, \mathbf{X}) \\ &= peak(dt, rt; dt_{max}, rt_{max}, h_{max}, x_1, x_2, x_3, x_4, x_5, x_6) \ . \end{aligned} \tag{3.33}$$

[1] $\mathbf{X}$ beschreibt an dieser Stelle nicht die Menge der Merkmalsvektoren

Dann ergibt sich der Fehler E eines Individuums $\mathbf{X}$ als:

$$E(\mathbf{X}) = \sum_{k=1}^{|B|} \sum_{i=i_1}^{i_2} \sum_{j=j_1}^{j_2} \frac{\left(B_{k_{i,j}}^{(p)} - peak(normdt_i, rt_j; normdt_{max}^{B_k^{(p)}}, rt_{max}^{B_k^{(p)}}, h_{max}^{B_k^{(p)}}, \mathbf{X}) \right)^2}{|B| \cdot (i_2 - i_1 + 1) \cdot (j_2 - j_1 + 1)}. \quad (3.34)$$

Mit diesen Angaben wird die Evolutionsstrategie gestartet. Das letzte Elter-Individuum beschreibt die Wahl der festen Belegung der Funktionsparameter, womit jetzt die Formel 3.31 zur Berechnung des erwarteten Peakverlaufs angewendet werden kann.

3.3.3 Extraktion mit Hilfe der Peakdatenbank

Dieser Abschnitt beschreibt die Generierung des Merkmalsvektors mit Hilfe der Peakdatenbank. Dazu wird die Lösung der Aufgabe, eine „chromatogrammerklärende“ Kombination der Peaks der Datenbank zu finden, vorgestellt. Diese Lösung ist in der hier entwickelten Software implementiert.

Ausgangspunkt ist eine Peakdatenbank, wie sie im vorigen Abschnitt beschrieben wurde. Es wird folgende Notation verwendet:

- n : Anzahl Peaks in der Datenbank
- P_i : i-ter Peak der Datenbank, $i \in \{1, \ldots, n\}$
- $param_i = (\sigma_{dt}^{(i)}, b_{dt}^{(i)}, c_{dt}^{(i)}, \sigma_{rt}^{(i)}, b_{rt}^{(i)}, c_{rt}^{(i)})$: Funktionsparameter des i-ten Peaks (vgl. Formel 3.31)
- $\mathcal{W}_i = (dt_{i_{min}}, dt_{i_{max}}, rt_{i_{min}}, rt_{i_{max}})$: Fensterbereich der Lage des i-ten Peaks
- $\mathcal{T}_i = (\tau_i^{(dt_{min})}, \tau_i^{(dt_{max})}, \tau_i^{(rt_{min})}, \tau_i^{(rt_{max})})$: Toleranzfenster des i-ten Peakmaximums (vgl. Formeln 3.10 und 3.11)

Die Idee ist es, zu jedem Peak P_i der Datenbank die Chromatogrammdaten der einzelnen Fensterbereiche $\mathcal{W}_i$ durch die Funktion $peak_i(dt, rt; dt_{max}^{(i)}, rt_{max}^{(i)}, h_{max}^{(i)}, param_i)$ (vgl. Formel 3.3.2) möglichst gut abzubilden. Dazu werden geeignete Werte für $dt_{max}^{(i)}$, $rt_{max}^{(i)}$ und $h_{max}^{(i)}$ gesucht. Falls sich die Fensterbereiche von Peaks der Datenbank überschneiden,

könnte im Schnittbereich eine Überlagerung der Peaks liegen. Diese Bereiche werden daher zusammen durch die Summe der betroffenen $peak_i$ Funktionen „gefittet“. Die Anpassung der Funktionsparameter $dt_{max}^{(i)}$, $rt_{max}^{(i)}$ und $h_{max}^{(i)}$ an die Chromatogrammdaten erfolgt, wie schon die Parameterfindung der Peakfunktion, mittels einer Evolutionsstrategie, die im Folgenden genauer beschrieben wird.

Zunächst werden die Vorgaben der Evolutionsstrategie für die Anpassung der Peaks P_p, $p \in \{1, \ldots, n\}$ beschrieben, deren Fensterbereich $\mathcal{W}_p$ sich nicht mit anderen Fensterbereichen $\mathcal{W}_{p'}$ der Peaks $P_{p'}$, $p' \in \{1, \ldots, n\} \setminus \{p\}$ überschneidet.

Ein Individuum $\mathbf{X}$ besteht hier aus einer möglichen Belegung der drei Parameter:

$$\mathbf{X} = (x_1, x_2, x_3) \ . \tag{3.35}$$

mit

- x_1 entspricht der Belegung des Parameters $dt_{max}^{(p)}$,
- x_2 entspricht der Belegung des Parameters $rt_{max}^{(p)}$,
- x_3 entspricht der Belegung des Parameters $h_{max}^{(p)}$.

Initial werden diese Werte von mir mit den Mittelwerten des zugehörigen Toleranzbereiches bzw. einem klein gewählten Wert für die Höhe des Peaks belegt:

$$x_1 = (\tau_p^{(dt_{min})} + \tau_p^{(dt_{max})})/2 \tag{3.36}$$

$$x_2 = (\tau_p^{(rt_{min})} + \tau_p^{(rt_{max})})/2 \tag{3.37}$$

$$x_3 = 0,05 \ . \tag{3.38}$$

Der initiale Wert des Parameters $\Sigma = (\sigma_i)_{i \in \{1,2,3\}}$ wird von mir auf $\sigma_i = 0,01$ für alle $i \in \{1, 2, 3\}$ festgelegt, die Anzahl der Nachkommen λ beträgt 100. Die Wahl des einzigen neuen Elters einer Generation folgt der „Komma-Strategie“. Das Abbruchkriterium wird von mir als das Erreichen der 100. Generation definiert ($(1, \lambda)$-ES) mit 100 Generationen). Diese Parameterwahl stellte sich in Tests als geeignet heraus, diese Aufgabe des

Peakfittings zu lösen.

Die Güte eines Individuums ist umgekehrt proportional zum mittleren quadratischen Fehler $E(\mathbf{X})$ zwischen den Chromatogrammdaten im Fenster $\mathcal{W}_p$ und den Funktionswerten der Peakfunktion $peak_p(dt, rt; x_1, x_2, x_3, param_p)$. Gibt $\mathbf{M}_{i,j}$ die Signalintensität der anzupassenden Chromatogrammdaten zur Driftzeit dt_i und Retentionszeit rt_j an, und ist weiter

$$\begin{aligned} i_1 &= \min\{i \in \{0, \ldots, 500\} | \tau_p^{(dt_{min})} \leq dt_i \leq \tau_p^{(dt_{max})}\} && (3.39)\\ i_2 &= \max\{i \in \{0, \ldots, 500\} | \tau_p^{(dt_{min})} \leq dt_i \leq \tau_p^{(dt_{max})}\} && (3.40)\\ j_1 &= \min\{j \in \{0, \ldots, 1999\} | \tau_p^{(rt_{min})} \leq rt_j \leq \tau_p^{(rt_{max})}\} && (3.41)\\ j_2 &= \max\{j \in \{0, \ldots, 1999\} | \tau_p^{(rt_{min})} \leq rt_j \leq \tau_p^{(rt_{max})}\} \ , && (3.42) \end{aligned}$$

so berechnet sich der Fehler wie folgt:

$$E(\mathbf{X}) = \sum_{i=i_1}^{i_2} \sum_{j=j_1}^{j_2} \frac{(\mathbf{M}_{i,j} - peak_p(dt_i, rt_j; x_1, x_2, x_3, param_p))^2}{(i_2 - i_1 + 1) \cdot (j_2 - j_1 + 1)}. \quad (3.43)$$

Die Vorgaben der Evolutionsstrategie für sich überschneidende Fensterbereiche ergeben sich auf ähnliche Weise. Dafür seien $p_1, \ldots, p_K$ die Indizes der K Peaks, deren Fensterbereiche sich überschneiden. Ein Individuum $\mathbf{X}$ besteht in diesem Fall aus einer möglichen Belegung der $3 \cdot K$ Parameter:

$$\mathbf{X} = (x_{1,1}, x_{1,2}, x_{1,3}, \ldots, x_{K,1}, x_{K,2}, x_{K,3}) \ . \quad (3.44)$$

mit

- $x_{k,1}$ entspricht der Belegung des Parameters $dt_{max}^{(p_k)}$ für alle $k \in \{1, \ldots, K\}$,
- $x_{k,2}$ entspricht der Belegung des Parameters $rt_{max}^{(p_k)}$ für alle $k \in \{1, \ldots, K\}$,
- $x_{k,3}$ entspricht der Belegung des Parameters $h_{max}^{(p_k)}$ für alle $k \in \{1, \ldots, K\}$.

Initial werden diese Werte von mir analog zur Initialisierung auf Seite 44 mit

$$x_{k,1} = (\tau_{p_k}^{(dt_{min})} + \tau_{p_k}^{(dt_{max})})/2 \tag{3.45}$$

$$x_{k,2} = (\tau_{p_k}^{(rt_{min})} + \tau_{p_k}^{(rt_{max})})/2 \tag{3.46}$$

$$x_{k,3} = 0,05 \tag{3.47}$$

für alle $k \in \{1, \ldots, K\}$ belegt. Weiter wird $\Sigma = (\sigma_i)_{i \in \{1,\ldots,3 \cdot K\}}$ auf $\sigma_i = 0,01$ für alle $i \in \{1, \ldots, 3 \cdot K\}$ festgelegt, die Anzahl der Nachkommen λ beträgt 100. Die Wahl des einzigen neuen Elters einer Generation folgt der „Komma-Strategie“. Das Abbruchkriterium wird als das Erreichen der $100 \cdot K$-ten Generation definiert ($(1, \lambda)$-ES) mit $100 \cdot K$ Generationen). Diese Parameterwahl stellte sich in Tests als geeignet heraus, die Aufgabe des Peakfittings überlagerter Peaks zu lösen.

Die Indizes i_1, i_2, j_1 und j_2 des Fensterbereiches berechnen sich wie folgt:

$$i_1 = \min\{i \in \{0, \ldots, 500\} | \min_{k \in \{1,\ldots,K\}} \{\tau_{p_k}^{(dt_{min})}\} \leq dt_i \leq \max_{k \in \{1,\ldots,K\}} \{\tau_{p_k}^{(dt_{max})}\}\} \tag{3.48}$$

$$i_2 = \max\{i \in \{0, \ldots, 500\} | \min_{k \in \{1,\ldots,K\}} \{\tau_{p_k}^{(dt_{min})}\} \leq dt_i \leq \max_{k \in \{1,\ldots,K\}} \{\tau_{p_k}^{(dt_{max})}\}\} \tag{3.49}$$

$$j_1 = \min\{j \in \{0, \ldots, 1999\} | \min_{k \in \{1,\ldots,K\}} \{\tau_{p_k}^{(rt_{min})}\} \leq rt_j \leq \max_{k \in \{1,\ldots,K\}} \{\tau_{p_k}^{(rt_{max})}\}\} \tag{3.50}$$

$$j_2 = \max\{j \in \{0, \ldots, 1999\} | \min_{k \in \{1,\ldots,K\}} \{\tau_{p_k}^{(rt_{min})}\} \leq rt_j \leq \max_{k \in \{1,\ldots,K\}} \{\tau_{p_k}^{(rt_{max})}\}\}, \tag{3.51}$$

womit sich der Fehler $E(\mathbf{X})$ eines Individuums $\mathbf{X}$ ergibt:

$$E(\mathbf{X}) = \sum_{i=i_1}^{i_2} \sum_{j=j_1}^{j_2} \frac{\left(\mathbf{M}_{i,j} - \sum_{k=1}^{K} peak_{p_k}(dt_i, rt_j; x_{k,1}, x_{k,2}, x_{k,3}, param_{p_k})\right)^2}{(i_2 - i_1 + 1) \cdot (j_2 - j_1 + 1)}. \tag{3.52}$$

Für alle Peaks P_p der Datenbank, deren Fensterbereiche sich nicht mit denen anderer Peaks überschneiden, ergeben sich die Peakfunktionsparameter aus dem besten Individuum $\mathbf{X}$ der Evolutionsstrategie mit den zuerst beschriebenen Vorgaben wie folgt:

$$dt_{max}^{(p)} = x_1 \tag{3.53}$$

$$rt_{max}^{(p)} = x_2 \tag{3.54}$$

$$h_{max}^{(p)} = x_3\ . \tag{3.55}$$

Für alle Peakmengen $\{P_{p_1}, \ldots, P_{p_K}\}$ der Peaks der Datenbank, deren Fensterbereiche sich überschneiden, ergeben sich die Peakfunktionsparameter aus dem besten Individuum $\mathbf{X}$ der Evolutionsstrategie mit den zweiten beschriebenen Vorgaben wie folgt:

$$dt_{max}^{(p_k)} = x_{p_k,1} \text{ , für } k \in \{1, \ldots, K\} \tag{3.56}$$

$$rt_{max}^{(p_k)} = x_{p_k,2} \text{ , für } k \in \{1, \ldots, K\} \tag{3.57}$$

$$h_{max}^{(p_k)} = x_{p_k,3} \text{ , für } k \in \{1, \ldots, K\} \text{ .} \tag{3.58}$$

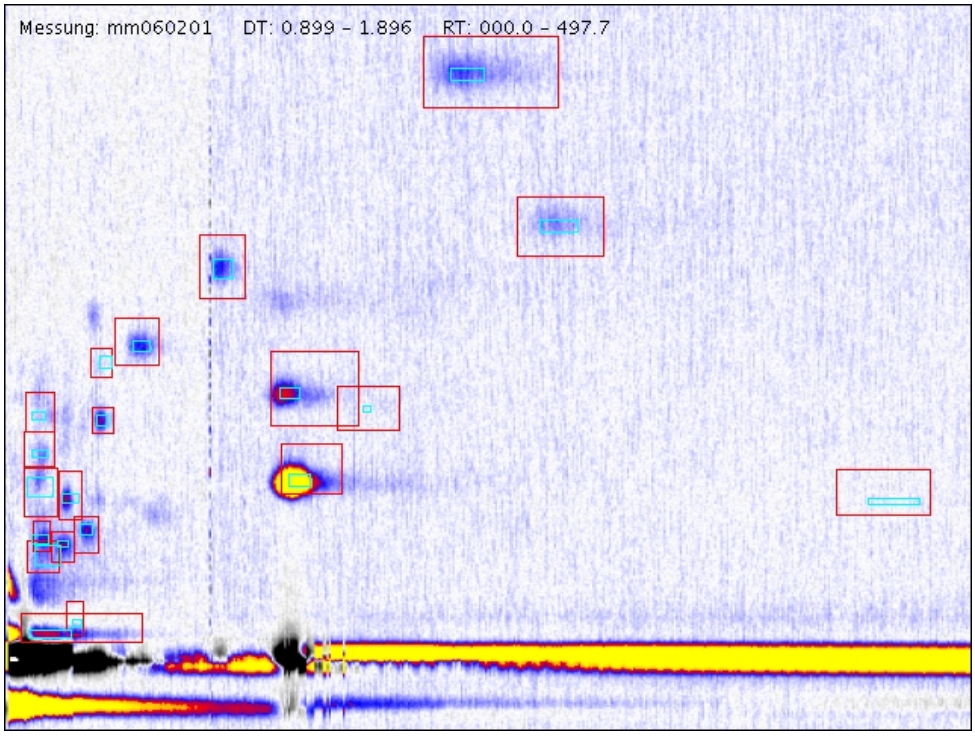
Für die Klassifikation sind nur die einzelnen Peakhöhen von Bedeutung, der Unterschied der Peaklagen eines bestimmten Peaks in verschiedenen Chromatogrammdaten ist durch Rauschen erklärbar. Damit ergibt sich der Merkmalsvektor $\mathbf{x}$ eines Chromatogramms bei einer vorgegebener Peakdatenbank mit n Peaks zu:

$$\mathbf{x} = (h_{max}^{(1)}, \ldots, h_{max}^{(n)}) \text{ .} \tag{3.59}$$

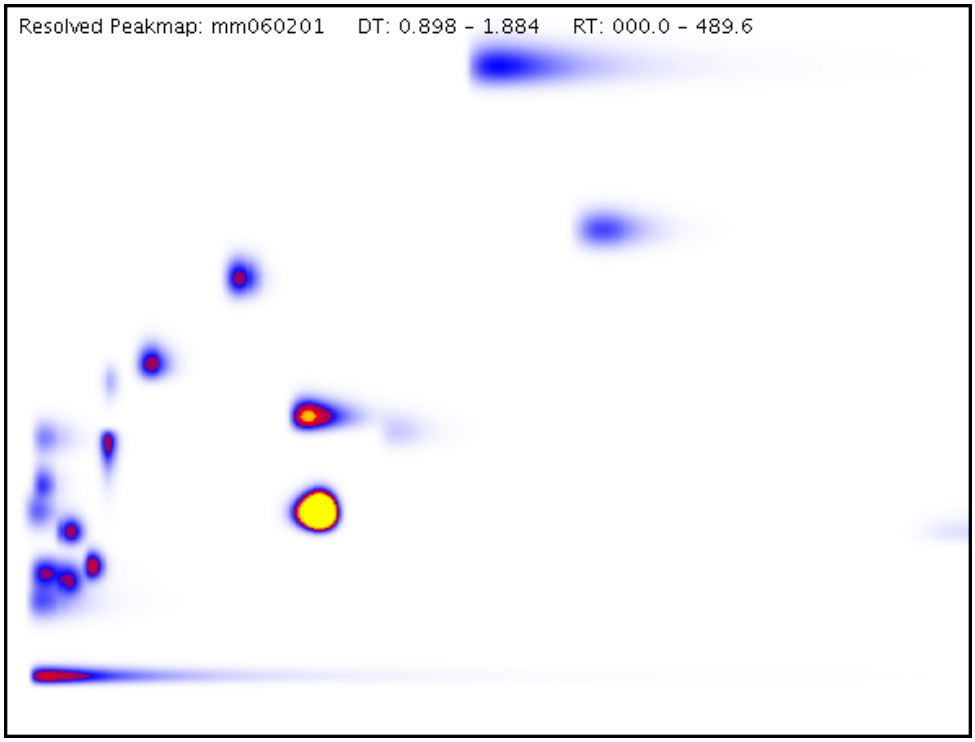
Zur Demonstration und Zusammenfassung, wie die Merkmalsextraktion mit realen Daten funktioniert, folgt ein Beispiel. Es wird zunächst eine Peakdatenbank, wie unter Abschnitt 3.3.1 beschrieben, erstellt. Dafür werden 40 vorverarbeitete Ausatemluft-Messdaten auf Peaks untersucht, wobei 20 Peakbereiche identifiziert werden, die mit unterschiedlich vielen, als relativ ungestört erachteten Beispiel-Messdaten in die Datenbank aufgenommen werden. Zu jedem der 20 Peaks werden die Funktionsparameter, wie unter Abschnitt 3.3.2 beschrieben, bestimmt.

Die Abbildung 3.14(a) zeigt ein Chromatogramm, dessen Merkmalsvektor bestimmt werden soll. Die rechteckigen Markierungen zeigen die Fensterbereiche der Peaks der Datenbank. Für diese Fensterbereiche, die sich teilweise überschneiden, werden mit dem oben beschriebenen Verfahren die Lage- und Höhenparameter der Peakfunktionen bestimmt. Die Summe aller 20 Peakfunktionen mit ihren jeweiligen Parametern ergibt in Heatmapdarstellung das Bild aus Abbildung 3.14(b).

Hier kann verglichen werden, wie gut die erkannten Peaks zu den Messdaten passen und wo eventuelle Ungenauigkeiten der Peakfunktionen auftreten.



(a) Messdaten mit Markierungen der Peakbereiche der Peakdatenbank



(b) Erkannte Peaks der Datenbank

Abbildung 3.14: Peakerkennung mittels Peakfunktionen der Datenbank

Kapitel 4

Entwurf einer Methode zur Klassifikation

4.1 Überblick

Dieses Kapitel behandelt den Entwurf einer Methode zur Klassifikation der IMS-Messdaten anhand der Merkmalsvektoren, die durch das im vorigen Kapitel beschriebene Verfahren generiert wurden. Es folgt eine Beschreibung der Struktur der vorhandenen Daten und eine anschließende Formulierung der Anforderungen an das Klassifikationsverfahren. Danach wird das für diese Arbeit gewählte Verfahren erläutert.

Die grundlegenden Daten des Klassifikationsverfahrens beim Training, Test und Einsatz sind durch Mengen von Merkmalsvektoren repräsentiert. Die Merkmalsvektoren enthalten hier ausschließlich numerische Attribute, die die Höhen der jeweiligen Peaks in den Messdaten beschreiben. Die Anzahl der Attribute entspricht der Anzahl der in der Peakdatenbank gespeicherten Peaks. Klassifiziert wird bezüglich einer endlichen Anzahl kategorischer Klassen, z.B. *„Proband hat Lungenkrebs“* und *„Proband hat keinen Lungenkrebs“*. Zum Training und zum Test wird eine Menge vorklassifizierter Merkmalsvektoren bereitgestellt, bzw. werden zur Validierung die Beispielinstanzen wie unter 2.1.5 mit Hilfe der „leave-one-out“ Methode in Trainings- und Testmenge unterteilt.

Die Anforderung an das Verfahren ist es, mit Hilfe der Trainingsinstanzen ein geeignetes Klassifikationsmodell zu erstellen, anhand dessen die vorgegebenen Klassen unterschieden werden können. Damit die Klassifikation von unbekannten Instanzen gelingt, sollte dieses Modell so allgemein wie möglich und so speziell wie nötig sein. Dafür sollte unter anderem eine geeignete Untermenge der aussagekräftigsten Attribute bzgl. der Klassifikationsaufgabe gewählt werden. Es handelt sich teilweise um Klassifikationsaufgaben, deren Ergebnis von großer Tragweite ist, beispielsweise falls eine Atemluftmessung bzgl. des Verdachtes auf Lungenkrebs klassifiziert wird. Daher sollte möglichst der Einfluss jedes Attributes an dem Klassifikationsergebnis erkennbar sein, um das erlernte Modell einer Plausiblitätsüberprüfung eines Experten unterziehen zu können. Weiter wäre ein Maß für die Sicherheit wünschenswert, mit der ein Messdatum einer Klasse zugeordnet wird.

4.2 Methodenbeschreibung

Als Klassifikationsverfahren wird für diese Arbeit der Naive Bayes Klassifizierer gewählt. Die Grundlagen der Methode sind unter Abschnitt 2.1.4 beschrieben.

Bei der Klassifikation von IMS-Daten existiert eine große Menge möglicher Attribute (Peaks), von denen jedoch vermutlich nur wenige für die Klassenzuordnung bedeutend sind. Daraus ergibt sich ein Grund für die Wahl der Naiven Bayes Methode. Diese fragmentiert den Instanzraum nicht, im Gegensatz zu Divide & Conquer-Methoden wie Entscheidungssbäumen oder Regelbasierten Systemen. Solche instanzraumteilenden Methoden sind besonders anfällig für irrelevante Attribute, da in tieferen Entscheidungsebenen nur wenige Beispielinstanzen für die Entscheidung zur Wahl des nächsten trennenden Attributes herangezogen werden. Auch Instanzbasierte Klassifikationsverfahren haben diesen Nachteil, da sie zur Klassifikation lediglich die lokale Nachbarschaft der wenigen nächstgelegenen Beispielinstanzen heranziehen.

Die hier verwendeten Attribute geben eine qualitative Beschreibung der Peaks an. Eine Reduktion auf die quantitative Angabe *„Peak vorhanden“*, bzw. *„Peak nicht vor-*

handen“ bedeutet einen erheblichen Informationsverlust. Der Naive Bayes Klassifizierer arbeitet natürlicherweise auf diesen numerischen Attributen. Angesichts der verrauschten Daten besteht die Hoffnung, dass eine solche Modellierung angemessen ist.

Falls eine universelle Peakdatenbank für mehrere Klassifikationsaufgaben verwendet wird, müssen nicht alle Peaks der Datenbank für die Klassifikation entscheidend sein. Weiter können sich Peaks, die in ihren Höhen stark korrelieren, negativ auf die Klassifikationsperformanz auswirken. Daher besteht die Möglichkeit eine Vorauswahl der Attribute zu treffen. Dies geschieht entweder manuell, indem ein Experte angibt, welche Peaks als irrelevant gelten und nicht betrachtet werden sollen, oder automatisch durch eine „greedy“-Vorwärtsauswahl. Dabei wird, ausgehend von einer leeren Attributmenge, jeweils das Attribut in die Attributmenge aufgenommen, welches die Leistung des Klassifizierers bzgl. der Trainingsmenge am deutlichsten verbessert. Zwar ist die Leistung des Modells bzgl. der Trainingsmenge kein verlässlicher Indikator für die Leistung bzgl. einer Testmenge, aber schon dieser einfache Ansatz zeigt in Experimenten eine Verbesserung der Gesamtleistung. Sobald die Performanz durch eine weitere Attributauswahl nicht verbessert werden kann, stoppt das Verfahren und gibt die Menge der gewählten Attribute aus.

Beim Einsatz eines Naiven Bayes Klassifizierers wird häufig eine Normalverteilung der Attributbelegung jedes Attributes innerhalb einer Klasse angenommen. Im Folgenden wird gezeigt, dass diese Annahme für die hier verwendete Art der Attributbelegung nicht haltbar ist. Es gelte, dass die Peakhöhe eines Peaks P_p direkt mit der Konzentration des zugehörigen Analyten korreliert und dieser Analyt in den Messdaten verschiedener Messungen entweder in einer Konzentration, die der Peakhöhe $h_p^{(1)}$ oder in einer Konzentration, die der Peakhöhe $h_p^{(2)}$ entspricht, vorkommt. Weiter seien die Wahrscheinlichkeiten für das Auftreten der beiden Konzentrationen gleich und der Messfehler kann durch einen normalverteilten Zufallswert beschrieben werden. Dann ergibt sich für eine Messreihe, bei der die Peakhöhe des Peaks P_p in vielen Beispielmessdaten bestimmt wird, eine Häufigkeitsverteilung in der Art von Abbildung 4.1. Zusätzlich wird die berechnete Normalverteilungsdichte abgebildet, deren Parameter sich aus dem Mittelwert und der Standardabweichung der gemessenen Peakhöhen ergibt.

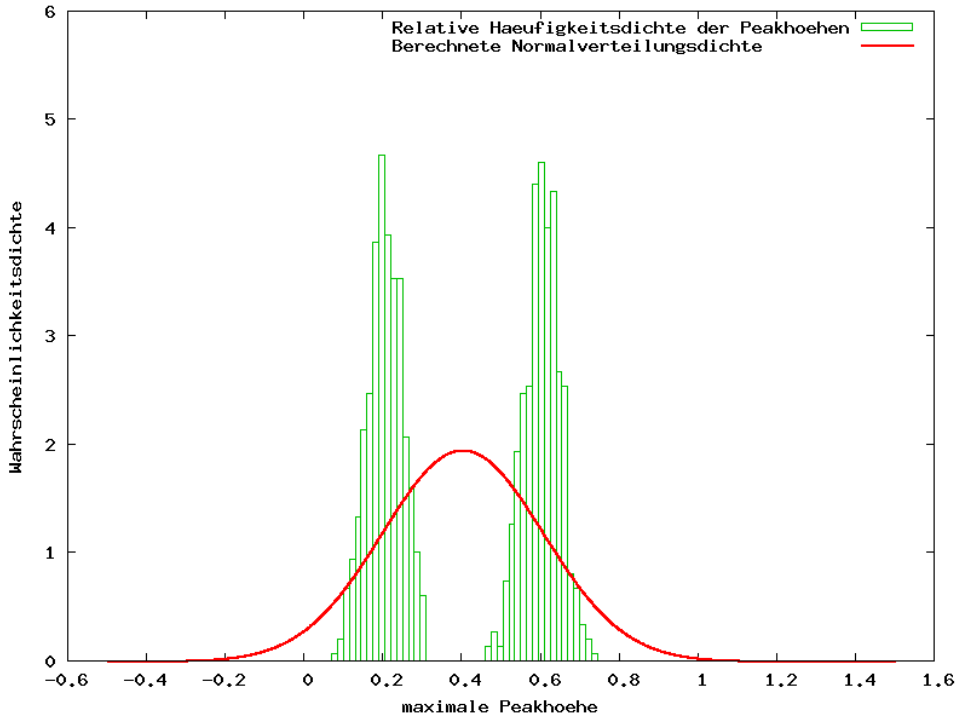


Abbildung 4.1: Häufigkeitsverteilung der gemessenen Peakhöhen und daraus berechnete Normalverteilungsdichte bei unangemessener Modellierung

Es zeigt sich deutlich, dass eine Modellierung der Wahrscheinlichkeitsdichte der Peakhöhe durch die Dichtefunktion der Normalverteilung in diesem Fall unangemessen ist.

Es wird eine Methode gesucht, die es erlaubt, die Wahrscheinlichkeitsdichten der Peakhöhen mit Hilfe der Trainingsdaten zu approximieren. Dies leistet die Parzen Fenster Methode, die nach der Festlegung der verwendeten Notation[1] beschrieben wird (vgl. [Duda01]).

Es sei:

- n : Anzahl der Trainingdaten
- x_i : Wert des i-ten Trainingsdatums, $i \in \{1, \ldots, n\}$
- $p_n(x)$: berechnete Wahrscheinlichkeitsdichte eines Wertes x

[1] Die Notation aus Abschnitt 2.1.4 gilt hier nicht.

- $\varphi(u)$: Kern.

Das Ziel ist es, die unbekannte Wahrscheinlichkeitsdichte, nach der die n Trainingsdaten generiert wurden, durch $p_n(x)$ zu approximieren. Dazu wird eine Funktion $\varphi(u)$ vorgegeben, der sogenannte Kern, der im Prinzip den Einflussbereich jedes Trainingswertes festlegt. Die Funktion sollte folgende Eigenschaften erfüllen:

$$\varphi(x) \geq 0 \tag{4.1}$$

und

$$\int \varphi(u)du = 1 \ . \tag{4.2}$$

Die Wahrscheinlichkeitsdichte eines Wertes x ergibt sich dann als Interpolation aus den Trainingsdaten, wobei der Einfluss des i-ten Trainingsdatums mit einem Abstand $(x - x_i)$ zu x durch den Funktionswert $\varphi(x - x_i)$ des Kerns beschrieben wird:

$$p_n(x) = \frac{1}{n} \sum_{i=1}^{n} \varphi(x - x_i) \ . \tag{4.3}$$

Zur Modellierung der Wahrscheinlichkeitsdichte der Peakhöhen wird als Kern die Normalverteilung mit dem Erwartungswert Null und der Standardabweichung σ_n vorgegeben:

$$\varphi(u) = \frac{1}{\sigma_n \sqrt{2\pi}} \cdot e^{-\frac{u^2}{2\sigma_n^2}} \ . \tag{4.4}$$

Dabei ist σ_n von der Anzahl der zur Verfügung stehenden Trainingsdaten n abhängig:

$$\sigma_n = \sigma_1 / \sqrt{n} \ , \tag{4.5}$$

der Parameter σ_1 wird vorgegeben. Wird diese Methode zur Berechnung der Wahrscheinlichkeitsdichte für die oben beschriebene Messreihe angewendet, ergibt sich für $\sigma_1 = 0,5$ die in Abbildung 4.2 dargestellte Approximation.

Im Gegensatz zur Modellierung durch die Dichtefunktion einer Normalverteilung wird durch die Parzen Fenster Methode eine gute Approximation der Peakhöhenverteilung erreicht.

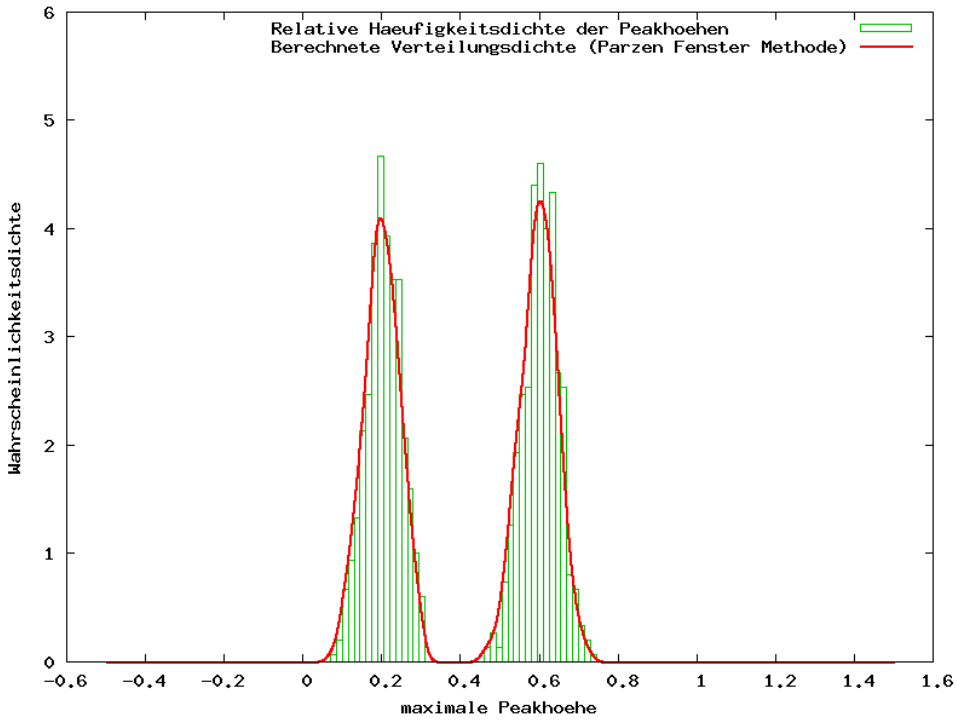


Abbildung 4.2: Häufigkeitsverteilung der gemessenen Peakhöhen und daraus berechnete Wahrscheinlichkeitsdichte mit der Parzen Fenster Methode

Zur weiteren Veranschaulichung ist in Abbildung 4.3 die mit der Parzen Fenster Methode berechnet Wahrscheinlichkeitsdichte der Peakhöhe eines Peaks aus 62 Atemluftmessdaten dargestellt.

Die größte Wahrscheinlichkeitsdichte besteht im Bereich einer Peakhöhe um Null. Das ist dadurch erklärbar, dass der untersuchte Peak in den meisten betrachteten Chromatogrammen nicht vorkommt. Es sind drei weitere Höhenebenen erkennbar, bei denen sich die gemessenen Peakhöhen um Werte von ca. $0,25$, $0,4$ und $0,6$ konzentrieren.

Das in den Anforderungen gewünschte Maß für die Sicherheit des Klassifizierers bei der Klassenzuordnung einer Instanz ist bei der Naive Bayes Methode durch die Angabe der Wahrscheinlichkeit der Klassenzugehörigkeit für alle Klassen gegeben. Diese berechneten Wahrscheinlichkeiten beziehen sich auf eine Übereinstimmung mit den Trainingsdaten. Ich möchte betonen, dass die Zuordnung eines Atemluftmessdatums zu der Klasse *„Proband hat Lungenkrebs“* mit einer Wahrscheinlichkeitsangabe von

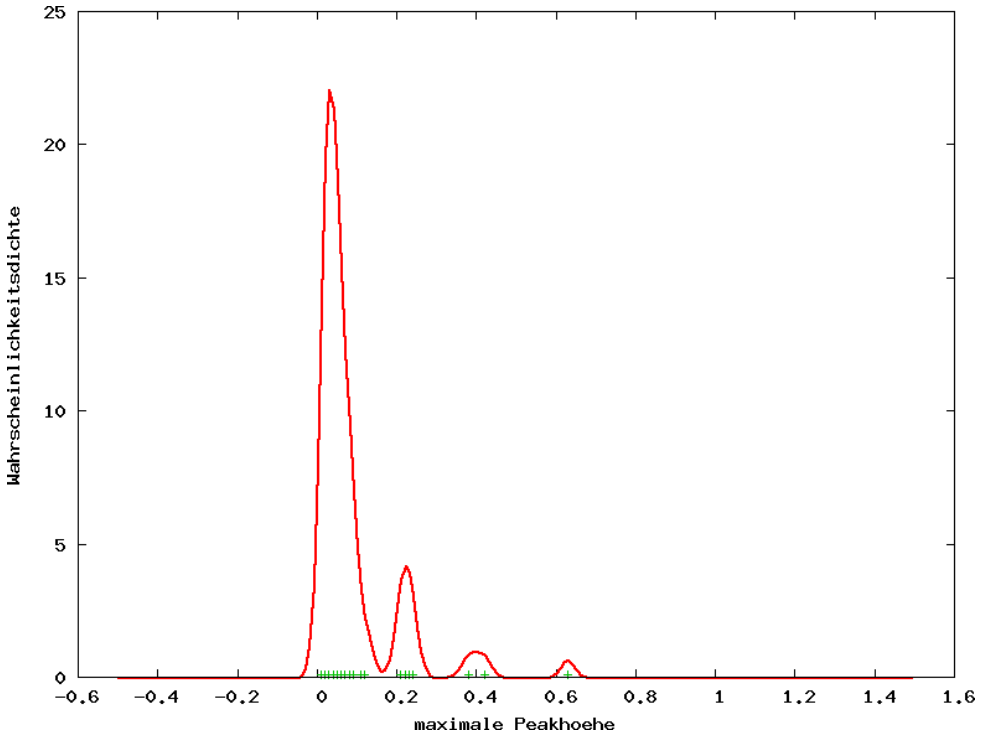


Abbildung 4.3: Mit der Parzen Fenster Methode berechnete Wahrscheinlichkeitsdichte der Peakhöhe eines Peaks

90 % nicht bedeutet, dass dieser Proband mit 90-prozentiger Sicherheit an Lungenkrebs erkrankt ist.

Die Anforderung, den Einfluss jedes Attributes auf die Klassifikation einer Instanz zu kennzeichnen, wird durch die Anzeige aller berechneten Wahrscheinlichkeiten $Pr(c_i|x_j)$ für $i \in \{1, \ldots m\}$ und $j \in \{1, \ldots n\}$ erfüllt. Dabei gibt $Pr(c_i|x_j)$ die Wahrscheinlichkeit der Klassenzugehörigkeit der Instanz zur Klasse c_i an, falls nur das Attribut x_j berücksichtigt wird. Die folgende Tabelle zeigt den Aufbau der Ausgabe:

	Klasse c_0	Klasse c_1	$\cdots$	Klasse c_m
Peak P_0	$Pr(c_0\|x_0)$	$Pr(c_1\|x_0)$	$\cdots$	$Pr(c_m\|x_0)$
Peak P_1	$Pr(c_0\|x_1)$	$Pr(c_1\|x_1)$	$\cdots$	$Pr(c_m\|x_1)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Peak P_n	$Pr(c_0\|x_n)$	$Pr(c_1\|x_n)$	$\cdots$	$Pr(c_m\|x_n)$
Alle Peaks	$Pr(c_0\|\mathbf{x})$	$Pr(c_1\|\mathbf{x})$	$\cdots$	$Pr(c_m\|\mathbf{x})$

Der letzten Zeile sind die Wahrscheinlichkeiten der Klassenzugehörigkeit der Instanz bei Betrachtung aller Attribute zu entnehmen. Die Klasse mit dem höchsten Wert wird als Ergebnisklasse des Klassifizierers ausgegeben.

Das vorgestellte Verfahren zur Klassifikation von Ionenmobilitätsspektrometerdaten mittels der Naiven Bayes Methode, gekoppelt mit der Bestimmung der Wahrscheinlichkeitsdichte mit Hilfe der Parzen Fenster Methode, wurde von mir in der Software „IMS-Analyse“ implementiert. Weiter kann der Anwender zwischen verschiedenen Methoden der Attributauswahl wählen. Die einfachste Möglichkeit ist es, alle Attribute zu verwenden, außerdem wird die oben beschriebene automatische Attributauswahl bereitgestellt. Zusätzlich kann eine manuelle Selektion der Attribute durchgeführt werden, wobei als Entscheidungshilfe die mit der Parzen Fenster Methode berechneten Wahrscheinlichkeitsdichten für jedes Attribut getrennt nach Klassen angezeigt werden (ähnlich Abb. 4.3). Beim Einsatz eines trainierten Klassifizierers ist die Anzeige der Wahrscheinlichkeiten für die Klassenzugehörigkeit einer Instanz zu jeder Klasse implementiert. Außerdem wird das Klassifikationsergebnis für eine getrennte Betrachtung jedes Attributes angezeigt. Die Anzeige entspricht der Form der vorigen Tabelle.

Kapitel 5

Beispielhafte Anwendung der Methode

Dieses Kapitel beschreibt die beispielhafte Anwendung des vorgestellten Verfahrens zur Klassifikation von realen IMS-Messdaten. Dazu kommt die im Rahmen dieser Arbeit entwickelte, prototypische Software „IMS-Analyse“ zum Einsatz. Eine weitergehende Funktionsbeschreibung des Programms wird im Anhang vorgestellt. Die Abbildung 5.1 zeigt grob den Ablauf des Verfahrens, alle notwendigen Schritte werden im Folgenden genauer erklärt.

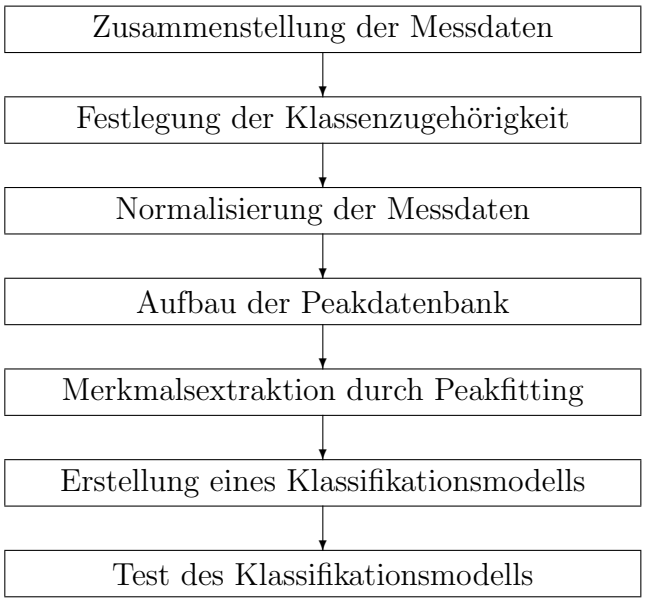


Abbildung 5.1: Verfahrensablauf zur Klassifikation von IMS-Daten

5.1 Zusammenstellung und Vorverarbeitung

Es wird zunächst ein Messdatensatz mit Ausatemluftmessdaten von 65 Patienten der Lungenklinik Hemer zusammengestellt. Die Messungen wurden von Meier im Rahmen seiner Diplomarbeit (siehe [Meier06]) durchgeführt. Bei 35 der 65 Patienten wurde mindestens eine Tumorart diagnostiziert, die anderen Patienten sind anderweitig erkrankt. Jedem Messdatum wird demnach die Klassenbezeichnung *„Tumor“* oder *„kein Tumor“* zugeordnet.

Der nächste Schritt besteht in der Untersuchung der Messdaten auf offensichtliche Fehler oder Ausreißer. Es werden 7 Messdaten identifiziert, die nach optischer Kontrolle fehlerhaft oder ungewöhnlich erscheinen. Einige davon zeigen ungewöhnlich starke Peaks und gelten daher als Ausreißer, andere enthalten sehr viele Peaks, sodass davon ausgegangen wird, dass der Patient kurz vor der Messung geraucht oder gegessen hat, womit die Messdaten für diesen Fall unbrauchbar sind. Nach der Bereinigung verbleibt ein Datensatz mit je 29 Messdaten der Klasse *„Tumor“* bzw. *„kein Tumor“*.

Alle Messdaten des Datensatzes werden, wie unter Abschnitt 3.2 auf Seite 25 beschrieben, normalisiert, d.h. die Driftzeitachsen werden am RIP normiert, der Feuchteeinfluss wird ausgeglichen und die Peaks durch die spektrenweise Anpassung der Signalintensität mittels eines Korrekturspektrums hervorgehoben.

Nach der Durchführung dieser grundlegenden Schritte besteht die nächste Aufgabe in der Erstellung einer Peakdatenbank zur anschließenden Extraktion der Merkmale.

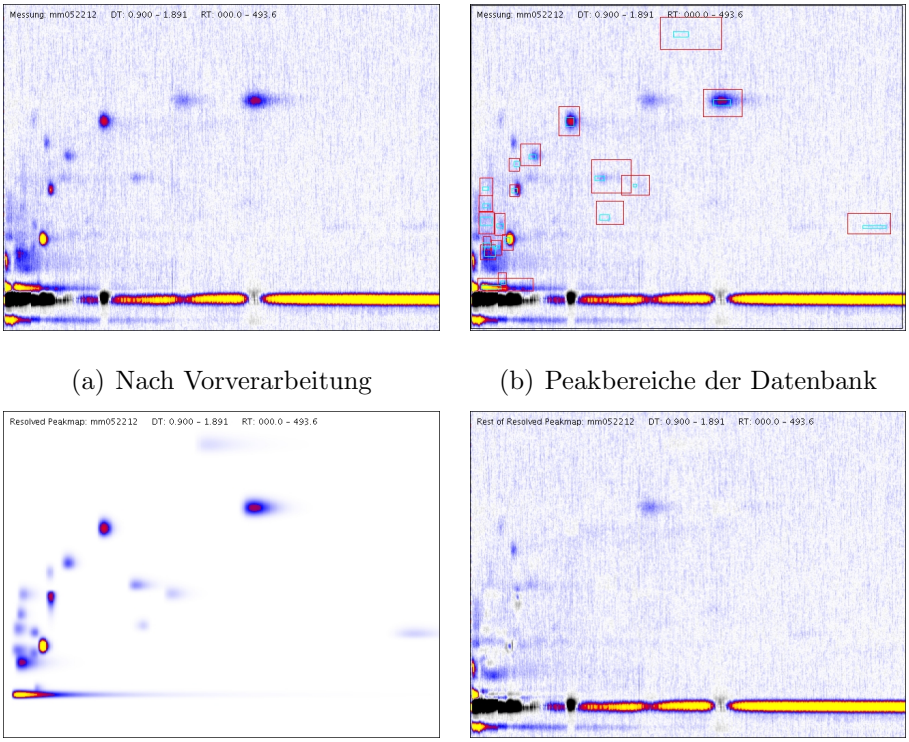
5.2 Peakdatenbank und Merkmalsextraktion

Mit dem unter Abschnitt 3.3.1 auf Seite 33 beschriebenen Verfahren werden nacheinander 20 Peaks in die Datenbank aufgenommen. Dafür wird zunächst ein Bereich markiert, in dem sich vermutlich ein interessanter Peak befindet. Dieser Bereich wird von der Software in allen 58 Messdaten untersucht, es werden diejenigen Bereiche als Chromatogramme

angezeigt, bei denen die maximale Signalintensität über einem vorgegebenen Grenzwert liegt. Die als repräsentativ für den jeweiligen Peak betrachteten Beispiele sind die Grundlage zur Fitnessbestimmung der Individuen der Evolutionsstrategie zur Ermittlung der Parameter der Funktion, die die dreidimensionale Kurvenform des Peaks beschreibt (vgl. Abschnitt 3.3.2, Seite 41).

Sind die Peakfunktionen durch ihre Parameter bestimmt, werden damit im nächsten Schritt die Peaks der 58 Messdaten wie unter Abschnitt 3.3.3 auf Seite 43 beschrieben aufgelöst.

Zur Veranschaulichung zeigt die folgende Abbildung unter (a) das Chromatogramm der Messdaten einer beispielhaften Atemluftmessung nach der Vorverarbeitung. Unter (b) sind zusätzlich die gewählten Fensterbereiche der 20 Peaks der Datenbank eingezeichnet. Unter (c) sind die gefundenen Peaks abgebildet und (d) zeigt die Differenz zwischen dem vorverarbeiteten Messdatum und den gefundenen Peaks.



(a) Nach Vorverarbeitung (b) Peakbereiche der Datenbank

(c) Gefundene Peaks (d) Differenz

Abbildung 5.2: Messdaten einer Ausatemluftmessung mit gefundenen Peaks

Es ist zu erkennen, wie akkurat die Peakbeschreibung der Peaks der Datenbank das Chromatogramm erklärt. Im Differenzbild zeigen die aufgelösten Peakbereiche lediglich ein Rauschen um die Nullebene. Außerdem gibt das Differenzbild Hinweise, welche Peaks bisher nicht in die Peakdatenbank aufgenommen wurden und daher im Merkmalsvektor nicht repräsentiert sind. Insgesamt erweist sich die in dieser Diplomarbeit erarbeitete Merkmalsbeschreibung der Chromatogramme als kompakt und präzise.

5.3 Klassifikation und Test

Im Folgenden wird die Erstellung und der Test der Performanz verschiedener Klassifikationsmodelle beschrieben. Zur Klassifikation der im vorigen Abschnitt bestimmten 58 Merkmalsvektoren wird das unter Abschnitt 4.2 auf Seite 50 spezifizierte Naive Bayes Verfahren mit einer Wahrscheinlichkeitsdichtebestimmung mittels der Parzen Fenster Methode verwendet. Dieses Verfahren benötigt zur Erstellung eines Klassifizierers folgende Vorgaben:

- Menge vorklassifizierter Instanzen zum Training und zur Bestimmung der Wahrscheinlichkeitsdichtefunktionen der Attribute,
- Menge der Attribute, die zur Klassifikation verwendet werden sollen (bei der automatischen Attributauswahl wird diese Menge beim Training evtl. weiter verfeinert) und
- den Parameter σ_1, der im Prinzip die Streuungsbreite der Attribute beschreibt.

Die Erstellung eines Klassifizierers ist in dem Sinne deterministisch, dass die gleichen (obigen) Vorgaben stets zum gleichen Klassifizierer führen. Die Performanz der Klassifizierer wird mit der Strategie der „leave-one-out“ Kreuzvalidierung (vgl. 2.1.5, Seite 11) getestet. Dabei stellen jeweils 57 der 58 Beispielinstanzen die Trainingsmenge dar, anhand der übrigen Instanz (Testinstanz) wird der Klassifizierer getestet. Bei fest gewähltem Verfahren zur Wahl der Attributmenge und fest gewähltem Parameter σ_1 wird die Performanz dieser Klasse von Klassifizierern durch den prozentualen Anteil der

richtig klassifizierten Testinstanzen der 58 aufgestellten Klassifizierer angegeben. Die Ergebnisse werden in Form einer Tabelle dargestellt, bei der jede Zeile ein Verfahren zur Wahl der Attributmenge und jede Spalte eine Belegung von σ_1 angibt.

Als Grundlage einer späteren Diskussion werden in den Abbildungen 5.3 und 5.4 die Graphen der mit der Parzen Fenster Methode berechneten Wahrscheinlichkeitsdichtefunktionen für $\sigma_1 = 0,1$ für alle Peaks dargestellt. Die roten Graphen zeigen die Wahrscheinlichkeitsdichte der Patienten mit einer Tumorerkrankung an, die blauen Graphen die Wahrscheinlichkeitsdichte der Patienten mit anderen Erkrankungen.

Als Verfahren der Attributauswahl kommt, neben der Wahl aller Attribute und der auf Seite 51 beschriebenen Vorwärtsauswahl, auch eine erschöpfende Suche im Attributraum nach den drei, vier bzw. fünf Attributen mit der besten Performanz zum Einsatz.

Die folgende Tabelle zeigt die Testergebnisse:

	$\sigma_1 = 0,01$	$\sigma_1 = 0,05$	$\sigma_1 = 0,1$	$\sigma_1 = 0,2$
Alle Attribute	48%	50%	44%	41%
Vorwärtsauswahl	48%	58%	50%	39%
Die besten 3 Attribute	70%	68%	63%	60%
Die besten 4 Attribute	74%	70%	67%	63%
Die besten 5 Attribute	72%	70%	68%	65%

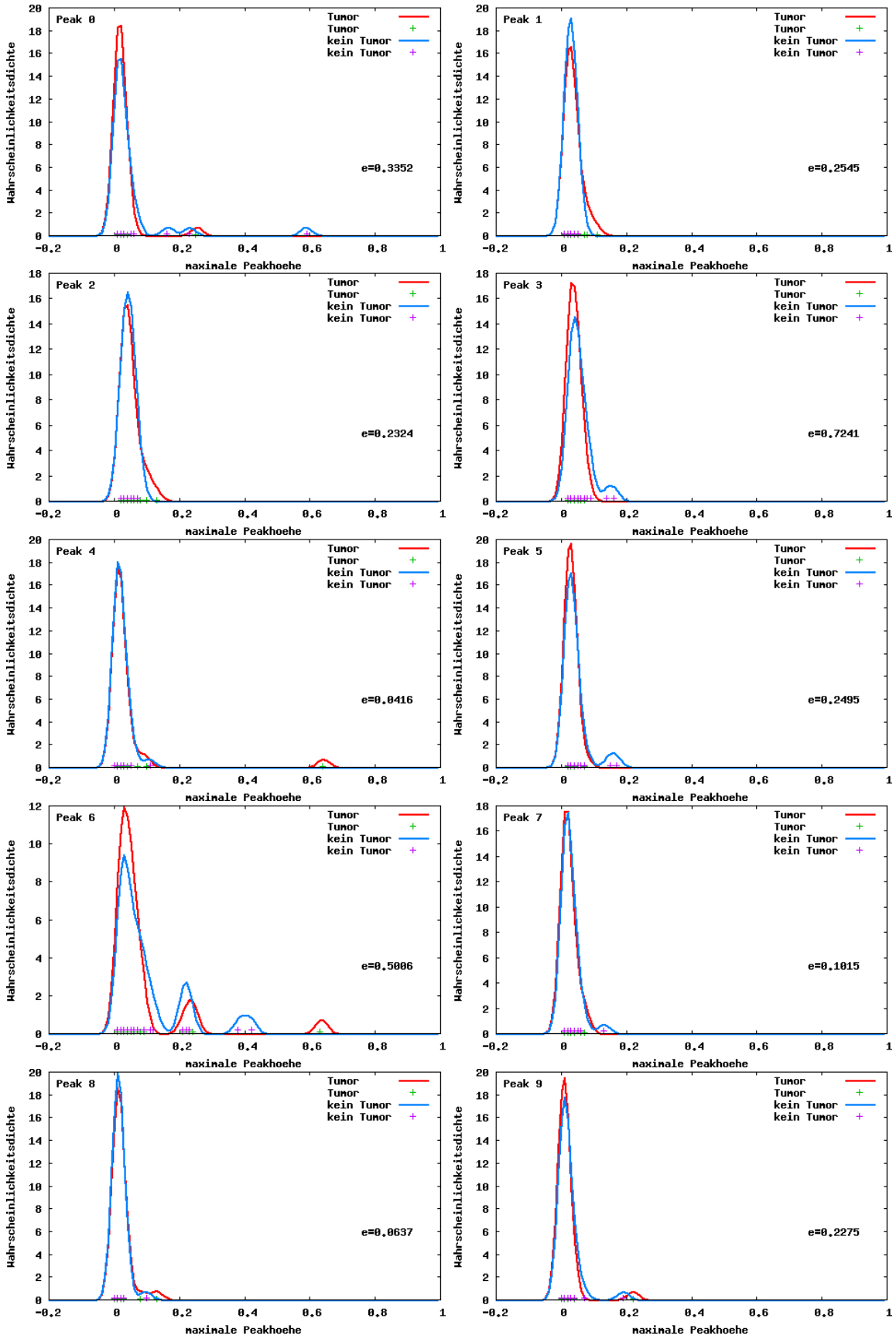


Abbildung 5.3: Wahrscheinlichkeitsdichte der Peaks, getrennt nach Klassen, $\sigma_1 = 0,01$, Teil 1

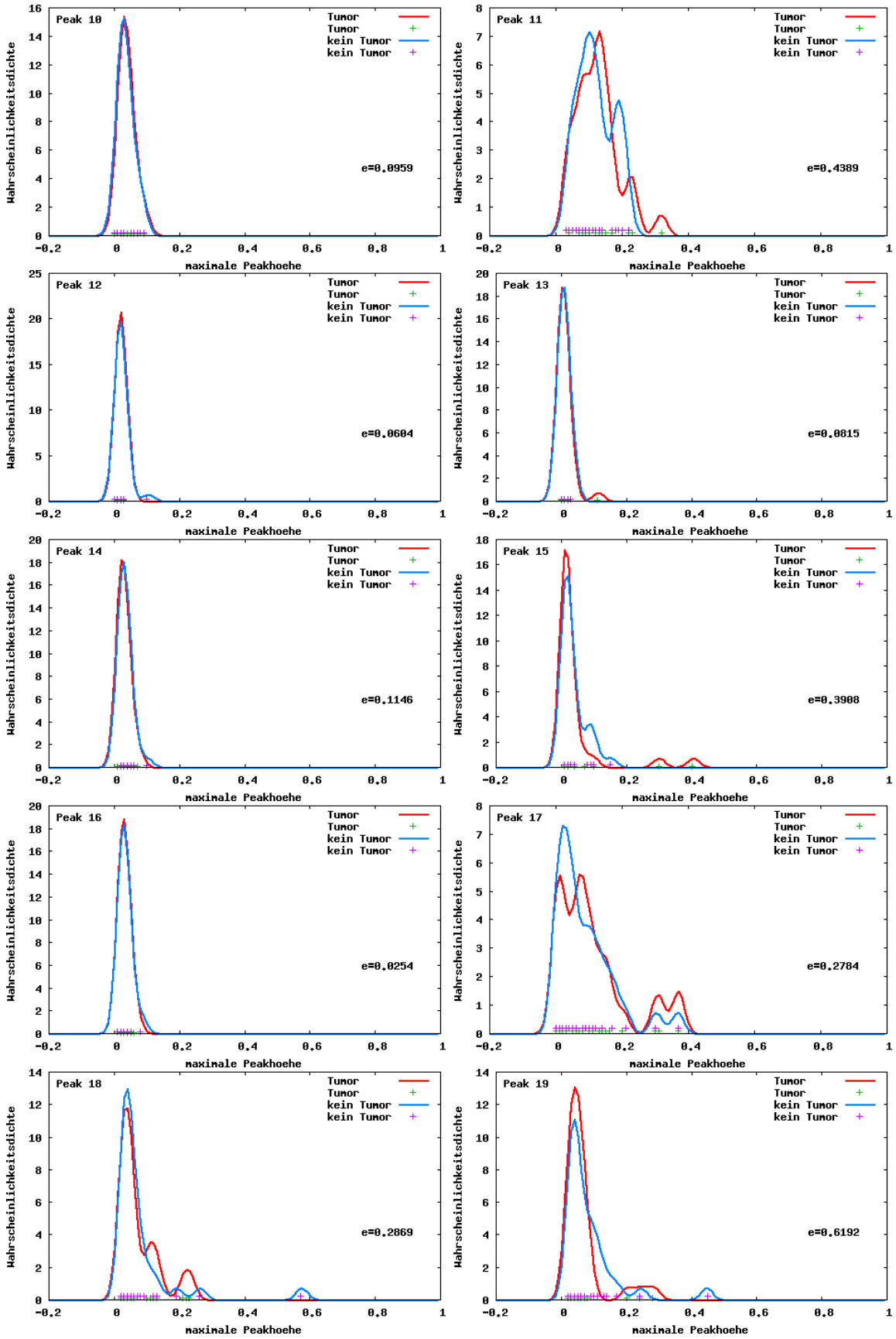


Abbildung 5.4: Wahrscheinlichkeitsdichte der Peaks, getrennt nach Klassen, $\sigma_1 = 0,01$, Teil 2

5.4 Interpretation der Ergebnisse

Die automatische Vorwärtsauswahl der Attribute bei der Klassifikation zeigt keine wesentlich bessere Performanz als die Berücksichtigung aller Attribute. Die beste Performanz mit 74% richtig klassifizierten Testinstanzen wird mit der Attributmenge $\{3, 11, 13, 16\}$ und $\sigma_1 = 0,01$ erzielt. Die besten Ergebnisse bei σ_1 Werten von $0,1$ und $0,2$ liefern häufig Kombinationen der Peaks 1,3,5,6 und 15 als Attributmenge.

Ein Blick auf die Wahrscheinlichkeitsdichtefunktionen der Attribute gibt erste Hinweise, warum sich die Klassifikation als schwierig herausstellt. Die Wahrscheinlichkeitsdichten beider Klassen liegen für alle Attribute sehr eng beieinander. Die meisten Peaks weisen die höchste Dichte im Höhenbereich $[0; 0,1]$ auf, was bedeutet, dass sie in den Trainigsinstanzen nur schwach und relativ selten auftreten. Lediglich der in jedem Chromatogramm vorkommende Peak 11 zeigt teilweise größere Höhenwerte an, aber auch diese trennen die beiden Klassen nicht.

Die hohe Klassenähnlichkeit der Attribute bietet dem Klassifikationsverfahren wenige Informationen, die es zur Trennung der Klassen verwenden kann. Wünschenswert wären Attribute, bei denen sich die Peakhöhen der verschiedenen Klassen um möglichst weit auseinanderliegende Punkte konzentrieren. Diejenigen Attribute, die sich noch am ehesten zur Unterscheidung der Klassen eignen, zeigen im Ansatz dieses Verhalten. Die Peaks 3 und 5 deuten z.B. bei höheren Werten darauf hin, dass keine Tumorbildung vorliegt, erhöhte Peak 1 Werte oder stark erhöhte Peak 15 Werte treten hingegen eher bei Patienten mit Tumoren auf. Die Peakhöhen dieser Peaks liegen in die meisten Trainingsbeispielen jedoch im unteren Bereich, sodass ihre Aussagekraft (in diesen Fällen) beschränkt ist.

Aufgrund des geringen Stichprobenumfangs wird auch die Glaubwürdigkeit einer Klassifikationsentscheidung in Frage gestellt. Bei den angesprochenen Attributen 1,3,5 und 15, die den größten Einfluss auf das Klassifikationsergebnis haben, wird die Entscheidung nur aufgrund weniger Trainigsbeispiele getroffen. Es kann keineswegs sichergestellt

werden, dass diese Beispiele lediglich Ausreißer der Norm darstellen. Daher wäre es interessant, das Klassifikationsverfahren mit einem größeren Stichprobenumfang zu testen. Dabei könnte es sich jedoch herausstellen, dass die gewonnenen Informationen aus den Atemluftmessungen nicht ausreichen, um Tumore sicher zu diagnostizieren.

5.5 Test des Verfahrens zur Merkmalsextraktion

Es stellt sich die Frage, ob die aufwendige Vorverarbeitung der Messdaten zur Gewinnung des Merkmalsvektors gerechtfertigt ist. Daher wird in diesem Abschnitt die Leistung der entwickelten Vorverarbeitung getestet und mit einem einfachen, aber oft eingesetzten Verfahren zur Peakdetektion verglichen.

Die Idee zur Leistungsbestimmung ist es, mit Hilfe der Peakfunktionen der Peakdatenbank künstliche Messdaten mit zufälligen Peaks zu erstellen und diese mit einem zusätzlichen Rauschen zu stören. Diese künstlichen Messdaten durchlaufen nun die Schritte der Vorverarbeitung, die schließlich zur Ausgabe der Merkmalsvektoren führen. Im Anschluss werden die berechneten Merkmalsvektoren mit den vorher zufällig gewählten Peakhöhen verglichen, um die Güte der Vorverarbeitungsschritte abzuschätzen.

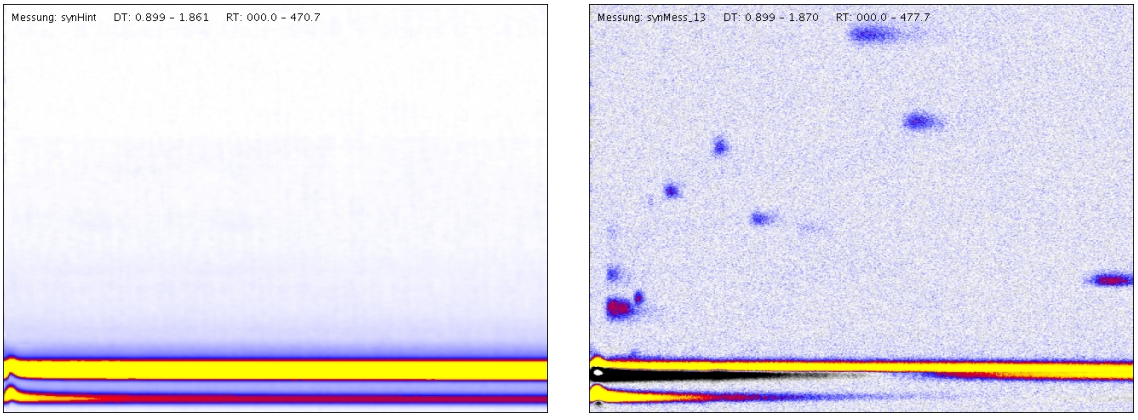
Der erste Schritt besteht in der Generierung der künstlichen Messdaten. Zunächst wird ein Messdatum benötigt, das außer dem RIP und Feuchtepeak keine weiteren Peaks enthält. Dazu wird aus 20 am RIP normierten Raumluftmessungen, die generell neben dem RIP und Feuchtepeak nur wenige schwache Peaks aufweisen, ein neues Messdatum $M_{Hintergrund}$ zusammengestellt. Die Signalintensität dieses Messdatums erhält für jede Position (Driftzeit, Retentionszeit) die fünftschwächste Signalintensität dieser Position in den 20 Raumluftmessungen. Damit wird zum einen nicht die ausreißeranfällige Methode der Minimumsbildung verwendet und zum anderen wird gegen die Übertragung von schwachen, aber häufig vorkommenden Peaks der Raumluftmessungen vorgegangen. Die verbleibenden schwachen Peakbereiche werden anschließend manuell durch Kopien eines passenden Bereichs ohne Peaks ersetzt. Das entstandene peakleere Messdatum bildet den Hintergrund für alle künstlich erzeugten Messdaten.

Mit diesem Hintergrund werden 40 synthetische Messdaten generiert. Dazu werden für jedes Messdatum die 20 Peakfunktionen der Peakdatenbank mit zufälligen Werten belegt und die Funktionswerte an allen Positionen des Ausgangsmessdatums zum Hintergrund addiert. Der i-ten Peakfunktion mit $i \in \{0, \ldots, 19\}$ werden dabei zufällige Parameterwerte aus dem jeweiligen Toleranzbereich der Peaklage bzw. normalverteilt um die mittlere vorkommende Höhe zugewiesen (vgl. Abschnitt 3.3.1, Seite 33):

- Die Driftzeit wird uniform zufällig aus $[\tau_i^{(dt_{min})}; \tau_i^{(dt_{max})}]$ gewählt,
- die Retentionszeit wird uniform zufällig aus $[\tau_i^{(rt_{min})}; \tau_i^{(rt_{max})}]$ gewählt und
- die maximale Höhe wird auf einen normalverteilten, zufälligen Wert mit $\mu = 0,03$ und $\sigma = 0,02$ gesetzt, wobei ein Wert kleiner als Null auf Null gesetzt wird.

Die gewählten Höhen werden für einen späteren Vergleich gespeichert. Anschließend wird das Messdatum mit einem normalverteilt zufälligen Rauschen mit $\mu = 0$ und $\sigma = 0,01$ für jede Signalposition gestört.

Die Abbildung 5.5 zeigt unter (a) das erstellte „leere“ Chromatogramm und unter (b) ein daraus künstlich erzeugtes Messdatum.



(a) Hintergrund (b) Messdatum, Peaks + Hintergrund

Abbildung 5.5: Chromatogramme künstlich erzeugter Messdaten

Mit den 40 erstellten Messdaten wird nun das hier entwickelte Verfahren zur Merkmals-

extraktion aus IMS-Chromatogrammen getestet. Dazu durchlaufen die 40 Messdaten die einzelnen Schritte der Vorverarbeitung, die 40 berechneten Merkmalsvektoren werden gespeichert. Der Merkmalsvektor eines Chromatogramms enthält die 20 maximalen Höhen der Peakfunktionen, die den jeweiligen Peaks angepasst wurden.

Eine einfache, aber dennoch häufig verwendeten Methode zur Peakdetektion in Chromatogrammen sucht in vorgegebenen Fensterbereichen nach dem Signalmaximum. Liegt dieses über einem Grenzwert, wird ein Peak an dieser Stelle angenommen.

Diese Methode dient nun dem Vergleich und der Einschätzung der Leistung des neu entwickelten Verfahrens. Dazu werden zu allen synthetischen Messdaten die Signalmaxima der Fensterbereiche aller Peaks der Peakdatenbank berechnet und als Merkmalsvektor gespeichert.

Interessant ist nun die Differenz zwischen den berechneten Peakhöhen und den bei der Erstellung der künstlichen Messdaten vorgegebenen Peakhöhen. Sicherlich spricht eine kleine berechnete Differenz für die Qualität eines Verfahrens. Noch wichtiger ist es aber, dass die Schwankungsbreite der berechneten Differenz für verschiedene Beispiele möglichst gering ist, da sie Aufschluss über die Reproduzierbarkeit der Ergebnisse gibt. Bei einer geringen Schwankungsbreite und einer hohen Differenz zum Sollwert ist der vom Verfahren produzierte Fehler für die Messdaten ähnlich, sodass eine Vergleichbarkeit dennoch gegeben ist.

Die Abbildung 5.6 zeigt das Ergebnis des Tests beider Verfahren auf Basis der 40 Beispielmessdaten. Es ist der Bereich der Differenzen zum Sollwert der Peakhöhen jedes Peaks für beide Verfahren dargestellt. Die Boxen zeigen dabei die unteren Quantile (0,25-Quantile) und die oberen Quantile (0,75-Quantile), die unteren und oberen Linien die Minima bzw. die Maxima der Differenzen zum Sollwert jedes Peaks an.

Er ist zu erkennen, dass das hier entwickelte Verfahren eine deutlich geringere Schwankungsbreite der Differenzen aufweist als das Verfahren der Maximumsbestimmung. Auch die mittlere Differenz liegt beim neuen Verfahren meist nahe Null. Das spricht dafür, dass

das entwickelte Verfahren zur Merkmalsextraktion durch Fitting von Peakfunktionen vergleichsweise gute und reproduzierbare Ergebnisse für die Beschreibung der Peaks in Chromatogrammen erzeugt.

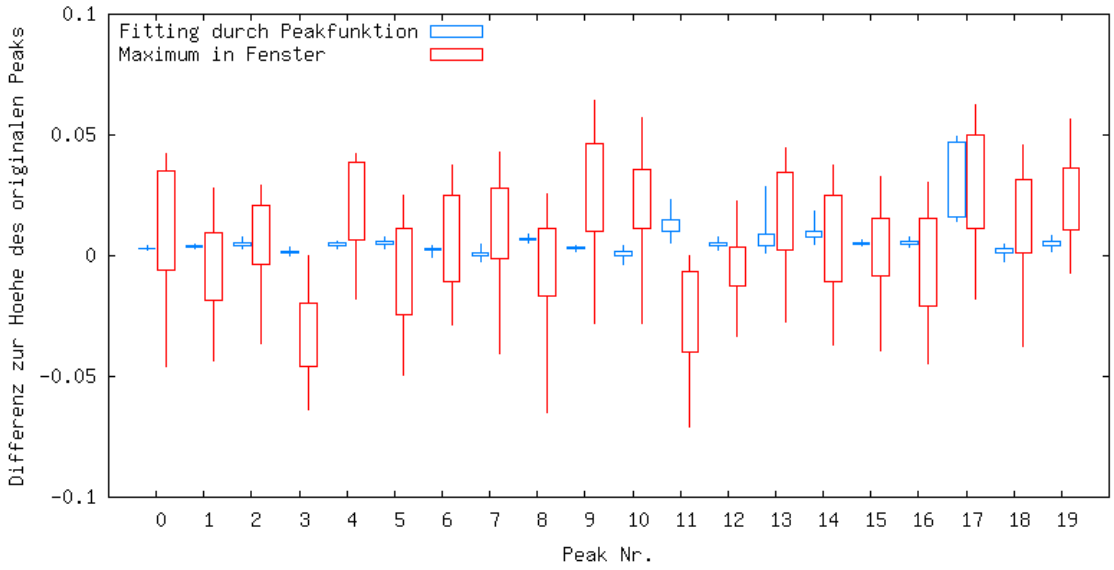


Abbildung 5.6: Quantile der Differenzen zum Sollwert beider Verfahren

Kapitel 6

Zusammenfassung und Ausblick

Die Ionenmobilitätsspektrometrie ist ein leistungsfähiges Analyseverfahren zur Identifizierung von Analyten in Gasgemischen. Bei der Messung einer Gasprobe mit dem für diese Arbeit verwendeten MCC / ^{63}Ni - IMS entstehen bis zu einer Million Datenpunkte, die Informationen über die Zusammensetzung des Gasgemisches und der Konzentration der einzelnen Analyten enthalten. Diese Arbeit beschreibt ein Verfahren zur Analyse und Klassifikation dieser Daten. Aus einer Menge vorklassifizierter Beispielmessdaten werden einzelne Analyten, die sich in Form von Signalausschlägen in den Messdaten zeigen, erkannt und extrahiert. Dafür wurde eine vollkommen neue Methode entwickelt, um die Form der Signalausschläge durch eine analytische Funktion zu beschreiben. Die Parameter dieser Funktion werden anhand von Beispieldaten mittels einer Evolutionsstrategie erlernt und in einer Datenbank gespeichert. Mit Hilfe der parametrisierten Funktion können die Signalausschläge in zuvor unbekannten Messdaten wiedererkannt werden. Weiter hilft die mathematische Beschreibung überlagerte Signale zu trennen und sie den richtigen Analyten zuzuordnen.

Durch die Erkennung der Signalausschläge eines Messdatums kann dieses durch den Vektor seiner Merkmale beschrieben werden. Anhand der Merkmalsvektoren der Beispielmessdaten werden Eigenschaften gesucht, die charakteristisch für die Zugehörigkeit eines Messdatums zu einer Klasse sind. Die Wahrscheinlichkeitsdichten jedes Attributes und jeder Klasse können mittels der Methode der Parzen Fenster bestimmt werden. Die

Informationen der Wahrscheinlichkeitsdichten werden von der Naiven Bayes Methode zur Berechnung der Wahrscheinlichkeiten der Zugehörigkeit eines Messdatums zu den vorgegebenen Klassen verwendet.

Das entwickelte Klassifikationsverfahren wurde mit realen Messdaten getestet, wobei Ausatemluftmessungen von Patienten der Lungenklink in Hemer verwendet wurden. Bei einer Gruppe wurde mindestens eine Tumorart diagnostiziert, die andere Gruppe war anderweitig erkrankt. Es wurde keine eindeutige Klassentrennung erreicht, da die extrahierten Peaks der verschiedenen Klassen keine signifikanten Unterschiede aufwiesen.

Die größte Herausforderung beim Entwurf des Verfahrens bestand in der Entwicklung einer geeigneten Vorverarbeitung der Datenmengen. Durch die hohe Sensibilität des Ionenmobilitätsspektrometers, die für eine Analyse kleinster Bestandteile in Gasen notwendig ist, reagieren die Geräte äußerst empfindlich auf die verschiedensten Einflüsse. Diese Einflüsse mussten so gut wie möglich ausgeglichen werden, um die Vergleichbarkeit der verschiedenen Messdaten sicherzustellen. Als besonders schwierig stellte sich die Zuordnung eines Signalausschlages zu einem festen Analyten heraus, da die Position im Chromatogramm Schwankungen unterliegt und die Signalmaxima teilweise sehr dicht beieinander liegen. Schließlich konnte für die gefundene Lösung eine erhebliche Verbesserung im Vergleich zu der Methode der Peakdetektion durch Maximumsbestimmung in festgelegten Fenstern bzgl. der Genauigkeit und Reproduzierbarkeit der aufgelösten Merkmale gezeigt werden.

Eine Bedingung zur Anwendung der vorgestellten Methode ist es, dass die verwendeten Messdaten alle mit der gleichen Gerätekonfiguration aufgenommen wurden. Ein weiterer Schritt ist die Entwicklung eines Standards für Messdaten, der zusätzlich eine Vergleichbarkeit der Messdaten, die mit verschiedenen Gerätekonfigurationen aufgenommen wurden, ermöglicht. Dafür müssten die zahlreichen Einflüsse auf das Messergebnis bei der Messung genauer erforscht und entsprechende Transformationen gefunden werden. Ist das erreicht, könnte eine umfassende Peakdatenbank erstellt werden, die eine genauere Analyse der Analyten zulässt. Dazu wäre es auch ein spannender Versuch, die molekularen Eigenschaften eines Analyten mit der Peakfunktion in Verbindung zu

bringen. Eine Methode, die aus der Struktur der Moleküle die Peakfunktionsparameter und den Lagebereich der oder des Peaks ableitet, wäre ein deutlicher Fortschritt.

Das erarbeitete Verfahren bedarf jedoch noch weiterer Tests. Interessant ist die Anwendung mit einem größeren Stichprobenumfang, der leider zum Zeitpunkt dieser Arbeit noch nicht zur Verfügung stand. Derzeit wird die Ausatemluft vieler Patienten mit unterschiedlichen Krankheitsbildern in der Lungenklinik in Hemer gemessen. Die Untersuchung dieser Messdaten ergibt bestimmt spannende Ergebnisse. Weiter ist der Einsatz dieses Verfahrens in anderen Bereichen der Gasprobenanalyse denkbar.

Anhang A

Funktionsbeschreibung „IMS-Analyse“

A.1 Überblick

Die in dieser Arbeit beschriebenen Methoden wurden in der prototypischen Software „IMS-Analyse“[1] implementiert. Dafür wurde die Programmiersprache Java in der Version 1.5.0_06 verwendet, als Entwicklungsumgebung kam Eclipse SDK unter der Linux-Distribution Ubuntu 6.06 LTS zum Einsatz. Dank der Plattformunabhängigkeit von Java ist „IMS-Analyse“ auch unter Microsoft Windows lauffähig. Aufgrund der Verwendung neuerer Java-Funktionen ist die Laufzeitumgebung der Version 1.5 oder höher unbedingt erforderlich. Für die Nutzung der 3D Funktionen wird zusätzlich die Java 3D API[2] benötigt.

Im Folgenden soll ein Überblick über den Aufbau und Funktionsumfang dieser Software gegeben werden. Am einfachsten ist die Struktur an einem beispielhaften Ablauf der Anwendung erklärbar.

[1] Kontakt: bboedeker@gmx.de

[2] Java 3D^{TM} Homepage: https://java3d.dev.java.net/

A.2 Hauptfenster

Nach dem Start von „IMS-Analyse“ öffnet sich das Fenster aus der Abbildung A.1. Auf

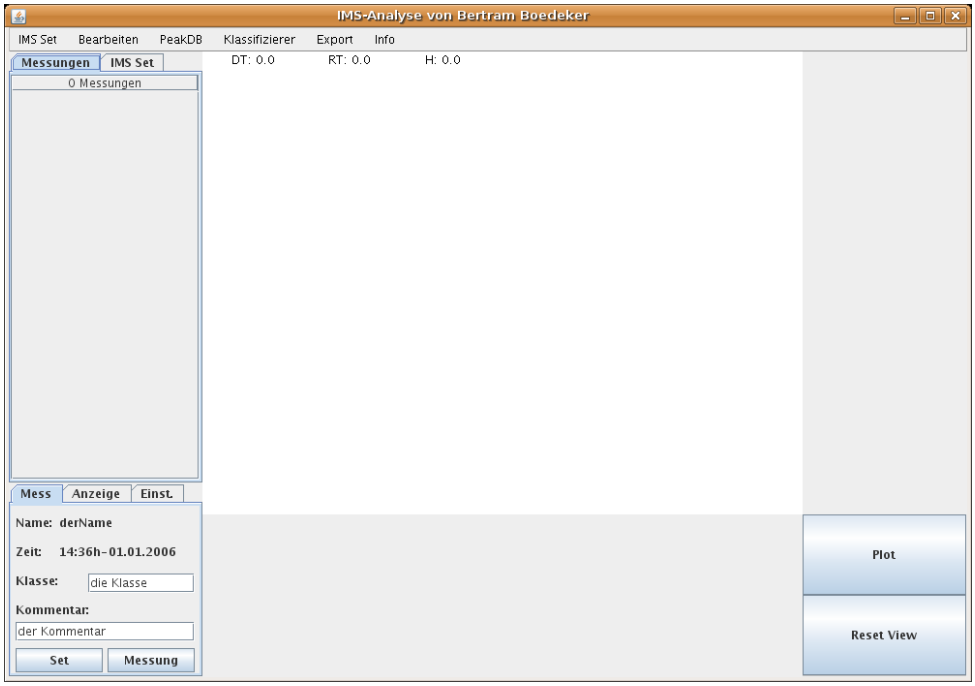


Abbildung A.1: Startfenster „IMS-Analyse“

der linken Seite befinden sich zwei Informationfelder. Im oberen Feld werden wahlweise die zur Verfügung stehenden Messdaten oder der Datensatz der zusammengestellten Beispielmessdaten angezeigt. Im unteren Feld werden wahlweise weitere Informationen einer gewählten Messung oder die Parameter zur Darstellung von Chromatogrammen angezeigt. Der gesamte Teil rechts der Informationsfelder dient der Darstellung der Messdaten in Form von Chromatogrammen. Des Weiteren ist am oberen Rand die Menüleiste erkennbar.

A.3 Konvertierung von Messdaten

Der erste Schritt besteht in der Konvertierung der rohen Messdaten in das programmeigene Format. Die verwendeten rohen IMS-Messdaten liegen als Menge von bis zu 501 Dateien (mit den Endungen *„.000“* bis *„.500“*) vor. Jede Datei beinhaltet die Daten ei-

nes Spektrum zu einer bestimmten Retentionszeit. Die Retentionszeitinformation befindet sich im Kopfteil der Datei, danach folgen 2000 Zeilen, die jeweils die Driftzeit und das zu dieser Zeit gemessene Signal angeben. Ein solches Messdatum muss vor der weiteren Verwendung in „IMS-Analyse“ zunächst in ein programmeigenes Format für Messdaten umgewandelt werden. Bei dieser Umwandlung werden gleichzeitig weitere Informationen, die die Vorverarbeitung betreffen, generiert. Beispielsweise werden die RIP-Positionen der Spektren ermittelt, daraus die am RIP normierte Driftzeitskala berechnet und das Korrekturspektrum zur Hervorhebung der Peaks bestimmt. Der Konvertierungsprozess kann aus dem Programm heraus über den Menüeintrag *„Messungs Konverter“* im Menü *„IMS-Set“* aufgerufen werden. Es öffnet sich ein Fenster der lokalen Verzeichnishierarchie in dem lokale Messdatendateien im Rohformat angezeigt werden. Zur besseren Übersicht werden jeweils nur die Dateien mit der Endung *„.000“* dargestellt. Hieraus wählt der Anwender diejenigen Messdaten aus, die er im Folgenden weiter verwenden möchte. Daraufhin werden die gewählten Messdaten umgewandelt und in einem vorher festgelegten Verzeichnis als serialisierte Java-Klassen mit der Dateiendung *„.imsjava“* abgelegt. Alternativ können auch ganze Verzeichnisse mit Messdaten durch Aufruf des Menüeintrages *„Konvertiere Verzeichnis“* im Menü *„IMS-Set“* umgewandelt werden. Nach der Konvertierung werden die Messdaten im Informationsfeld *„Messungen“* angezeigt.

A.4 Anzeige von Chromatogrammen

Durch Auswahl eines Messdatums wird das Chromatogramm im Hauptbereich des Fensters dargestellt (siehe Abbildung A.2).

Bei der Bewegung des Mauszeigers über dem Chromatogramm wird die derzeitige Driftzeit, die Retentionszeit und die Signalintensität dieser Position im oberen Bildbereich angezeigt. Durch einen Klick mit der linken Maustaste im Chromatogramm wird an dieser Position ein schwarzes Fadenkreuz eingeblendet. Dazu erscheint im rechten Bereich das Spektrum zur markierten Retentionszeit, im unteren Bereich wird die Höhenlinie zur markierten Driftzeit angezeigt.

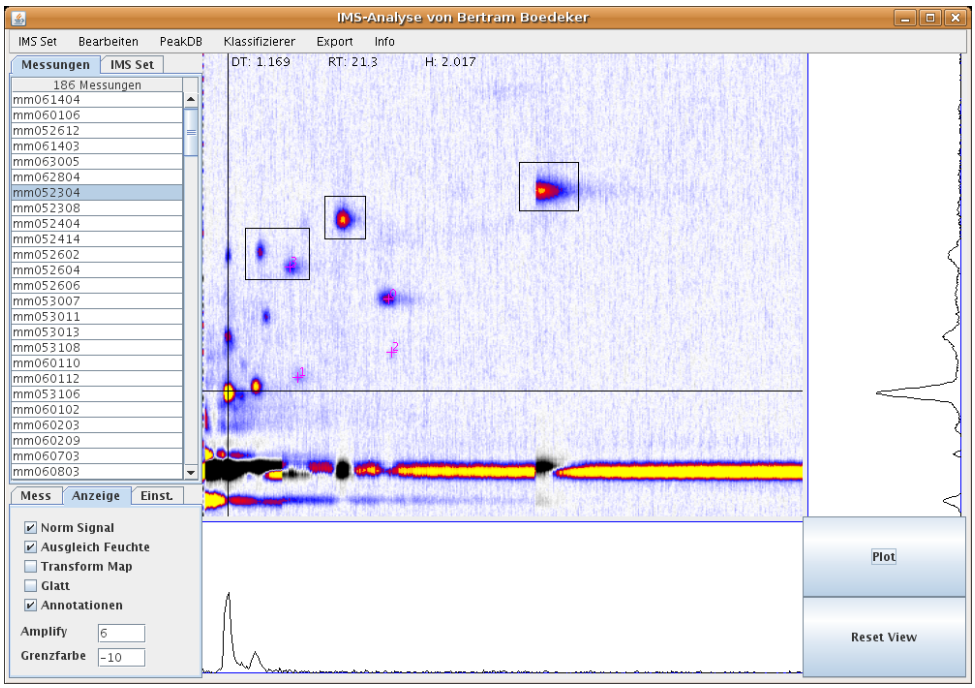


Abbildung A.2: Anzeige eines Chromatogrammes

Zieht man bei gedrückter linker Maustaste ein (imaginäres) Fenster auf, so wird der Bildausschnitt auf diesen Bereich festgelegt. Damit können auch kleinere Details deutlich dargestellt werden. Der Schaltknopf *„Reset View“* unten rechts im Fenster setzt den Bildausschnitt auf die Gesamtansicht zurück. Weiter besteht die Möglichkeit im Chromatogramm Markierungen (sogenannten Annotationen) einzufügen. Ein Klick der rechten Maustaste markiert diese Stelle mit einem magentafarbenen Kreuz. Ein Fenster, welches mit gedrückter rechter Maustaste aufgezogen wird, erscheint als schwarzes Rechteck. Besonders diese Fenstermarkierungsfunktion wird für verschiedene Aufgaben benötigt, die später im Text erklärt werden. Ein Klick mit der mittleren Maustaste löscht die jeweils nächstliegendste Annotation wieder. Bei Fenstern ist dafür der untere linke Punkt ausschlaggebend. Die erstellten Annotationen sind in allen Chromatogrammen sichtbar. So kann z.B. eine auffällige Peakkonstellation eines Chromatogrammes markiert werden, um anschließend weitere Chromatogramme mit dieser Markierung vergleichen zu können.

Im unteren Informationsfeld *„Anzeige“* stehen einige Optionen zur Anzeige bereit.

Die Änderungen werden durch die Betätigung des Schaltknopfes *„Plot“* übernommen und das Chromatogramm wird neu dargestellt.

- *„Norm Signal“*: Bei Auswahl wird die Signalintensität mittels berechnetem Korrekturspektrum korrigiert, ansonsten ist die Signalintensität um eine Konstante reduziert (vgl. Abschnitt 3.2.3, Seite 29).
- *„Ausgleich Feuchte“*: Bei Auswahl wird der Einfluss der Feuchte im Chromatogramm korrigiert (vgl. Abschnitt 3.2.2, Seite 28).
- *„Transform Map“*: Bei Auswahl werden die Chromatogrammintensitäten passend zu einer Standardskala für Drift- und Retentionszeit transformiert. Diese Funktion wurde zu Testzwecken benutzt und sollte im Allgemeinen nicht verwendet werden.
- *„Glatt“*: Bei Auswahl werden die Chromatogrammintensitäten mit einem Tiefpassfilter geglättet.
- *„Annotationen“*: Gibt an, ob die eingefügten Markierungen (Annotationen) im Chromatogramm angezeigt werden sollen.
- *„Amplify“*: Dieser Wert ist ein Faktor der Signalintensität. Bei Werten über eins wird die Signalintensität verstärkt dargestellt, bei Werten kleiner als eins geschwächt. In der Praxis hebt ein Faktor von sechs bis sieben bei gleichzeitiger Wahl von *„Norm Signal“* die Peaks im Chromatogramm deutlich hervor.
- *„Grenzfarbe“*: Dieser Wert gibt eine Grenze für die Darstellung der Intensität an, Bereiche mit einer niedrigeren Intensität erscheinen im Chromatogramm weiß. So kann der Rauschbereich ausgeblendet werden.

A.5 Beispieldatensatz

Der nächste Schritt besteht in der Zusammenstellung eines Beispieldatensatzes vorklassifizierter Messdaten aus der Sammlung der konvertierten Messdaten. Dazu wird zuerst über den Menüpunkt *„IMS-Set“* ➔ *„Neu“* ein leeres Set erstellt. Für das Einfügen der

Beispiele gibt es verschiedene Möglichkeiten. Einmal können verschiedene Messdaten im Informationsfeld selektiert werden, die über den Menüpunkt *„IMS-Set“* ➔ *„Füge Auswahl hinzu“* zum Set hinzugefügt werden. Wählt man nun den Informationsfeld-Reiter *„IMS-Set“* werden die Messdaten im Set angezeigt. Nach Auswahl eines Beispiels wird dieses im Hauptteil des Fensters als Chromatogramm dargestellt.

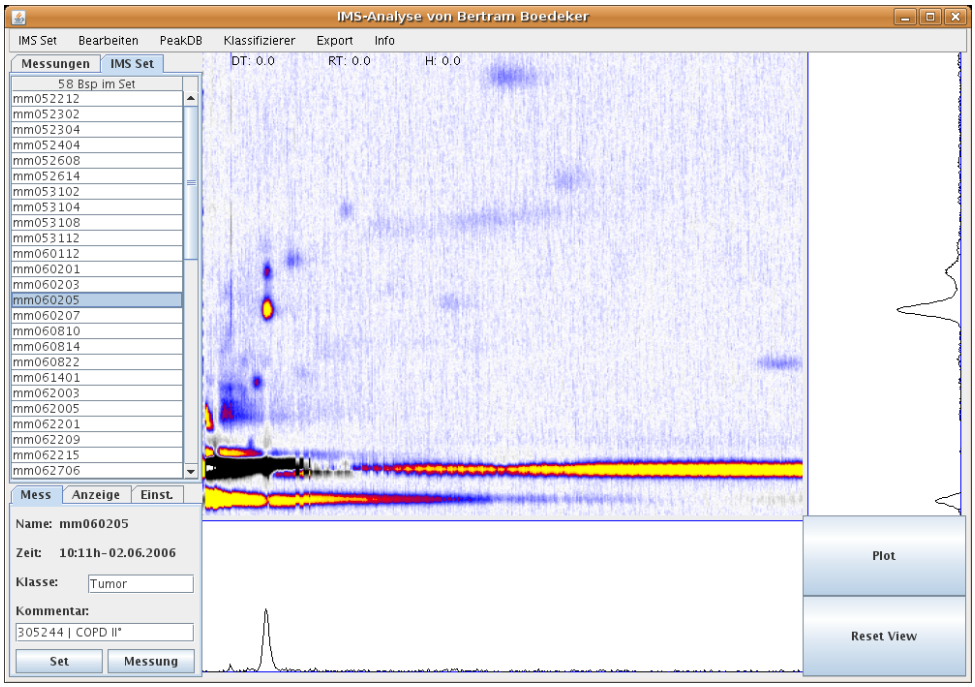


Abbildung A.3: Weitere Informationen zum Messdatem stehen unten links. Hier können die Klasse und ein Kommentar festgelegt werden.

Im unteren Informationsfeld unter Reiter *„Mess“* werden die zusätzlichen Informationen: Name, Datum und Uhrzeit der Messdatenaufnahme, Klasse des Messdatums und Kommentar sichtbar (siehe Abb. A.3). Hier kann die Klasse, der das jeweilige Beispiel angehören soll, eingetragen und ein Kommentar hinzugefügt werden. Mit den darunterliegenden Schaltknöpfen *„Set“* und *„Messung“* wird die Eingabe zum Messdatum im Set bzw. im Messdatum selbst gespeichert. Die Klassenzugehörigkeit sollte nur im Set gespeichert werden, da das Messdatum je nach Klassifikationsaufgabe zu unterschiedlichen Klassen gehören könnte. Als Kommentar eines Messdatums wäre z.B. der Vermerk, dass es sich um eine Atemluftmessung handelt und die Angabe der entsprechenden Patientennummer denkbar. Diese Angabe könnte im Messdatum gespeichert werden, da

sie unabhängig von der Klassifikationsaufgabe gilt.

Eine andere Möglichkeit Messdaten in das Set einzufügen ist die Wahl des Menüpunktes *„Importiere aus Textdatei“* im Menü *„IMS-Set“*. Hier kann eine bereits erstellte Textdatei geöffnet werden, bei der jede Zeile den Namen eines Messdatums enthält. Diese Messdaten werden ins Set übernommen. Noch komfortabler können Messdaten über den Menüpunkt *„Importiere aus .csv Datei“* im Menü *„IMS-Set“* in das Set geladen werden. Eine solche Datei sollte tabellenförmig die Angaben der ins Set aufzunehmenden Messdaten enthalten, die Spalten sind dabei durch Kommata getrennt. In der ersten Zeile stehen die Spaltentitel. Eine Spalte mit dem Titel *„Messung“* ist Pflicht, weitere optionale Spaltentitel können *„Klasse“*, *„Kommentar“*, *„PatNr“* oder *„NebenDiagnose“* sein. Die Schreibweise ist dabei genau einzuhalten. Falls die Spalte *„Klasse“* vorhanden ist, wird ihr Wert dem jeweiligen Messdatum als Klasse zugeordnet. Aus den Werten der anderen Spalten (falls vorhanden) wird der Kommentar zusammengesetzt. Viele Tabellenkalkulationsprogramme unterstützen den Export in das *„.csv“* Format, womit sie sich zum bequemen Erstellen dieser Dateien qualifizieren.

Um Messdaten aus dem Datensatz zu entfernen, werden diese mit der Maus markiert und im Menü *„IMS-Set“* wird *„Lösche Auswahl“* gewählt. Ein weiterer Unterpunkt des Menüs *„IMS-Set“* ist *„Zeige Alle Beispiele“*, womit ein Textfenster geöffnet wird, welches die Kenndaten aller Beispiele im Datensatz anzeigt. Die Unterpunkte *„Sortiere nach Klassen“* und *„Sortiere nach Namen“* sortieren die Beispiele im Datensatz nach Klassenzugehörigkeit bzw. Messdatennamen.

A.6 Peakdatenbank

Der nächste wichtige Schritt ist der Aufbau einer Peakdatenbank. Zunächst wird über den Menüpunkt *„Zeige PeakDB“* im Menü *„PeakDB“* das Hauptfenster der Peakdatenbank geöffnet (siehe Abb. A.4).

Hier stehen folgende Optionen zur Auswahl:

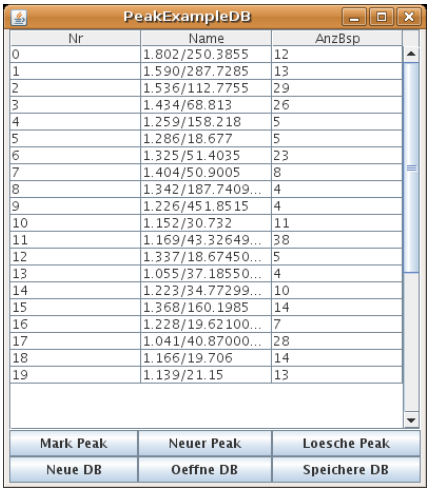


Abbildung A.4: Hauptfenster der Peakdatenbank mit 20 Peaks

- *„Neue DB“* legt eine neue leere Peakdatenbank an.
- *„Öffne DB“* öffnet eine bestehende Peakdatenbank, die über ein Dialogfenster ausgewählt wird.
- *„Speichere DB“* speichert die aktuelle Peakdatenbank unter dem Namen, der in einem Dialogfeld bestimmt wird.
- *„Mark Peak“* zeigt den Fensterbereich des gewählten Peaks im Chromatogramm an.
- *„Neuer Peak“* nimmt einen neuen Peak in die Peakdatenbank auf. Der Fensterbereich ist durch das zuletzt markierte schwarze Fenster im Chromatogramm (siehe Annotationen) definiert. Dieser Bereich wird aus allen Beispielen des Sets, deren Maximum in diesem Bereich über einem anzugebenden Grenzwert liegt, in die Peakdatenbank kopiert. Der Name des neuen Peaks ergibt sich aus dem Mittelpunkt des Fensterbereiches.
- *„Lösche Peak“* löscht den markierten Peak mit den zugehörigen Beispielen aus der Peakdatenbank.

Der Menüeintrag *„Neuen Peak einfügen“* im Menü *„PeakDB“* hat die selbe Funktionalität wie der Schaltknopf *„Neuer Peak“*. Weiter kann der Anwender mehrere neue Peaks gleichzeitig in die Peakdatenbank einfügen, in dem er die gewünschten Peakbereiche durch

Fenster markiert (siehe Annotationen) und im Menü *„Neue Peaks einfügen“* wählt. Durch das Anklicken eines Peaks der Peakdatenbank mit der linken Maustaste wird ein weiteres Fenster geöffnet, das die zu diesem Peak gespeicherten Beispiele zeigt (siehe Abb. A.5).

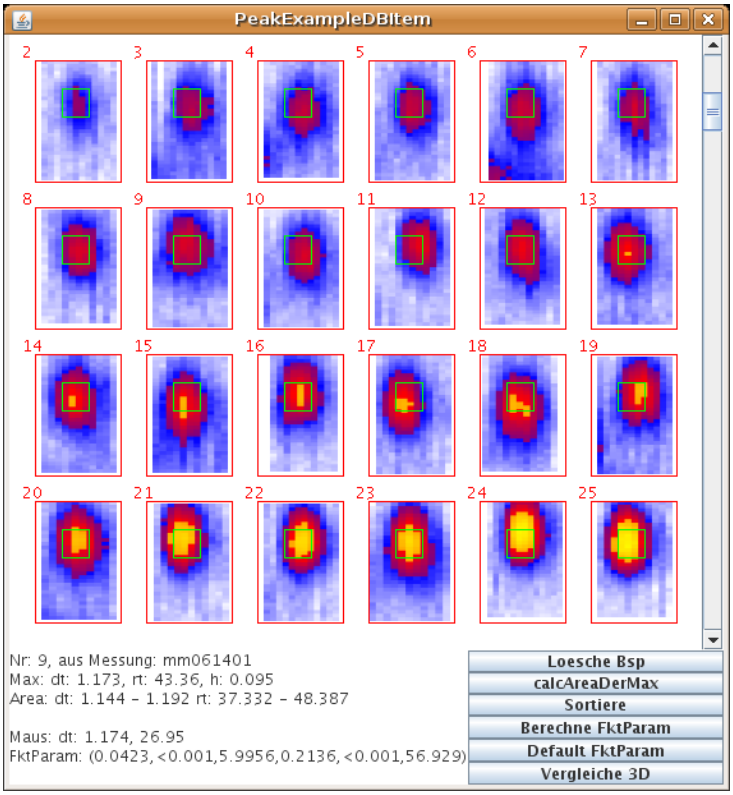


Abbildung A.5: Beispiele eines Peaks der Peakdatenbank, sortiert nach Intensität

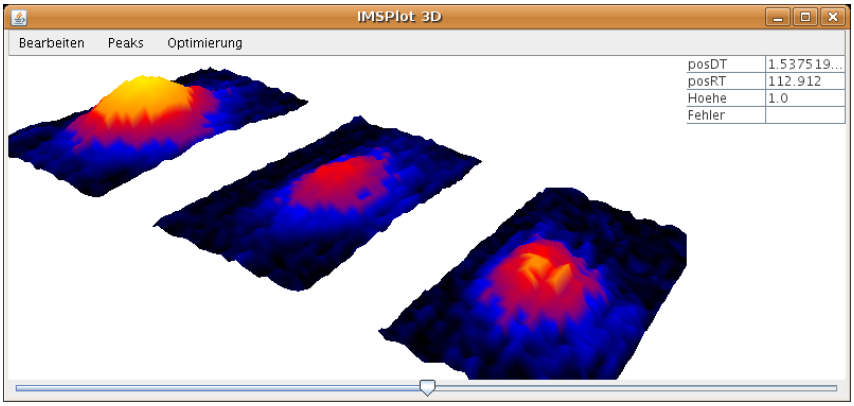
Hier stehen folgende Befehle zur Verfügung:

- Beispiele können mit der Maus gewählt werden. Dadurch werden weitere Informationen zu diesem Beispiel angezeigt.
- *„Lösche Bsp“* löscht das ausgewählte Beispiel zum Peak.
- *„calcAreaDerMax“* berechnet den Bereich, der alle Maxima der Beispiele einschließt (grünes Fenster) neu. Dies sollte vor jeder Berechnung der Funktionsparameter durchgeführt werden.
- *„Sortiere“* sortiert die Beispiele aufsteigend nach maximaler Intensität.

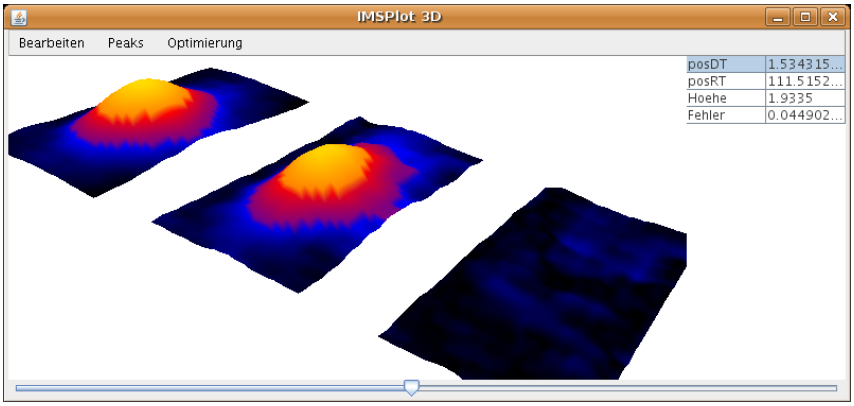
- *„Berechne FktParameter“* startete die Ermittlung der Peakfunktionsparameter des Peaks mit Hilfe einer Evolutionsstrategie und den aktuellen Beispielen zu diesem Peak.

- *„Default FktParameter“* setzt die Peakfunktionsparameter auf ihre Ausgangswerte zurück.

- *„Vergleiche 3D“* öffnet eine dreidimensionale Ansicht um zwei Peakbeispiele miteinander vergleichen zu können (siehe Abb. A.6). Dadurch kann z.B. festgestellt werden, ob zwei Peakbeispiele durch die gleiche Peakfunktion darstellbar sind.



(a) Frage: Kann das erste Peakbeispiel (Links) durch das zweite Peakbeispiel (Mitte) durch Lageverschiebung und Intensitätsskalierung dargestellt werden?



(b) Antwort: Ja, es kann, die Differenz ist nahe Null. Die beiden Peakbeispiele wurden zusätzlich mit einem Tiefpassfilter geglättet.

Abbildung A.6: Vergleich zweier Beispiele eines Peaks der Peakdatenbank. Der rechte Bereich zeigt jeweils die Differenz zwischen den beiden Peaks.

Ist die Peakdatenbank mit repräsentativen Beispielen jedes Peaks gefüllt, kann über den Menüeintrag *„Berechne Funktionsparameter“* im Menü *„PeakDB“* die evolutionäre Bestimmung aller Peakfunktionsparameter angestoßen werden. Danach ist es sinnvoll die Peakdatenbank zu speichern. Nun können die Peaks der Messdaten mit Hilfe der Peakdatenbank aufgelöst werden. Dazu ist es einmal möglich, die Peaks eines einzelnen Messdatums zu bestimmen. Dies geschieht durch den Aufruf von *„Resolve Peaks ES“* im Menü *„PeakDB“*. Nach der Peakbestimmung wird ein Fenster geöffnet, welches die gefundenen Peaks als Chromatogramm darstellt. Dieses kann später über den Menüpunkt *„Show Resolved Map“* jederzeit wieder geöffnet werden. Der Menüpunkt *„Show Rest Map“* öffnet hingegen ein Fenster, in dem das Chromatogramm angezeigt wird, das entsteht wenn die Intensitätswerte der aufgelösten Peaks von den originalen Intensitätswerten abgezogen werden. Hier sind Bereiche, deren Peaks nicht in der Datenbank enthalten sind, identifizierbar. Die Peaks aller Beispiele im IMS-Set werden durch die Wahl des Menüpunktes *„Resolve All Peaks ES“* berechnet. Dies kann besonders auf älteren Rechnern viel Zeit in Anspruch nehmen. Nach dieser Berechnung empfiehlt es sich das IMS-Set zu speichern, dadurch werden auch die gefundenen Peaks gesichert.

A.7 Klassifikation

Nachdem zu allen Beispielen des IMS-Sets die Peaks mit Hilfe der Peakfunktionen bestimmt sind, kann die Klassifikation getestet werden. Dafür stehen zwei Methoden bereit. Über den Menüpunkt *„Klassifizierer“* ➔ *„Teste mit MyNaiveBayes“* wird ein Klassifikationsmodell erstellt, dem alle Beispiele im Set als Trainingsmenge dienen. Mit diesem Klassifikationsmodell werden im Anschluss alle Beispiele des Sets klassifiziert und es wird überprüft, ob die Zuordnung richtig ausfällt. Natürlich sagt dieses Vorgehen nichts über die Leistung des Modells bzgl. unbekannter Messdaten aus. Die zweite Testmethode wird über den Menüpunkt *„Klassifizierer“* ➔ *„Teste mit MyNaiveBayes leave-one-out“* aufgerufen. Hier werden nach dem Prinzip der *„Leave-One-Out“* Methode genausoviele Klassifikationsmodelle erstellt, wie es Beispiele im Set gibt, wobei jedes Messdatum einmal nicht zum Training verwendet wird, sondern zum anschließenden Test des Klassifikationsmodells bereitsteht. Bei beiden Testmethoden werden zunächst zwei Parameter abgefragt:

Zum einen muss der Parameter der Streuungsbreite (entspricht σ_1 in Abschnitt 4.2) angegeben werden, zum anderen wird die Selektionsmethode der Attribute in einem Auswahldialog festgelegt. Dafür stehen diese Methoden zur Wahl:

- *„keine Selektion (alle Attribute)“*: Es werden alle Attribute verwendet.
- *„manuell“*: Über ein Textfeld werden die gewünschten Attributnummern eingegeben.
- *„manuell mit Anzeige der berechneten Wahrscheinlichkeitsdichten“*: Es öffnet sich ein Fenster mit einer Übersicht über die Wahrscheinlichkeitsdichten aller Attribute, getrennt nach Klassen (siehe Abb. A.7). Hier können die gewünschten Attribute mit der rechten Maustaste markiert werden. Durch einen Klick der linken Maustaste auf ein Attribut wird ein Fenster mit der Detailansicht der Verteilungsdichte geöffnet (siehe Abb. A.8).
- *„automatische Vorwärts-Selektion (greedy)“*: Die Attribute werden bei der Erstellung eines Klassifizierers automatisch gewählt (siehe Seite 51).

Mit diesen Parametern und den Beispielen im Set wird die gewählte Testmethode des Klassifikationsverfahrens angewendet und das Ergebnis in einem Fenster angezeigt (siehe Abb. A.9).

Um weitere Details zur Klassifikation eines einzelnen Beispieles zu erhalten reicht ein Klick mit der linken Maustaste auf die entsprechende Zeile. Es öffnet sich das Fenster aus Abbildung A.10.

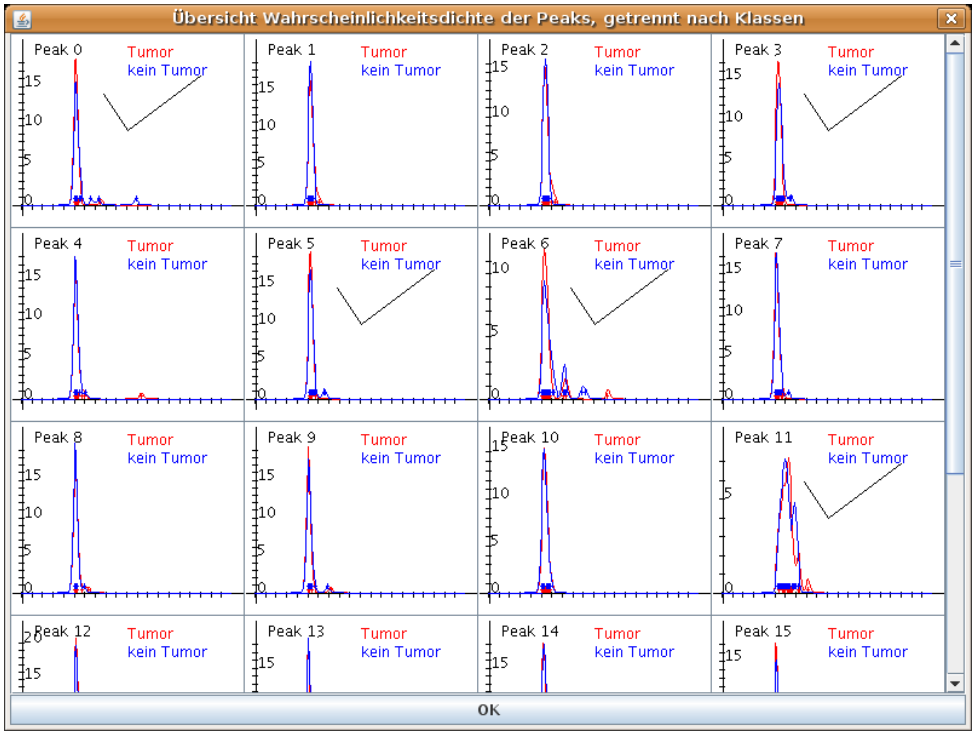


Abbildung A.7: Übersicht über die Wahrscheinlichkeitsdichten aller Attribute, getrennt nach Klassen. Markierte Attribute werden für die Klassifikation verwendet.

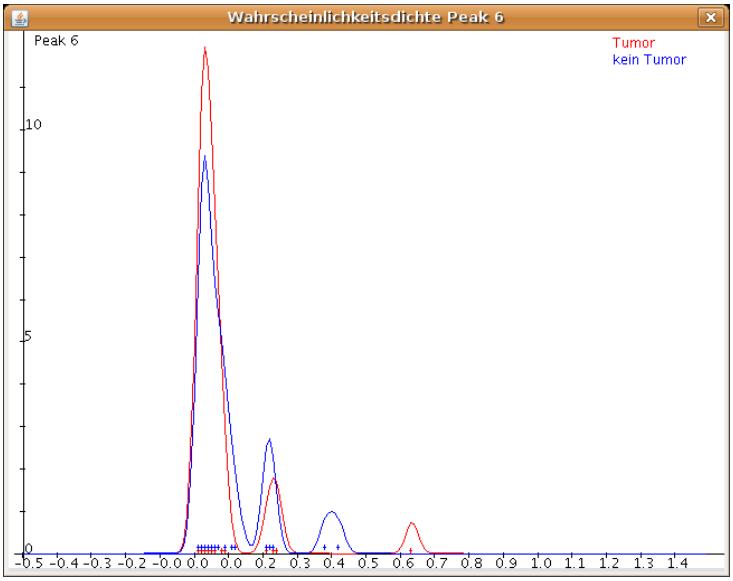


Abbildung A.8: Detailansicht der Wahrscheinlichkeitsdichte eines Attributs, der rote Graph zeigt die Verteilungsdichte der Peakhöhe von Peak 6 der Patienten mit einem Tumor, der blaue Graph die der Patienten ohne Tumor.

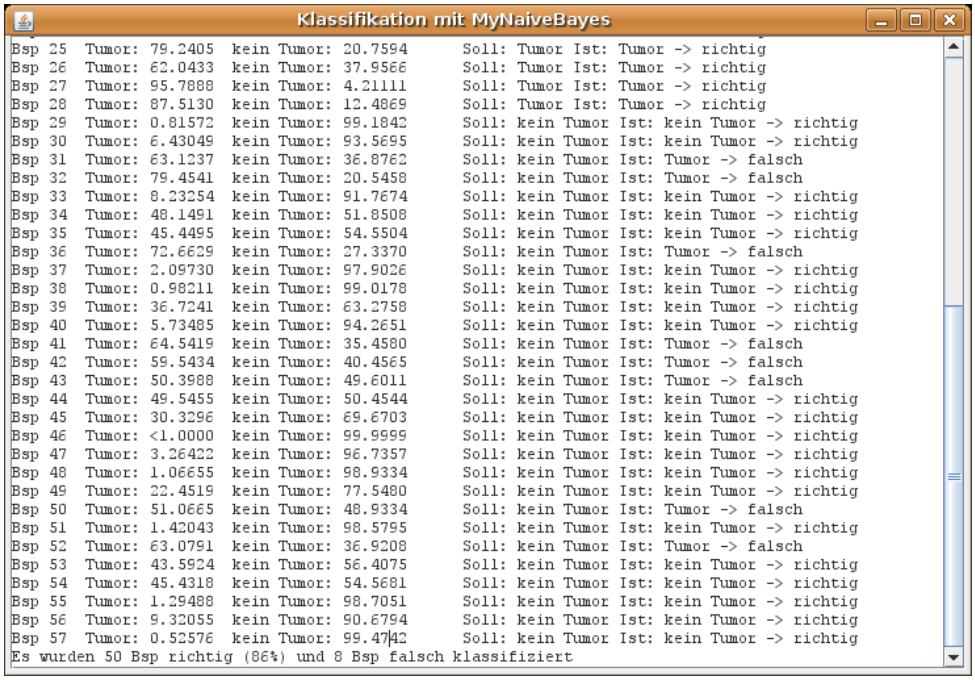


Abbildung A.9: Ergebnis eines Klassifikationstests, der Klick der linken Maustaste auf eine Zeile öffnet Details zur Klassifikation des jeweiligen Beispiels

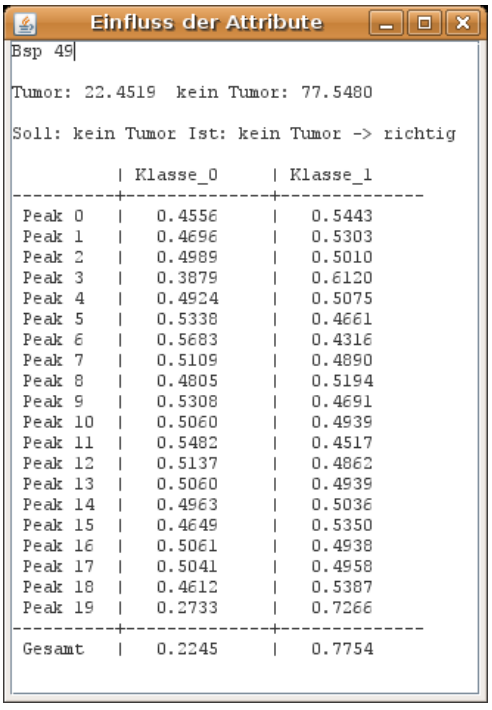


Abbildung A.10: Details zur Klassifikation eines Messdatums, dargestellt wird der Einfluss jedes Attributes am Gesamtklassifikationsergebnis

A.8 Weitere Funktionen

- *„Plot 3D“* zeigt den Chromatogrammbereich des zuletzt markierten Fensters in einer dreidimensionalen Darstellung (siehe Abb. A.11). Die Kameraposition kann mit der Maus frei bestimmt werden.

- *„Compare 3D“* ermöglicht den Vergleich eines Bereiches zweier Chromatogramme in einer dreidimensionalen Darstellung (ähnlich zu der Funktion *„Vergleiche 3D“* zum Vergleich von Peakbeispielen). Der Bereich wird durch das zuletzt markierte Fenster im Chromatogramm bestimmt, die zwei Beispiele durch Selektion in der Liste des Sets. Ein Beispiel zeigt die Abbildung A.12.

- *„Export Image“* speichert den aktuellen Bildausschnitt des angezeigten Chromatogramms in einer zu wählenden Datei im „jpeg“ Format.

- *„Export All Images“* speichert den Bildausschnitt des zuletzt markierten Fensters von allen Beispielen in einem Verzeichnis der Wahl.

- *„Export All ResolvedMap Images“* speichert die Chromatogramme der gefundenen Peaks der Peakdatenbank aller Beispiele im Bereich des zuletzt markierten Fensters in einem Verzeichnis der Wahl.

- *„Export All RestMap Images“* speichert analog zu *„Export All ResolvedMap Images“* die Chromatogramme der Differenzen zwischen Messdatenintensität und gefundenen Peaks.

- *„Export All Wdichte Images“* speichert die Bilder der berechneten Wahrscheinlichkeitsdichtefunktionen getrennt nach Klassen für alle Attribute in einem Verzeichnis der Wahl.

- *„Export All Wdichte Images ohne Klassentrennung“* speichert die Bilder der berechneten Wahrscheinlichkeitsdichtefunktionen ohne Berücksichtigung der Klasse für alle Attribute in einem Verzeichnis der Wahl.

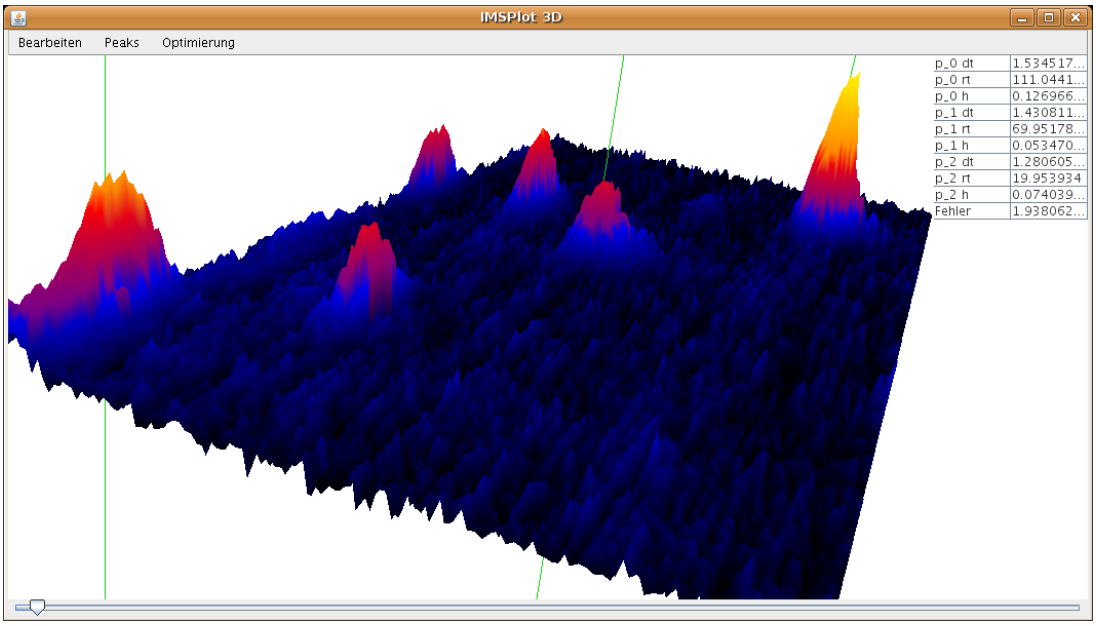


Abbildung A.11: Dreidimensionale Darstellung der Signalintensitäten eines Bereichs. Der Blickwinkel und Zoomfaktor kann beliebig eingestellt werden.

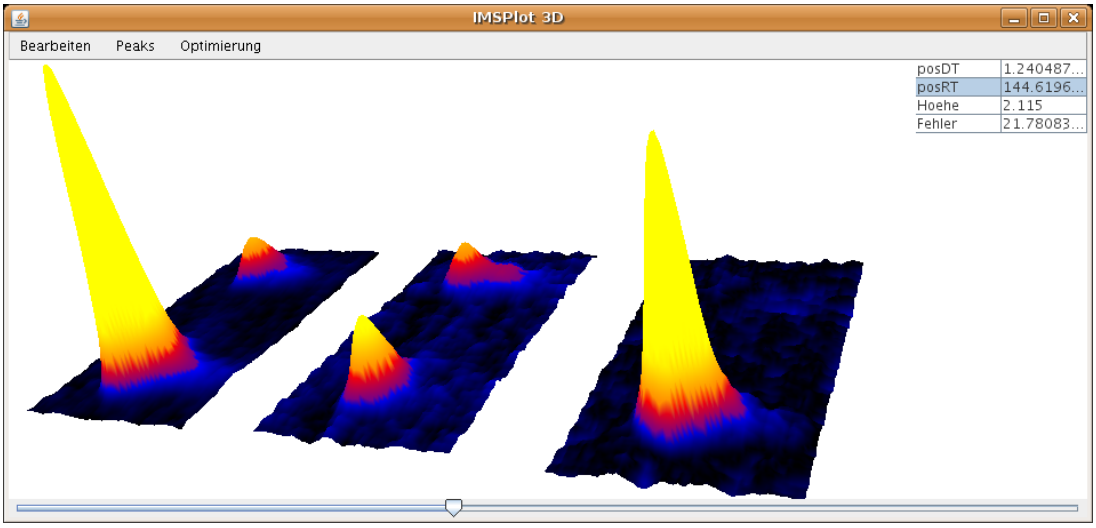


Abbildung A.12: Vergleich zweier Chromatogramme durch dreidimensionale Darstellung des gleichen Bereiches. Links ist das erste Beispiel, in der Mitte das zweite Beispiel und rechts die Signaldifferenz zwischen dem ersten und zweiten Beispiel dargestellt. Die Intensität beim zweiten Beispiel wurde um den Faktor $2,1$ verstärkt, sodass der hintere Peak im Differenzbild verschwindet. Der vordere Peak bleibt trotz der Verstärkung des zweiten Beispiels im ersten Beispiel eindeutig höher.

Abbildungsverzeichnis

Literaturverzeichnis

[Duda01] R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification, Wiley 2001

[Witten01] I. H. Witten, E. Frank: Data-Mining, Hanser 2001

[Langley94] P. Langley,S. Sage: Induction of Selective Bayesian Classifiers, In Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence (S. 399-406), Morgan Kaufmann 1994

[Eiceman05] G.A. Eiceman, Z. Karpas: Ion Mobility Spectrometry 2nd ed., Taylor & Francis, 2005

[Baumbach05a] J. I. Baumbach: Metabolomics - Charakterisierung wesentlicher Stoffwechselprodukte von Zellen, Bakterien und beim Menschen, BIOforum 10/2005, S. 45-47

[Baumbach05b] J. I. Baumbach: Metabolomics - Spektrometerentwicklung für die medizinische Analytik beispielsweise in Ausatemluft, GIT Labor Fachzeitschrift 11/2005, S. 2-4

[Ruzsányi05] V. Ruzsányi: Analyse flüchtiger Metaboliten von der Ausatemluft mittels Ionenmobilitätsspektrometer, Dissertation, Fachbereich Bio- und Chemieingenieurwesen, Universität Dortmund, 2005

[Meier06] M. Meier: Atemluftanalyse mittels Ionenmobilitätsspektrometrie in klinischer Anwendung, Diplomarbeit, Fachhochschule Gelsenkirchen, 2006

Danksagung

Herrn Prof. Dr. Rudolph danke ich sehr für die Vergabe des Themas, die Herstellung des Kontaktes zum ISAS und die Betreuung der Arbeit. Sie haben es mir damit ermöglicht, die wissenschaftlichen Arbeitsmethoden des ISAS kennenzulernen.

Herrn PD Dr. Baumbach danke ich sehr für seine stetige Betreuung, die weit über die wöchentlichen Diskussionsrunden hinausging. Sie haben mich stets mit wertvollen Ratschlägen unterstützt und die Grundlage für mein Verständnis des komplexen Themenbereiches der IMS-Analyse geschaffen.

Frau Sabine Bader danke ich sehr für die kompetente und freundliche Betreuung. Oft warst Du meine erste Anlaufstelle bei Fragen, Deine Anregungen haben mich auch in schwierigen Abschnitten oft weitergebracht.

Meinen Mitstudenten der „Montagsrunde“ (Ihr wisst wer, Ihr seid) danke ich für die netten Kontakte, die geführten Diskussionen und das gemeinsame Durchstehen der Rückschläge.

Dem gesamten Team des ISAS danke ich für die freundliche Aufnahme und Unterstützung während der Zeit meiner Arbeit.