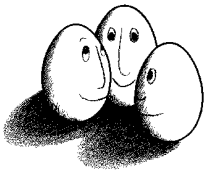
Diplomarbeit

Die Relevanz bestimmter Exons für die Überlebensprognose beim Neuroblastom

Miriam Bützken



Diplomarbeit
am Fachbereich Informatik
der TU Dortmund

29. September 2009

Betreuer:

Prof. Dr. Katharina Morik
Dipl.-Inform. Benjamin Schowe

Danksagung

An dieser Stelle möchte ich mich ganz herzlich bei Prof. Dr. Katharina Morik und Dipl.-Inf. Benjamin Schowe für die freundliche und ermutigende Betreuung während dieser Arbeit bedanken.
Weiterhin danke ich allen, die diese Arbeit Korrektur gelesen haben.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einleitung

1.1. Einführung

Zu den häufigsten bösartigen Krebserkrankungen im Kindesalter zählt das Neuroblastom. Neuroblastome haben einen sehr variablen Krankheitsverlauf, dieser reicht vom Tod (ca. 40% der erkrankten Kinder versterben trotz intensiver Behandlung innerhalb der ersten Jahre der Erkrankung) bis zu sogenannten „Spontanheilungen“, bei denen sich der Tumor ohne Behandlung zurückbildet.
Ein Hauptziel der gegenwärtigen Neuroblastom-Forschung besteht darin, eine möglichst genaue Einschätzung des Tumor-Verhaltens vorhersagen zu können. Dazu wird versucht, ein Screening-Programm zu entwickeln, das „harmlosere“ Formen der Erkrankung von Tumoren abgrenzt, die schlechtere Prognosen bieten. Mittels dieser Abgrenzung könnten sowohl ein „Overtreatment“ (Überbehandlung) von spontan regredierenden Tumoren wie auch ein „Undertreatment“ (Unterbehandlung) von aggressiven Formen des Tumors verhindert werden.
Eine weitere erstrebenswerte Zukunftsvision ist, für jeden einzelnen Patienten bei Diagnosestellung eine Risikoabschätzung abgeben zu können und so ein möglichst effektives, auf den einzelnen Patienten maßgeschneidertes Therapiekonzept zu erstellen.
Für diese Zielsetzungen ist von großer Bedeutung, Informationen der Genaktivitäten von Neuroblastom-Patienten zu erfassen, weil die Aktivität bestimmter Gene in einer Zelle einen Indikator für den Zellzustand und für die Funktionen, die diese Zelle wahrnimmt, darstellt. Ein wichtiges Hilfsmittel dabei ist die *DNA-Microarray-Technologie*. Mit dieser Technologie können die Genaktivitäten von mehreren tausend Genen simultan ermittelt werden. In der Tumor-Forschung, insbesondere in der Neuroblastom-Forschung, erhofft man sich durch den Einsatz solcher Arrays einen Einblick in die Mechanismen des Tumorgeschehens. Microarrays sind daher aus der molekularbiologischen Grundlagenforschung nicht mehr wegzudenken.
Neben den klassischen Expressions-Microarrays, die den Expressionswert (=Aktivität) eines Gens ermitteln, werden auch sogenannte „Exon-Arrays“ für eine detailliertere Auswertung herangezogen. Diese sind eine Weiterentwicklung der Expressions-Microarrays, die nicht mehr „Gen-basiert“ arbeiten, sondern die Feinstruktur der Gene einbeziehen. Dabei werden die Gen-Untereinheiten, die „Exons“, gezielt berücksichtigt und deren Expressionen getrennt ermittelt. Ein Ziel

von Microarray-Experimenten ist die Untersuchung von Unterschieden in der Expression zwischen Proben aus verschiedenen Gruppen/Populationen, um z.B. diejenigen Gene oder Exons zu bestimmen und zu identifizieren, die zu Unterschieden in der Erkrankung und deren Verlauf beitragen.

Durch Microarray-Experimente werden enorme Datenmassen generiert. Die Herausforderung in der statistischen Analyse besteht daher in der hohen Zahl von Genabschnitten mit gemessener Expression im Vergleich zur meist geringen Anzahl untersuchter Proben. Das heißt, es werden viele Merkmale (Attribute), aber nur wenige Beispiele (Patienten) erhoben. Durch die Weiterentwicklung der Microarray-Technologie und der Einführung der Exon-Microarrays hat sich die Anzahl der zu messenden Merkmale noch weiter erhöht. So werden z.B. bei den in dieser Arbeit verwendeten Microarrays pro Patient ca. 1,4 Millionen Merkmale gemessen, von denen ungefähr 280.000 für die weitere Analyse verwendet werden - bei nur 131 Beispielen (Patienten).

Bei der Analyse der Expressionswerte des Neuroblastom-Datensatzes ergeben sich somit zwei Fragestellungen:

- Kann ein (guter) Klassifikator zur Vorhersage des Krankheitsverlaufes gefunden werden?
- Können signifikante Exons ermittelt werden, die biologische Rückschlüsse auf die Überlebensprognose der Patienten zulassen und die für die weitere Neuroblastom-Forschung von Bedeutung sind?

In dieser Arbeit wird untersucht, ob und wie gut sich die Patienten anhand ihrer Exon-Expressionsprofile klassifizieren lassen. Hierbei soll die Hypothese untersucht werden, dass nicht ein bestimmtes Exon die Klassifikation determiniert, sondern dass das Zusammenspiel einer Anzahl von Exons für die Zugehörigkeit zu einer Klasse verantwortlich ist.
Zur Bestimmung differentiell exprimierter Exons werden Methoden zur Merkmalsauswahl eingesetzt. Mit derart ausgewählten Merkmalen können oftmals Klassifikatoren konstruiert werden, die eine gute Vorhersage für neue Patientenprofile liefern. Jedoch konnte gezeigt werden, dass auch auf ausgewählten Merkmalsmengen, die wenige (bis keine) Übereinstimmungen aufweisen, ähnlich gute Klassifikationsergebnisse erreicht werden können. Aus biologischer Sicht sind solche Instabilitäten meistens nicht wünschenswert. Das Interesse besteht hier in der Auswahl von Merkmalen (Genen und/oder Exons), deren Ausprägung auf eine genetische Grundlage zurückgeführt werden kann. Kann solch eine Grundlage vermutet werden, sollten diese Merkmale von den Auswahlmethoden auch bei verschiedenen Datensätzen immer unter den relevant gefundenen Merkmalen sein. Solche Exons könnten helfen, evtl. Rückschlüsse auf die Prognose von Neuroblastom-Patienten

(Überleben ohne Rezidiv oder Rückfall der Erkrankung) zuzulassen und zum Verständnis der Entstehung der Erkrankung beizutragen. Aus biologischer Sicht ist somit eine stabile Auswahl von Merkmalen wünschenswert.

1.2. Struktur dieser Diplomarbeit

Kapitel 2 enthält zunächst einen Überblick über die Krebserkrankung des Neuroblastom.

In Kapitel 3 werden für das Verständnis dieser Arbeit benötigte Grundlagen vorgestellt. Im ersten Teil wird ein Einblick in die Molekularbiologie gegeben. Danach erfolgt eine Beschreibung der DNA-Microarray-Technologie. Im letzten Abschnitt werden die vorliegenden Daten und deren Gewinnung erläutert.

Kapitel 4 beschäftigt sich mit dem *maschinellen Lernen*. Neben einer Einführung in das Thema und der Definition wichtiger Begriffe werden verschiedene Lernverfahren, die in dieser Arbeit zum Einsatz kommen, vorgestellt und anschließend aufgezeigt, wie die Lernerfolge dieser Verfahren bewertet werden können. Der letzte Abschnitt des Kapitels beschreibt die Anwendung des maschinellen Lernens auf die vorliegenden Daten.

Kapitel 5 widmet sich der Merkmalsauswahl. Im ersten Abschnitt erfolgt eine Einführung in diesen Themenbereich und eine Erläuterung wichtiger statistischer Kennwerte. Danach werden die in dieser Arbeit verwendeten Methoden zur Merkmalsauswahl vorgestellt. Im Anschluss daran wird auf die wichtige Kenngröße „Robustheit“ der Auswahl von Merkmalen eingegangen und in diesem Zusammenhang das spezielle Verfahren der sog. *Ensembles* erläutert.

In Kapitel 6 werden die durchgeführten Experimente mit den vorliegenden Daten beschrieben. Zu Beginn wird die verwendete Lernumgebung *RapidMiner* vorgestellt und die notwendige Vorverarbeitung der Daten erläutert. Anschließend werden die einzelnen Experimentreihen beschrieben und deren Ergebnisse bewertet.

In Kapitel 7 werden die ausgewählten Merkmalsmengen auf Übereinstimmungen sowie auf Korrelationen zwischen ihren Merkmalen untersucht und die Patientenprofile anhand ihrer Merkmale analysiert. Anschließend erfolgt eine Einführung in die Klassifikation von Texten und die Vorstellung des von Joachims [34] entwickelten statistischen Lernmodells für die Textklassifikation. Es wird untersucht, ob die Eigenschaften dieses Modells auf die vorliegenden Neuroblastom-Daten übertragen und daraus Ähnlichkeiten zwischen den Lernaufgaben festgestellt werden können.

Kapitel 8 enthält eine abschließende Zusammenfassung der vorliegenden Arbeit und einen Ausblick auf mögliche weitere Untersuchungen.

2. Das Neuroblastom

Dieses Kapitel gibt einen Überblick über das „Neuroblastom“, einer Krebserkrankung im Kindesalter und über die Besonderheiten dieser Tumoren.

2.1. Allgemeines

Das Neuroblastom ist ein maligner (bösartiger) Tumor, der vorwiegend im Säuglings- und Kleinkindalter auftritt. Er geht aus entarteten, unreifen (embryonalen) Zellen des sympathischen Nervensystems hervor, welches als ein Teil des autonomen Nervensystems die unwillkürlichen Funktionen, wie Herz- und Kreislaufsystem, Darm- und Blasentätigkeit, usw. steuert. Die Fehlentwicklung dieser noch nicht ausgereiften Nervenzellen beginnt bereits vor der Geburt und kann eine Folge von Chromosom-Veränderungen und/oder fehlerhaften Genregulationen sein. Neuroblastome sind überall dort lokalisiert, wo sich sympathisches Nervengewebe befindet. Am häufigsten entstehen sie im Nebennierenmark sowie im Bereich des Nervengeflechts auf beiden Seiten der Wirbelsäule, dem sogenannten Grenzstrang.
Die Therapie des Neuroblastoms gestaltet sich oftmals schwierig, da einige dieser Tumoren sich häufig erst in fortgeschrittenen Stadien manifestieren, sehr aggressiv wachsen und häufig Resistenzen gegenüber gängigen Chemotherapeutika aufweisen.

2.2. Epidemiologie

Neuroblastome machen ungefähr 7-10% aller Krebserkrankungen im Kindes- und Jugendalter aus [12]. Sie sind die häufigsten soliden Tumore in der Kindheit [7]. Etwa eines von 100.000 Kindern ist pro Jahr betroffen. In Deutschland erkranken jährlich etwa 150 Kinder neu an einem Neuroblastom. Der Altersmedian zum Zeitpunkt der Diagnose beträgt 18 Monate; am häufigsten betroffen sind Neugeborene und Säuglinge vor dem ersten Lebensjahr (etwa 40%), 75% bis 4 Jahre und 98% bis maximal 10 Jahre [12]). Das Geschlechterverhältnis ist mit 1,1:1 für Jungen zu Mädchen annähernd gleich. Es wird vermutet, dass eine nicht unerhebliche Anzahl von Erkrankungen im Säuglingsalter unentdeckt regrediert (sich zurückbildet) [53].

2.3. Stadieneinteilung nach INSS

Mit Hilfe des *International Neuroblastoma Staging System* (INSS) lassen sich die Patienten hinsichtlich ihres Krankheitsstadiums in unterschiedliche Gruppen einteilen [13]. Die Stadieneinteilung richtet sich nach der Ausbreitung des Tumors zum Zeitpunkt der Diagnosestellung. Diese Unterscheidung in verschiedene Stadien ist wichtig für die exakte Therapie und gibt Anhaltspunkte für die Prognose und das Ansprechen der Therapie.
Folgende Stadien werden unterschieden:

- *Lokalisierte Stadien (Stadien 1-3)*: In diesen Stadien steht die Operation im Vordergrund, kombiniert mit einer Chemotherapie. Prognostisch ungünstig ist Stadium 3 und ein Alter >1 Jahr.
  - *Stadium 1*: Der Primärtumor ist auf das Ursprungsorgan beschränkt und lässt sich chirurgisch komplett entfernen; verdächtige Lymphknoten sind negativ.
  - *Stadium 2A*: Der Primärtumor konnte nicht komplett entfernt werden. Die Größenausdehnung überschreitet die Wirbelsäule nicht; kein Lymphknotenbefall in der Umgebung des Tumors.
  - *Stadium 2B*: Die Größenausdehnung des Tumors überschreitet die Wirbelsäule nicht; Lymphknoten der gleichen Seite sind befallen.
  - *Stadium 3*: Der Primärtumor ist weit über den Ursprungsort mit Befall der Lymphknoten hinausgewachsen und lässt sich nicht komplett entfernen.

- *Stadium 4*: Der Primärtumor und Metastasen sind an vielen Stellen des Körpers, z.B. in Leber, in Lymphknoten, im Knochen, im Knochenmark, in der Haut und/oder in anderen Organen.
  Die Behandlung erfolgt mittels einer Kombination aus intensiver Chemotherapie und einer Operation. Prognostisch ungünstig zeigen sich auch hier wiederum ein Alter >1 Jahr sowie das Vorliegen einer MYCN-Amplifikation (siehe Kap. 2.5).

- *Stadium 4S*: Vorhandener Primärtumor wie beim Stadium 1 und 2, Metastasen sind nur in der Haut, Leber und/oder Knochenmark. Der Knochenmarkbefall ist gering. Das Stadium *4S* wird nur bei Säuglingen im ersten Lebensjahr benannt.
  Patienten, die dem Stadium *4S* zugeordnet werden, müssen bei Erfüllung bestimmter Kriterien nicht behandelt, sondern können klinisch beobachtet werden. In das Stadium 4S fallen Säuglinge, die klinisch stabil sind. Zeigt der Tumor dann eine nicht bedrohliche Progression oder Regression sowie

eine fehlende MYCN-Amplifikation (siehe Kap. 2.5), ist eine Behandlung vorerst nicht erforderlich.

Anmerkung: Bei Neuroblastomen des Stadiums *4S*, die nur im Säuglingsalter beobachtet werden, können als Besonderheit Spontanregressionen ohne jegliche Therapie auftreten.

2.4. Therapie/Prognose

Die Behandlung richtet sich nach der örtlichen Ausdehnung des Tumors (Stadium) und dem Ausmaß der Bösartigkeit. Grundsätzlich besteht sie aus einer Kombination aus Chemotherapie, Operation und manchmal auch Bestrahlung.
Die Heilungsaussichten lassen sich bei einem Neuroblastom für den Einzelfall nur schwer abschätzen. Sowohl das Ausmaß der Erkrankung als auch die Aggressivität des Tumors und das Alter des Patienten spielen eine entscheidende Rolle.
Eine Prophylaxe ist bisher nicht bekannt. In einer Studie mit dem sogenannten „Windeltest“ wurde untersucht, ob ein Urintest zur Früherkennung des Neuroblastoms geeignet ist. Hierbei wurde der Urin auf tumorspezifische Substanzen getestet. In einer bundesweiten epidemiologischen Studie [53] konnte gezeigt werden, dass der Test ungeeignet ist. Zwar konnte mit dem Test bei einigen Kindern ein frühes Tumorstadium identifiziert werden, aber weder die Sterblichkeit noch die Zahl der fortschreitenden Erkrankungen konnten durch die Früherkennungsmaßnahme gesenkt werden.
Eine gute Prognose besteht bei Kindern mit dem Neuroblastom-Stadium 4*S* sowie in der Regel bei begrenzten Tumoren und bei jüngeren Kindern. Bei älteren Kindern mit metastasiertem Neuroblastom (Stadium 4) sind die Heilungsaussichten trotz intensiver Therapie noch immer ungünstig.
Die Prognose der Einzelstadien nach der Neuroblastomstudie ‘97 [6] basierend auf 5-Jahres-Überlebenswahrscheinlichkeiten ist wie folgt:

- Stadium 1: 99%
- Stadium 2: 93%
- Stadium 3: 83%
- Stadium 4: 31%
- Stadium 4S: 77%

Trotz neuerer und intensiverer Therapiemodalitäten zeigt Stadium 4 ein drastisches Absinken der Überlebensrate.

2.5. Onkogen MYCN

Das Onkogen (Krebsgen) MYCN liegt in etwa 20% der Neuroblastome amplifiziert vor. Betrachtet man Patienten mit lokalisiert wachsenden Tumoren über einen längeren Zeitraum (10 Jahre), so ergibt sich eine Überlebensrate von Kindern mit Tumoren ohne MYCN-Amplifikation von über 90%, mit MYCN-Amplifikation hingegen von nur etwa 40-50 % [7]. Zum Zeitpunkt der Diagnose haben jedoch etwa 70 % der Kinder bereits Metastasen, was ihre Prognose weiter verschlechtert. In dieser Kategorie beträgt für Kinder in einem Alter von unter 12 Monaten die Wahrscheinlichkeit für „ereignisfreies Überleben“ über 3 Jahre bei Tumoren ohne Amplifikation von MYCN 93% und bei Tumoren mit Amplifikation des Onkogens hingegen nur 10% [12].

3. Grundlagen

Dieses Kapitel enthält für das Verständnis dieser Arbeit benötigten Grundlagen der Molekularbiologie, eine Beschreibung der Microarray-Technologie und der durch diese Technologie gewonnenen Daten.

3.1. Molekularbiologie

In diesem Abschnitt werden o.g. Grundlagen erläutert. Es werden die wichtigsten biologischen Moleküle vorgestellt und der Ablauf der Genexpression beschrieben. Der Inhalt ist angelehnt an [56][15] [66][69].

3.1.1. Proteine

Proteine (umgangssprachlich „Eiweiße“) gehören zu den Grundbausteinen aller Zellen und sind die wichtigsten Baustoffe des Körpers. Proteine übernehmen in allen organischen Bereichen wichtige Funktionen:

- Strukturproteine, wie z.B. Keratin, sind Basis-Baustoffe für Haut, Haare und Nägel.
- Enzyme (auch diese zählen zu den Proteinen) katalysieren chemische Reaktionen.
- Transportproteine transportieren z.B. Hämoglobin (roter Blutfarbstoff).
- Bewegungsproteine, wie z.B. Actin und Myosin, ermöglichen Muskelkontraktionen.
- Verteidigungsproteine helfen bei Schutz und Abwehr.
- Regulatorische Proteine steuern Prozesse wie Wachstum, Schlaf/Wach -Rhythmus, Hungergefühl, etc.

Proteine bestehen aus kleineren Bausteinen, den Aminosäuren. Durch Peptidbindungen werden die einzelnen Aminosäuren zu langen Ketten aneinandergeknüpft und sind dadurch zu einer spezifischen dreidimensionalen Raumstruktur gefalten

Abbildung 3.1.: Protein (Bildquelle [65])

(siehe Abbildung 3.1). Durch die unterschiedlichen Kombinationen dieser Bausteine entsteht eine große Vielfalt unterschiedlicher Proteine. Die Aminosäurenabfolge eines Proteins und damit der Aufbau eines Proteins ist im genetischen Code, der DNA (Kap. 3.1.2) festgelegt.

3.1.2. DNA (Desoxyribonukleinsäure)

Die DNA (Desoxyribonukleinsäure) ist der Träger der Erbinformation eines Organismus. Die DNA kodiert nicht nur die gesamte Information für den strukturellen Aufbau des Lebewesens, sondern auch alle Mechanismen für die Erhaltung und Replikation der DNA in den Zellen. Die grundlegenden Bausteine der DNA sind sogenannte Nukleotide. Nukleotide bestehen aus einem Pentose (Zucker), einer Phosphatgruppe und einer stickstoffhaltigen Base. Die Pentose der DNA ist die Desoxyribose. Das „Rückgrat“ der DNA besteht aus alternierenden Zuckermolekülen und Phosphatgruppen. Die Basen sind an Zucker gebunden. Durch Phosphodiesterbindungen zwischen dem Zuckern des einen Nukleotids und dem Phosphat des nächsten Nukleotids werden die Nukleotide verbunden. Die vier Basen der DNA sind:

- Adenin (A)
- Cytosin (C)
- Guanin (G)
- Thymin (T)

Die DNA ist doppelsträngig. Sie besteht aus zwei Polynukleotidsträngen, die durch Wasserstoffbrücken zwischen zwei komplementären Basen miteinander verbunden sind (siehe Abbildung 3.2). Die Reihenfolge der Basen eines Stranges

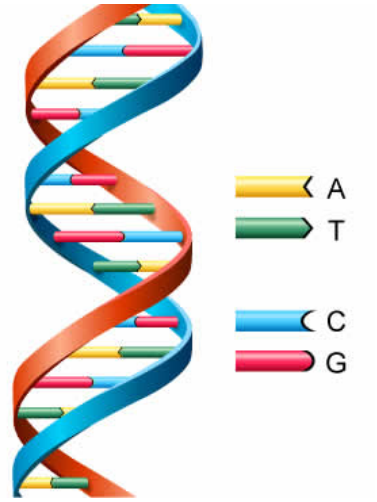


Abbildung 3.2.: Doppelhelixstruktur
(Bildquelle [9])

bestimmt vollständig die Basensequenz des anderen Stranges. Die zwei Polynukleotidstränge laufen in entgegengesetzter Richtung. In der DNA ist Adenin immer mit Thymin und Cytosin immer mit Guanin gepaart. Die dreidimensionale Struktur der DNA wurde 1953 von Watson und Crick [62] entschlüsselt. Nach ihrem bekannten Strickleitermodell bilden die DNA-Doppelstränge eine gewundene Helix, wobei die Basen in das Innere der Spirale hineinragen. Die gewundene Spiralstruktur verleiht der DNA eine hohe Stabilität und Widerstandsfähigkeit gegenüber DNA spaltenden Enzymen. Aufgrund der komplementären Struktur der DNA kann sie sich bei der Zellteilung verdoppeln. Wie bei einem Reißverschluss werden die zwei Stränge aufgespalten und können sich dann wieder zu einem neuen Doppelstrang ergänzen. Auch bei der Übertragung genetischer Information wird die DNA an einer Stelle aufgespalten und kann von der RNA (Kap. 3.1.3) „abgeschrieben“ werden.

3.1.3. RNA (Ribonukleinsäure)

Neben der DNA gibt es eine weitere Nukleinsäure, die in den Zelle vorkommt, die RNA (Ribonukleinsäure). Die RNA ist chemisch weitgehend ähnlich zur DNA mit zwei Ausnahmen:

- Die Base Thymin (T) ist in der RNA durch die Base Uracil (U) ersetzt.
- Als Zuckergrundgerüst wird nicht wie bei der DNA Desoxyribose, sondern Ribose verwendet.

Der Zuckeraustausch macht die RNA weniger stabil. Der wichtigste Unterschied zur DNA besteht darin, dass die RNA als Einzelstrang vorliegt, also nicht wie die

DNA als Doppelhelix. Es gibt verschiedene RNA-Typen: mRNA, tRNA, rRNA, siRNA und miRNA. Hier beschränke ich mich auf die für diese Arbeit wichtige Boten-RNA (mRNA, *messenger*-RNA); ihre Funktion besteht darin, genetische Informationen aus dem Zellkern zu den Ribosomen, dem Ort, wo die Proteine gebildet werden, zu transportieren (siehe Kap. 3.1.5).

3.1.4. Gene

Ein Gen ist ein kodierender Sequenzabschnitt auf der DNA (siehe Abbildung 3.3), der die Formel für die chemische Komposition und Herstellung eines Proteins enthält. Ein bestimmtes Gen kann mit Hilfe komplexer Regulierungsmechanismen

Abbildung 3.3.: Gene, Abschnitte auf der DNA

verschiedene Proteine codieren. Der Sequenzabschnitt lässt sich in zwei funktionell unterschiedliche Bereiche aufteilen, in Exons (von engl. **ex**pressed regi**ons**) und Introns (von engl. **int**ervening regi**ons**).

- Exons bezeichnen kodierende Abschnitte der DNA; in diesen Abschnitten ist die Information über die Aminosäuresequenz eines Proteins enthalten.
- Introns hingegen stellen nichtkodierende Abschnitte der DNA dar; diese nichtkodierenden Bereiche enthalten u.a. Regulationsregionen, die die Intensität der Genexpression steuern.

Abbildung 3.4 zeigt ein Gen mit der Unterteilung in Exons und Introns.

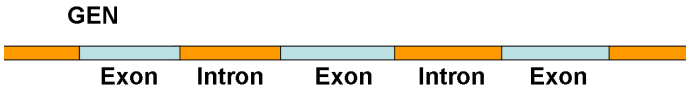


Abbildung 3.4.: Gen mit Unterteilung in Exons und Introns

3.1.5. Genexpression

Der Begriff *Genexpression* bezeichnet die Umsetzung genetischer Information in Proteine. Die enthaltene Information ist als lineare Abfolge der Basen der DNA darstellbar. Die Basensequenz bestimmt die Abfolge der Aminosäuren in einem Protein und bildet die Primärstruktur des Proteins. Diese Primärstruktur bestimmt die letztlich dreidimensionale Form des Proteins. Die DNA ist im Zellkern

enthalten. Die Synthese der Proteine findet in den Ribosomen statt. Ribosome kann man im Aufbau und in der Funktionsweise mit Maschinen vergleichen. Sie lesen die chemische Formel und synthetisieren dann gemäß den Anleitungen Proteine. Damit die genetische Information zu den Ribosomen gelangt, bedarf es einer Informationsübertragung der genetischen Information vom Zellkern zu den Ribosomen. Diese Aufgabe übernimmt die mRNA (siehe Kap. 3.1.3).
Der Weg von der DNA zum synthetisierten Protein besteht im Wesentlichen aus zwei Schritten:

1. Bestimmte Teile eines DNA-Strangs werden durch spezielle Enzyme (RNA Polymerase) abgelesen, was als *Transkription* bezeichnet wird. Dabei entsteht mRNA als exakte, komplementäre Kopie eines Teilstücks des Stranges.

2. Aus der mRNA werden mit Hilfe von Ribosomen Proteine synthetisiert; dieser Vorgang wird als *Translation* bezeichnet.

Das Modell dieses Informationsflusses lässt sich vereinfacht wie folgt darstellen:

$$\text{DNA} \overset{Transkription}{\rightarrow} \text{mRNA} \overset{Translation}{\rightarrow} \text{Proteine}$$

Dieses Modell des Informationflusses wird als *zentrales Dogma der Molekularbiologie* bezeichnet.
Der erste Schritt, die Transkription, besteht in der lokalen Aufspaltung der Wasserstoffbrücken. Dadurch liegt die DNA abschnittsweise in zwei einzelnen Strängen vor. Man unterscheidet den informationstragenden Sinnstrang und den komplementären Antisinnstrang, der für das Ablesen benutzt wird. Das Abschreiben des DNA-Stranges erfolgt durch die Anlagerung komplementärer Basen und die Verknüpfung dieser Basen mittels Enzyme. Die dadurch entstandene Sequenz enthält dann eine genaue Kopie aller Exons und Introns des abgeschriebenen DNA-Abschnitts. Durch Entfernen der Introns und anschließender Wiedervereinigung der Exons entsteht die mRNA. Diesen Vorgang des Entfernens nennt man *Spleißen*. Abbildung 3.5 zeigt diese Entstehung der mRNA. Die genetische Information der DNA wird durch die Transkription auf die neu gebildete mRNA übertragen.
Die mRNA transportiert die Information als Bote zu den Ribosomen, wo der zweite Schritt, die Translation, erfolgt. Bei der Translation wird die in der Basenabfolge der mRNA liegende genetische Information in eine Abfolge von Aminosäuren übersetzt, die nach ihrer Verknüpfung ein bestimmtes Protein bilden.
Die Menge der Proteine wird durch die mRNA gesteuert. Kennt man die Menge einer bestimmten mRNA in einer Zelle, kann man daraus folgern, wie stark

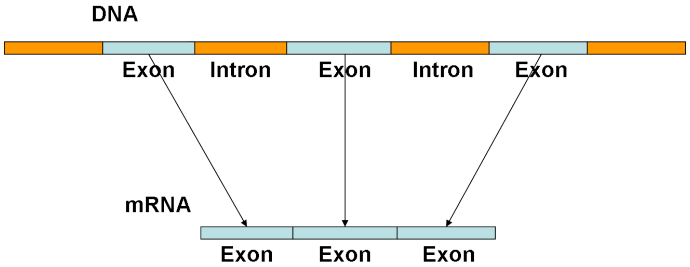


Abbildung 3.5.: Spleißen, Entstehung der mRNA.

ein entsprechendes Protein synthetisiert wird. Ist demnach ein Gen (oder Genabschnitt) aktiv, so wird die mRNA dieses Gens gebildet. Somit kann die Genexpression mit einem DNA-Microarray (siehe Kap. 2.2) bestimmt werden, indem die Menge der gebildeten mRNA ermittelt wird, was dem ersten Schritt, der Transkription, entspricht.

3.1.6. Genetischer Code

Mit Hilfe der Ribosome wird die genetische Information in Proteine umgesetzt. Dabei kodiert eine bestimmte Abfolge von drei Nukleotiden der DNA und daraus abgeleitet die mRNA eine bestimmte Aminosäure. Das „Alphabet“ der DNA besteht aus 4 Buchstaben (Basen): A (Adenin), C (Cytosin), G (Guanin) und T (Thymin) - während das „Proteinalphabet“ 20 Buchstaben (Aminosäuren) besitzt. Drei aufeinanderfolgende Basen (Triplett oder Codon genannt) sind nötig, um eindeutig eine Aminosäure zu bestimmen. Mit einem Triplett hat man 4^3=64 eindeutige Zuordnungen, es gibt aber nur 20 Aminosäuren. Da alle möglichen Codons auch verwendet werden, ist der genetische Code zwar eindeutig, aber degeneriert. So können bis zu sechs verschiedene Tripletts eine Aminosäure kodieren. In Abbildung 3.6 ist die Kodierung der Aminosäuren durch die Basentripletts zu erkennen.

3.2. DNA-Microarrays

Dieser Abschnitt gibt eine Beschreibung in die Microarray-Technologie und der durch diese Technologie gewonnenen Daten.

3.2.1. Einführung

Der Begriff „Microarray“ ist eine Sammelbezeichnung für moderne molekularbiologische Untersuchungssysteme. Es existieren verschiedene Formen von Microarrays:

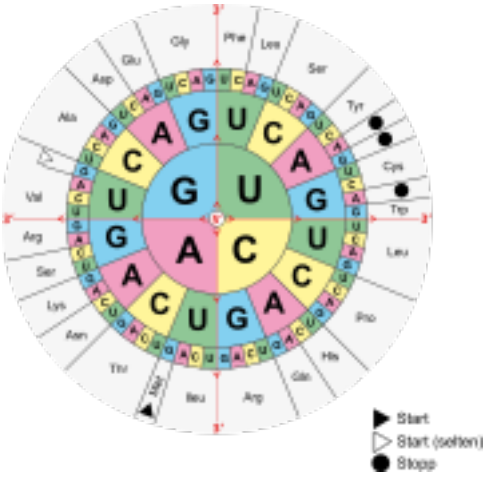


Abbildung 3.6.: Genetische Code-Sonne
(Bildquelle [65])

- DNA-Microarrays
- Protein-Microarrays
- Transfektions-Microarrays
- Tissue-Microarrays

Im Weiteren wird ausschließlich auf die Technologie der DNA-Microarrays fokussiert, da die verwendeten Daten mittels dieses Array-Typs ermittelt wurden. Für die anderen Array-Formen wird auf entsprechende Literatur verwiesen [48], [35]. DNA-Microarrays haben sich in den letzten Jahren bewährt und etabliert, weil sie die simultane Messung von mehreren tausend Genen in einem einzigen „Experiment“ ermöglichen. Untersuchungen zu Veränderungen in der Genexpression mit Hilfe von DNA-Microarrays sind ein wichtiger Bestandteil in der Forschung für die Bereiche Pharmazie, Medizin, Biochemie, Genetik und Molekularbiologie geworden. Die Aktivität bestimmter Gene in einer Zelle gibt wichtige Informationen über den Zellzustand und die Funktionen der Genprodukte innerhalb dieser Zelle. Microarrays bestehen aus einem Trägermaterial, auf dem Nukleotidsequenzen (Oligonukleotide oder cDNA-Fragmente) in genau definierter Anordnung immobilisiert sind. Wegen seiner günstigen Eigenschaften wird als Trägermaterial sehr häufig z.B. Glas verwendet, weil Glas eine nur sehr geringe Eigenfluoreszenz besitzt. Die immobilisierten Nukleotidsequenzen, oft als probes bezeichnet, sind komplementär zu derjenigen Sequenz, die im Untersuchungsmaterial nachgewiesen werden soll. Die Nukleotidsequenzen des zu untersuchenden Materials bezeichnet man als target. Die Microarray-Technologie basiert auf dem Prinzip

der Hybridisierung. Als Hybridisierung wird die Bindung/Anlagerung einzelsträngiger Nukleotidsequenzen (DNA oder RNA) an ihre komplementäre, ebenfalls einzelsträngige Zielsequenz bezeichnet. Durch Markierung der Nukleotidsequenzen des zu untersuchenden Gewebes mittels eines fluroszierenden Farbstoffes führt dies bei Bindung mit komplementären Sequenzen auf dem Array zu einem hellen Signal. Ist die Zielsequenz in der RNA nicht vorhanden, so findet keine Bindung an den probes statt und es wird kein Helligkeitssignal abgegeben. Die aufgebrachten probes weisen somit unterschiedliche Helligkeitssignale auf. Je mehr Bindung stattgefunden hat, also je mehr RNA eines Gens/Exons im Probematerial vorhanden ist, desto intensiver sind die einzelnen Signale. Die Intensität eines Signals ist somit ein Maß für die Expression eines Gens/Exons. Durch optisches Scannen der Arrays wird die Menge der hybridisierungsspezifischen RNA durch Ermittlung der Helligkeitsverteilung bestimmt.

3.2.2. Microarray-Typen und ihre Herstellung

Grundsätzlich lassen sich zwei Typen von DNA-Microarrays unterscheiden. Beide können zur Untersuchung der Expression verwendet werden. Es bestehen jedoch Unterschiede im Herstellungsprozess und den verwendeten Sequenzen.

1. cDNA-Microarrays

cDNA-Microarrays (auch als „Spotted"-Microarrays bezeichnet) verwenden als probes, synthetisierte cDNA-Sequenzen die aus biologischem Material mittels gentechnischer Methoden gewonnen und auf den Array-Träger platziert werden. Die Auftragung der cDNA kann durch zwei unterschiedliche Verfahren [42] erfolgen:

- *Microspotting*
  Bei diesem Verfahren werden die probes mit einer Kapillare direkt auf den Träger appliziert. Der Nachteil dieses Verfahren besteht darin, dass ein unmittelbarer Kontakt zur Oberfläche des Arrays gegeben ist, was zu Ungleichmäßigkeiten führen kann.

- *Microspraying*
  Eine Alternative zum *Microspotting* stellt das *Microspraying* dar; hierbei wird ohne Berührung des Objektträgers gearbeitet. Die probes werden mittels eines Roboters direkt auf das Trägermaterial aufgedruckt. Diesen Vorgang kann man sich wie beim Tintenstrahldrucker vorstellen, nur wird eben keine Tinte, sondern DNA verwendet.

Nach der Auftragung erfolgt die Immobilisierung der probes auf dem Träger mittels UV-Bestrahlung. Dieser Typ von Microarrays ist in der Herstellung relativ kostengünstig.

Abbildung 3.7.: Affymetrix GeneChip®
(Bildquelle [1])

2. Oligonukleotid-Microarrays

Oligonukleotid-Microarrays bestehen aus synthetisch hergestellten Oligonukleotiden. Oligonukleotide sind kurze einsträngige Nucleinsäure-Moleküle, die aus wenigen bis zu einigen hundert Nukleotiden aufgebaut sind. Der bekannteste Hersteller von Oligonukleotid-Microarrays ist Affymetrix [49]. Affymetrix bezeichnet diese Microarrays als GeneChips® (Abbildung 3.7 zeigt einen GeneChip® von Affymetrix).
Die probes werden nicht wie bei cDNA-Microarrays auf den Träger aufgebracht, sondern *in situ*, d.h. direkt auf der Oberfläche des Trägers synthetisiert. Dazu wird ein photolitographisches Verfahren verwendet [24]. Durch die Photolithographie, die der Halbleiterindustrie entlehnt ist, wird die gezielte schrittweise Synthese der Oligonukleotide, Base für Base, an festgelegter Positionen auf dem Träger erlaubt. Abbildung 3.8 zeigt das Prinzip der Herstellung von Affymetrix-Arrays durch das Verfahren der Photolithographie. Auf dem Trägermaterial werden Verbin-

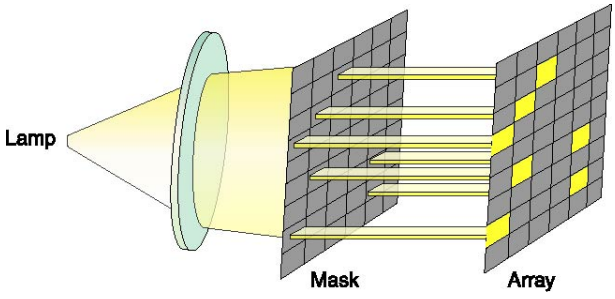


Abbildung 3.8.: Herstellung von Oligonukleotidarrays durch Photolithographie
(Bildquelle [49])

dungsmoleküle aufgebracht und befestigt; diese sind mit einer photochemischen

„Schutzgruppe“ gegenüber ungewollten Bindungsreaktionen geschützt. Mit Hilfe von Lichtmasken werden die „Schutzgruppen“ an verschiedenen Stellen entfernt. An diesen nun freien Stellen können sich die zugesetzten gewünschten Nukleotide binden. Die Enden dieser zugesetzten Nukleotide sind wieder mit einer photoreaktiven Gruppe vor weiteren Ablagerungen geschützt. Durch mehrere Durchläufe werden die einzelnen Nukleotide an die entsprechenden Sequenzenden angelagert. Nach Abschluss dieses Verfahrens sind somit beliebige Oligonukleotidsequenzen auf dem Array synthetisiert.
Mit dieser Methode kann eine viel höhere Dichte als bei den cDNA-Microarrays erzielt werden. Die erzeugten probes sind in ihrer Sequenz begrenzt (bei den Affymetrix GeneChips® sind die immobilisierten Oligonukleotide 25 Nukleotide lang); Gene oder Exons können daher auf mehrere probes verteilt sein. Affymetrix fasst diese probes zu sogenannten *probeSets* zusammen. Diese Form der Microarrays sind weniger anfällig gegenüber räumlichen Fehlern, jedoch ist ihr Herstellungsprozess recht kostspielig.
Die in dieser Arbeit verwendeten Daten wurden mit dem **Human Exon Array GeneChip®** von Affymetrix gewonnen. Bei dieser Art von Array werden keine Gene, sondern Exons auf ihre Expression untersucht. Das Array enthält ca. 1.4 Milionen probeSet(Exons) und pro probeSet liegen ca. vier „Exonteilstücke“ (probes) vor.

3.2.3. Stufen eines Microarray-Experiments

Im Folgenden werden die verschiedenen Schritte bei der Durchführung eines Microarray-Experiments erläutert.

Probengewinnung

Die Gewinnung der zu untersuchenden Probe erfolgt bei beiden Array-Typen ähnlich: Zuerst wird mRNA aus dem zu untersuchendem Gewebe isoliert und aufgereingt. Steht nicht genug RNA zur Verfügung, kann diese durch geeignete Verfahren vervielfältigt werden. Da die gewonnene mRNA äußerst instabil ist, wird sie nicht direkt als target verwendet, sondern im Labor durch reverse Transkiptasen in die zur mRNA komplementäre cDNA übersetzt. Die cDNA enthält keine Intron-Sequenzen mehr. Die Proben für die cDNA-Arrays werden mit fluoreszierenden Farbstoffen markiert und anschließend auf das Array gegeben.
Bei den GeneChips wird die cDNA noch weiter in cRNA umgewandelt und der Markierungsschritt erfolgt erst nach der Hybridisierung.

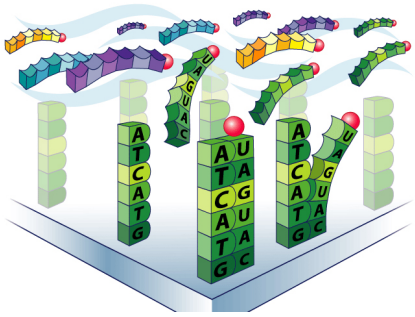


Abbildung 3.9.: Hybridisierung der Sequenzen eines Arrays mit Fluoreszenz-markierten Sequenzen des zu untersuchenden Gewebes
Bildquelle [1]

Hybridisierung der Arrays

Unter *Hybridisierung* versteht man den Prozess der Bindung zweier komplementärer DNA-Stränge bzw. eines DNA-Stranges mit seiner komplementären RNA. Um Genexpression zu quantifizieren, werden die zu hybridisierenden RNA-Sequenzen zunächst fluoreszierend markiert. Anschließend folgt die Hybridisierung an die immobilisierten Sequenzen auf dem Array. Dabei gilt: Ist viel Zielsequenz in der aufgetragenen Gewebeprobe vorhanden, so findet viel Bindung an der zur Zielsequenz komplementären Sequenz statt. Um unerwünschte Bindungen auf dem Array zu vermeiden, wird die Oberfläche zuerst mit einem Hybridisierungspuffer vorbehandelt. Wegen der grundlegenden Eigenschaft der komplementären Basenpaarung können nur entsprechende Sequenzen hybridisieren. Zur Hybridisierung ist eine bestimmte Umgebungstemperatur erforderlich. Viele Faktoren wie Luftfeuchtigkeit, Temperaturschwankungen, Lösungsmittel, etc. können die Qualität der Hybridisierung maßgeblich beeinflussen. Nach der Hybridisierung werden die Arrays gewaschen. Dies ist ein sehr wichtiger Schritt, um ungebundene targets zu entfernen. Abbildung 3.9 zeigt das Prinzip der Hybridisierung; die freien markierten Sequenzen des zu untersuchenden Materials binden an den Sequenzen auf dem Array.

Scannen der Arrays

Nach dem Waschvorgang werden die Arrays gescannt. Die Fluoreszenzfarbstoffe werden durch Bestrahlung mit einem Laser zur Lichtemission angeregt. Nach dem Abtasten liegt somit ein Bild aus den verschiedenen Helligkeitswerten (Intensitäten) der probes eines Arrays vor.
Abbildung 3.10 zeigt den Scan eines Affymetrix GeneChip®. Die Intensitäten spiegeln die Anzahl der gebundenen targets an den komplementären probes wider. Findet viel Bindung an der Zielsequenz statt, führt dies zu einem hellen

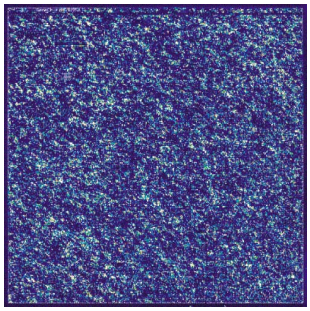


Abbildung 3.10.: Scan eines Affymetrix GeneChip®, (Bildquelle [1])

Signal. Ist im zu untersuchenden Material keine komplementäre Sequenz enthalten, findet keine Bindung statt und es wird kein Helligkeitssignal abgegeben. Die weitere eigentliche Datengewinnung aus diesen „Rohwerten“ erfolgt mittels entsprechender Software.

3.3. Datenbasis

3.3.1. Datengewinnung

Die Erhebung von Expressionsdaten ist ein aufwendiger Vorgang. Die experimentellen Schritte zur Datengewinnung, die Hybridisierung und das Scannen des Arrays wurden bereits in Kapitel 3.2.3 vorgestellt. Die weiteren Schritte beziehen sich auf die Datengewinnung bei den Affymetrix-GeneChips®. Das gescannte Pixelbild liegt zunächst als DAT-Datei vor.
Diese entstandenen Bilddaten werden dann weiter in Expressionsdaten überführt. Diesen Prozess der Überführung bezeichnet man als „Low-Level-Analyse“. Er gliedert sich in die folgenden Teile:

- Bildanalyse
- Hintergrundkorrektur
- Normalisierung
- Aggregation zu einem Expressionswert pro Exon

Bildanalyse

Eine Probe wird in der DAT-Datei durch mehrere Bildpunkte repräsentiert. Aus diesen Werten wird bei der Bildanalyse eine Gesamtintensität pro Probe/Spot ermittelt; diese Werte werden in einer CEL-Datei abgespeichert. Die CEL-Dateien sind pro Experiment etwa 65 MB groß.

Hintergrundkorrektur

Bei der optischen Signalerfassung entsteht in der Regel ein Hintergrundsignal, z.B. durch Streulicht. Sinn der Hintergrundkorrektur ist es, dieses „Rauschen“ zu schätzen und von den ermittelten Werten zu eliminieren.Ein Chip, bei dem keine Bindung stattgefunden hat, sollte auch kein Signal zeigen [25].

Normalisierung

Bei der Herstellung der Arrays und der Durchführung von Experimenten können Varianzen nichtbiologischen Ursprungs entstehen, z.B. durch unterschiedliche experimentelle Bedingungen, wie unterschiedliche Mengen an RNA, an Farbstoff, andere Belichtung, etc. Um solche Varianzquellen zu vermeiden und verschiedene Arrays in einer Experimentreihe miteinander vergleichen zu können, werden verschiedene Normalisierungstechniken angewendet. Eine Übersicht zu Normalisierungstechniken für Affymetrix GeneChips® wird in [32] gegeben. Die Normalisierungstechniken können in zwei Typen unterteilt werden. Einige Techniken verwenden zur Normalisierung ein „Baseline-Array“, anhand dessen die Werte der anderen Arrays normalisiert werden. Als Baseline-Array kann z.B. das Array benutzt werden, welches eine mittlere Helligkeitsintensität im Vergleich zu den anderen Arrays aufweist. Bei der zweiten Technik werden die Informationen aller Arrays zur Normalisierung herangezogen. Eine Form dieser Normalisierung stellt z.B. die Quantil-Normalisierung dar, die auf der Annahme basiert, dass die Histogramme aller Arrays einer Experimentreihe vergleichbar sind. (Für weitere Details siehe z.B. [31]).

Aggregation

Die hintergrundkorrigierten und normalisierten *probe*-Intensitätswerte werden zu einem probeSet-Wert zusammengefasst. Dieser Wert entspricht der Expression eines Exons.

Nach der Low-Level-Analyse liegen die Werte der Exon-Expressionen in einer Matrix M der Dimension (m x n) vor. Hierbei entspricht m der Anzahl der probeSet und n der Anzahl der Array-Experimente.
Die weitere Analyse dieser Daten (evtl. mit zusätzliche bekannten Parametern, wie z.B. der Klassenzugehörigkeit) bezeichnet man als „High-Level-Analyse“.

3.3.2. Datenerhebung für die vorliegende Arbeit

Die in dieser Arbeit untersuchten Exon-Expressionsdaten von Neuroblastom-Patienten stammen vom Universitätsklinikum Essen aus dem Onkologischen La-

bor der Kinderklinik und wurden mit dem „GeneChip® Human Exon Array“ [1] der Firma Affymetrix (Santa Clara, USA) gewonnen. Es liegen CEL-Dateien (Textdateien, Tabulator-getrennt) von 135 Patienten vor.
Zu jedem Patienten sind darüber hinaus weitere Informationen (Angabe des Geschlechts, Alter bei Diagnose, klinisches Stadium des Tumors, Gesamtüberleben, ereignisfreies Überleben und MYCN-Amplifikation) vorhanden.
Die Hintergrundkorrektur, Normalisierung und Aggregation der 135 CEL-Dateien erfolgte mittels RMA (robust multiarray average, siehe [31]), das in der Bioconductor-Plattform [8] implementiert ist. RMA führt eine Korrektur des Hintergrundrauschens über ein globales Modell der Intensitätsverteilung für jedes einzelne Array durch, die Normalisierung über alle Arrays geschieht mit einer Quantil-Normalisierung und die Aggregation erfolgt über eine Medianglättung. Das so ermittelte Expressionsmaß für jedes Exon wird auf der log2-Skala angegeben. Durch die Logarithmierung der Expressionswerte werden die Daten in eine Normalverteilung „gezwungen“.
Danach liegen die 135 Patientenprofile als Matrix vor. Für jeden Patienten liegen Expressionswerte von 287.329 Exons vor. Als zusätzliche Klassenzugehörigkeit wird „ereignisfreies Überleben“ angegeben:

- Eine „0“ als Eintrag bedeutet „Patient lebt ohne Ereignis“.
- „1“ steht für „Patient hatte ein Rezidiv“.

82 Patienten können der Klasse „ereignisfreies Überleben = 0“ und 49 Patienten der Klasse „ereignisfreies Überleben = 1“ (Rezidiv) zugeordnet werden. 4 Patienten können aus biologischer Sicht nicht zur weiteren Analyse verwendet werden.

4. Maschinelles Lernen

In diesem Kapitel erfolgt ein einführender Einblick in das *maschinelle Lernen*. Dabei werden Begriffe und Definitionen aus diesem Bereich erläutert, die in dieser Arbeit verwendet werden oder für das Verständnis von Bedeutung sind. Darüber hinaus werden verschiedene Lernverfahren, die im Rahmen dieser Arbeit zum Einsatz kommen, und deren Bewertung vorgestellt. Abschließend wird die Anwendung des maschinellen Lernens auf die vorliegenden Daten beschrieben.

4.1. Einführung

Nach Wrobel et al. [70] beschäftigt sich das *maschinelle Lernen*

> [...] *mit der computergestützten Modellierung und Realisierung von Lernphänomenen.*

Eine präzise inhaltliche Definition für das intuitive Verständnis des „Lernen“ zu geben, erweist sich nach [70] als äußerst schwierig. In der Literatur sind eine Vielzahl von unterschiedlichen Definitionen für das „Lernen“ zu finden. Eine bekannte und oft zitierte Definition ist die von Herbert Simon [54]:

> *Lernen ist jeder Vorgang, der ein System in die Lage versetzt, bei der zukünftigen Bearbeitung derselben oder ähnlicher Aufgaben diese besser zu erledigen.*

Ähnlich wie für das „Lernen“ wurden auch für das „maschinelle Lernen“ verschiedene Definitionen vorgeschlagen und diskutiert (siehe [5]). Eine allgemein akzeptierte Definition des maschinellen Lernens gibt es jedoch bisher nicht.
Nach [64] wird maschinelles Lernen wie folgt definiert:

> *Maschinelles Lernen ist ein Oberbegriff für die künstliche Generierung von Wissen aus Erfahrung. Ein künstliches System lernt aus Beispielen und kann nach Beendigung der Lernphase verallgemeinern.*

Da eine intensionale Definition recht unpräzise ist, ziehen Wrobel et al. [70] eine alternative Betrachtungsweise heran. Sie versuchen, maschinelles Lernen extensional über einzelne Typen von Lernaufgaben, die bisher Gegenstand des Forschungsgebietes gewesen sind, zu definieren. Die Schwierigkeit solch einer extensionalen

Definition besteht darin, alle bisher bekannten Lernaufgaben mit einzubeziehen, insbesondere da sich sowohl die Forschung wie auch die Lernaufgaben stetig weiterentwickeln und redefiniert werden.
Die Definition einer Lernaufgabe im Allgemeinen, wie in [70] gegeben, ist die Folgende:

Definition 4.1 (Lernaufgabe)
Eine Lernaufgabe wird definiert durch

- *eine Beschreibung der dem lernenden System zur Verfügung stehenden Eingaben (ihrer Art, Verteilung, Eingabezeitpunkte, Darstellung und sonstigen Eigenschaften),*
- *die vom lernenden System erwarteten Ausgaben (ihrer Art, Funktion, Ausgabezeitpunkte, Darstellung und sonstigen Eigenschaften) und*
- *den Randbedingungen des Lernsystems selbst (z.B. maximale Laufzeiten oder Speicherverbrauch).*

Ein System löst die Lernaufgabe genau dann erfolgreich, wenn es in der Lage ist, bei Eingaben, die den Spezifikationen entsprechen, unter den geforderten Randbedingungen Ausgaben mit den gewünschten Eigenschaften zu erzeugen.

4.2. Einsatzgebiete des maschinellen Lernens

Maschinelles Lernen kommt überall dort zum Einsatz, wo versucht wird, in einer Menge von Daten Muster und Gesetzmäßigkeiten zu entdecken. Für diese Aufgabenstellung werden Methoden entwickelt, mit denen es Computern ermöglicht wird „zu lernen“.
Oftmals sind die betrachteten Daten so umfangreich, dass sie durch den Menschen allein nicht mehr effizient analysiert werden können. Daher ist maschinelles Lernen in vielen unterschiedlichen Anwendungsbereichen zu finden, zum Beispiel im Bereich des Marketings, des Investments, der Astronomie, der Geowissenschaften, der Betrugserkennung, der Biologie, der Medizin, der Bioinformatik, usw.

4.3. Überwachtes und unüberwachtes Lernen

Eine wichtige Unterscheidung beim maschinellen Lernen ist die zwischen überwachtem und unüberwachtem Lernen.

Überwachtes Lernen (engl. supervised learning)

Beim überwachten Lernen wird anhand einer Menge von Eingabewerten und Ausgabewerten (Eingabe-Ausgabe-Paaren (x, y)) versucht, eine Funktion f zu erlernen bzw. abzuschätzen. Die erlernte Funktion soll für jede Eingabe x einen Funktionswert $f(x) = \hat{y}$ bestimmen mit dem Ziel, dass $\hat{y}$ möglichst dem zugehörigen y entspricht. Es soll also eine Abbildung der Eingabewerte auf die Ausgabewerte erlernt werden, mit der für jedes Beispiel anhand seiner Eingabewerte eine Vorhersage der Ausgabewerte prognostiziert werden kann. Dabei soll die erlernte Funktion die unbekannte Funktion, gemäß der die Beispiele berechnet wurden, möglichst gut approximieren. Die Menge der Eingabe-Ausgabe-Paare, die zum Lernen verwendet wird, bezeichnet man als Beispielmenge (Synonyme: Trainingsmenge, Eingabemenge). Die Beispielmenge besteht aus n Beispielen $x_1, x_2, ..., x_n$. Jedes Beispiel x_i kann als d-dimensionaler Vektor $x_i = (x_{i1}, ..., x_{id})$ von Merkmalsausprägungen einer Merkmalsmenge X = (X_1,...,X_d) aufgefasst werden. Zu jedem Beispiel x_i existieren Ausgabewerte y_i. Somit haben die Eingabe-Ausgabe-Paare die Form $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. Durch die auf der Eingabemenge erlernte Funktion können Vorhersagen für neue Beispiele x_j (deren Ausgabewerte nicht bekannt sind) getroffen werden.
Ein lernendes System kann sich während des Lernvorgangs selbst kontrollieren, indem es die bereits bei der Eingabe gegebenen Ausgabewerte mit den erlernten Ausgaben überprüft und sich so selbst in seinem Lernvorgang verbessern kann.
Die Verfahren Regression [2] und die in dieser Arbeit verwendete Klassifikation (siehe Kap. 4.6) gehören zum Bereich des überwachten Lernen. Die Regression stellt die Vorhersage kontinuierlicher Ausgabewerte vor, während die Klassifikation die Vorhersage diskreter Ausgabewerte darstellt.

Unüberwachtes Lernen (engl. unsupervised learning)

Auf diese Form des Lernens wird hier nur kurz eingegangen, da sich die vorliegende Arbeit ausschließlich mit dem überwachten Lernen beschäftigt.
Beim unüberwachten Lernen liegen nur unklassifizierte Beispiele (die erwarteten Ausgabewerte sind nicht bekannt) vor. Das Ziel dieser Form des Lernens besteht darin, Regelmäßigkeiten und Zusammenhänge innerhalb der Eingabebeispiele selbstständig zu entdecken, insbesondere Ähnlichkeiten in den Merkmalen von Beispielen und diese geeignet zu präsentieren. Ein typisches Verfahren des unüberwachten Lernens ist die Clusteranalyse, deren Ziel es ist, Daten so in Gruppen (Cluster) zusammenzufassen, dass sich diese innerhalb eines Clusters möglichst ähnlich sind und die Daten unterschiedlicher Cluster möglichst unähnlich zueinander sind. Für weitere Details wird auf [70],[2],[20],[29] verwiesen.

4.4. Lernverfahren

Zur Lösung unterschiedlicher Lernaufgaben existieren im maschinellen Lernen verschiedene Lernverfahren; diese lassen sich dem überwachten oder dem unüberwachtem Lernen zuordnen. Da diese Arbeit sich mit dem überwachtem Lernen beschäftigt, wurden folgende Verfahren aus diesem Bereich verwendet:

- k-Nächste-Nachbarn (k-Nearest-Neighbors)
- Naive Bayes
- Stützvektormethode (SVM)

Alternative Lernverfahren wie Entscheidungsbäume oder Regel-Lerner wurden nicht verwendet, da durch die Dimension der vorliegenden Daten eine Interpretation der Ergebnisse dieser Verfahren sehr schwierig bis unmöglich ist.

4.4.1. k-Nächste-Nachbarn (k-Nearest-Neighbors)

Das k-Nächste-Nachbarn Verfahren (kurz: *kNN*) ist ein einfaches, aber häufig sehr effizientes Lernverfahren. Das Verfahren geht zurück auf Arbeiten von Fix und Hodges [23]. Die Idee des Lernverfahrens besteht darin, dass einander ähnliche Daten der gleichen Klasse angehören. Zur Bestimmung der Klassenzugehörigkeit eines unbekannten Beispiels werden die k nächsten Nachbarn bestimmt und das Beispiel wird der Klasse zugeordnet, zu der die meisten dieser nächsten Nachbarn zählen (Mehrheitsvotum). Im einfachsten Fall, k = 1, wird nur der nächste Nachbar betrachtet und dessen Klassenzugehörigkeit für das neue Beispiel übernommen. Bei dem Verfahren wird davon ausgegangen, dass jedes Beispiel x_i, beschrieben als d-dimensionaler Vektor $x_i = (x_{i1}, ..., x_{id})$ von Merkmalsausprägungen einer Merkmalsmenge X = (X_1,...,X_d), als Punkt in einem d-dimensionalen Vektorraum V dargestellt werden kann. Zur Bestimmung des nächsten Nachbarn werden Distanzfunktionen benutzt. Am häufigsten wird hierzu die *Euklidische Distanzfunktion* verwendet. Der Euklidische Abstand zwischen zwei Beispielvektoren x_i und x_j ist gegeben durch:

$$dis(x_i, x_j) = \sqrt{\sum_{r=1}^{d} (x_{ir} - x_{jr})^2} \tag{4.1}$$

Im einfachstem Fall, $k = 1$, wird ein unbekanntes Beispiel x_j derjenigen Klasse C zugewiesen, zu der das von x_j am wenigsten entfernte Beispiel x_i (nächster Nachbar) angehört:

$$C(x_j) = C(\arg\min_{x_i} dis(x_j, x_i)) \tag{4.2}$$

Bei der Betrachtung mehrerer nächster Nachbarn wird das neu zu klassifizierende Beispiel x_j der Klasse der Mehrzahl seiner k nächsten Nachbarn zugeordnet. Für das neue Beispiel x_j wird also die Klasse gewählt, die unter den k nächsten Nachbarn am häufigsten auftritt.

$$C(x_j) = \arg\max_{l \in C} \sum_{i=1}^{k} \delta(l, f(x_i)) \tag{4.3}$$

mit

$$\delta(a,b) = \begin{cases} 0, \text{ wenn } a \neq b \\ 1, \text{ wenn } a = b \end{cases}$$

Durch unterschiedliche Werte von k ist es möglich, auch unterschiedliche Klassifizierungsergebnisse für ein neues Beispiel zu erhalten. Durch die Wahl eines ungeraden Wertes für k kann eine eindeutige Mehrheit sichergestellt werden.

4.4.2. Naive Bayes

Der *Naive-Bayes-Klassifizierer* [46] ist ein Lernverfahren, das auf dem Bayes-Theorem (oder auch Satz von Bayes [3]) - benannt nach dem Mathematiker Thomas Bayes (1702-1761) - beruht. Mit diesem Theorem kann eine Abschätzung der Wahrscheinlichkeit jeder Klasse C_i für ein zu klassifizierendes Beispiel gegeben werden. Anhand dieser Wahrscheinlichkeiten lässt sich das Beispiel klassifizieren, indem die Klasse vorausgesagt wird, für die die höchste Wahrscheinlichkeit geschätzt wurde. Für zwei Ereignisse A und B lautet das Theorem:

$$P(A \mid B) = \frac{P(B \,|A) \cdot P(A)}{P(B)} \tag{4.4}$$

Die bedingten Wahrscheinlichkeiten $P(A \,|B)$ und $P(B \,|A)$ werden *a-posteriori Wahrscheinlichkeiten* genannt. Hingegen werden die Wahrscheinlichkeiten $P(A)$ und $P(B)$ als *a-priori Wahrscheinlichkeiten* bezeichnet.
Angenommen, es existieren d Merkmale $A_1, A_2, ..., A_d$ und a_i sei der Merkmalswert des Merkmals A_i. Es seien k Klassen $C_1, C_2, ..., C_k$ gegeben. X sei ein zu klassifizierendes Beispiel, was durch die Merkmalswerte $(a_1, a_2, ..., a_d)$ beschrieben ist. Zur Klassifikation eines Beispiels X ist die Klasse C_j gesucht, für die die a-posteriori Wahrscheinlichkeit, gegeben X, maximal ist. Setzt man C_j und X in Formel 4.4 ein, ergibt sich:

$$P(C_j \mid X) = \frac{P(X \,|C_j) P(C_j)}{P(X)} \tag{4.5}$$

Die Zielfunktion lässt sich formal beschreiben durch:

$$\begin{aligned} & \arg\max_{C_j \in C} P(C_j \,|X) \\ &= \arg\max_{C_j \in C} \frac{P(X \,|C_j) \cdot P(C_j)}{P(X)} \\ &= \arg\max_{C_j \in C} P(X \mid C_j) \cdot P(C_j) \end{aligned} \tag{4.6}$$

Der Schätzwert $\widehat{P}(C_j)$ für $P(C_j)$ kann aus den Trainingsbeipielen berechnet werden, die sich in den entsprechenden Klassen befinden.

$$\widehat{P}(C_j) = \frac{|C_j|}{\sum_{C' \in C} |C'|} \tag{4.7}$$

Hierbei ist $|C_j|$ die Anzahl der Beispiele in Klasse C_j.

Einen Schätzwert für die Wahrscheinlichkeit $P(X \,|C_j)$ zu erhalten, ist schwieriger. Erst durch die Annahme der Unabhängigkeit der Merkmale wird eine Bestimmung des Schätzwertes möglich; die Multiplikation von Wahrscheinlichkeiten ist nur dann erlaubt, wenn die Ergebnisse voneinander unabhängig sind. Obwohl diese naive Annahme, die dem Verfahren seinen Namen einbrachte, in der Praxis selten zutrifft, erzielt das Verfahren bei vielen realen Anwendungen gute Ergebnisse [67].

$P(X \,|C_j)$ kann als Produkt aus den bedingten Wahrscheinlichkeiten für die in X vorkommenden Merkmale berechnet werden:

$$P(X \mid C_j) = \prod_{i=1}^{d} P(a_i \mid C_j) \tag{4.8}$$

Die Wahrscheinlichkeit $P(a_i \mid C_j)$ kann unabhängig von den anderen Merkmalen abgeschätzt werden. Die relativen Häufigkeiten der Beispiele, für die das Merkmal A_i den Wert a_i hat und die Klassenzugehörigkeit C_j ist, werden durch die relative Häufigkeit der Klasse C_j ($\widehat{P}(C_j)$, siehe 4.7) geteilt.

$$P(a_i \mid C_j) = \frac{P(a_i \cap C_j)}{P(C_j)} \tag{4.9}$$

In Formel 4.6 eingesetzt ergibt dies zusammenfassend:

$$\arg\max_{C_j \in C} P(C_j) \prod_{i=1}^{d} P(a_i \mid C_j) \tag{4.10}$$

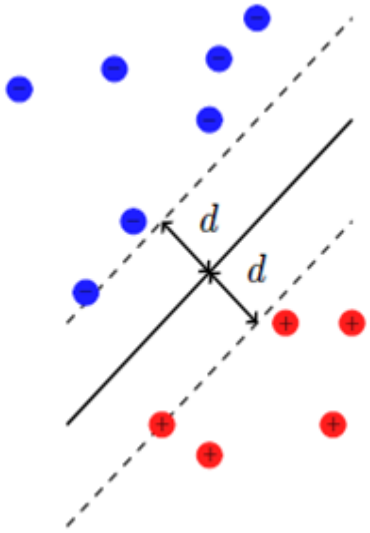


Abbildung 4.1.: Hyperebene bei linear separierbaren Daten (Bildquelle [47])

4.4.3. Support Vector Machine (SVM - Stützvektormethode)

Eine *Support Vector Machine* (SVM), im Deutschen als *Stützvektormethode* bezeichnet, ist eine Entwicklung aus der statistischen Lerntheorie. Sie wurde ursprünglich entwickelt von Vladimir Vapnik [61]. Die Methode beruht auf der Idee der strukturellen Risikominimierung [60]. Die folgenden Darstellungen basieren wesentlich auf [61], [16] und [34], auf die für eine tiefergehende Einführung in diese Theorie verwiesen wird.
Eine Support Vector Machine ist ein Algorithmus, der versucht, für eine Menge von Trainingsbeispielen $E = \{(x_1, y_1), ..., (x_n, y_n)\}$ mit Klassenzugehörigkeiten $y_i \in \{+1, -1\}$ und Beispielvektoren $x_i \in^d$ eine Hyperebene H so zu bestimmen, dass die Beispielmenge korrekt getrennt wird und dabei ein größtmöglicher Abstand zu den Beispielen besteht. Die Maximierung des Abstands ist wichtig für eine gute Generalisierungsfähigkeit, um neue (noch nicht betrachtete) Beispiele möglichst korrekt zu klasssifizieren. Im Allgemeinen gibt es mehrere mögliche Hyperebenen, die die Trainingsbeispiele richtig voneinander trennen. Die SVM wählt genau die Hyperebene, die den Abstand zu den nächstgelegenen Beispielen maximiert. Dieser Abstand, also der Bereich um eine Hyperebene, in dem keine Beispiele liegen, ist der sogenannte *Margin*. Eine Hyperebene, welche die Anforderung des maximalen Abstands erfüllt, wird als *optimale Hyperebene* bezeichnet. In Abbildung 4.1 ist eine solche Hyperebene H gezeigt. H_1 und H_2 sind die korrespondierenden Ränder (Hyperebenen) zu H und der Abstand von H_1 zu H_2 ist der Margin. H_1 und H_2 sind parallel zueinander. Die gesuchte Hyperebene H hat

die Form

$$(\vec{w} \cdot \vec{x} + b) = 0 \tag{4.11}$$

Dabei ist $\vec{w}$ der Normalenvektor der Hyperebene, $\vec{x}$ ein zu klassifizierender Beispielvektor und b die Verschiebung zum Ursprung. Eine Änderung von b bewirkt eine Parallelverschiebung der Hyperebene entlang des Normalenvektors. Der Abstand der Hyperebene zum Ursprung ist $|b| / \|\vec{w}\|$, wobei $\|\vec{w}\| = \sqrt{\langle \vec{w}, \vec{w} \rangle}$ die Euklidische Norm von $\vec{w}$ ist. Die Beispiele, die der Hyperebene am nächsten liegen, werden als Stützvektoren (Support Vectors) bezeichnet. Nur diese Beispiele bestimmen die Lage der trennenden Hyperebene und definieren den Margin.
Die Hyperebene separiert die Beispiele in zwei Klassen; daher wird gefordert, dass die folgenden Bedingungen gelten:

$$\vec{w} \cdot \vec{x_i} + b \geq +1 \quad \forall i \quad \text{mit } y_i = +1 \tag{4.12}$$

$$\vec{w} \cdot \vec{x_i} + b \leq -1 \quad \forall i \quad \text{mit } y_i = -1 \tag{4.13}$$

Diese Bedingungen können zu folgender Bedingung zusammengefasst werden:

$$y_i(\vec{w} * \vec{x_i} + b) \geq 1 \qquad \forall i \tag{4.14}$$

Die Beispiele, für die die Bedingung 4.14 zu einer Gleichung wird, liegen auf der Hyperebene $H1 : \vec{w} \cdot \vec{x_i} + b = +1$ mit Normalenvektor $\vec{w}$ und einem Abstand $|1 - b| / \|\vec{w}\|$ vom Ursprung. Ebenso gibt es auf der gegenüberliegende Seite eine Hyperebene $H2 : \vec{w} \cdot \vec{x_i} + b = -1$ mit einem Abstand $|-1 - b| / \|\vec{w}\|$. Daraus folgt für den Margin $= \frac{2}{\|\vec{w}|}$. Der Margin wird maximiert, wenn man $\|\vec{w}\|$ minimiert.

Zusammenfassend muss also das folgende Optimierungsproblem gelöst werden, um eine optimale Hyperebene zu finden: Minimiere $\|\vec{w}|$ (dadurch wird der Margin maximiert), so dass für alle i die Bedingung 4.14 gilt. Mit Bedingung 4.14 wird sichergestellt, dass die Beispiele korrekt getrennt werden. (Für weitere Details zur Lösung des Optimierungsproblems siehe [16]). Nachdem eine Lösung für das Optimierungsproblem gefunden wurde, können neue Beispiele mit der Funktion

$$h(\vec{x}) = sign(\vec{w} * \vec{x} + b) = y \tag{4.15}$$

klassifiziert werden. Die Funktion $sign :\rightarrow \{+1, -1\}$ hat den Wert $+1$, wenn a $>$ 0 ist, sonst nimmt sie den Wert -1 an.

Oft sind die Trainingsdaten nicht eindeutig durch eine Hyperebene voneinander zu trennen. Es können Beispiele innerhalb der Margin oder auf der falschen Seite der Hyperebene liegen. Zur Trennung solcher Beispiele wurden sogenannte *Schlupfvariablen* ξ_i eingeführt, die die Nebenbedingungen 4.14 „auflockern“. Ein positiver Wert der Schlupfvariable ξ_i zeigt eine Verletzung der Nebenbedingung

für das Beispiel i an.
Ziel ist es, den Margin so weit wie möglich zu maximieren und dabei die Fehlplatzierungen ($\xi_i > 0$) so gering wie möglich zu halten. Einen sogenannten „weichen Rand“ (soft margin) erreicht man, indem das Optimierungsproblem um die Schlupfvariablen erweitert wird:

$$\text{Minimiere } \|\vec{w}\| + C \sum_{i=1}^{n} \xi_i \tag{4.16}$$

unter den Nebenbedingungen

$$y_i(\vec{w} * \vec{x_i} + b) \geq 1 - \xi_i, \ \forall i : \xi_i \geq 0 \tag{4.17}$$

Zum minimierenden Wert werden „Zusatzkosten“ hinzuaddiert. Der Parameter C stellt einen Kostenfaktor dar, mit dem justiert werden kann, wie hoch eine Fehlplatzierung „bestraft“ werden soll; je höher C, desto stärker wird also ein Fehler bestraft.

Oft sind Beispiele nicht linear voneinander separierbar. Eine Möglichkeit, solche nicht linear separierbaren Daten zu trennen, ist, die Eingabedaten mittels einer nichtlinearen Abbildung Φ in einen Merkmalsraum H höherer Dimension zu transformieren.

$$\Phi : \mathbb{R}^d \rightarrow H \tag{4.18}$$

In diesem Merkmalsraum wird die optimale Hyperebene bestimmt und es erfolgt eine Rücktransformation in den Ursprungsraum. Dabei wird die Hyperebene zu einer nicht-linearen Trennfläche im Eingaberaum. Der neue Merkmalsraum kann unendlich dimensional werden, wodurch die Berechnungen der Abbildungen sehr komplex und rechenlastig werden können. Um die Abbildung nicht explizit berechnen zu müssen, werden sogenannte *Kernfunktionen* K eingeführt, die das Skalarprodukt zweier Vektoren im hochdimensionalen Abbildungsraum explizit berechnen. Eine Kernfunktion K wird allgemein folgendermaßen beschrieben:

$$K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) \tag{4.19}$$

Es kann nicht jede beliebige Funktion gewählt werden, sondern die Abbildungen sind an bestimmte Bedingungen, das *Mercer-Theorem* [44], gebunden. Häufig eingesetzte Funktionen, die das Mercer-Theorem erfüllen, sind die folgenden vier Kernfunktionen (entnommen aus [30]):

- Linearer Kern: $K(x_i, x_j) = x_i^T x_j$
- Polynomialer Kern: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$

- Radiale Basisfunktion (RBF): $K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2), \gamma > 0$
- Sigmoide Kernfunktion: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$

dabei sind γ, r und d Kernparameter.

4.5. Bewertung des Lernerfolgs

Nach dem Trainieren der Lernverfahren ist man an dem Lernerfolg interessiert, d.h. wie erfolgreich die Lernverfahren gelernt haben. Man möchte wissen, wie gut die Verfahren neue unbekannte Beispiele klassifizieren, d.h. wie gut die einzelnen Lernverfahren generalisieren, um dadurch verschiedene Lernverfahren miteinander vergleichen zu können.

4.5.1. Gütekriterien

Die Ergebnisse einer Klassifikation lassen sich im Zwei-Klassen-Fall gemäß den vier Kombinationsmöglichkeiten von tatsächlicher und vorhergesagter Klasse in einer 2×2 Matrix anordnen. Diese Matrix nennt sich *Konfusionsmatrix*. In Abbildung 4.1 wird diese Matrix dargestellt; hierbei bezeichnet TP (true positiv -

		tatsächliche Klasse	
		positiv	**negativ**
vorhergesagte Klasse	**positiv**	TP	FP
	negativ	FN	TN

Tabelle 4.1.: Konfusionsmatrix: Gegenüberstellung der wahren und prognostizierten Klasse

richtig positiv) die Anzahl der positiven Beispiele, die durch die Klassifikation korrekt klassifiziert wurden. FP steht für die Anzahl der fälschlicherweise als positiv klassifizierten Beispiele, die aber in der Realität zu den negativen Beispielen gehören. FN ist die Anzahl der positiven Beispiele, die fälschlicherweise als negativ erkannt wurden und TN gibt die Anzahl der korrekt als negativ eingestuften Beispiele an.
Gütekriterien, die zu diesen Bewertungen herangezogen werden, sind z.B. *Accuracy* (Klassifikationsgenauigkeit, d.h. der Anteil der Beispiele, der korrekt klassifiziert wurde), *Precision* (Präzision, d.h. der Anteil der korrekten Entscheidungen für eine Klasse gemessen an allen gegebenen Entscheidungen für diese Klasse) und *Recall* (Trefferquote, d.h. der Anteil der korrekten Entscheidungen für eine Klasse gemessen an allen möglichen korrekten Antworten). Das häufigste verwendete Gütekriterium ist die Accuracy.

Anhand der Matrixeinträge lassen sich diese Gütekriterien herleiten [29]. Dabei wird angenommen, K sei ein Klassifikator, $TR \subseteq D$ die Traingsmenge, $TE \subseteq D$ die Testmenge und $C(x)$ die tatsächliche Klasse eines Beispiels x. Die einzelnen Gütekriterien werden im Folgenden präzisiert.

Accuracy (Klassifikationsgenauigkeit)

Der Wert für Accuracy ist der Anteil der korrekten Klassifikationen dividiert durch die Gesamtzahl der Klassifikationen. Für ein Klassifikationsproblem mit zwei Klassen ergibt sich folgende Formel:

$$Accuracy_{TE}(K) = \frac{|\{x \in TE \,|K(x) = C(x)\}|}{|TE|} \tag{4.20}$$

Bezogen auf die Konfusionsmatrix ergibt dies:

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{4.21}$$

Die Accuracy beschreibt also die Leistungsfähigkeit eines Verfahrens im Ganzen.

Precision (Präzision)

Der Wert für Precision ergibt sich aus dem Verhältnis der Anzahl der Beispiele, die tatsächlich zur Klasse i gehören, zu den Beispielen, die vom Klassifikator der Klasse i zugeordnet wurden.

$$Precision_{TE}(K, i) = \frac{|\{x \in K_i \,|K(x) = C(x)\}|}{|K_i|} \tag{4.22}$$

Bezogen auf die Konfusionmatrix ergibt dies bzgl. der positiven Klasse:

$$Precision = \frac{TP}{TP + FP} \tag{4.23}$$

Bezogen auf die negative Klasse ergibt sich:

$$Precision = \frac{TN}{TN + FN} \tag{4.24}$$

Recall (Trefferquote)

Der Wert für Recall einer Klasse i gibt den Anteil der korrekt dieser Klasse zugeordneten Beispiele gemessen an der Gesamtheit aller tatsächlich dieser Klasse i zugehörigen Beispiele an und berechnet sich fogendermaßen:

$$Recall_{TE}(K, i) = \frac{|\{x \in C_i \,|K(x) = C(x)\}|}{|C_i|} \tag{4.25}$$

Bezüglich der Konfusionsmatrix ergibt dies bezogen auf die positive Klasse:

$$Recall = \frac{TP}{TP + FN} \tag{4.26}$$

Bezogen auf die negative Klasse ergibt sich:

$$Recall = \frac{TN}{TN + FP} \tag{4.27}$$

Precison und Recall beziehen sich auf die einzelnen Klassen und ermöglichen so, Rückschlüsse für eine einzelne Klasse zu ziehen.

4.5.2. Validierung

In dem vorherigen Abschnitt wurden Gütekriterien vorgestellt, die verwendet werden, um den Lernerfolgt von Lernverfahren nach dem Test an Beispielen zu bewerten.
Erfolgt ein Test des Klassifikators auf den Trainingsdaten (also den Beispielen, auf denen das Lernverfahren auch gelernt hat), führt dies zu viel optimistischeren Ergebnissen, als es bei neuen unbekannten Daten in Wirklichkeit zu erwarten wäre. Man spricht in diesem Fall von einer Überanpassung (Overfitting) auf die Trainingsbeispiele. Um eine realistische Schätzung zu erhalten, teilt man die Menge der Beispiele auf.

Trainings- und Testmenge (Hold-Out-Methode)

Eine einfache Methode zur Messung der Leistungsfähigkeit ist die Aufteilung in eine Trainings- und eine Testmenge. In der Praxis teilt man die Menge der Beispiele meistens im Verhältnis zwei Drittel der Trainingsmenge und ein Drittel der Testmenge zu [67]. Bei ausreichend großen Datenmengen kann mit dieser Aufteilung eine gute Schätzung der Leistungsfähigkeit vorgenommen werden.
In der Realität stehen oft nur wenige Beispieldaten zur Verfügung, so dass bei einer Einteilung in eine Trainings- und eine Testmenge nur wenige Beispiele zum Trainieren eines Klassifikators gegeben sind. Dies steht im Gegensatz dazu, dass,

um einen guten Klassifizierer zu erhalten, so viele Daten wie möglich zum Trainieren genutzt werden sollen [67].

Kreuzvalidierung

Um den Nachteil der geringen Beispielmenge und die daraus resultierenden Probleme zu vermeiden, wird in vielen Fällen eine *k-fache Kreuzvalidierung* (engl. cross validation) angewendet. Hierbei wird der Datensatz in k disjunkte Teilmengen $K_1, K_2, ..., K_k$ von möglichst gleicher Größe zerlegt. Nach der Zerlegung, wird k-mal trainiert und getestet. Bei jeder Iteration wird die - wechselnde - k-te Teilmenge als Testmenge zurückgelegt und auf den restlichen $k-1$ Teilmengen trainiert. Nach den k Durchläufen werden die Ergebnisse der einzelnen Läufe gemittelt.

4.6. Klassifikation als Anwendung der Microarray-Technologie

Wie in Kapitel 4.3 erwähnt, zählt die Klassifikation zum überwachten Lernen. Ziel der Klassifikation ist, jedes Beispiel einer Klasse zuzuordnen. Bei mit Micorarrays gewonnenen Daten stellen die pro Patient ermittelten Expressionswerte der Microarray-Experimente die Eingabewerte und die zugeordnete Klassenzugehörigkeit des Patientens den Ausgabewert dar. Bei den gegebenen Daten entspricht also ein Patientenprofil einem Beispiel; ein Beispiel wird als Vektor $\vec{x_i}$ aufgefasst. Die Merkmalsausprägungen des Vektors entsprechen den Expressionswerten der Exons eines Patienten. Somit stellt ein Patientenprofil einen 287.329-dimensionalen Vektor (Merkmalsvektor) $\vec{x_i}$ dar. Jedem Patienten ist ein Label y_i für die Klassenzugehörigkeit "Rezidiv“ oder „kein Rezidiv“ zugeordnet.
Da die Patienten in zwei Klassen unterteilt werden können, handelt es sich um ein binäres Klassifikationsproblem. Zur Lösung der Klassifikationsaufgabe werden die vorgestellten Lernverfahren (siehe Kap. 4.4) verwendet und die Lernergebnisse mit den beschriebenen Gütekriterien (siehe Kap. 4.5.1) bewertet.

Die SVM wurde für diese Aufgabe gewählt, da diese sich zur Klassifikation hochdimensionaler Daten (hohe Anzahl beschreibender Merkmale) nach [70] gut eignet. Eine Übersicht der allgemeinen Anwendungsmöglichkeiten der SVM ist in [27] zu finden. Die geeignete Verwendung der SVM auf Gen-Expressiondaten ist in der Literatur z.B. in [14], [26] und [28] beschrieben. Auch die Verwendung des k-Nächsten-Nachbarn Verfahrens auf Gen-Expressiondaten ist in der Literatur zu finden, z.B. in [19], [26].
In der Literatur werden die Lernverfahren auf Expressionswerte von Genen eingesetzt. Die Verwendung der Lernverfahren für Daten, die mittels der neueren

Exonarray-Technologie gewonnen wurden, konnte bei der Literaturrecherche bisher nicht gefunden werden.

5. Merkmalsauswahl (feature selection)

Dieses Kapitel widmet sich der Merkmalsauswahl (Synonym: Merkmalsselektion). Nach einer Einführung werden für diese Arbeit wichtige statistische Kennwerte beschrieben und die verwendeten Methoden zur Merkmalsauswahl vorgestellt.

5.1. Einführung

Die Anzahl der Merkmale, die pro Beispiel (im vorliegendem Fall pro Patient) erhoben werden, ist in den letzten Jahren durch Neu- und Weiterentwicklungen der Systeme zur Datenerhebung rasant angestiegen, so dass die einzelnen Beispiele durch eine enorme Anzahl von Merkmalen repräsentiert (hochdimensionale Datensätze) werden. Beispielsweise werden, wie bereits beschrieben, bei Exon-Microarray-Experimenten bis zu 1,4 Millionen Merkmale pro Patient ermittelt.
Dieser Entwicklung folgend wird wohl auch in Zukunft für viele praktische Anwendungen die Anzahl von Merkmalen weiter ansteigen. Bei dieser Menge von Daten tritt häufig der sog. „Fluch der Dimensionen“ [4] auf. Dieses Problem beschreibt den Zusammenhang, dass mit zunehmender Anzahl von Merkmalen (Dimensionen) die für eine akzeptable Leistung des Systems benötigte Anzahl an Beispielen exponentiell steigt.
Der vorliegende Neuroblastom-Datensatz enthält Expressionswerte von 131 Patienten, die zur weiteren Analyse verwendet werden; dies ist im Vergleich zur hohen Dimension an Merkmalen eine sehr geringe Anzahl. Da in vielen Bereichen durch eine Reduktion der Dimension (geringere Anzahl an Merkmalen) ein besseres Lernergebnis erzielt werden konnte, sollte auch für die vorliegenden Neuroblastom-Daten eine Überprüfung, ob eine Reduktion der Merkmale eine Verbesserung der Lernergebnisse bewirkt, versucht werden.
Ein weiterer Nachteil, der sich aus dem großen Datenvolumen ergibt, ist die Laufzeit der Lernverfahren auf diesen Daten. Durch eine Selektion von Merkmalen und der damit verbundenen Dimensionsreduktion kann die Laufzeit erheblich reduziert werden.
Wie bereits erläutert, ist man aus biologischer Sicht zum einen an einer genauen Vorhersage der Klassenzugehörigkeit der Patienten interessiert und zum anderen auf der Suche nach signifikanten Merkmalen, also Merkmalen, die zwischen den

Patientengruppen stark unterschiedlich oder innerhalb der Gruppen sehr ähnlich sind, um aus diesen signifikanten Merkmalen Rückschlüsse auf biologische Zusammenhänge zu gewinnen. Dazu ist diese hohe Anzahl an Merkmalen (ein Merkmal entspricht einem Gen/Exon) problematisch; die Interpretierbarkeit der Ergebnisse bei solchen Dimensionen ist oft sehr schwierig bis unmöglich. Durch eine geschickte Merkmalsauswahl müssten bei einer medizinischen Anwendung dann nicht mehr 1,4 Millionen Merkmale auf einem Array untersucht werden, sondern nur noch die reduzierte Anzahl von Merkmalen. Aus den erwähnten Gründen ist der Versuch einer Dimensionsreduktion auch aus dieser Sicht sinnvoll.
Ein wichtiger Aspekt, der noch erwähnt werden soll, ist, dass durch die Methoden der Merkmalsauswahl die Original-Repräsentation der Merkmale nicht verändert, sondern nur eine Untermenge der Merkmale ausgewählt wird [50]. Andere Dimensionsreduktions-Verfahren (Merkmalextraktion), wie z.B. die Hauptkomponentenanalyse (principial component analysis, für weitere Details dieses Verfahren siehe [29]) führen zu der erwähnten Veränderung der Repräsentation.

Zusammenfassend lässt sich festhalten, dass das Ziel der Merkmalsauswahl darin besteht, eine Teilmenge der gegebenen Merkmale auszuwählen, wodurch folgende Vorteile angestrebt werden:

- Datenreduktion zur Verringerung des Speicherplatzes und zur Verbesserung der Laufzeit der Lernverfahren
- Verbesserung der Lernergebnisse
- Erleichterung des Datenverständnisses,tieferer Einblick in den unterliegenden Prozess, durch den die Daten generiert wurden.
- bessere Visualisierung der Daten

Klassen von Methoden zur Merkmalsauswahl

Es existieren unterschiedliche Methoden zur Merkmalsauswahl. Diese werden in der Literatur in die drei folgenden Klassen unterteilt (vgl. [10]):

- Filter-Methoden
- Wrapper-Methoden
- Embedded-Methoden

Die verschiedenen Methoden unterscheiden sich deutlich im Ablauf und ihrem Konzept. Daher werden die wesentlichen Eigenschaften im Folgenden aufgeführt.

Filter-Methoden

Die Merkmalsselektion mit einem Filter-Ansatz erfolgt unabhängig vom angewendeten Lernverfahren. Filter-Methoden können somit als Vorverarbeitungsschritt für die danach angewendeten Lernverfahren dienen. Mit einer Bewertungsfunktion (Relevanzkriterium) wird die Güte für jedes einzelne Merkmal unabhängig voneinander bewertet. Es gibt viele unterschiedliche Bewertungsfunktionen, einige werden im Kap. 5.3 vorgestellt. Nach der Bewertung erfolgt meistens die Erstellung einer Rangliste der Merkmale auf der Basis ihrer Bewertung. Die Merkmale der höchsten Ränge werden für die weitere Analyse ausgewählt und/oder den Lernverfahren als Eingabe übergeben. Die Anzahl der auszuwählenden Merkmale kann vom Nutzer selbst bestimmt werden. Es werden die ersten k Merkmale des Rankings gewählt oder alternativ nur die Merkmale weiter berücksichtigt, deren Bewertungen einen bestimmten Schwellwert erreichen.
Ein entscheidender Vorteil der Filter-Methoden ist, dass diese Methoden einen vergleichsweise geringen Rechenaufwand benötigen (die Komplexität ist linear in der Anzahl der Merkmale) und diese Methoden unabhängig vom späteren Lernverfahren arbeiten. Die Unabhängigkeit vom Lernverfahren kann sich allerdings auch als Nachteil erweisen, da die durch die Bewertungsfunktion gewichteten Merkmale und die darauf begründete geeignete ausgewählte Untermenge von Merkmalen für das nachgeschaltete Lernverfahren nicht optimal geeignet sein kann.
Die Filter-Methoden lassen sich in univariate und multivariate Filter unterteilen. Multivariate Filter berücksichtigen im Unterschied zu univariaten Filtern die Abhängigkeit der Merkmale untereinander. Die multivariaten Filter sind in ihrer Anwendung zeitintensiver. Der t-Test (siehe Kap. 5.3.1) ist ein Vertreter der univariaten Filter. Beispiele für die multivariaten Filter sind Relief (siehe Kap. 5.3.3) und der „Markov blanket filter“ [38].

Wrapper-Methoden

Bei den Wrapper-Methoden erfolgt die Merkmalsselektion wie bei den Filter-Methoden in einem Vorverarbeitungsschritt, jedoch wird ein Lernverfahren als *Blackbox* genutzt, um die Selektion auszuführen. Über das angewendete Lernverfahren ist kein Wissen notwendig. Nach einem bestimmten Schema werden verschiedene Teilmengen gebildet; diese Teilmengen werden evaluiert bis ein zuvor definiertes Stoppkriterium erreicht ist. Als Bewertungsfunktion wird hier die Modellgüte einer Teilmenge herangezogen. Ein Nachteil dieser Methoden ist, dass sie sehr rechenaufwendig sind und im Vergleich zu den Filter-Methoden ein höheres Risiko zum „Overfitting“ haben [50]. Durch ihre hohe Komplexität werden diese Methoden meist nur bei kleinen Merkmalsmengen (< 100) eingesetzt.
Wrapper-Methoden benutzen oft die folgenden Selektionverfahren:

- *Sequential Forward Selection*(SFS): Hierbei wird mit einer leeren Menge von Merkmalen gestartet und in jedem Schritt wird sukzessiv immer das „beste“ Merkmal hinzugefügt.

- *Sequential Backward Elimination* (SBE): Zu Beginn ist die Menge aller Merkmale gegeben und in jedem Schritt wird sukzessiv das „schlechteste“ Merkmal entfernt.

Embedded-Methoden

Im Gegensatz zu den beiden vorigen vorgestellten Ansätzen erfolgt die Merkmalsauswahl bei Embedded-Methoden nicht während eines Vorverarbeitungsschrittes, sondern die Suche nach der optimalen Untermenge an Merkmalen ist bei diesen Methoden in den Lernprozess integriert und kann als eine Suche in einem kombinierten Raum aus Merkmalen und Hypothesen aufgefasst werden [50]. Der Vorteil dieser Methoden besteht darin, dass eine Interaktion zu dem Lernverfahren besteht, die Methoden aber nicht so rechenzeitintensiv wie Wrapper-Methoden sind. Beispiele für diese Methode sind Entscheidungsbäume und die Verwendung des Gewichtsvektors der SVM (siehe Kap. 5.3.4.). Bei den Entscheidungsbäumen wird an jeder Stelle eines Entscheidungsvorgangs nach dem bedeutsamsten Merkmal für das weitere Vorgehen gesucht; dies ist das Merkmal, welches eine bestmögliche Trennung der Klassen erlaubt. Die Konstruktion des Baumes wird solange fortgesetzt bis ein Stoppkriterium erfüllt wird. Entscheidungsbäume suchen somit nach der Menge der bedeutsamsten Merkmale. Der von Guyon et al. [28] vorgeschlagene Algorithmus *Recursive Feature Elimination (RFE)* wendet eine lineare SVM an und bestimmt nach jedem Trainieren des Modells anhand der Einträge des Normalenvektors ein Merkmal, das im weiteren Verlauf des Algorithmus aufgrund seiner geringen Relevanz ausgeschlossen wird. Dieser Vorgang wird solange durchgeführt bis eine Reihenfolge der Wichtigkeit der Merkmale bestimmt ist.

5.2. Statistische Kennwerte

Bevor auf die verwendeten Methoden zur Merkmalsauswahl eingegangen wird, werden in diesem Abschnitt statistische Kennwerte definiert, die im weiteren Verlauf dieser Arbeit verwendet werden.

Für eine Indikatorfunktion I und eine Aussage a gilt:

$$I(a) = \begin{cases} 0, & \text{wenn } a \text{ falsch ist} \\ 1, & \text{wenn } a \text{ wahr ist} \end{cases} \tag{5.1}$$

Sei D ein Datensatz mit n d-dimensionalen Beispielen $x_1, ..., x_n$, ein i-tes Beispiel kann aufgefasst werden als ein Vektor $x_i = (x_{i1},...,x_{id})$ von Merkmalsausprägungen (Attribute) einer Merkmalsmenge X = $(X_1,...,X_d)$. Zu jedem Beispiel x_i existiert ein Klassenzugehörigkeitsmerkmal (Label) y_i. Bei einem binären Klassifikationsproblem, wie es in dieser Arbeit behandelt wird, liegen zwei Klassen C_1 und C_2 vor mit $x_i \in \{C_1, C_2\}$. Für die zwei Klassen C_1 und C_2 gilt $C_1 \cap C_2 = \emptyset$ und $|C_1 \cup C_2| = n$. Y = $(y_1, y_2, ..., y_n)$ sei der n-dimensionale Klassenvektor mit

$$y_i = \begin{cases} 0, & \text{wenn } y_i \in C_1 \\ 1, & \text{wenn } y_i \in C_2 \end{cases} \tag{5.2}$$

Der Mittelwert μ_{j1} für ein Merkmal (Exon) X_j in Klasse C_1 bzw. μ_{j2} für C_2 ist dann wie folgt definiert:

$$\mu_{j1} = \frac{1}{m_1} \sum_{i=1}^{n} x_{ij} I(y_i = 0) \tag{5.3}$$

$$\mu_{j2} = \frac{1}{m_2} \sum_{i=1}^{n} x_{ij} I(y_i = 1) \tag{5.4}$$

Dabei stehen die Variablen m_1 und m_2 für die Anzahl der Beispiele der Klassen C_1 und C_2.

$$m_1 = \sum_{i=1}^{n} I(y_i = 0) \tag{5.5}$$

$$m_2 = \sum_{i=1}^{n} I(y_i = 1) \tag{5.6}$$

Die empirischen Varianzen s_{j1}^2 und s_{j2}^2 lassen sich sich mit Hilfe der Mittelwerte μ_{j1} und μ_{j2} berechnen:

$$s_{j1}^2 = \frac{1}{m_1 - 1} \sum_{i=1}^{n} (x_{ij} - \mu_{j1})^2 I(y_i = 0) \tag{5.7}$$

$$s_{j2}^2 = \frac{1}{m_2 - 1} \sum_{i=1}^{n} (x_{ij} - \mu_{j2})^2 I(y_i = 1) \tag{5.8}$$

Weiter lassen sich die Standardabweichungen s_{i1} und s_{i2} ermitteln mit

$$s_{j1} = \sqrt{s_{j1}^2} = \sqrt{\frac{1}{m_1 - 1} \sum_{i=1}^{n} (x_{ij} - \mu_{j1})^2 I(y_i = 0)} \tag{5.9}$$

$$s_{j2} = \sqrt{s_{j2}^2} = \sqrt{\frac{1}{m_2 - 1} \sum_{i=1}^{n} (x_{ij} - \mu_{j2})^2 I(y_i = 1)} \tag{5.10}$$

5.3. Verwendete Methoden zur Merkmalsauswahl

In dem folgenden Abschnitt werden Methoden vorgestellt, die in dieser Arbeit zur Bewertung/Gewichtung der Merkmale herangezogen werden. Die Auswahl einer Merkmalsmenge erfolgt anhand der auf diesen Bewertungen erstellten Rangliste.

5.3.1. t-Statistiken

Die t-Statistiken zählen zu den im vorigen Abschnitt vorgestellten Filter-Methoden.

t-Test für gleiche Varianzen($s_{i1}^2 = s_{i2}^2$)

Eine oft verwendete Testmethode ist der sog. ***t*-Test**, der „Standardtest“ in der Statistik. Er gehört zu den parametrischen Tests[1].
Der t-Test kann als Einstichprobenfall oder als Zweistichprobenfall sowie für abhängige und unabhängige Stichproben durchgeführt werden. Im Weiteren wird ausschließlich der unabhängige Zweistichprobenfall behandelt. „Zweistichprobenfall“ bedeutet, dass Merkmalsunterschiede zwischen zwei Stichproben untersucht werden. Zwei Stichproben heißen voneinander *unabhängig*, falls ihre jeweiligen Stichprobenvariablen voneinander unabhängig sind.
Die dieser Arbeit zugrunde liegenden Expressionsdaten stellen einen unabhängigen Zweistichprobenfall dar, da Merkmale aus zwei verschiedenen Klassen (Rezidiv, kein Rezidiv) analysiert werden und die Stichproben unabhängig voneinander sind, da zur Erhebung der Merkmale unterschiedliche Patienten getestet wurden.
Vor Anwendung dieser Methode werden eine Nullhypothese H_0 und eine Alternativhypothese H_1 formuliert. Die Nullhypothese ist zumeist die Formulierung der Gleichheit, die Alternativhypothese die Formulierung eines Unterschieds bezüglich einer interessierenden Fragestellung. Man kann die Hypothesen zweiseitig (Gleichheit vs. Unterschied) oder einseitig (Gleichheit vs. positiver Effekt bzw. Gleichheit vs. negativer Effekt) formulieren. Mit Hilfe des t-Tests kann man die formulierten Hypothesen überprüfen.
In vielen praktischen Anwendungsfällen ist nur wenig über die tatsächliche Verteilung eines Merkmals bekannt. Die Annahme der Normalverteilung ist zwar oft begründet, dennoch ist die Varianz der Verteilung unbekannt. Die Idee des

[1] Parametrische statistische Tests setzen voraus, dass die Verteilung des untersuchten Merkmals in der Population bekannt ist.

t-Tests beruht darauf, die Varianz aus den Stichprobendaten zu schätzen. Bei den Exon-Expressionsdaten sind die Ausprägungen eines Merkmals die Expressionswerte eines Exons von mehreren Patienten (Beispielen). Für ein Exon liegen m_1 Expressionwerte der Klasse C_1 vor, $\mathrm{N_j}(\mu_{j1}, \sigma_{j1}^2)$ sei die angenommene Normalverteilung mit Mittelwert μ_{j1} und Varianz σ_{j1}^2. Ebenso liegen für das Exon m_2 Expressionwerte der Klasse C_2 vor, $\mathrm{N_j}(\mu_{j2}, \sigma_{j2}^2)$ sei die angenommene Normalverteilung mit Mittelwert μ_{j2} und Varianz σ_{j2}^2.
Die Hypothesen für die Exon-Expressionsdaten sind:

- **Nullhypothese**: H_0: $\mu_{j1} = \mu_{j2}$
  Bei betrachtetem Exon gibt es keinen Unterschied in den Expressionen der zwei zu vergleichenden Gruppen.

- **Alternativhypothese**: H_0: $\mu_{j1} \neq \mu_{j2}$
  Ein Exon kann als differentiell exprimiert angesehen werden, wenn die Nullhypothese H_0: $\mu_{j1} = \mu_{j2}$ abgelehnt und die Alternativhypothese akzeptiert wird.

Eine weitere Voraussetzung zur Annahme der Normalverteilung ist die Homogenität der unbekannten Varianzen ($s_{j1} = s_{j2}$), d.h. das untersuchte Merkmal hat in beiden Gruppen die gleiche unbekannte Standardabweichung s_j. Sind die Voraussetzungen erfüllt, kann die Teststatistik (Prüfwert) t berechnet werden: [43]

$$\text{Teststatistik } t = \frac{|\mu_{j1} - \mu_{j2}|}{sd_j} \sqrt{\frac{m_1 m_2}{m_1 + m_2}}$$

mit

$$sd_j = \sqrt{\frac{1}{(m_1 - 1) + (m_2 - 1)} \Big((m_1 - 1)s_{j1}^2 + (m_2 - 1)s_{j2}^2 \Big)}$$

(sd_j bezeichnet die gepoolte Standardabweichung der Stichproben (Schätzwert für die gemeinsame Standardabweichung))

Die Teststatistik t spiegelt das Verhältnis von Differenz der Stichprobenmittelwerte zum Standardfehler der Differenz der Stichprobenmittelwerte wider. Es können sowohl positive als auch negative Werte für t entstehen.
Die ermittelte Teststatistik t ist nicht mehr normalverteilt, sondern t-verteilt[1] mit

[1] Die t-Verteilung entsteht, wenn eine normalverteilte Zufallsgröße durch die Wurzel einer chi-quadratverteilten Zufallsgröße dividiert wird [43].

$m_1 + m_2 - 2$ Freiheitsgraden[2].

Zu der Teststatistik kann ein sogenannter p-Wert berechnet werden, welcher ausdrückt, wie wahrscheinlich oder unwahrscheinlich die Ergebnisse der gegebenen Stichproben sind, wenn von der Gültigkeit der Nullhypothese in der Grundgesamtheit ausgegangen wird. Liegt der ermittelte p-Wert unter einem vorgegebenem Signifikanzniveau α, dann wird die Nullhypothese abgelehnt. Der p-Wert ist somit ein Maß für die Signifikanz der Abweichung der Daten von der Nullhypothese. Je kleiner der p-Wert, desto weiter entfernt liegt die Teststatistik von ihrem erwarteten Wert und desto deutlicher wird demnach die Nullhypothese abgelehnt.

t-Test für ungleiche Varianzen ($s_{j1}^2 \neq s_{j2}^2$)

Auch für den Fall, dass die Varianzen der Verteilung ungleich sind oder Hinweise darauf vorliegen, also $s_{j1}^2 \neq s_{j2}^2$ ist, kann eine t-verteilte Teststatistik berechnet werden.
Dieser generalisierte t-Test für ungleiche Varianzen wurde von B.L. Welch [63] vorgeschlagen und trägt daher den Namen **Welch-Test**.
Bei aktiven Genen kann von einer größeren Variabilität in ihren Expressionswerten ausgegangen werden als bei inaktiven Genen. Somit kann eine Ungleichheit der Varianzen angenommen werden. Auch bei den Exon-Daten wird von dieser Variabilität der Expressionswerte ausgegangen. Wie der t-Test setzt auch der Welch-Test eine Normalverteilung voraus.
Die Hypothesen für die Exon-Expressionsdaten sind:

- **Nullhypothese**: H_0: $\mu_{j1} = \mu_{j2}$
  Die Nullhypothese H_0 ist identisch mit der des einfachen t-Tests.

- **Alternativhypothese**: H_0: $\mu_{j1} \neq \mu_{j2}$
  Ebenso stimmt die Alternativhypothese H_1 mit der Hypothese des obigen t-Tests überein.

Beim Welch-Test ist die Berechnung der Teststatistik etwas einfacher als beim einfachen t-Test. Die Berechnung der Freiheitsgrade für die t-Verteilung weicht von der des t-Tests ab.

$$\text{Teststatistik } t = \frac{|\mu_{j1} - \mu_{j2}|}{\sqrt{\frac{s_{j1}^2}{m_1} + \frac{s_{j1}^2}{m_2}}}$$

[2]Freiheitsgrade = die Anzahl unbestimmter Werte, aus denen sich eine Statistik zusammensetzt

Berechnung der Freiheitsgrade:

$$\text{(Hilfsgröße) } c = \frac{\frac{s_{j1}^2}{m_1}}{\frac{s_{j1}^2}{m_1} + \frac{s_{j2}^2}{m_2}}$$

$$\text{genäherte Anzahl Freiheitsgrade } v = \left\lfloor \frac{1}{\frac{c^2}{m_1 - 1} + \frac{(1-c)^2}{m_2 - 1}} \right\rfloor$$

(v ist i.a. keine ganze Zahl und wird auf die nächste ganze Zahl abgerundet)

Im Falle gleicher Stichprobenumfänge ($m_1 = m_2 = n$) ergeben sich folgende Vereinfachungen:

$$\text{Teststatistik } t = \frac{|\mu_{j1} - \mu_{j2}|}{\sqrt{\frac{s_{ij}^2 + s_{j2}^2}{n}}}$$

$$\text{Freiheitsgrade } v = \frac{(n-1)(s_{j1}^2 + s_{j2}^2)^2}{(s_{j1}^2)^2 + (s_{j2}^2)^2}$$

Die berechnete Teststatistik ist t-verteilt mit der entsprechend berechneten Anzahl Freiheitsgrade. Auch für den Welch-Test kann ein entsprechender p-Wert berechnet werden.

Die Teststatistik t kann als Bewertungsfunktion für ein Exon verwendet werden. Ein Exon bekommt einen hohen Wert, wenn die Expressionsmittelwerte der beiden Klassen für dieses Exon eine große Differenz bei geringer Varianz aufweisen. Somit kann mittels des t-Wertes ein Ranking für die Exons erstellt werden. Die Bewertung der Exons kann auch durch den ermittelten p-Wert erfolgen; hierzu wird ein Ranking der p-Werte erstellt (kleine p-Werte erhalten einen kleinen Rang) und eine entsprechende Merkmalsauswahl vorgenommen.

In dieser Arbeit wurde zur Bewertung der einzelnen Exons der Welch-Test angewendet, da von unterschiedlichen Varianzen ausgegangen wird. Es wurde sowohl eine Bewertung anhand der Teststatistik t wie auch anhand des ermittelten p-Wertes vorgenommen.

5.3.2. Significance Analysis of Microarrays (SAM)

Eine weitere statistische Methode zur Selektion differentiell exprimierter Exons ist die von Tusher, Tibshirani und Chu vorgeschlagene Methode SAM (Significant Analysis of Microarrays [59]).
Diese Teststatistik von SAM kann verwendet werden, um eine Bewertung der Exons vorzunehmen und ein Ranking aufzustellen. Bei den bisher vorgestellten t-Statistiken kann es vorkommen, dass kleine Varianzen eines Exons geringe Mittelwertsdifferenzen signifikant erscheinen lassen. Die SAM-Methode soll dieses Problem verhindern. SAM berechnet für jedes Exon (Merkmal) einen sogenannten „SAM score“ d_j nach der Formel:

$$\text{(Teststatistik) SAM score } d_j = \frac{\mu_{j1} - \mu_{j2}}{sd_j + s_0}$$

$$\text{mit } sd_j = \sqrt{\frac{(m_1 - 1)s_{j1}^2 + (m_2 - 1)s_{j2}^2}{m_1 + m_2 - 2}\left(\frac{1}{m_1} + \frac{1}{m_2}\right)}$$

(sd_j entspricht dem Schätzwert für die gemeinsame Standardabweichung)

Die Gleichung für sd_j ist identisch mit der Gleichung, die vom t-Test (mit gleichen Varianzen) zur Berechnung der Standardabweichung benutzt wird. Die Konstante s_0 soll ein zu-groß-Werden des SAM-score verhindern, falls die Standardabweichungen sd_j nahe bei Null liegen.
s_0 ist für alle Exons gleich und wurde in den Experimenten auf das Quantil $Q_{0,5}$ der gepoolten Standardabweichungen gesetzt.

5.3.3. Relief

Diese Methode wurde entwickelt von Kira et al. [36][37]; seitdem wurden einige Modifikationen und Erweiterungen an dem Verfahren vorgenommen (siehe z.B.[40]).
Der Relief-Algorithmus wird in vielen Bereichen des maschinellen Lernens eingesetzt, so z.B. auch in der Bioinformatik zur Gen-Selektion [71]. Das Relief-Verfahren zählt zu den multivariaten Filteransätzen. Multivariate Methoden bewerten ein Merkmal nicht isoliert, sondern es wird das Zusammenwirken mehrerer Merkmale zugleich betrachtet. Auf diese Weise werden Abhängigkeiten zwischen den Merkmalen berücksichtigt.
Der Algorithmus basiert auf Instanz-basiertem Lernen (Instanz = Beispiel). Die Grundlagen des Relief-Algorithmus sind eine heuristische Suchfunktion und ein Distanzmaß.

Im Folgenden wird die Arbeitsweise des Original-Algorithmus [36][39] beschrieben:
Zuerst wird zufällig eine Instanz ausgewählt, dann wird zu dieser ausgewählten Instanz der nächste Nachbar (Near-hit) derselben Klasse gesucht, also die Instanz, welche mit den meisten Merkmalen der gewählten Instanz übereinstimmt und der gleichen Klasse angehört. Ebenso wird zu der gewählten Instanz der nächste Nachbar (Near-miss) gewählt, also die Instanz, die in möglichst vielen Merkmalen mit der ausgewählten Instanz übereinstimmt, aber der anderen Klasse angehört. Als Distanzmaß für die Suche der nächsten Nachbarn (Near-hit, Near-miss) wird der Euklidische Abstand (siehe Kap. 4, Formel 4.1) verwendet.
Die Atttributgewichte werden basierend auf den Differenzen zum „Near-hit“ bzw. „Near-miss“ angepasst. Gewichte von Merkmalen, die zwischen der ausgewählten Instanz und dem nächsten Nachbarn derselben Klasse übereinstimmen oder mit dem nächsten Nachbarn der anderen Klasse nicht übereinstimmen, werden erhöht. Merkmale, die zwischen der ausgewählten Instanz und dem nächsten Nachbarn derselbe Klasse unterscheiden oder mit dem nächsten Nachbarn der anderen Klasse übereinstimmen, werden verringert. Relief ignoriert Korrelationen zwischen Merkmalen; somit kann es vorkommen, dass ausgewählte Merkmale stark korreliert sind.
In den Experimenten dieser Arbeit wurde der erweiterte Algorithmus von Kononenko [40] benutzt; hierbei werden zur Gewichtung der Merkmale mehrere nächste Nachbarn einbezogen.

5.3.4. SVM-Gewichtung

Im Kapitel 4.4.3 wurden Support Vector Machines (SVMs) als Lernverfahren vorgestellt. Wie schon in diesem Kapitel beschrieben, versuchen SVMs eine Hyperebene zu finden, die zwei Klassen optimal voneinander trennen, also die Trainingsbeispiele möglichst fehlerfrei in die positive und die negative Klasse aufzuteilen und dabei den Abstand zu den benachbarten Trainingspunkten zu maximieren.
Die Entscheidungsfunktion, die die Klassenzugehörigkeit eines neuen Beispiels bestimmt, kann durch die folgende Formel dargestellt werden:

$$f(x_i) = sign(\vec{w} * \vec{x_i} + b)$$

Dabei ist $\vec{w}$ der Normalenvektor, b ist die Verschiebung vom Ursprung und $\vec{x_i}$ ist die Vektorrepräsentation eines Beispiels.
Bei einer linearen Trennung fließen die einzelnen Merkmale mit unterschiedlichen Gewichten in die Trennung ein. Je höher der Betrag eines Eintrags des Normalenvektors $\vec{w} = (w_1, ..., w_n)$, desto größer ist der Einfluss des Merkmals i auf die vorgenommene Trennung. Ist ein Merkmal irrelevant für die Trennung, fällt der Betrag des entsprechenden Eintrags der Normalenvektors relativ gering aus. Somit kann eine Gewichtung der Merkmale anhand der Beträge der entsprechenden

Einträge des Normalenvektors vorgenommen werden. Die Gewichte können dann zur Bewertung der Merkmale verwendet und ein entsprechendes Ranking erstellt werden. Anhand des Rankings kann die Merkmalsauswahl erfolgen.

5.4. Robustheit/Stabilität von Merkmalsauswahl-Methoden

Ein wichtiger Aspekt bei der Merkmalsauswahl ist die Robustheit/Stabilität, mit der bestimmte Merkmalsmengen von den Methoden ausgewählt werden. Die Robustheit einer Merkmalsauswahl-Methode kann beschrieben werden durch die Abweichungen der ausgewählten Merkmalsmengen auf leicht veränderten Datensätzen, die alle der gleichen Grundgesamtheit angehören. Werden immer dieselben Merkmale auf den variierenden Datensätzen ausgewählt, resultiert dies in einer hohen Robustheit.
Liegt das Interesse einer Untersuchung im Finden eines guten Klassifikators und wird dabei eine ausgewählte Merkmalsmenge verwendet, spielt die Robustheit der ausgewählten Merkmalsmenge oft nur eine untergeordnete Rolle. In Studien zur Analyse von Microarray-Daten konnte gezeigt werden, dass eine Methode zur Merkmalsauswahl unterschiedliche Untermengen an Merkmalen bei Veränderungen der Trainingsdaten selektiert. Mit den selektierten Untermengen konnten fast identische Lernergebnisse erreicht werden [21]. Für den vorliegenden Neuroblastom-Datensatz bedeutet dies, dass eine Auswahlmethode auf Untermengen des Datensatzes unterschiedliche Mengen von Merkmalen als signifikant bewerten kann und die einzelnen ausgewählten Merkmalsmengen dadurch wenige bis gar keine Übereinstimmungen bei den Merkmalen aufzeigen können. Solche Instabilitäten mindern das Vertrauen der jeweiligen Gebietsexperten. Besonders im Bereich der Biologie und Genetik liegt das Interesse an einer robusten Merkmalsauswahl, da - wenn von einer genetischen Grundlage für eine differentielle Expression eines Merkmals (Gens/Exons) zwischen zwei Patientengruppen ausgegangen werden kann - dieses Merkmal von den Methoden hoch bewertet werden und so in der Auswahl, auch auf unterschiedlichen oder variierenden Datensätzen, immer (oder häufig) vorhanden sein sollte. Um ein Maß für die Robustheit/Stabilität angeben zu können, ist zunächst eine Definition von Ähnlichkeitsmaßen notwendig.

5.4.1. Ähnlichkeitsmaße

Um die Ähnlichkeiten selektierter Merkmalsmengen vergleichen zu können, existieren unterschiedliche Ähnlichkeitsmaße. Ein Beispiel für ein solches Maß ist der

Jaccard-Koeffizient (benannt nach dem Schweizer Botaniker Paul Jaccard (1868-1944)):

$$\text{Jaccard}(v_i, v_j) = \frac{|v_i \cap v_j|}{|v_i \cup v_j|} = \frac{\sum_{t=1}^{u} I(v_i^t = v_j^t = 1)}{\sum_{t=1}^{u} I(v_i^t + v_j^t > 0)} \tag{5.11}$$

v_i ist der d-dimensionale Vektor einer selektierten Merkmalsmenge i; hierbei ist $v_i^t = 1$, wenn das Merkmal f_t in der ausgewählten Merkmalsmenge i vorhanden ist, ansonsten sei $v_i^t = 0$. $I(\cdot)$ ist eine Indikatorfunktion (siehe Formel 5.1) und u ist die Anzahl der zu vergleichenden Merkmalsmengen.

5.4.2. Bewertung der Robustheit/Stabilität

Je ähnlicher sich die selektierten Mengen sind, desto höher ist die Gesamtstabilität. Die Gesamtstabilität kann definiert werden als:

$$\text{Stabilität} = \frac{2\sum_{i=1}^{u}\sum_{j=i+1}^{u} S(v_i, v_j)}{u(u-1)} \tag{5.12}$$

$S(v_i, v_j)$ entspricht dem Ähnlichkeitsmaß zweier Mengen i und j.
Am Beispiel des Jaccard-Koeffizient ergibt sich folgendes Maß:

$$\text{Stabilität} = \frac{2\sum_{i=1}^{u}\sum_{j=i+1}^{u} Jaccard(v_i, v_j)}{u(u-1)} \tag{5.13}$$

Die Gesamtstabilität ist der Durchschnitt der Ähnlichkeiten aller Kombinationen der u Merkmalsmengen.

5.5. Ensembles

Um das Risiko einer instabilen Auswahl an Merkmalen zu reduzieren, wird im weiterem Verlauf die Merkmalsauswahl mit Ensembles vorgestellt. Die zugrunde liegende Idee von Ensembles besteht darin, verschiedenartige Methoden, Klassifikatoren oder Regressions-Modelle auf irgendeine Art zusammenzufassen, um von jedem das Beste zu erhalten. Die einzeln Ergebnisse werden zu einem Gesamtergebnis zusammengefasst. Die Kombination der Ergebnisse kann auf mehrere Arten erfolgen, z.B. durch einen Mehrheitsentscheid oder durch eine Mittelwert-Berechnung.

Vertreter der Ensemble-Methoden sind z.B. Boosting und Bagging.

- **Boosting**, ursprünglich von Schapire [52] vorgestellt, ist ein iterativer Prozess. Schrittweise werden einfache Klassifikatoren trainiert. Hierbei werden Beispiele, die im vorherigen Schritt falsch klassifiziert wurden, höher gewichtet und bekommen so mehr Einfluss im Training. Beim Boosting wird somit jeder erzeugte Klassifikator durch die Leistung des vorherigen Klassifikators beeinflusst. Zusätzlich wird jeder Klassifikator nach dem Beitrag seiner Leistung gewichtet. Der Gesamtklassifikator bildet dann eine gewichtete Mehrheitsentscheidung.

- **Bagging** (auch „Bootstrap-Aggregation“ genannt) wurde von Breiman [11] vorgestellt. Bei dieser Methode werden die einzelnen Klassifikatoren getrennt voneinander erzeugt. Jeder Klassifikator wird auf einer eigenen Trainingsmenge trainiert. Zur Erzeugung dieser Trainingsmengen wird das Verfahren des „Bootstrappings“ verwendet: Aus der Beispielmenge wird so oft (mit Zurücklegen) ein Beispiel gezogen, bis die neue Beispielmenge die selbe Kardinalität wie die Ursprungsmenge hat. Bagging kann für die Klassifikation wie auch für die Regression verwendet werden. Bei der Klassifikation wird jene Klasse vorausgesagt, die am meisten vorausgesagt wurde (Mehrheitsentscheidung). Bei der Regression wird der Durchschnitt aller gemachten Vorhersagen benutzt.

Thomas G. Dietrich deutet in [18] auf drei Problembereiche hin, bei denen der Einsatz eines Ensembles angemessen ist und somit das Ergebnis des Ensembles die Resultate der einzelnen Modelle übertrifft:

- Ein statistisches Problem tritt auf, wenn die Trainingsdaten eine zu kleine Untermenge der Grundgesamtheit ausmachen. So können mehrere Klassifikatoren erzeugt werden, die die Daten mit einer ähnlichen Fehlerrate klassifizieren, aber unterschiedlich generalisieren. Es existieren also mehrere verschiedene, aber gleich gute Hyperebenen. Die Wahrscheinlichkeit, durch die Kombination mit einem Ensemble bessere Ergebnisse als durch einen zufällig ausgewählten Klassifikator zu bekommen, ist hoch.

- Ein zweiter Problembereich ist, dass verschiedene einzelne Klassifikatoren bei ihrer Suche in einem lokalen Maximum landen können. Bei Verwendung eines Ensembles ist es wahrscheinlicher, eine bessere Approximation des Optimums zu erreichen.

- Der dritte Problembereich ist, dass die wahre Funktion nicht durch eine Hyperebene aus dem Hypothesenraum angenähert werden kann. Mithilfe

einer Aggregation der einzelnen Klassifikatoren lässt sich der Hypothesenraum erweitern.

5.5.1. Merkmalssauswahl mit Ensembles

Eine weitere Anwendungsmöglichkeit für Ensemble-Verfahren ist, die Stabilität der Auswahl von Merkmalen zu verbessern. Wie in Abschnitt 5.4 bereits angesprochen ist man in Bereichen, wie beispielsweise der Biologie, an einer stabilen Auswahl an Merkmalen interessiert, um z.B. biologische Rückschlüsse ziehen zu können. Hier kann der Einsatz von Ensemble-Verfahren vorteilhaft sein.
Bei der Merkmalsauswahl können verschiedene einzelnen Mengen von ausgewählten Merkmalen zu einer Gesamtmenge kombiniert werden. Für die Kombination dieser Mengen existieren wie bei der Kombination von Klassifikatoren im vorherigen Abschnitt verschiedene Möglichkeiten. Saeys et al. z.B. setzten in [51] Ensemble-Verfahren zur stabileren Merkmalsauswahl ein.

5.5.2. Anwendung von Ensembles

In der vorliegenden Arbeit wird ein Ensemble benutzt, indem die auf Untermengen (Untermengen der Beispielsmenge) selektierten Mengen von Merkmalen so kombiniert werden, dass immer (oder häufig) vertretene Merkmale der einzelnen Mengen in der kombinierten Menge einen hohen Rangplatz erhalten. Durch diese Kombination soll eine stabilere Merkmalsmenge entstehen.
Die Aufteilung des Datensatzes in Teilmengen, die Kombination der Merkmalsmenge und die Beurteilung der Robustheit erfolgt dabei auf folgende Weise:

Aufteilung

Ein Datensatz D wird in n Untermengen D_i möglichst derselben Größe zerlegt, es gilt $D_1 \cup D_2 \cup ... \cup D_n = D$ und $D_i \cap D_j = \emptyset$, $\forall i, j \, \{1, .., n\}$.

Kombination

Es erfolgen n Iterationen. Pro Iteration wird für die Menge D/D_i $(i = 1, .., n)$ eine Bewertung $v_{i,j}$ der Merkmale X_j, $j = 1, ..., d$ vorgenommen; diese Bewertung erfolgt mit einer vorher ausgewählten Methode (z.B. t-Statistik, Relief, etc.). Die Merkmale werden gemäß ihrer Bewertungen sortiert und die Bewertungen der ersten k Merkmale werden auf $v_{i,j} = 1$ gesetzt, alle weiteren Merkmale erhalten als neues Gewicht $v_{i,j} = 0$. Das Gesamtgewicht v_j für ein Merkmal X_j im Ensemble wird ermittelt, indem aufsummiert wird, in wie vielen Iterationen das

entsprechende Merkmal auf den ersten k Rängen landet, also ein Gewicht $v_{i,j} = 1$ hat.

$$v_j = \sum_{i=1}^{n} (v_{i,j})$$

Aus den sortierten Gesamtgewichten v_j, $j = 1, .., d$ wird ein Gesamt-Ranking erstellt und die Merkmale der k höchsten Ränge werden als die „signifikanten" Merkmale zurückgegeben und können für weitere Untersuchungen verwendet werden.

Robustheit der Merkmalsauswahl

Zur Bewertung der Robustheit der Merkmalsauswahl mit der Ensemble-Methode werden u Merkmalsmengen, die anhand von Gesamt-Rankings durch Ensembles ausgewählt wurden, mit dem im Formel 5.11 vorgestellten Jaccard-Koeffizient auf ihre Ähnlichkeit überprüft. Die Gesamtrobustheit wird dann nach Formel 5.12 berechnet.

6. Durchgeführte Experimente

In diesem Kapitel werden die im Rahmen dieser Arbeit durchgeführten Experimente vorgestellt. Nach einer Beschreibung der verwendeten Lernumgebung wird zunächst die Vorverarbeitung der Patienten-Daten erläutert. Danach werden die einzelnen Experimentreihen und deren Ergebnisse präsentiert.

6.1. Verwendete Lernumgebung

Alle Experimente wurden mit der frei verfügbaren Lernumgebung *RapidMiner* (ehemals *YALE*) [45] durchgeführt. RapidMiner stellt verschiedene Lernverfahren und Operatoren zur Datenvorverarbeitung, Merkmalsselektion und Generierung, Parameteroptimierung, Validierung, Visualisierung, etc. zur Verfügung. Dieses Java-basierende System wurde am Lehrstuhl 8 der Fakultät Informatik der Universität Dortmund entwickelt. Als Open-Source-Projekt beinhaltet es die Möglichkeit, Erweiterungen/Veränderungen am Quellcode vorzunehmen und eigene Implementierungen einzubinden.

6.2. Vorverarbeitung der Daten

Damit die Patienten-Daten von *RapidMiner* verarbeitet werden können, müssen zunächst einige grundsätzliche Vorverarbeitungsschritte erfolgen, so muss z.B. die Matrix der Eingabedaten transponiert werden, die Label (Klassenmerkmale) umbenannt werden, etc. Da einige Patientenprofile aus biologischer Sicht nicht für die Untersuchungen geeignet sind - weil diese Patienten in ihrem Krankheitsbild völlig atypisch sind und wahrscheinlich einer anderen Tumorerkrankung zuzuordnen sind - werden diese Beispiele zunächst entfernt.
Nach diesen Vorverarbeitungsschritten verbleiben 131 Patientenprofile, aufgeteilt in 82 Beispiele der Klasse „EFS“ (ereignisfreies Überleben) und 49 Beispiele der Klasse „Event“ (Rezidiv). Pro Patient sind Werte für 287.329 Merkmale (Exons) und ein Label vorhanden.
Im Verlauf der Experimente werden noch zwei weitere Vorverarbeitungsschritte verwendet.

- **Normalisierung**

  Eine Normalisierung der Eingabedaten kann in vielen Anwendungsfällen nützlich sein. „Normalisierung“ bezeichnet das Transferieren von Werten einer Variablen in ein festes Intervall. Bei nicht normalisierten Daten können Merkmale aufgrund ihres Wertebereichs bevorzugt oder benachteiligt werden. Dieser Effekt tritt insbesondere bei stark unterschiedlichen Wertebereichen auf, z.B. bei Werten für Alter und Größe. Erstrecken sich die Attributausprägungen eines Merkmals über ein sehr großes Intervall, fallen diese Merkmale mehr ins Gewicht, als wenn ihre Ausprägungen sich nur über ein sehr kleines Intervall erstrecken. Um diesen Effekt zu vermeiden, werden die Daten normalisiert.
  Zwei gebräuchliche Verfahren zur Normalisierung sind:

  Min-Max-Normalisierung

  Skalierung der Werte x_i, mit $x_i \in \mathbb{R}$ und $i = 1, ..., n$, eines Merkmals X mit Wertebereich $[min_X, max_X]$ in ein vorgegebenes Intervall $[min_{new}, max_{new}]$

  $$x_i^{new} = \frac{x_i - min_X}{max_X - min_X} \ (max_{new} - min_{new}) + min_{new}$$

  z-Score Normalisierung

  Die Werte x_i, mit $x_i \in \mathbb{R}$ und $i = 1, ..., n$, eines Merkmals X werden so transferiert, dass der Mittelwert bei 0 liegt und eine Standardabweichung von 1 erreicht wird:

  $$x_i^{new} = \frac{x_i - \mu}{\sigma}$$

  wobei μ der Mittelwert und σ die Standardabweichung über alle x_i ist.

- **Merkmal-Filter**

  Um bestimmte Merkmale von der weiteren Analyse auszuschließen, wird ein Merkmal-Filter verwendet. Aus biologischer Sicht können Merkmale, die nicht bei mindestens 30% der Patienten einen logarithmierten Expressionswert größer 5 aufzeigen, entfernt werden, weil die meisten Ausprägungen für diese Merkmale aus biologischer Sicht mit hoher Wahrscheinlichkeit auf „Rauschen“ zurückzuführen sind.
  Die Anzahl der Merkmale reduziert sich durch Einsatz dieses Filters von 287.329 Merkmale auf 184.985 Merkmale.

6.3. Experimente ohne Merkmalsauswahl

Die im Folgenden beschriebenen Experimente wurden mit der Zielsetzung durchgeführt, eine Einschätzung darüber abgeben zu können, wie gut sich die Patienten anhand ihrer Expressionsprofile klassifizieren lassen.
Bei den Experimenten wurden die in Kapitel 4.4 vorgestellten Lernverfahren (Stützvektormethode, Naive Bayes und k-Nächster-Nachbar) zur Klassifikation der Patienten anhand ihrer Expressionswerte verwendet. Die Ergebnisse der verschiedenen Lernverfahren wurden anhand der in Kapitel 4.5.1 erläuterten Gütekriterien (Accuracy, Precision und Recall) bewertet.

Zur Vergleichbarkeit der Expressionswerte verschiedener Arrays wurde bereits eine Normalisierung mittels RMA (siehe Kap. 3.3.2) durchgeführt. Die Wertebereiche der einzelnen Merkmale sind nicht gravierend verschieden; ihre logarithmierten Expressionswerte erstrecken sich in einem Wertebereich von 1 bis 14. Durch eine weitere Normalisierung könnten daher Werte erwünschter unterschiedlicher Ausprägungen unterdrückt werden, was einen negativen Einfluss auf die Ergebnisse ausüben könnte.
Um zu bewerten, ob eine Normalisierung der verwendeten Daten sinnvoll ist und welche Normalisierung für diese Daten am geeignetsten ist, wurden die Experimente sowohl ohne Normalisierung als auch mit den beiden vorgestellten Normalisierungen durchgeführt. Bei der Min-Max-Normalisierung wurde der Wertebereich der Merkmale auf das Intervall $[0, 1]$ beschränkt.

Für einen ersten Überblick über die Güte der verschiedenen Lernverfahren wurden zunächst folgende Parametereinstellungen verwendet:
Für den Naive-Bayes Klassifizierer wurden die Voreinstellungen von RapidMiner übernommen; bei dem kNN Verfahren wurde $k = 3$ gesetzt und bei der SVM wurde ein linearer Kern und für den Kostenparameter $C = 1000$ gewählt. Ein linearer Kern ist lt. [17] im Falle einer geringen Beispielmenge aus hochdimensionalen Merkmalsvektoren (wie beim Neuroblastom-Datensatz gegeben) am geeignetsten. Bei allen Experimenten wurden die Resultate 5-fach kreuzvalidiert. Die Beschränkung auf 5-fach wurde gewählt, da die Experimente auf Grund der hohen Anzahl von Merkmalen sehr rechenzeitintensiv sind.

Für sämtliche in diesem Kapitel abgebildeten Ergebnis-Tabellen gelten folgende Bezeichnungen:

- Die Stützvektormethode wird in den Tabellen mit „SVM“ abgekürzt, Naive Bayes mit „NB“ und k-Nächster-Nachbar mit „kNN“ (wobei k durch die entsprechende Anzahl einbezogener Nachbarn ersetzt wird).

- Die Werte der verschiedenen Gütekriterien sind Prozentangaben. Für die Gütekritierien gelten folgende Abkürzungen:
  - „R-EFS“ für Recall der Klasse EFS,
  - „R-Event“ für Recall der Klasse Event,
  - „P-EFS“ für Precision der Klasse EFS,
  - „P-Event“ für Precision der Klasse Event
- „Norm.“ steht für Normalisierung.

Die Ergebnisse der **ersten Experimentreihe** können Tabelle 6.1 entnommen werden.

Lerner	Norm.	R-EFS	R-Event	P-EFS	P-Event	Accuray
SVM	keine	85.71	63.89	80.60	71.88	**77.84** (+/- 8.60)
	0-1-N.	93.90	42.86	73.33	80.77	74.79 (+/- 3.22)
	z-score-N.	85.71	61.11	79.41	70.97	76.84 (+/- 7.32)
3NN	keine	86.59	53.06	75.53	70.27	**74.05** (+/- 6.14)
	0-1-N.	85.37	53.06	75.27	68.42	73.25 (+/- 5.01)
	z-score-N.	82.93	48.98	73.12	63.16	70.17 (+/- 7.22)
NB	keine	100	0.0	62.60	0.0	**62.62** (+/- 2.41)
	0-1-N.	96.34	4.08	62.07	40.00	61.85 (+/- 3.13)
	z-score-N.	100.00	0.00	62.60	0.00	62.56 (+/- 10.80)

Tabelle 6.1.: Erste Experimentreihe mit verschiedenen Normalisierungen.

Die besten Resultate erzielte bei diesen Lerndurchläufen eindeutig die SVM - ohne eine vorherige Normalisierung der Daten. Auch der 3NN-Lerner konnte bei diesen Experimenten ohne eine vorherige Normalisierung bessere Ergebnisse erzielen als mit einer Normalisierung. Der Naive Bayes Lerner lieferte im Vergleich zu den beiden anderen Methoden relativ schlechte Ergebnisse.

Bei den nächsten Experimenten wurde der Merkmal-Filter eingesetzt. Die Einstellungen waren ansonsten identisch mit denen der ersten Experimentreihe. Die Ergebnisse dieser **zweiten Experimentreihe** sind Tabelle 6.2 zu entnehmen.

Aus dieser Tabelle ist ersichtlich, dass der Einsatz des Merkmal-Filters zu einer Verbesserung der Ergebnisse führte. Diese Verbesserung unterstützt die biologische Annahme, dass die gefilterten Werte durch biologische und technische Störgrößen entstanden und somit für die Analyse unbrauchbar sind.
Wie schon in den ersten Experimenten zeigte auch in dieser Reihe die SVM die

Lerner	Norm.	R-EFS	R-Event	P-EFS	P-Event	Accuray
SVM	keine	87.80	73.47	84.71	78.26	**82.51** (+/- 7.69)
	0-1-N.	89.02	67.35	82.02	78.57	80.97 (+/- 7.10)
	z-score-N.	89.02	65.31	81.11	78.05	80.20 (+/- 6.86)
3NN	keine	85.37	57.14	76.92	70	74.84 (+/- 6.07)
	0-1-N.	81.71	61.22	77.91	66.67	73.99 (+/- 9.02)
	z-score-N.	86.59	57.14	77.17	71.79	**75.67** (+/- 8.88)
NB	keine	100	0.0	62.60	0.0	**62.62** (+/- 2.41)
	0-1-N.	95.12	6.12	62.90	42.86	61.77 (+/- 10.78)
	z-score-N.	100.00	0.00	62.60	0.00	62.56 (+/- 10.80)

Tabelle 6.2.: Zweite Experimentreihe mit Einsatz des Merkmal-Filters und verschiedener Normalisierungen.

besten Ergebnisse - ohne eine vorherige Normalisierung der Daten.
Zusammenfassend lässt sich festhalten, dass die Ergebnisse für die einzelnen Lernverfahren mit einer vorherigen Normalisierung der Daten sich nicht stark von den Ergebnissen ohne vorherige Normalisierung unterscheiden. Durch die z-Score-Normalisierung werden etwas bessere Ergebnisse erzielt als mit der Min-Max-Normalisierung. Ohne Normalisierung wurden im Durchschnitt etwas bessere Ergebnisse erreicht als mit einer Normalisierung.

6.3.1. Parameteroptimierung

Eine Veränderung der Parameterwerte der einzelnen Lernverfahren kann eine bedeutende Verbesserung der Lernergebnisse bewirken. Daher erfolgte in der **dritten Versuchsreihe** die Bestimmung der Parameterwerte für die verschiedenen Lernverfahren mittels einer Parameteroptimierung. Durch diese Optimierung sollte eine Verbesserung der Lernergebnisse erreicht werden.

- Für den Naive-Bayes-Lerner kann keine Parameteroptimierung vorgenommen werden, da diese Methode keine Parameter zur Variation anbietet.
- Bei der SVM wurde der Kostenparameter C optimiert. Der Wertebereich für die Optimierung des Parameters C wurde von 0-1000 beschränkt.
- Der zu optimierende Parameter beim kNN-Verfahren ist der Parameter k, welcher die Anzahl der zu betrachtenden Nachbarn angibt. Hierbei wurde der Wertebereich für k bei der Optimierung auf $1, 3, 5, 7$ und 9 eingeschränkt.

Die Ermittlung der Parameter erfolgte unter Verwendung des Merkmal-Filters und ohne vorherige Normalisierung der Daten. Für den Kostenparameter C der SVM wurde bei der Optimierung $C = 50$ ermittelt, für die Anzahl der Nachbarn

des kNN-Lerners wurde ein Wert von $k = 5$ ermittelt.
Die erzielten Ergebnisse der Lernverfahren bei Einstellung der ermittelten Parameterwerte nach 10-facher Kreuzvalidierung (10-fache Kreuzvalidierung bedingt zwar eine längere Rechenzeit, führt aber meistens zu aussagekräftigeren Ergebnissen) sind in Tabelle 6.3 dargestellt.

Lerner	Parameter	R-EFS	R-Event	P-EFS	P-Event	Accuray
SVM	C=50	87.80	77.55	86.75	79.17	83.96 (+/- 8.75)
5NN	k=5	87.80	55.10	76.60	72.97	75.89 (+/- 9.73)
NB		100.00	0.00	62.60	0.00	62.62 (+/- 2.41)

Tabelle 6.3.: Ergebnisse nach Parameteroptimierung

Zum einheitlichen Vergleich wurde auch das Naive-Bayes-Verfahren in einer 10-fachen Kreuzvalidierung validiert. Die Ergebnisse zeigen, dass durch die Optimierung der Parameter eine Steigerung der Lernergebnisse erreicht wird. Das beste Lernergebnis liefert wie auch in den vorherigen Experimenten die SVM.

Aufgrund der bisherigen Erkenntnisse erfolgen die weiteren Experimente ohne eine vorherige Normalisierung der Daten und unter Verwendung der ermittelten Parameterwerte für die entsprechenden Lernverfahren.

6.4. Experimente mit Merkmalsauswahl

6.4.1. Experimente zur Klassifikation

In Kapitel 5 wurden Methoden zur Merkmalsauswahl vorgestellt. Insbesondere wurde das biologische Interesse hinsichtlich einer möglichst kleinen Anzahl von relevanten Merkmalen erläutert. Für die Auswahl dieser relevanten Merkmale wurden Methoden eingesetzt, mit denen die Merkmale eine Bewertung erhalten, so dass ein Ranking erstellt werden und anhand dessen eine Auswahl von Merkmalen getroffen werden kann. Diese Merkmale können sowohl zur Konstruktion eines Klassifikators verwendet werden als auch zum Verständnis von genetischen Vorgängen dienen. Zur Selektion von Merkmalen wurden in den weiteren Experimenten folgende Methoden verwendet (siehe Kap. 5.3):

- SVM-Gewichtung
- Relief
- SAM

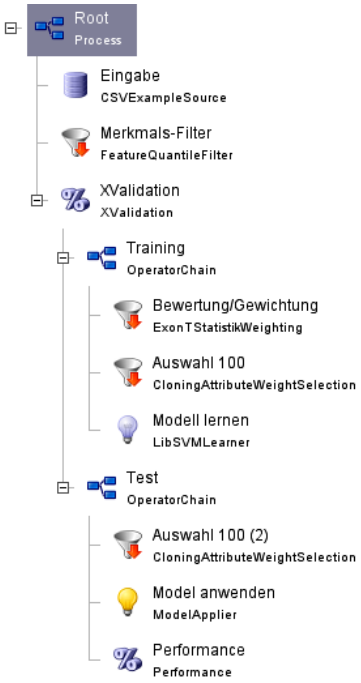


Abbildung 6.1.: Experimentaufbau Merkmalsauswahl

- Welch-Test
- t-Statistik

In den Experimenten wurden die Anzahl der auszuwählenden Merkmale zunächst auf 100 gesetzt. Diese Anzahl wurde gewählt, da aus biologischer Sicht wünschenswert ist, für weitere medizinische Tests und Analysen eine reduzierte Merkmalsmenge (ca. 80-200 Merkmale) zu erreichen.
Wird die Auswahl der Merkmale vor der Kreuzvalidierung vorgenommen, führt dies meist zu zu optimistischen Ergebnissen [55], da auf Beispielen validiert wird, die bereits zur Auswahl der Merkmale verwendet wurden. Trifft man die Auswahl der Merkmale innerhalb der Kreuzvalidierung, können aufgrund der unterschiedlichen Trainingsmengen unterschiedliche Merkmale ausgewählt werden. Bei den folgenden Experimenten wurde die Bewertung und die Auswahl der Merkmale sowie das Trainieren auf diesen Merkmalen in die Kreuzvalidierung eingebunden, um eine Überanpassung zu vermeiden. Der dabei verwendete Experiment-Aufbau ist vereinfacht in 6.1 abgebildet.
Um anschließend vergleichen zu können, ob und wie die ausgewählte Merkmalsmenge das Lernergebnis der Lerner beeinflusst, wurde jeder Lerner auf 100 zufällig ausgewählten Merkmalen trainiert. Die Ergebnisse sind in Tabelle 6.4 aufgeführt.

Lerner	R-EFS	R-Event	P-EFS	P-Event	Accuray
SVM	73.17	61.22	75.95	57.69	68.69 (+/- 8.60)
5NN	82.93	53.06	74.73	65.00	71.74 (+/- 2.08)
NB	76.83	32.83	65.62	45.71	60.28 (+/- 5.33)

Tabelle 6.4.: Ergebnisse mit zufälliger Merkmalsauswahl

Die Ergebnisse der Lerner auf die ersten 100 ausgewählten Merkmale bzgl. der Auswahlmethoden zeigt Tabelle 6.5.

Lerner	Methode	R-EFS	R-Event	P-EFS	P-Event	Accuray
SVM	SVM-Ge.	85.37	51.02	74.47	67.57	**72.48** (+/- 7.57)
	Relief	81.71	59.18	77.01	65.91	57.26 (+/- 20.11)
	SAM	73.17	67.33	78.95	60.00	70.94 (+/- 5.95)
	Welch-Test	78.05	42.86	69.57	53.85	64.81 (+/- 10.32)
	t-Statistik	74.39	46.94	70.11	52.27	64.13 (+/- 10.16)
5NN	SVM-Ge.	90.24	55.10	77.08	77.14	**77.09** (+/- 4.88)
	Relief	86.59	59.18	78.02	72.50	76.30 (+/- 3.95)
	SAM	73.17	46.94	69.77	51.11	63.30 (+/- 10.06)
	Welch-Test	98.78	4.08	63.28	66.67	63.33 (+/- 11.63)
	t-Statistik	87.80	34.69	69.23	62.96	67.89 (+/- 7.69)
NB	SVM-Ge.	82.93	61.22	78.16	68.18	74.79 (+/- 6.77)
	Relief	81.71	71.43	82.72	70.00	**77.86** (+/- 7.45)
	SAM	86.59	30.61	67.62	57.69	65.58 (+/- 13.22)
	Welch-Test	76.83	30.61	64.95	44.12	59.52 (+/- 11.63)
	t-Statistik	78.05	18.37	61.54	33.33	55.70 (+/- 10.85)

Tabelle 6.5.: Experimentreihe mit Einfacher-Merkmalsauswahl

Die Ergebnisse der Experimente variieren im Vergleich zu den Experimenten ohne eine vorherige Merkmalsauswahl. Je nach Wahl der Methode werden bessere oder schlechtere Ergebnisse erreicht. Auch im Vergleich zu den Ergebnissen mit einer zufälligen Merkmalsauswahl sind die Ergebnisse sehr wechselhaft. SAM, SVM-Gewichtung und Relief liefern im Schnitt bessere Ergebnisse im Vergleich zur zufälligen Auswahl, die Auswahl des Welch-Tests und der t-Statistik sind schlechter als die zufällige Auswahl. Die SVM erreicht auf keiner der ausgewählten Merkmalsmengen ihr Ergebnis ohne eine vorherige Auswahl. Der Naive-Bayes-Lerner und 5-Nächster-Nachbar können je nach eingesetzter Methode etwas bessere Ergebnisse als ohne eine Merkmalsauswahl erzielen.
Die wechselhaften Ergebnisse können an durch die Methoden schlecht ausgewählten (bewerteten) Merkmalen liegen oder die Anzahl der ausgewählten Merkmale von 100 ist zu gering oder evtl. auch zu hoch. Evtl. deuten die Ergebnisse auch

darauf hin, dass eine Auswahl von bestimmten Merkmalen nicht sinnvoll ist und stattdessen alle Merkmale für eine gute Vorhersage benötigt werden.
Im Anschluss an diese Experimentreihe wurden dieselben Experimente mit einem Ensemble (siehe Kap. 5.5) durchgeführt, um zu überprüfen, ob eine stabilere Auswahl der Merkmale zu besseren Ergebnissen führt. Die Anzahl der „Runden" innerhalb des Ensembles wurde auf 10 gesetzt und die Anzahl der zu betrachtenden Merkmale pro Ranking auf 100 (es wurden also 10 Rankings erstellt, wobei die Merkmale der ersten 100 Ränge pro Ranking beim Mehrheitsvoting des Gesamt-Rankings berücksichtigt wurden). Die ersten 100 Merkmale des Ensembles wurden für die Lernläufe verwendet. Der Experimentaufbau ist Abbildung 6.2 zu entnehmen. Tabelle 6.6 listet die erreichten Ergebnisse auf den ausgewählten Merkmalen auf.

Lerner	**Methode**	**R-EFS**	**R-Event**	**P-EFS**	**P-Event**	**Accuray**
SVM	SVM-Ge.	82.93	51.02	73.91	64.10	70.97 (+/- 4.73)
	Relief	80.49	48.98	72.53	60.00	68.69 (+/- 5.16)
	SAM	78.05	69.39	81.01	65.38	**74.87** (+/- 4.77)
	Welch-Test	81.71	59.18	77.01	65.91	73.25 (+/- 6.98)
	t-Statistik	79.27	55.10	74.71	61.36	70.17 (+/- 7.62)
5NN	SVM-Ge.	87.80	59.18	78.26	74.36	**77.09** (+/- 5.45)
	Relief	85.37	63.27	79.55	72.09	**77.09** (+/- 2.46)
	SAM	82.93	53.06	74.73	65.00	71.79 (+/- 6.93)
	Welch-Test	80.49	59.18	76.74	64.44	72.48 (+/- 5.80)
	t-Statistik	82.93	63.27	79.07	68.89	75.50 (+/- 9.09)
NB	SVM-Ge.	82.93	63.27	79.27	68.89	75.56 (+/- 7.58)
	Relief	81.71	71.43	82.72	70.00	77.86 (+/- 7.45)
	SAM	84.15	69.39	82.14	72.34	**78.66** (+/- 9.19)
	Welch-Test	80.49	71.43	82.50	68.63	77.07 (+/- 7.77)
	t-Statistik	80.49	71.43	82.50	68.63	77.07 (+/- 10.09)

Tabelle 6.6.: Experimentreihe mit Ensemble-Merkmalsauswahl

Durch die Ensemble-Merkmalsauswahl verbessern sich die Lernergebnisse für alle Lerner bis auf eine Ausnahme (SVM mit SVM-Gewichtung) im Vergleich zu der einfachen Merkmalsauswahl (siehe Tabelle 6.5). Auch sind alle Ergebnisse besser als die Ergebnisse der zufälligen Auswahl von Merkmalen (siehe Tabelle 6.4). Die Ergebnisse für eine ausgewählte Merkmalsmenge sind für den Naive-Bayes-Lerner immer besser als ohne eine vorherige Auswahl. Auch der Nächste-Nachbar-Lerner übertrifft mit einer ausgewählten Merkmalsmenge überwiegend seine Ergebnisse ohne Merkmalsauswahl. Die Ergebnisse der SVM konnten durch eine geringere Anzahl von Merkmalen die Ergebnisse ohne vorherige Auswahl nicht übertreffen.

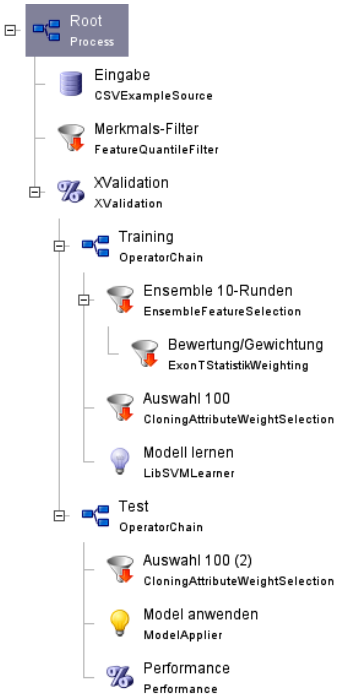


Abbildung 6.2.: Experimentaufbau Ensemble-Merkmalsauswahl

Lerner	Methode	R-EFS	R-Event	P-EFS	P-Event	Accuray
SVM	SVM-Ge.	86.59	71.43	83.53	76.09	80.99 (+/- 7.58)
5NN	SVM-Ge.	89.02	65.31	81.11	78.05	80.11 (+/- 9.29)
NB	SVM-Ge.	81.71	71.43	82.72	70.00	77.86 (+/- 8.03)

Tabelle 6.7.: Einfache-Merkmalsauwahl 10-fach kreuzvalidiert

Die Ergebnisse der Experimente ohne Merkmalsauswahl und optimierten Parametern (siehe Tabelle 6.3) wurden 10-fachen kreuzvalidiert. Die Ergebnisse mit Merkmalsauswahl dagegen sind nur 5-fach kreuzvalidiert. Daher wurden die Experimente zum besseren Vergleich mit Merkmalsauswahl exemplarisch anhand der SVM-Gewichtung auch 10-fach kreuzvalidiert. Tabelle 6.7 fasst die Ergebnisse der Merkmalsauawahl ohne Ensemble nach 10-facher Kreuzvalidierung zusammen.

Die ausgewählten Experimente wurden zum Vergleich ebenso mit der Ensemble-Merkmalsauswahl gestartet. Tabelle 6.8 enthält die Ergebnisse.

Die Experimente zeigen, dass durch die 10-fache Kreuzvalidierung eine höhere Accuracy erzielt wird. Eine Erklärung dafür könnte sein, dass bei einer 10-fachen Kreuzvalidierung mehr Trainingsbeispiele pro Durchgang verwendet werden (bei der 5-fachen Kreuzvalidierung wird pro Durchgang auf ca. 105 Beispielen gelernt,

Lerner	Methode	R-EFS	R-Event	P-EFS	P-Event	Accuray
SVM	SVM-Ge.	84.15	59.18	77.53	69.05	74.84 (+/- 8.35)
5NN	SVM-Ge.	87.80	63.27	80.00	75.61	78.63 (+/- 8.27)
NB	SVM-GE.	82.93	71.43	82.93	71.43	78.68 (+/- 10.03)

Tabelle 6.8.: Ensemble-Merkmalsauwahl 10-fach kreuzvalidiert

bei der 10-fachen auf ca. 118 Beispielen), bei mehr Trainingsbeispielen können die Lernverfahren besser lernen.

Kombination von Ensembles

Ensembles wurden in den bisherigen Experimenten dazu eingesetzt, ein Gesamt-Ranking der Merkmalsmengen auf Basis einer einzelnen Auswahlmethode zu erstellen. Nun wurde ein Ensemble verwendet, um die Gesamt-Rankings verschiedener Ensembles zu einem „(Gesamt-)Gesamt-Ranking“ zu kombinieren.
Innerhalb eines Ensembles wurden die Merkmale pro Runde mit einer Methode (z.B. Welch-Test, Relief, etc.) bewertet und daraus ein Ranking erstellt; anhand dieser Rankings wurde ein Gesamt-Ranking durch einen Mehrheitsentscheid generiert. Diese Gesamt-Rankings wurden dann kombiniert, indem die einzelnen Werte für ein Merkmal bei allen Gesamt-Rankings verschiedener Ensembles aufsummiert wurden. Auf den Merkmalen der ersten 100 Ränge des „(Gesamt-)Gesamt-Rankings“ wurden dann die Lernverfahren trainiert. Da sich der Welch-Test und die Methode der t-Statistik in ihrer Bewertung sehr ähnlich sind, wurde die t-Statistik bei diesen Experimenten nicht miteinbezogen, da durch die Ähnlichkeit der zwei Methoden diese einen zu starken Einfluss auf das Gesamt-Ranking hätten. Tabelle 6.9 enthält die Ergebnisse der Experimente.

Lerner	R-EFS	R-Event	P-EFS	P-Event	Accuray
SVM	78.05	48.98	71.91	57.14	67.15 (+/- 8.00)
5NN	84.15	59.18	77.53	69.05	74.84 (+/- 6.54)
NB	80.49	69.39	81.48	68.00	76.38 (+/- 10.29)

Tabelle 6.9.: Kombination von Ensembles

Durch die Kombination der Rankings konnte keine Verbesserung der Lernergebnisse erreicht werden. Alle Verfahren konnten mit einer einzelnen Auswahlmethode bessere Ergebnisse erzielen.

Bei diesen Experimenten mit Merkmalsauswahl wurde die Anzahl der ausgewählten Merkmale bisher konstant auf 100 Merkmale gesetzt. Zur Überprüfung, wie

sich die Lernerfolge der Lerner auf eine höhere oder geringere Anzahl ausgewählter Merkmale verändern, wurde eine Experimentreihe durchgeführt, in der die Anzahl der ausgewählten Merkmale in 10er-Schritten von 10 bis auf 200 Merkmalen erhöht wurde. Diese Experimente wurden sowohl für die Einfache- als auch für die Ensemble-Version durchgeführt. Als Bewertungsmethode wurde die SVM-Gewichtung herangezogen. Die Resultate dieser Experimente sind in den Grafiken 6.3 - 6.8 dargestellt.

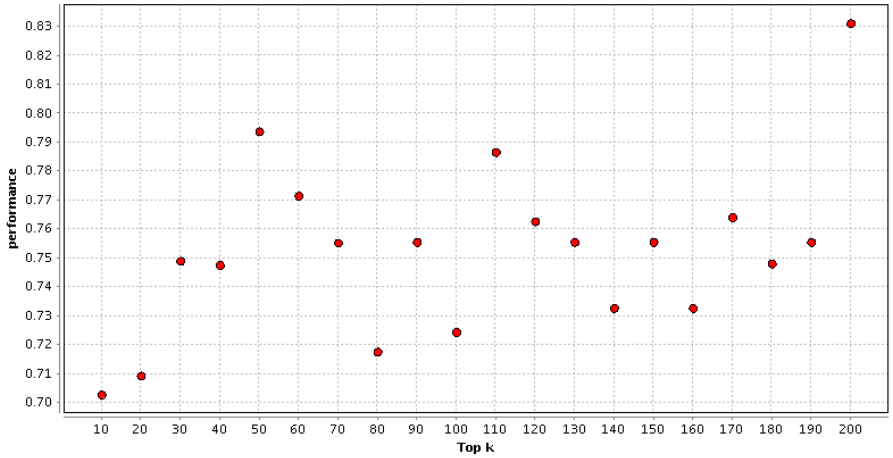


Abbildung 6.3.: Erreichte Accuracy bei einer Merkmalsmenge (Einfache-Auswahl) mit k Merkmalen für die SVM

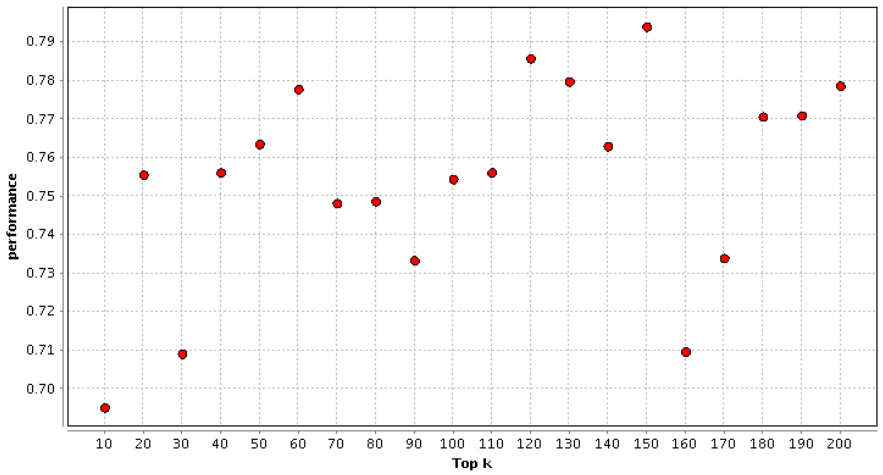


Abbildung 6.4.: Erreichte Accuracy bei einer Merkmalsmenge (Ensemble-Auswahl) mit k Merkmalen für die SVM

Tabelle 6.10 fasst die besten Ergebnisse und die dafür benötigte Anzahl ausgewählter Merkmale für jedes eingesetzte Lernverfahren zusammen.

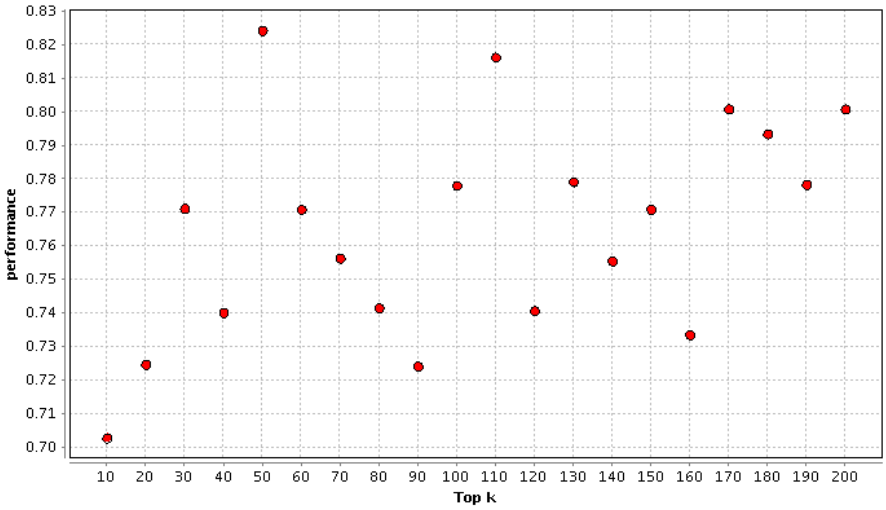


Abbildung 6.5.: Erreichte Accuracy bei einer Merkmalsmenge (Einfache-Auswahl) mit k Merkmalen für 5NN

Lerner	Auswahl	Methode	k	Accuray
SVM	Einfache	SVM-Ge.	200	83.11 (+/- 8.68)
SVM	Ensemble	SVM-Ge.	150	79.40 (+/- 5.70)
5NN	Einfache	SVM-Ge.	50	82.42 (+/- 3.17)
5NN	Ensemble	SVM-Ge.	150	83.22 (+/- 4.55)
NB	Einfache	SVM-Ge.	170	80.17 (+/- 7.78)
NB	Ensemble	SVM-Ge.	40	80.23 (+/- 5.76)

Tabelle 6.10.: Ergebnisse für die Auswahl von k Merkmalen

Aus den Grafiken 6.3 - 6.8 ist zu entnehmen, dass die Ergebnisse für die Accuracy für eine variierende Anzahl (k) an Merkmalen stark schwanken und es daher schwierig ist, eine Aussage darüber zu treffen, mit welcher oder ab welcher Anzahl an Merkmalen ein gutes Ergebnis erreicht werden kann oder ab wann sich eine Stabilität der Ergebnisse einstellt. Die Schwankungen können auf mehrere Faktoren zurückgeführt werden: auf die Kreuzvalidierung (es wurde 5-fach kreuzvalidiert), auf Inkonsistenzen innerhalb der Daten (herstellungsbedingte Faktoren), beim Einsatz des Ensembles auf die interne Aufteilung und auf die geringe Anzahl an Beispielen, die zum Lernen verwendet werden konnten.
Allen Lernern ist gemein, dass sie für eine kleine Anzahl von ausgewählten Merkmalen schlechtere Ergebnisse liefern. Erst ab einem k von ca. 20-30 (Merkmalen) werden die Ergebnisse besser.

Mit den gewonnenen Erkenntnissen kann die zu überprüfende Hypothese aus

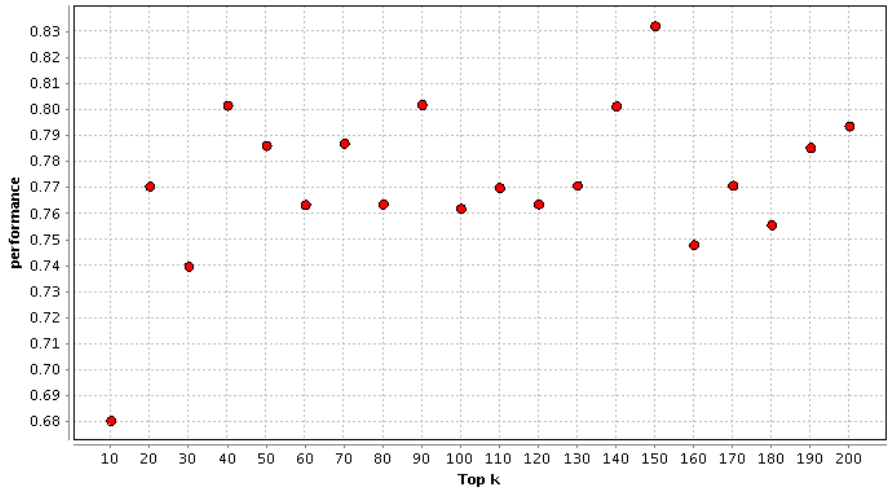


Abbildung 6.6.: Erreichte Accuracy bei einer Merkmalsmenge (Ensemble-Auswahl) mit k Merkmalen für 5NN

der Einführung - dass nicht ein bestimmtes Exon die Klassifikation determiniert, sondern das Zusammenspiel einer Anzahl von Exons für die Zugehörigkeit zu einer Klasse verantwortlich ist - insoweit bestätigt werden, dass durch eine geringe Anzahl von Merkmalen schlechtere Klassifikationsergebnisse erzielt werden. Dieses spricht dafür, dass mehrere Exons (Merkmale) an der Ausprägung der unterschiedlichen Krankheitsverläufe beteiligt sind.

Die Ergebnisse der SVM deuten daraufhin, dass dieses Verfahren für die vorliegenden Daten bessere Ergebnisse erzielt, wenn eine größere Anzahl von Merkmalen zum Lernen eingesetzt wird. Abbildung 6.4 zeigt, dass die SVM mit der einfachen Merkmalsauswahl bei $k = 200$ eine deutliche Steigerung des Lernerfolgs erreicht. Ob eine weitere Erhöhung von k sich durch eine Steigerung des Lernerfolgs auszeichnet, wurde überprüft, indem in einem weiteren Experiment die Anzahl der Merkmale schrittweise von 100 auf 5000 Merkmale gesteigert wurde. Zur Merkmalsauswahl wurde die SVM-Gewichtung benutzt. Die Ergebnisse sind grafisch in 6.9 abgebildet. Die SVM zeigt bessere Ergebnisse auf einer größeren Menge von Merkmalen. Mit 1000 Merkmalen konnte eine Acurracy von 83.19% erreicht werden. Die Ergebnisse sind jedoch sehr schwankend, so dass nicht unbedingt die Accuracy mit der Anzahl an Merkmalen steigt.

Bei der Ensemble-Version wurde pro Runde intern ein Ranking erstellt und abschließend aus den verschiedenen Rankings ein Gesamt-Ranking mittels eines Mehrheitsentscheids erstellt.
Eine weitere Möglichkeit besteht darin, zu fordern, dass nur solche Merkmale in

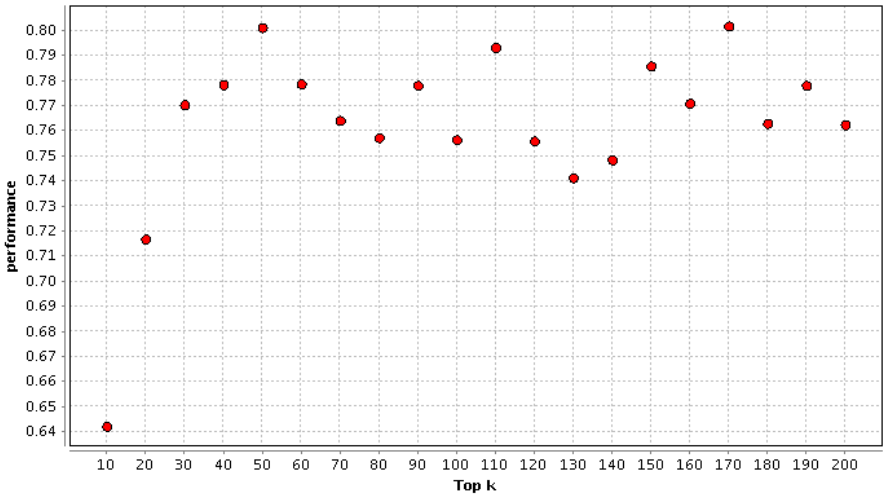


Abbildung 6.7.: Erreichte Accuracy bei einer Merkmalsmenge (Einfache-Auswahl) mit k Merkmalen für NB

das Gesamt-Ranking einbezogen werden, die in mindesten p Prozent der einzelnen Rankings enthalten sind. Insbesondere aus biologischer Sicht ist man an Merkmalen interessiert, die möglichst oft in den einzelnen Rankings enthalten sind. Um solche Merkmale zu entdecken, wurde zunächst für jede Bewertungsfunktion überprüft, ob überhaupt Merkmale existieren, die in jedem einzelnen Ranking enthalten sind. Tabelle 6.11 zeigt die Anzahl der Merkmale, die in jedem internen Ranking des Ensembles unter den ersten 100 Rängen enthalten sind. Als Grundlage für die Auswahl dient der komplette Datensatz.

Methode	**Anzahl Runden**	**Anzahl Merkmale**
SVM-Gewichtung	10	11
Relief	10	27
SAM	10	57
Welch	10	33
t-Statistik	10	32

Tabelle 6.11.: Anzahl immer enthaltener Merkmale

In den folgenden Experimenten wurden nur die Merkmale zum Lernen berücksichtigt, die in jeder Runde des Ensembles unter die ersten 100 Ränge gewählt wurden. Wie bereits erläutert, können die Ergebnisse zu optimistisch ausfallen, wenn die Auswahl der Merkmale außerhalb der Kreuzvalidierung erfolgt, da die Testbeispiele dann schon bei der Auswahl der Merkmale mitgewirkt haben. Wird die Auswahl der Merkmale in die Kreuzvalidierung eingeschlossen, kann

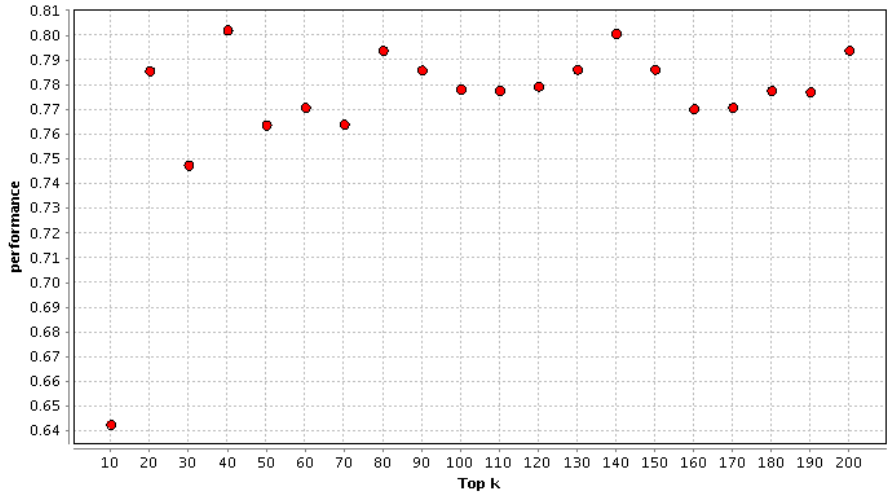


Abbildung 6.8.: Erreichte Accuracy bei einer Merkmalsmenge (Ensemble-Auswahl) mit k Merkmalen für NB

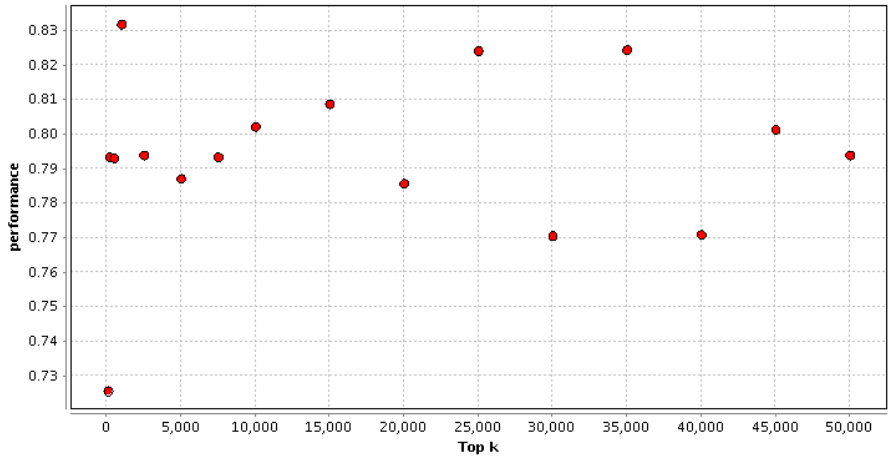


Abbildung 6.9.: Erreichte Accuracy der SVM auf k ausgewählten Merkmalen

die Anzahl der ausgewählten Merkmale variieren und es können pro Ensemble unterschiedliche Merkmale in der Auswahl vorhandenen sein. Beide Varianten wurden getestet. Da SAM die meisten „immer enthaltenen“ Merkmale liefert und die SVM-Gewichtung die geringste Anzahl dieser Merkmale, die in jeder Runde enthalten sind, wurden diese beiden Methoden zur Merkmalsbewertung und anschließender Auswahl bei den nachfolgenden Experimenten eingesetzt. In Tabelle 6.12 sind die Ergebnisse mit Merkmalsauswahl innerhalb der Kreuzvalidierung und in Tabelle 6.13 mit Merkmalsauswahl außerhalb der Kreuzvalidierung aufgezeigt.

Lerner	Methode	R-EFS	R-Event	P-EFS	P-Event	Accuray
SVM	SAM	76.83	59.18	75.90	60.42	70.26 (+/- 3.48)
5NN	SAM	81.71	57.14	76.14	65.12	72.54 (+/- 6.53)
NB	SAM	86.59	73.47	84.52	76.60	81.74 (+/- 7.64)
SVM	SVM-Ge	78.05	48.98	71.91	57.14	67.15 (+/- 8.00)
5NN	SVM-Ge.	81.71	57.14	76.14	65.12	72.56 (+/- 5.35)
NB	SVM-Ge.	80.49	67.35	80.49	67.35	75.58 (+/- 8.24)

Tabelle 6.12.: Ergebnisse Merkmalsauswahl in jeder Runde innerhalb Kreuzvalidierung

Lerner	Methode	R-EFS	R-Event	P-EFS	P-Event	Accuray
SVM	SAM	76.17	59.18	75.90	60.42	70.17 (+/- 7.22)
5NN	SAM	82.93	65.31	80.00	69.57	76.27 (+/- 5.83)
NB	SAM	85.37	73.47	84.34	75.00	80.88 (+/- 5.52)
SVM	SVM-Ge.	81.71	57.14	76.14	65.12	72.48 (+/- 6.74)
5NN	SVM-Ge.	86.59	71.43	83.53	76.09	80.91 (+/- 7.30)
NB	SVM-Ge.	90.24	81.63	89.16	83.33	87.04 (+/- 4.56)

Tabelle 6.13.: Ergebnisse Merkmalsauswahl in jeder Runde außerhalb der Kreuzvalidierung

Durch die SVM-Gewichtung wurden bei den Experimenten mit Merkmalsauswahl innerhalb der Kreuzvalidierung ca. 10 Merkmale ausgewählt und zum Lernen verwendet. Für diese ausgewählten Merkmale zeigen 5NN und SVM etwas schlechtere Ergebnisse als auf 100 ausgewählte Merkmale (siehe Tabelle 6.6). Im Vergleich zu den Ergebnissen aus den Experimenten ohne Merkmalsauswahl (siehe Tabelle 6.3) schneidet besonders das Ergebnis der SVM schlecht ab (Abweichung von 16.81%). Naive Bayes zeigt für die ca. 10 Merkmale der SVM-Gewichtung ungefähr die gleiche Accuracy wie auf 100 ausgewählte Merkmale. Aus den Ergebnissen kann gefolgert werden, dass die durchschnittlich 10 ausgewählten Merkmale zwar im gewissen Maß eine Trennung der Patientenprofile zulassen, aber durch Hinzunahme weiterer Merkmale eine Verbesserung möglich ist.
Biologisch betrachtet, können die ausgewählten Merkmale von Interesse sein; zwar lassen sich die Patienten anhand dieser Merkmale von den Lernverfahren nicht zu 100% richtig klassifizieren, dennoch ist eine Zuteilung der Patienten zu den Klassen in einem gewissen Maß möglich. So könnten die Ausprägungen dieser Merkmale evtl. auf eine genetische Grundlage zurückgeführt werden. Bei SAM wurden ca. 50 Merkmale „immer“ ausgewählt und so zum Lernen verwendet. Die auf diesen Merkmalen erreichten Ergebnisse sind im Vergleich zu den bisher erreichten Ergebnissen meist besser. So übertreffen Naive Bayes und kNN ihre Ergebnisse der Auswahl von 100 Merkmalen (Naive Bayes zeigt eine Steigerung

von 16,16%, kNN von 9,24%). Das Ergebnis der SVM ist wesentlich schlechter als bei der Auswahl von 100 Merkmale und im Vergleich ohne Merkmalsauswahl. Auch diese Merkmale sollten biologisch genauer betrachtet werden.

6.4.2. Experimente zur Stabilität/Robustheit

In dem folgenden Abschnitt wird die Stabilität (siehe Kap. 5.4) der Auswahl der Merkmalsmengen betrachtet. Die durchgeführten Experimente wurden sowohl für die Merkmalsauswahl mit der Ensemble-Version als auch für die Einfache-Auswahl durchgeführt und miteinander verglichen.
Um eine Aussage bzgl. der Stabilität einer ausgewählten Merkmalsmenge treffen zu können, wurde die Beispielmenge des Neuroblastom-Datensatzes in 10 Teilmengen möglichst gleicher Größe aufgeteilt. In 10 Durchgängen wurde auf 9 von diesen 10 Teilmengen (abwechselnd wird eine Teilmenge nicht betrachtet) eine Bewertung/Gewichtung der Merkmale unter Hinzunahme einer Bewertungsmethode (siehe Kap. 5.3) vorgenommen und daraus ein Ranking erstellt. Die Merkmale der ersten k Ränge wurden als selektierte Merkmale betrachtet. Die verschiedenen selektierten Mengen von Merkmalen wurden mittels des Jaccard Koeffizienten (siehe Kap. 5.4.1, Formel 5.11) auf ihre Ähnlichkeit überprüft und die Gesamtstabilität wurden mit der Formel 5.12 (siehe Kap. 5.4.2) gemessen. Bei der Ensemble-Version wurde somit auf 9 von 10 Teilmengen ein Gesamt-Ranking (durch 10 intern ausgewählte Merkmalsmengen) erstellt. Die nach den 10 Durchläufen auf Basis der Gesamt-Rankings ausgewählten 10 Merkmalsmengen wurden dann auf ihre Ähnlichkeit verglichen und eine Gesamtstabilität berechnet.
Die Ergebnisse aus Tabelle 6.11, welche die Anzahl der Merkmale angibt, die in jeder Runde unter den besten 100 Merkmalen sind, lässt vermuten, dass kein Auswahlverfahren eine Stabilität von 1 (komplett identisch) aufweist.

Der Parameter k für die Anzahl der pro Durchlauf berücksichtigen Ränge wurde in einer ersten Experimentreihe von $k = 100$ in einer weiteren auf $k = 200$ erhöht. Zusätzlich wurde untersucht, ob eine Vorab-Normalisierung der Eingabedaten einen Einfluss auf das Ergebnis der Robustheit hat. Dazu wurden die Experimente einmal mit den absoluten Werten und einmal mit normalisierten Werten durchgeführt. Zur Normalisierung wurden die in Abschnitt 6.1 vorgestellte Min-Max-Normalisierung und z-Score-Normalisierung verwendet. Tabelle 6.14 zeigt die Ergebnisse der Experimente. Es lässt sich feststellen, dass die Art der Normalisierung nur einen sehr geringen Einfluss auf die Ergebnisse ausübt. Diese ist für die Methoden, die jedes Merkmal einzeln gewichten (wie SAM und Welch-Test), nicht verwunderlich. Auf eine Normalisierung kann wie auch schon bei den Experimenten zur Klassifikation somit verzichtet werden. Weiter ist zu beobach-

k	Norm.	t-Statistik		Welch-Test		SAM		Relief		SVM-Ge.	
		Ein.	**Ens.**	**Ein.**	**Ens.**	**Ein.**	**Ens.**	**Ein.**	**Ens.**	**Ein.**	**Ens.**
100	keine	0.459	0.450	0.445	0.467	0.644	0.651	0.489	0.504	0.306	0.325
200	keine	0.479	0.479	0.487	0.482	0.700	0.703	0.513	0.524	0.298	0.324
100	0-1-N.	0.449	0.459	0.455	0.467	0.626	0.629	0.490	0.515	0.288	0.308
200	0-1-N.	0.479	0.479	0.487	0.482	0.647	0.649	0.530	0.554	0.318	0.337
100	z-score-N.	0.459	0.450	0.455	0.467	0.404	0.444	0.497	0.514	0.305	0.325
200	z-score-N.	0.479	0.479	0.487	0.482	0.418	0.448	0.536	0.559	0.299	0.327

Tabelle 6.14.: Robustheit verschiedener Merkmalsauswahl-Methoden für die Einfache und die Ensemble-Auswahl. Bei der t-Statistik wurde das Ranking anhand der berechneten t-Werte vorgenommen. Beim Welch-Test erfolgte das Ranking anhand der p-Werte.

ten, dass durch ein Ensemble die Stabilität der ausgewählten Merkmalsmenge bei allen Methoden zunimmt und dass durch $k = 200$ nur eine sehr geringe Verbesserung der Stabilität erreicht wird. Diese Verbesserung fällt so minimal aus, dass man die Stabilität für $k = 100$ und $k = 200$ als (nahezu) gleich bezeichnen kann. Die Stabititätswerte der von der t-Statistik und von dem Welch-Test ausgewählten Mengen sind beinahe identisch. Dass diese Verfahren keine besseren Werte erzielen, kann darauf zurückgeführt werden, dass beide bei der Bewertung der Merkmale (Exons) den Mittelwert und die geschätzten Varianzen verwenden. Ein hoher t-Wert kann durch eine geringe Varianz hervorgerufen werden. Weisen Merkmale durch Zufall innerhalb einer Teilmenge eine geringe Varianz auf (was bei der hohen Anzahl an Merkmalen und der geringen Anzahl an Beispielen zu vermuten ist) würden diese Merkmale fälschlich als signifikant betrachtet werden. In einer anderen Teilmenge könnte die Varianz für dieses Merkmal größer sein und somit würde der t-Wert geringer ausfallen und das Merkmal aus diesem Grund nicht unter den ersten 100 Rängen zu finden sein.
Mit SAM werden die stabilsten Merkmalsmengen erzeugt. SAM addiert zum Ausgleich kleiner Varianzen eine Konstante zum Nenner (siehe Kap. 5.3.2) worauf das gute Abschneiden zurückgeführt werden kann. Die SVM-Gewichtung schneidet hinsichtlich der Stabilität der Merkmalsmengen deutlich am schlechtesten ab. Dies könnte darauf zurückgeführt werden, dass sich mehrere Hypothesen finden lassen, die eine gute Separation zulassen.
Grundsätzlich kann davon ausgegangen werden, dass zuverlässigere Aussagen bzgl. der Stabilität/Robustheit gemacht werden können, wenn eine höhere Anzahl an Beispielen (Patienten) vorliegen würde.

6.5. Zusammenfassung

In den durchgeführten Experimenten zur Klassifikation und zur Robustheitsbewertung wurde vorrangig überprüft, ob eine vorherige Merkmalsauswahl durch verschiedene Methoden zu einer Verbesserung der Lernergebnisse führt und wie stabil die ausgewählten Merkmalsmengen sind. Für den Naive-Bayes-Lerner und den kNN-Lerner konnten durch die Auswahl bestimmter Merkmale verbesserte Ergebnisse im Vergleich zu den Ergebnissen ohne vorherige Auswahl erreicht werden. Die SVM liefert die besten Ergebnisse ohne vorherige Selektion von bestimmten Merkmalen. Dies zeigt, dass die SVM auch mit einer hohen Anzahl von Merkmalen zurecht kommt und so ein Arbeiten in hohen Dimensionen ermöglicht (siehe [33]).
Ohne eine Merkmalsauswahl weisen die Ergebnisse der Lernverfahren teilweise starke Abweichungen auf. So erreicht die SVM eine Accuracy von 83.96%, wohingegen Naive Bayes nur eine Accuracy von 62.53% erreicht. Die Ergebnisse der Lernverfahren auf die ausgewählten Merkmalsmengen sind abhängig davon,

wie die Merkmalsauswahl erfolgt. Bei einer einfachen Merkmalsauswahl zeigen die Ergebnisse der Lernverfahren starke Schwankungen, auf eine einfache Merkmalsauswahl durch Relief liefert die SVM z.B. eine Accuracy von nur 57,26%, Naive Bayes hingegegen eine Accuracy von 77,86%. Bei einer Merkmalsauswahl mittels eines Ensembles weichen die erlangten Resultate der Lernverfahren nicht sonderlich stark voneinander ab, die SVM schneidet hierbei etwas schlechter ab als Naive Bayes und kNN (siehe Kap.6, Tabelle 6.6). Die Wahl des Lernverfahrens für ausgewählte Merkmalsmengen des Ensembles hat somit keinen allzu großen Einfluss auf die Güte des Ergebnisses. Für alle Lernverfahren ist es etwa gleich schwierig, die Beispiele (Patienten) anhand der gegebenen Merkmale zu trennen. Die erreichte Accuracy der Lernverfahren auf den Patientendaten kann man als „akzeptabel“ bezeichnen. Aus medizinischer Sicht reichen diese Ergebnisse aber nicht, um neue Patienten anhand ihrer Expressionsprofile so exakt einer Klasse zuzuordnen, dass man für jeden Patienten die richtige Therapie anhand der Vorhersage einleiten könnte. Für beide Klassen werden in den Experimenten Fehlvorhersagen getroffen; Fehlvorhersagen können sich auf beide Patientengruppen sehr negativ auswirken. Wird einem Patienten fälschlicherweise eine gute Prognose in Aussicht gestellt, so erhält dieser Patient keine oder eine falsche Therapie. Ebenso gravierend kann der komplementäre Fall sein: einem Patienten wird fälschlicherweise ein ungünstiger Krankheitsverlauf (Rezidiv) prognostiziert. Der Patient erhält daraufhin eine Therapie, deren Nebenwirkungen gravierend sein können (im schlimmsten Fall verstirbt der Patient an den Nebenwirkungen der Therapie).
Bei den Experimenten zur Robustheit zeigte sich, dass durch ein Ensemble eine stabilere Auswahl an Merkmalen erreicht werden konnte als durch eine einfache Merkmalsauswahl (siehe Tabelle 6.14). In den Experimenten mit solch einer stabileren Auswahl von Merkmalen (siehe Tabelle 6.6) konnten bessere Lernergebnisse errungen werden als bei einer einfachen Auswahl (siehe Tabelle 6.5). Mit SAM konnte die höchste Robustheit erreicht werden, die SVM hingegen zeigte die geringste Robustheit.

7. Auswertung und Analyse

In diesem Kapitel werden zunächst die ausgewählten Merkmalsmengen genauer betrachtet, um zu analysieren, ob es Übereinstimmungen zwischen den ausgewählten Merkmalsmengen verschiedener Methoden gibt und ob innerhalb der Merkmalsmengen Korrelationen zwischen den Merkmalen (Exons) bestehen.

Anschließend werden Ähnlichkeiten zur Textklassifikation untersucht. Textklassifikationsaufgaben zeichnen sich durch spezielle statistische Eigenschaften aus. Ein aus der Textklassifikation bekanntes statistisches Lernmodell, welches auf diesen Eigenschaften beruht, ist das von Joachims [34] entwickelte *TCat-Modell*. Da die vorliegenden Daten der Neuroblastom-Patienten einige dieser statistischen Eigenschaften aufzeigen, wird der Versuch unternommen, die Daten als TCat-Konzept zu modellieren, um ggfs. Aussagen, dieses Modell für die Textklassifikation für die vorliegenden Daten zu übernehmen.

7.1. Analyse der ausgewählten Merkmalsmengen

Zunächst wurden die von den untersachiedlichen Methoden ausgewählten Merkmalsmengen auf Übereinstimmungen untersucht. Anschließend wurde untersucht, ob innerhalb einer Merkmalsmenge Korrelationen zwischen den Merkmalen vorhanden sind.

7.1.1. Übereinstimmungen ausgewählter Merkmalsmengen

In Kapitel 5.3 wurden verschiedene Methoden zur Bewertung von Merkmalen eingeführt, auf deren Basis eine Merkmalsuswahl erfolgen kann. In den Experimenten zur Robustheit (siehe Kap. 6.4.2) konnte gezeigt werden, dass durch ein Ensemble eine stabilere Merkmalsmenge selektiert wurde. In den durchgeführten Experimenten (siehe Kapitel 6.4.1) wurden die mit dem Ensemble ausgewählten Merkmale zum Lernen eingesetzt. Die Ergebnisse der einzelnen Lernverfahren auf diesen ausgewählten Merkmalen zeigen eine Accuracy zwischen 68,69% und 78,66%. Mit dem Ensemble wurde ein Gesamt-Ranking aus intern ausgewählten Merkmalsmengen generiert. Die ersten 100 Merkmale dieses Gesamt-Rankings werden nun genauer betrachtet und auf ihre Übereinstimmungen untersucht. Hierbei soll

überprüft werden, ob es Merkmale (Exons) gibt, die von verschiedenen Methoden als signifikant bewertet wurden. Daraus könnte ein biologischer Hintergrund abgeleitet werden. Die Anwendung des Ensembles erfolgte auf dem kompletten Datensatz. Im weiteren Verlauf wird nur noch von ausgewählten Merkmalsmengen verschiedener Auswahlmethoden gesprochen, dies impliziert hier immer die Verwendung eines Ensembles, auch wenn dies nicht mehr explizit erwähnt wird. Tabelle 7.1 gibt einen Überblick über die Anzahl der Übereinstimmungen zwischen den ausgewählten Mengen.

Methode-1	**Methode-2**	**Anzahl Übereinstimmungen**
SVM-Gewichtung	SAM	5
SVM-Gewichtung	Welch-Test	11
SVM-Gewichtung	t-Statistik	12
SVM-Gewichtung	Relief	8
t-Statistik	Welch-Test	89
t-Statistik	SAM	21
t-Statistik	Relief	33
Welch-Test	SAM	20
Welch-Test	Relief	31
SAM	Relief	31

Tabelle 7.1.: Vergleich der ersten 100 Ränge der Gesamt-Rankings von unterschiedlichen Bewertungsmethoden

Die höchste Übereinstimmung weisen die Merkmalsmengen der t-Statistik und des Welch-Test auf. Dies ist darauf zurückzuführen, dass der p-Wert auf Basis des t-Wertes berechnet wird. SAM ist eine Erweiterung der t-Statistiken, die Anzahl der Übereinstimmungen von SAM mit der t-Statistik und dem Welch-Test sind fast identisch. Die Gesamt-Rankings mittels SAM, Welch-Test und t-Statistik stimmen in 17 Merkmalen überein. Diese 17 Merkmale sind im Anhang (siehe A.1) mit zusätzlichen Informationen aufgelistet. Auffällig bei den Übereinstimmungen ist, dass viele Exons zu gemeinsamen Genen gehören. Drei dieser Gene (NTRK1, CHD5 und PLXNA4) sind für die Neuroblastombiologie relevante Gene. Inwieweit die anderen Übereinstimmungen von biologischer Bedeutung sind, sollte weiter untersucht und bewertet werden.
Die auf Basis der SVM-Gewichtung ausgewählte Merkmalsmenge weist mit den anderen Verfahren nur eine durchschnittliche Übereinstimmung von etwa 10% auf. Das kann darauf zurückgeführt werden, dass die SVM-Gewichtung Interaktionen der Merkmale bei ihrer Bewertung miteinbezieht. Die anderen Methoden (außer Relief) bewerten jedes Merkmal unabhängig voneinander. Die 8 Übereinstimmungen der Mengen der SVM-Gewichtung und Relief können von Interesse

sein, da beide Methoden Interaktionen berücksichtigen. Die entsprechenden Merkmale sind im Anhang (siehe A.1) zu finden. Auffällig ist hierbei, dass von den 8 Exons 5 Exons zu einem Gen gehören. Dieses Gen (PLXNA4) wurde biologisch bereits in Verbindung mit der Erkrankung gebracht.
Allen Merkmalsmengen ist gemeinsam, dass sie in 4 Merkmalen (Exons) übereinstimmen. Diese 4 Exons wurden also von den Bewertungsfunktionen immer unter die ersten 100 Ränge eingestuft. Von den 4 Exons gehören 3 zu einem gemeinsamen Gen (PLXNA4). Auch diese Exons sind im Anhang zu finden (sieh A.1).
Inwieweit die anderen Übereinstimmungen von biologischer Bedeutung sind, sollte weiter untersucht und bewertet werden.

Aus Tabelle 7.1 ergibt sich, dass im Durchschnitt die Übereinstimmungen zwischen den verschiedenen Merkmalsmengen bei ca. 25% liegen. Wie auch in anderen Untersuchungen (z.B. siehe [21]) zeigt dies, dass mit unterschiedlich ausgewählten Merkmalen ähnlich gute Lernergebnisse erzielt werden können. So weisen z.B. die SVM-Gewichtung und SAM kaum Übereinstimmungen innerhalb der ausgewählten Merkmale auf, liefern aber bei den Lerndurchläufen ähnlich gute Ergebnisse (siehe Kapitel 6.5, Tabelle 6.6) z.B. erzielte das Naive-Bayes-Verfahren auf den ausgewählten anhand von SAM bewerteten Merkmalen eine Acuraccy von 78.66% und auf die anhand der SVM-Gewichtung ausgewählten Merkmale eine Accuracy von 75.56%.
Im vorherigen Kapitel wurden mit dem Ensemble für jede Auswahlmethode die Merkmale ermittelt, die in jeder Runde des Ensembles unter den ersten 100 Rängen sind (siehe 6.11). Mit diesen Merkmalen wurde eine Klassifikation der Patienten vorgenommen (siehe Kap. 6 , Tabellen 6.12 und 6.13). Im Weiteren wurden diese Merkmalsmengen nun auf ihre Übereinstimmungen überprüft. Die Übereinstimmungen sind in Tabelle 7.2 angegeben. Die gefundenen Übereinstimmungen sind natürlich auch in den Übereinstimmungen der ersten 100 Merkmale enthalten. Durch die zusätzliche Einschränkung kann bei gefundenen Übereinstimmungen die evtl biologische Relevanz weiter gestützt werden.

Die Schnittmenge aller Merkmalsmengen enthält nur ein einzelnes Merkmal (siehe 7.3).

Dieses Gen ist biologisch bekannt und wird meistens mit einer guten Prognose der Erkrankung assoziiert. Weitere Exons dieses Gens wurden bei den Übereinstimmungen der Merkmalsmengen mit 100 Merkmalen gefunden.

Die Lernverfahren liefern auf die von SAM bewerteten Merkmale die besten Ergebnisse. Zugleich zeigt die auf Basis von SAM ausgewählte Merkmalsmenge die höchste Stabilität (siehe Tabelle 6.14). Die Robustheit stellt einen wichtigen Aspekt für die biologische Betrachtungsweise und die Relevanz bestimmter

Methode-1	Methode-1	Anzahl Übereinstimmungen
SVM-Weighting	SAM	1
SVM-Weighting	Welch	2
SVM-Weighting	t-Statistik	2
SVM-Weighting	Relief	2
t-Statistik	Welch	28
t-Statistik	SAM	8
t-Statistik	Relief	10
Welch	SAM	7
Welch	Relief	10
SAM	Relief	9

Tabelle 7.2.: Vergleich der Merkmale, die in jeder Runde des Ensembles enthalten sind

probeset-id	transkript-id	gene-assignment
3073484	3073267	NM-020911//PLXNA4//plexin A4//7q32.3//91584

Tabelle 7.3.: Schnittmenge der „immer“ enthaltenen Merkmale

Merkmale dar, somit könnten die auf Bassis von SAM selektierte Merkmalsmenge biologisch interessant sein; die Merkmale sind im Anhang (siehe A.4) aufgeführt. Von den 100 Merkmalen (Exons) gehören viele Exons gemeinsamen Genen an. Insgesamt sind nur Exons von 34 verschiedenen Genen enthalten. Unter diesen Genen befinden sich Gene (NTRK1, CHD5, PLXNA4 und CACNA2D3 [58]), für die eine biologische Relevanz im Zusammenhang mit der Erkrankung bekannt ist. Dies kann aus biologischer Sicht eine Motivation sein, auch andere selektierte Merkmale in dieser Menge einer genaueren biologischen Analyse zu unterziehen.

7.1.2. Merkmalskorrelationen

Der Begriff der *Korrelation* beschreibt die Stärke des linearen Zusammenhangs zwischen zwei Merkmalen [22]. Sind zwei Merkmale korreliert, so sind diese zwei Merkmale in jedem Fall linear voneinander abhängig. Im umgekehrten Fall, dass zwei Merkmale unkorreliert sind, bedeutet dies nicht, dass die Merkmale tatsächlich unabhängig sind. Es können auch nichtlineare Abhängigkeiten zwischen Merkmalen bestehen. Das Vorzeichen der Korrelation bestimmt, ob es sich um einen linearen Zusammenhang mit positiver oder negativer Steigung handelt. Bei linear unabhängigen Merkmalen ist die Korrelation = 0. Zur Messung der Korrelation gibt es verschiedene Maße. Das bekannteste Maß für die Korrelation zwischen Merkmalen ist der Korrelationskoeffizient nach Pearson (siehe [22]):

Definition 7.1 (Korrelationskoeffizient)
Für zwei Merkmale X und Y ist die Korrelation gegeben durch

$$Corr(X,Y) = \rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Seien $x_1, ..., x_d$ und $y_1, ..., y_d$ die Beobachtungen für X und Y, dann kann die Korrelation durch die Stichprobenkorrelation (Pearsonscher Korrelationskoeffizient) geschätzt werden. Die Berechnung erfolgt durch

$$r_{XY} = \frac{\sum_{i=1}^{d}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{d}(x_i - \overline{x})^2 \sum_{i=1}^{d}(y_i - \overline{y})^2}}$$

Der Korrelationskoeffizient r liegt immer zwischen -1 und $+1$. Mit den berechneten Korrelationswerten kann eine Korrelationsmatrix erstellt werden.

Im Bereich des maschinellen Lernens besteht das Interesse oft darin, Merkmale, die stark mit anderen Merkmalen korrelieren, zu identifizieren und zu reduzieren. Der Grund liegt darin, dass diese Merkmale meistens einen fast identischen Informationsgehalt aufweisen und dementsprechend durch Hinzunahme eines Merkmals, welches hoch mit einem bereits verwendeten Merkmal korreliert, für das Lernen keinen weiteren Nutzen bringt.
Aus biologischer Sicht kann jedoch sehr wohl ein Interesse an korrelierten Merkmalen bestehen, da diese Merkmale eine biologische Abhängigkeit zwischen einzelnen Genen/Exons (Merkmalen) widerspiegeln können. Aus biologischer Betrachtungsweise könnten sogar durch die Reduzierung der Redundanz wichtige Informationen verloren gehen.
Bei der Betrachtung der Korrelationsmatrixen der verschiedenen Merkmalsmengen (mit jeweils 100 Merkmalen) zeigt sich, dass SAM, t-Statistik und Welch-Test sehr viele hoch korrelierte Merkmale aufweisen. Abbildung 7.1 zeigt eine Visualisierung der berechneten Korrelationsmatrix für SAM; dabei stellt ein Punkt die Korrelation zweier Merkmale dar. Aus der Abbildung wird ersichtlich, dass SAM viele hoch korrelierte Merkmale aufweist und keine unkorrelierten Merkmale vorhanden sind.
Die Korrelationen der von Relief bewerteten Merkmale hat eine andere Verteilung; hier sind überwiegend unkorrelierte bis mäßig korrelierte Merkmale enthalten. Die SVM-Gewichtung besitzt kaum hoch korrelierte Merkmale. Die meisten Wert für die Korrelation zweier Merkmale liegen in einem Bereich von -0,1 und +0,3. Abbildung 7.2 zeigt eine Visualisierung der Korrelationsmatrix der Merkmale der

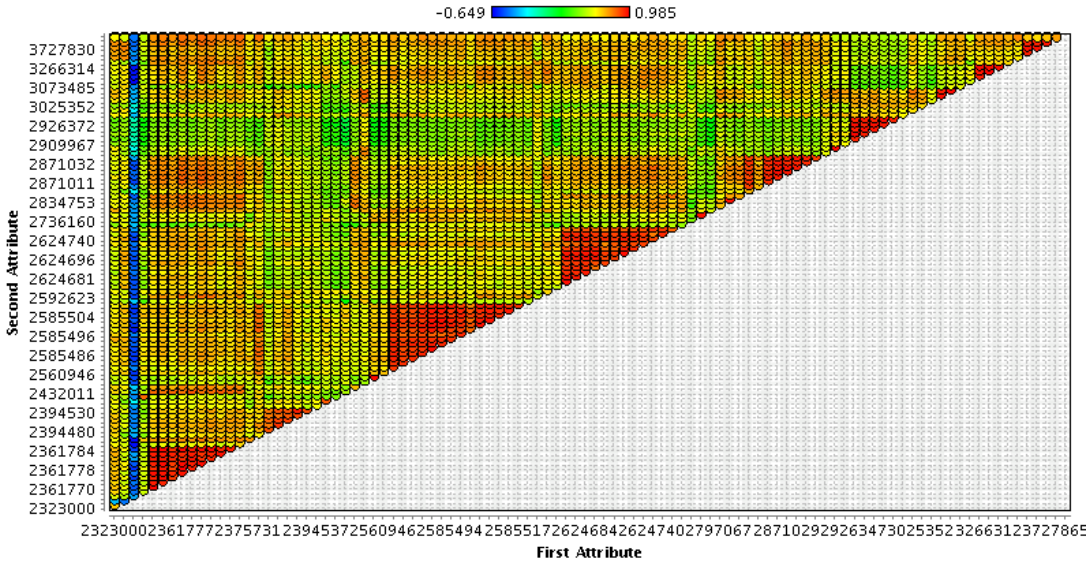


Abbildung 7.1.: Visualisierung Korrelationsmatrix SAM

SVM-Gewichtung.
Da SAM die meisten korrelierten Merkmale liefert und die SVM-Gewichtung die wenigsten, werden die Merkmale dieser Mengen im Folgenden genauer auf ihre Korrelation untersucht. In Abbildung 7.1 und 7.2 ist zu erkennen, dass sich die korrelierten Merkmale zu Gruppen zusammenfassen lassen (rote Bereiche in den Grafiken). Für SAM zeigen beide Klassen (EFS und Event) die selben korrelierten Merkmale und Gruppen auf. Es können 8 Gruppen hochkorrelierter Merkmale (Korrelation $< 0,9$) identifiziert werden. Auffallend dabei ist, dass die Exons (Merkmale) innerhalb einer Gruppe immer einem Gen zugeordnet werden können. Die Anzahl der Merkmale pro Gruppe und die entsprechenden Gen-Informationen (in Klammern angegeben) zu diesen Genen sind wie folgt:

- 1. Gruppe: 10 Merkmale (NM-001007792 // NTRK1 // neurotrophic tyrosine kinase)
- 2. Gruppe: 16 Merkmale (NM-002976 // SCN7A // sodium channel)
- 3. Gruppe: 13 Merkmale (NM-018398 // CACNA2D3 // calcium channel)
- 4. Gruppe: 7 Merkmale (NM-022140 // EPB41L4A // erythrocyte membrane protein band 4.1 like 4A // 5q22.2 // 64097)
- 5. Gruppe: 6 Merkmale (NM-004100 // EYA4 // eyes absent homolog 4 (Drosophila) // 6q23 // 2070)

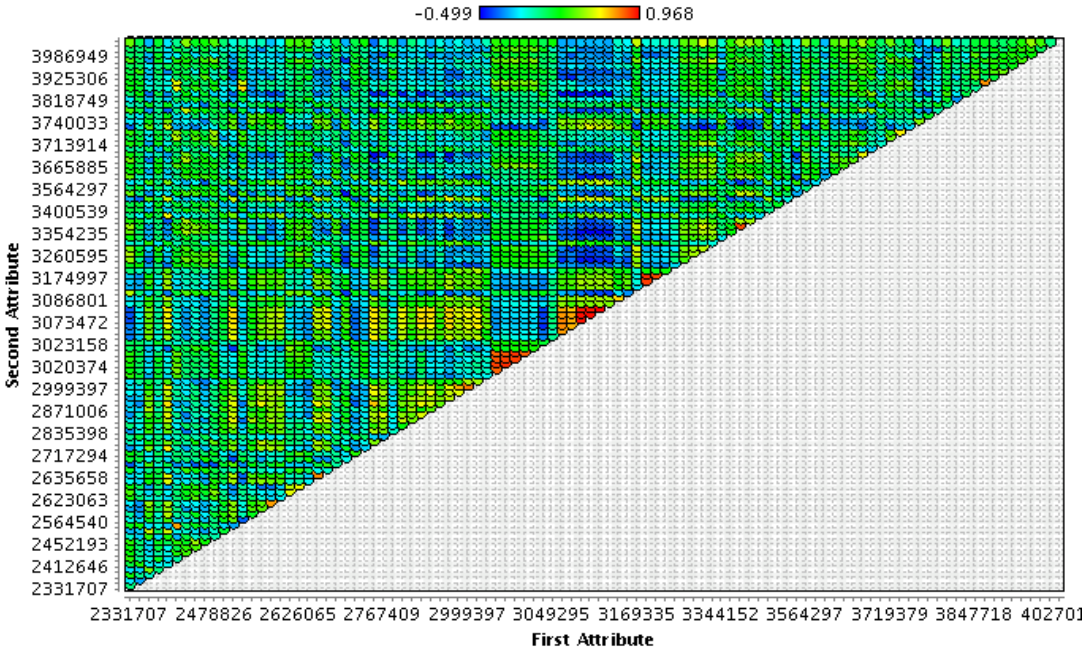


Abbildung 7.2.: Visualisierung Korrelationsmatrix SVM-Gewichtung

- 6. Gruppe: 3 Merkmale (NM-020911 // PLXNA4 // plexin A4 // 7q32.3 // 91584)
- 7. Gruppe: 4 Merkmale (NM-003054 // SLC18A2 // solute carrier family 18 (vesicular monoamine))
- 8. Gruppe: 4 Merkmale (NM-153228 // ANKFN1 // ankyrin-repeat and fibronectin type III domain containing 1 // 17q22 // 162282)

Die Merkmale der SVM-Gewichtung zeigen nur wenige Korrelationen. Dies ist darauf zurückzuführen, dass die SVM-Gewichtung Interaktionen zwischen den Merkmalen berücksichtigt und somit korrelierte Merkmale niedrig bewertet werden und daher meistens nicht in der Auswahl vorhanden sind. Wie auch schon bei SAM zeigen die zwei Klassen (EFS und Event) die selben Gruppen korrelierter Merkmale. Es lassen sich 2 Gruppen bilden, wobei - wie auch bei SAM beobachtet - die Merkmale zu einem Gen gehören:

- 1. Gruppe: 4 Merkmale (NM-020911 // PLXNA4 // plexin A4 // 7q32.3 // 91584)
- 2. Gruppe: 3 Merkmale (NM-006914 // RORB // RAR-related orphan receptor B // 9q22 // 6096)

Biologisch betrachtet zeigen die hoch korrelierten Merkmale innerhalb der Gruppen keine möglichen Zusammenhänge auf, da alle Gruppen einem Gen zugeordnet werden können. Hier könnte man die vage Vermutung äußern, dass es nicht nötig ist, diese Gene auf Exon-Ebene zu betrachten, sondern es ausreicht (oder den selben Nutzen erfüllt), die Gen-Ebene zu berücksichtigen.
Um beurteilen zu können, ob die hoch korrelierten Merkmale auch für die Lernverfahren ohne Bedeutung sind (keinen weiteren Informationsgewinn liefern), wurden mehrere Paare dieser hoch korrelierten Merkmale ausgewählt und mit jedem einzelnen Merkmal „gelernt“ und auf der Trainingsmenge validiert sowie mit beiden Merkmalen zusammen „gelernt“ und validiert. Die Ergebnisse wurden anhand der Accuracy miteinander verglichen. Hierbei zeigte sich, dass durch Kombination dieser Merkmale keine weitere Verbesserung der Accuracy erreicht werden kann. Ebenso wurden Paare von nicht korrelierten (oder wenig korrelierten) Merkmalen ausgewählt und die Ergebnisse auf den einzelnen Merkmalen und auf der Kombination dieser Merkmale verglichen. Hierbei zeigte sich, dass die Kombination dieser Merkmale eine Verbesserung der Accuracy bewirkte.
Da diese Experimente nur auf einigen manuell selektierten Merkmalspaaren durchgeführt wurden, muss die folgende Bewertung als sehr optimistisch angesehen werden: für das Lernen ist es ausreichend, eines der korrelierten Merkmale zu verwenden. Die Hinzunahme weiterer korrelierter Merkmale bringt keine Verbesserung des Lernergebnisses. Für die getesteten Merkmale gilt dies auch aus biologischer Sicht, da sie gemeinsamen Genen angehören.

7.2. Vergleiche von Patientenprofilen

Die Korrelationsanalyse identifizierte die SVM-Gewichtung und SAM-Bewertung als die Vertreter, deren ausgewählte Merkmale hinsichtlich der Korrelation der Merkmale untereinander im Vergleich zu den anderen Merkmalsauswahl-Methoden die extremsten Werte aufzeigten. Die Merkmale der SVM-Gewichtung zeigten wenig Korrelationen untereinander. Die Merkmale der SAM-Bewertung hingegen zeigten viele Korrelationen unter den 100 Merkmalen.
Anhand der von diesen zwei Methoden ausgewählten Merkmale wurden die einzelnen Patientenprofile anhand ihrer Expressionswerte verglichen. Ziel dieser Analyse war:

- Patienten innerhalb einer Klasse (EFS / Event) zu identifizieren, die sich stark in ihren Expresssionswerten unterscheiden

- Patienten unterschiedlicher Klassenzugehörigkeit zu ermitteln, die ähnliche Expressionswerte aufweisen

- Gruppen von Exons zu identifizieren, die innerhalb der Klassen „gleich“ exprimiert sind
- Gruppen von Exons zu identifizieren, die innerhalb der Klassen verschieden exprimiert sind

Für diese Untersuchung wurden pro Klasse alle Beispiele bezüglich der 100 Merkmale miteinander verglichen. Für jedes Paar von Beispielen wurde pro Merkmal (Exon) ermittelt, wie stark sich die Expressionwerte dieses Exons zwischen den zu vergleichenden Patienten voneinander unterscheiden. Als Maß für den Unterschied wurde der Betrag der Differenz gewählt.
Problematisch an diesem Ansatz ist die Frage, bis zu welcher Differenz Expressionswerte als „gleich ausgeprägt“ angesehen werden können. Aus biologischer Sicht ist die Beantwortung dieser Frage besonders bei Exon-Expressionen (hierzu liegen noch nicht genug Erfahrungsberichte vor) nicht einfach. Es ist keine Definition für „nicht gleich exprimiert“ gegeben, da die durch die Technik bedingte Messungsungenauigkeit nicht klar definiert ist. Eine Differenz < 1 kann in jedem Fall als „gleich exprimiert“ aufgefasst werden. In dieser Arbeit wird daher als zulässige Differenz für die Gleichheit $< 1,5$ gewählt.

Die Anzahl der Patienten-Vergleiche lässt sich mit dem Binomialkoeffizienten berechnen:

Definition 7.2 (Bionomialkoeffizient)
Seien n und k natürliche Zahlen einschließlich der Null und $n \geq k$. Dann ist der ***Bionomialkoeffizient*** $\binom{n}{k}$ *(„n über k“) definiert als:*

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Die Klasse „ereignisfreies Überleben“ (EFS) enthält 82 Patienten. Somit ergibt sich nach Definition 7.2 für die Vergleiche zwischen $k = 2$ Patienten aus einer Menge von $n = 82$ verschiedenen Patienten eine Anzahl von 3321 Vergleichen. Der Klasse „Rezidiv“ (Event) können 49 Patienten zugeordnet werden. Nach Definition 7.2 ergibt dies 1176 Vergleichsmöglichkeiten. Die zulässige Differenz (um von „gleich exprimiert“ zu sprechen) zwischen den Expressionswerten für ein Exon bei zwei verschiedenen Patientenprofilen ist auf 1,5 festgesetzt.

7.2.1. Analyse innerhalb der Klassen für SAM

Klasse EFS

Im Folgendem werden die 100 Merkmale betrachtet, welche auf Basis der SAM-Bewertung ausgewählt wurden, für die Klasse EFS betrachtet. Zuerst Patientenbezogen und anschließend Exon-bezogen.

1. **Patienten-bezogene Analyse**
   Innerhalb der Klasse EFS schwankt die Anzahl der unterschiedlich exprimierten Exons bei den einzelnen Patienten-Vergleichen von nur 2 unterschiedlich exprimierten Exons bis hin zu allen (=100) Exons. Es gibt keine zwei Beispiele, die in allen Expressionswerten vollständig übereinstimmen. Aber es existieren Beispiele, die sich in nur zwei Exon-Werten voneinander unterscheiden:

   - Patient 12[1] und Patient 116
   - Patient 70 und Patient 74

   Patienten-Vergleiche, die in allen Werten eine Differenz $>$ 1,5 aufzeigen, sind z.B.:

   - Patient 103 und Patient 44
   - Patient 18 und Patient 52

   In der Grafik 7.3 ist die Anzahl der Abweichungen (unterschiedlich exprimierte Werte) für die Patientenvergleiche abgebildet. Auf den x/y-Achsen sind die Patienten-Gegenüberstellungen aufgetragen. Ein Punkt steht für den Vergleich zweier Patienten. Die Farbwahl gibt die Anzahl der Differenz an (rot: Differenzen in allen 100 Exons, blau: Differenzen in nur 2 Exons). Die Grafik verdeutlicht noch einmal, dass sich die Patienten bei den Vergleichen in ihren Expressionswerten häufig voneinander unterscheiden.
   Der Patient mit der Id 18 weist bei den Vergleichen mit den anderen Patienten die meisten Abweichungen auf (im Anhang siehe A.3 Abbildung A.1 sind die Abweichungen der Patienten zu den anderen Patienten innerhalb der Klasse grafisch dargestellt).
   Jedem Patienten der Klasse EFS lässt sich ein Krankheitsstadium zuordnen. Aus der Anzahl der Abweichungen zwischen den Vergleichen zweier Patienten folgern zu können, dass beide Patienten in ihrem Krankheitsstadium übereinstimmen oder nicht übereinstimmen, kann nicht bestätigt werden. Sowohl Patienten des gleichen Stadiums wie auch Patienten unterschiedlichen Stadiums können innerhalb der Vergleiche wenige oder viele Abweichungen ihrer Expressionswerte aufweisen.
   Eine genauere Betrachtung der 5 Patienten, welche in über 80% der Vergleiche von den anderen Patienten bezüglich ihrer Expressionswerte abwichen, ergab, dass diese Patienten in ihrem Krankheitsstadium übereinstimmen. Die Patienten werden Stadium 4 (siehe Kapitel 2.3) zugeordnet, was mit einer schlechten Prognose einhergeht. Eine allgemeingültige Aussage über die Anzahl der Abweichungen kann aber nicht gegeben werden.

[1] Patienten-Ids laufen nicht nur von 1 bis 131, sondern von 1 bis 153, nicht jede Id ist belegt.

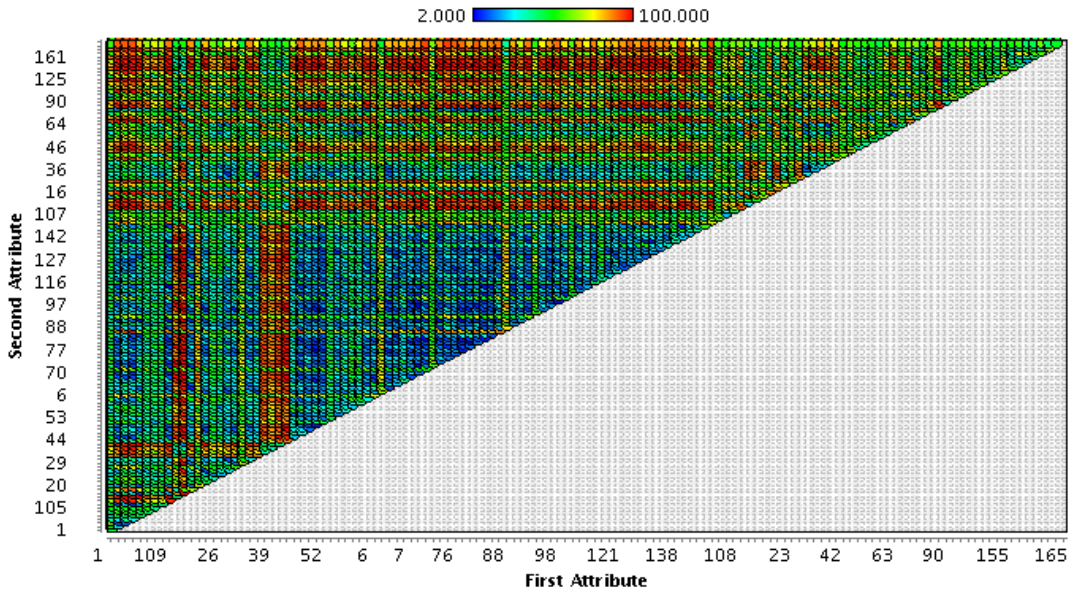


Abbildung 7.3.: Anzahl unterschiedlich exprimierter Exons bei Patienten-Vergleichen für SAM innerhalb der Klasse EFS

2. **Exon-bezogene Analyse**
   Für die 82 Patienten konnte kein Exon ermittelt werden, das innerhalb der Patientenvergleiche immer „gleich“ exprimiert ist. Exons mit den wenigsten/meisten Abweichungen innerhalb der Vergleiche können zu folgenden Gruppen zusammengefasst werden:

   - Eine Gruppe von 7 Exons, die bei den Vergleichen weniger als 30% Abweichungen aufzeigten, also innerhalb der Klasse ziemlich ähnlich exprimiert sind.
   - Eine Gruppe von 9 Exons, die in den Vergleichen Abweichungen von über 55% haben und somit innerhalb der Klasse differentiel exprimiert sind.

Die Exons innerhalb dieser Gruppen können oft gemeinsamen Genen zugeordnet werden. Im Anhang ist eine Abbildung (siehe A.2) eingefügt, die die Anzahl der Abweichungen pro Exon bei den Vergleichen visualisiert.

Klasse Event

Im Folgenden wird die Klasse Event (Patienten erleiden einen Rückfall) anhand der 100 Merkmale von SAM betrachtet. Zu dieser Klasse zählen 49 Patienten. Die Anzahl der aller Vergleiche zwischen den Patienten dieser Klasse beträgt 1172.

1. **Patienten-bezogene Analyse**
   Bei den Patienten-Vergleichen innerhalb der Klasse Event variiert die Anzahl der zwischen den Patienten unterschiedlich exprimierten Merkmale von 7 unterschiedlich exprimierten Exons bis zu 100 (von 100!) Exons, die eine Differenz $> 1{,}5$ betragen. Patienten-Vergleiche mit wenigen/vielen Abweichungen sind:

   - Patient 38 und Patient 65 mit nur 7 Abweichungen
   - Patient 10 und Patient 19 mit den meisten Abweichungen (100%)

   Patient 10 und Patient 83 zeigen bei den Vergleichen zu allen anderen Patienten die höchste Anzahl an Abweichungen (ca. 74% Abweichungen). Diese zwei Patienten weisen hinsichtlich ihrer Krankheitsstadien keine Gemeinsamkeiten auf. Bezüglich ihrer Expressionswerte stimmen die beiden Patienten aber bis auf 14 Abweichungen miteinander überein. Beide Patienten haben sehr hoch exprimierte Werte.

2. **Exon-bezogene Analyse**
   Anhand der Anzahl der Abweichungen pro Exon bei den Vergleichen lassen sich folgende Gruppen zusammenfassen:

   - Bei den Patienten-Vergleichen konnte kein Exon selektiert werden, welches innerhalb der Vergleiche relativ konstant „gleich exprimiert“ war. Die zwei Exons, die innerhalb der Gruppe am wenigsten unterschiedlich exprimiert waren, sind Exon 2327561 und Exon 2736160.
   - Zur Gruppe der Exons, welche sich bei den Vergleichen als sehr unterschiedlich ausgeprägt zeigten, zählen 14 Exons.

   Auch hier zeigt sich, dass die gefundenen Exons häufig gemeinsamen Genen zugeordnet werden können.

Zusammenfassung: Analyse innerhalb der Klassen für SAM

Innerhalb der 100 Merkmale ließen sich für die Klassen EFS und Event jeweils Gruppen von Exons bilden, deren Expressionswerte zwischen den Patienten wenig variieren und Gruppen von Exons, deren Expressionswerte zwischen den Patienten-Vergleichen stärker variieren.
Bei der Klasse EFS konnte eine Gruppe von 7 Exons selektiert werden, bei denen die Patientenprofile bei den Vergleichen für diese Exons wenig Abweichungen zeigten. Die Werte der Patienten für diese gefundenen Exons sind sich nur innerhalb der Klasse EFS ähnlich; innerhalb der Klasse Event zeigen die Patienten für diese Exons bei über 50% der Vergleiche Abweichungen auf. Einige dieser 7 Exons gehören zu gemeinsamen Genen (siehe Anhang Tabelle A.5). Inwieweit

diese Gene eine biologisch Relevanz aufzeigen, sollte weiter abgeklärt werden; eines dieser Gene (CACNA2D3) ist in anderen Untersuchungen bereits als für die Neuroblastomforschung interessant eingestuft worden (siehe [58]).
Es konnten für beide Klassen Patienten identifiziert werden, deren Expressionswerte sich von den anderen Patienten innerhalb einer Klasse häufig unterschieden. Durch diese heterogene Verteilung innerhalb einer Klasse ist es für die Lernverfahren schwierig, ein Modell mit einer guten Vorhersageprognose zu generieren. Eine Möglichkeit ist, Patienten, die sich stark von allen anderen Patienten innerhalb einer Klasse absetzen, nicht zum Lernen zu verwenden. Aus biologischer Sichtweise ist es problematisch, einen Zusammenhang zwischen den Patienten einer Klasse und der Erkrankung herzustellen, wenn die Patienten hinsichtlich ihrer Expressionsprofile sehr unterschiedlich sind.

7.2.2. Analyse zwischen den Klassen für SAM

Bei der Analyse der Patientenprofile zwischen verschiedenen Klassen soll der Frage nachgegangen werden, ob es Patienten aus verschiedenen Klassen gibt, die sich in ihren Profilen gleich bzw. ähnlich sind.
Der Klasse EFS gehören 82 Patienten an und der Klasse Event 49 Patienten. Die Patienten der unterschiedlichen Klassen werden miteinander verglichen. Es entstehen $82 \cdot 49 = 4018$ Vergleichsmöglichkeiten. Bei dem Vergleich zweier Patienten unterschiedlicher Klassen werden - wie in den vorherigen Analysen - die Ausprägungen der einzelnen Merkmalswerte einzeln miteinander verglichen und eine Differenz zwischen zwei Merkmalen $\leq 1,5$ als „gleich“ exprimiert angesehen. Bei den Vergleichen zwischen den Patienten konnten keine zwei Patienten aus unterschiedlichen Klassen identifiziert werden, die in all ihren Expressionswerten vollkommen übereinstimmen. Die meisten Übereinstimmungen in ihren Werten haben:

- Patient 27 (EFS) und Patient 36 (Event)
- Patient 10 (EFS) und Patient 142 (Event)

Beide Vergleiche weisen von 100 möglichen Abweichungen nur 6 Abweichungen zwischen den Expressionswerten auf. Diese Patienten sind sich in ihren Profilen also sehr ähnlich, gehören aber unterschiedlichen Klassen an.
Es konnte ermittelt werden, dass die Expressionswerte von Patienten unterschiedlicher Klassen nie vollständig übereinstimmen, sich aber in vielen Vergleichen nur in einigen Merkmalsausprägungen voneinander unterscheiden.

7.2.3. Analyse innerhalb der Klassen für SVM-Gewichtung

Wie zuvor mit den Merkmalen der SAM-Bewertung sollen nun die Patientenprofile anhand der 100 Merkmale der SVM-Gewichtung genauer betrachtet werden.

1. **Patienten-bezogene Analyse**
   Bei der SVM-Gewichtung schwankt die Anzahl der unterschiedlich exprimierten Exons von 4 unterschiedlich exprimierten Exons bis zu 48 unterschiedlich exprimierten Exons innerhalb der Patienten-Vergleiche. Ebenso wie bei der vorherigen SAM Auswertung gibt es auch bei den Merkmalen der SVM-Gewichtung keine zwei Patienten, die in allen Expressionswerten eine Differenz $< 1{,}5$ zeigen und somit vollständig miteinander übereinstimmen. Bei den von der SVM-Gewichtung ausgewählten Merkmalen zeigen die meisten Patienten-Vergleiche Abweichungen von unter 30%. Im Vergleich zu SAM ist die Anzahl der Abweichungen insgesamt viel geringer. Patienten, die viele Übereinstimmungen in den ausgewählten Attributen aufweisen, sind z.B.:

   - Patient 91 und Patient 103
   - Patient 77 und Patient 79

   Die meisten Abweichungen von über 45% zeigen die Vergleiche zwischen:

   - Patient 12 und Patient 153
   - Patient 44 und Patient 53

   Wie auch schon bei der SAM-Auswahl beobachtet, gibt es kein Patientenprofil, welches sich von allen anderen Profilen in seinen Expressionswerten abweichend zeigt. Auch hier lässt sich keine Patienten-Einteilung zu ihren Krankheitsstadien anhand der Anzahl der Abweichungen feststellen.
   Die Patienten 44 und 153, welche sich in ihren Expressionswerten deutlich von den anderen Patienten unterschieden, gehören Stadium 4 an. Schon bei SAM wurde beobachtet, dass die Patienten mit den meisten Abweichungen Stadium 4 zuzuordnen sind. Bei diesen Patienten ließen sich jedoch keine Gemeinsamkeiten feststellen.

2. **Exon-bezogene Analyse**
   Exons, für die wenig/viele Abweichungen bei den Vergleichen vorliegen, sind:

   - Exon 2394707 und Exon 3996449. Bei diesen Exons zeigen die Patienten bei den Vergleichen keine Abweichung $> 1{,}5$. Diese 2 Exons sind also innerhalb der Klasse EFS für alle Patienten „gleich“ exprimiert.
   - Exon 3174996 und Exon 24331606. Diese Exons zeigen die meisten Abweichungen bei den Vergleichen, von 3321 Vergleichen zeigen die

Expressionswerte der Patienten für die beiden Exons Abweichungen bei über 57% der Vergleiche.

Auch hier können Exons zu Gruppen anhand ihrer Anzahl Abweichungen zusammengefasst werden:

- 12 Exons zeigen sich innerhalb der Klasse als relativ „gleich exprimiert“
- 5 Exons, die innerhalb der Klasse unterschiedlich exprimiert sind, können zu einer Gruppe zusammengefaßt werden.

Im Vergleich zu SAM gehören innerhalb dieser Gruppen nur wenige Exons gemeinsamen Genen an.

Klasse Event

Nun wurden die Exons der SVM-Gewichtung für die Klasse Event untersucht.

1. **Patienten-bezogene Analyse**
   Auch die von der SVM-Gewichtung ermittelten 100 Merkmale wurden für die Patientenprofile untersucht. Wie auch bei der Betrachtung der von SAM selektierten Merkmale gibt es keine zwei Patienten, bei denen alle Exonwerte vollkommen übereinstimmen. Patienten-Vergleiche mit wenig/vielen Abweichungen sind:

   - Patient 108 und Patient 155. Sie zeigen mit 56 von 100 möglichen Abweichungen die meisten Abweichungen.
   - Patient 10 und Patient 83. Sie zeigen mit 5 Abweichungen die geringste Anzahl auf.

   Diese beiden Patienten wurden bei der vorherigen Analyse der von SAM ermittelten Merkmale als die Patienten identifiziert, die die meisten Abweichungen zu den übrigen Patienten zeigten. Bei den von der SVM-Gewichtung ermittelten Werten weisen diese beiden Patienten jedoch nicht die meisten Unterschiede zu den übrigen Patienten auf. Hier liegen sie mit 33,25% und 27,75% Abweichungen im mittlerem Bereich. Patienten mit den meisten/wenigsten Abweichungen zu allen anderen Patienten der Klasse:

   - Patient 106 hat mit 23,35% Abweichungen bei den Vergleichen mit den anderen Patienten die wenigsten Abweichungen
   - Patient 165 ist mit 41,46% Abweichungen der Patient, mit den meisten Abweichungen den anderen Patienten gegenüber.

   Innerhalb der Klasse Event gibt es drei Patienten, deren Expressionsprofile sich von den anderen Patienten absetzen Diese drei Patienten sind:

   - Patient 165, Patient 159 und Patient 155.

2. **Exon-bezogene Analyse**
   Es kann ein Exon gefunden werden, welches bei allen Patienten „gleich“ exprimiert ist: Exon 3996949. Exon 3936523 ist in nur einem Vergleich unterschiedlich exprimiert.
   Gruppen-Bildung:
   - Es können 16 Exons zu einer Gruppe zusammengefasst werden, die bei weniger als 10% der 1176 Vergleiche zwischen den Patienten Abweichungen zeigen.
   - Bei über 52% der Vergleiche zeigen die Patienten für 11 Exons Abweichungen. Diese 11 werden zu einer Gruppe zusammengefasst.

Auch hier können nur wenige Exons innerhalb der Gruppen zu gemeinsamen Genen zusammengefasst werden.

Zusammenfassung: Analyse innerhalb der Klassen - SVM

Aus den besten hundert Merkmalen der SVM-Gewichtung ließen sich innerhalb der Klassen EFS und Event Gruppen von Exons bilden, deren Expressionswerte zwischen den Patienten wenig variieren und Gruppen von Exons, deren Expressionswerte zwischen den Patienten-Vergleichen stark variieren.
Innerhalb der Klasse Event konnten 16 Exons zu einer Gruppe zusammengefasst werden. Für diese Exons zeigten die Expressionswerte der Patienten bei den Vergleichen unter 10% Abweichungen. Innerhalb der Klasse EFS konnten 12 solcher Exons selektiert werden.
Die beiden Gruppen stimmen in 9 der Exons miteinander überein. Das bedeutet, dass sich für diese Exons die Expressionswerte sowohl in der Klasse EFS als auch in der Klasse Event unter den Patienten nur gering unterscheiden. Wenn diese Exons innerhalb der Klassen immer sehr ähnlich exprimiert sind und zwischen den Klassen sehr unterschiedlich, wären diese Exons gute Merkmale zum „Lernen“ und könnten auch biologisch von Interesse sein. Zur Überprüfung der Unterschiede zwischen den Klassen wurde der Welch-Test angewendet. Dieser ergab, dass nur für 2 der 9 Exons ein signifikanter Unterschied zwischen den zwei Klassen besteht, diese zwei Exons sind 3688897 und Exon 3688909. Beide Exons werden einem Gen (SLC6A8) zugeordnet. Für die anderen 7 Exons konnte kein signifikanter Unterschied zwischen den Klassen festgestellt werden. Dies bedeutet, dass diese Exons in beiden Klassen immer ähnlich exprimiert sind und biologisch wahrscheinlich nicht mit der Vorhersageprognose in Verbindung zu bringen sind. Diese Vermutung müsste jedoch biologisch weiter untersucht werden. Die den 7 Exons zugeordneten Gene weisen keine Übereinstimmungen mit bereits bekannten Genen der Forschung auf.
Von den 3 verbleibenden Exons in den Gruppen EFS könnte Exon 3333654 (zugehöriges Gen „TAF6L“) von Bedeutung sein, da dieses innerhalb der Klasse sehr

ähnlich exprimiert ist und in der Klasse Event die Werte für dieses Exon bei den Vergleichen sehr schwankend sind.
Allgemein lässt sich beobachten, dass bei den von der SVM-Gewichtung ausgewählten Merkmalen die Werte der Patienten-Vergleiche viel weniger Abweichungen aufweisen als bei den von der SAM-Bewertung selektierten Merkmalen.
Im Gegensatz zu den von SAM gewählten Merkmalen zeigen für die anhand der SVM-Gewichtung ausgewählten Merkmale keine zwei Patienten in allen 100 Expressionswerten Abweichungen. Aber auch hier gibt es keine zwei Patienten, die in allen Expressionswerte vollkommen übereinstimmen.
Die Patienten innerhalb einer Klassen zeigen sich sind zwar bzgl. ihrer Expressionsprofile nicht so heterogen wie bei unter den von SAM ausgewählten Merkmalen, aber dennoch machen die vielen Unterschiede ein gutes „Lernen“ sehr schwierig. Auch eine Selektion von wirklich relevanten Merkmalen ist bei den Verteilungen sehr problematisch.

7.2.4. Analyse zwischen den Klassen für SVM-Gewichtung

Anhand der 100 ausgewählten Merkmale durch die SVM-Gewichtung konnten auch keine zwei Patienten gefunden werden, die in allen Merkmalswerten identisch sind. Jedoch gibt es auch hier Patienten unterschiedlicher Klassenzugehörigkeit, die sich in nur wenigen Expressionswerten voneinander unterscheiden. Z.B. unterscheiden sich:

- Patient 81 (EFS) und Patient 10 (Event) in nur 5 Ausprägungen voneinander

Die Ähnlichkeiten der Patienten zwischen den Klassen erschweren eine gute Trennung der Patienten.

7.3. Textklassifikation

Die Lernaufgabe bei der Textklassifikation besteht darin, Dokumente (Beispiele) in vorgegebene Kategorien zu klassifizieren. Es wird meistens die Annahme getroffen, dass die Reihenfolge der Wörter in einem Dokument unberücksichtigt bleibt. Der Inhalt eines Dokuments wird durch einen Dokumentenvektor $\vec{x_i}$ beschrieben; dies ist ein Vektor von Merkmalswerten. Die dabei häufig verwendete Art der Repräsentation ist die „bag-of-words“ Darstellung. Ein Merkmal entspricht einem Wort. Der Wert eines Merkmals für ein Dokument ist die Häufigkeit, mit der das Wort im Dokument auftritt. Die Häufigkeit des Vorkommens eines Wortes w in einem Dokument d wird auch als *Termfrequenz TF(w,d)* bezeichnet.
In Abbildung 7.4 ist diese Repräsentation eines Dokuments illustriert.

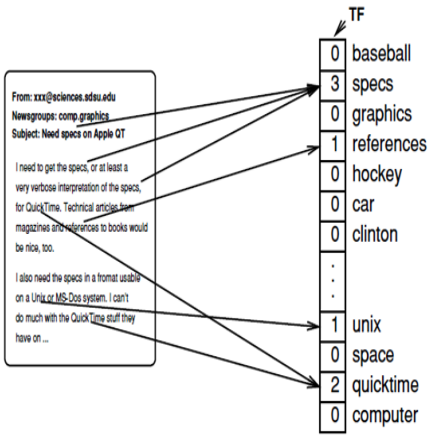


Abbildung 7.4.: Repräsentation eines Dokuments (Text) als Merkmals-Vektor (entnommen aus [34].

7.3.1. TCat-Modell

Das **T**ext-**Cat**egorisierungs-Modell (TCat) von Joachims [34] beweist die Anwendbarkeit von SVMs auf Textklassifikationsaufgaben und beruht auf fünf statistischen Eigenschaften:

- Hochdimensionaler Merkmalsraum
  Bei Textklassifikationsproblemen liegt ein hochdimensionaler Merkmalsraum vor. Jedes Wort der Trainings-Dokumente wird als Merkmal betrachtet; abhängig von der Größe der Dokumenten-Menge entstehen schnell Dimensionen von 30.000 und mehr Merkmalen.

- Heterogener Gebrauch von Worten
  Dokumente, die derselben Kategorie angehören, haben manchmal nur wenige bis gar keine Überschneidungen der in ihnen vorkommenden Wörter. So können Dokumente aus der gleichen Kategorie aus völlig verschiedenen Wörtern bestehen. Es gibt keine relevanten Wörter, die in allen Dokumenten einer Klasse vorkommen. So kann es vorkommen, dass ein Dokument A mit einem anderen Dokument B Gemeinsamkeiten aufweist und Dokument A auch Gemeinsamkeiten mit Dokument C hat, aber alle drei Dokumente A, B und C keine Übereinstimmungen zeigen. Dieses Phänomen wird von Wittgenstein [68] als „Familienähnlichkeit“ bezeichnet.

- Redundanz von Merkmalen
  Die überwiegenden Dokumente enthalten mehrere Wörter, die die Klassenzugehörigkeit bestimmen. Auch nach Entfernen dieser charakteristischen Wörter kann anhand der verbleibenden Wörter der Inhalt des Dokuments in einem gewissen Maß beschrieben werden.

- Spärlich besetzte Dokumentenvektoren
  Es sind viele Wörter im Lexikon enthalten. Jedes Dokument enthält aber nur eine kleine Anzahl unterschiedlicher Worte. Daraus folgt, dass der Dokumentenvektor an vielen Positionen Nullen enthält und somit spärlich besetzt ist.

- Zipfs Gesetz
  Das Zipfsche Gesetz ist nach dem Linguisten George Kingsley Zipf benannt und befasst sich mit der Häufigkeitsverteilung von Worten in Texten [72]. Das Gesetz besagt, dass, wenn ein Ranking der Wörter nach ihrer Häufigkeit (r) aufgestellt wird, dann für die Worthäufigkeit TF_r (Termfrequenz) eines Wortes vom Rang r gilt:

  $$TF_r = \frac{1}{r} \cdot TF_{rmax} \tag{7.1}$$

  TF_{rmax} bezeichnet die Häufigkeit des häufigsten Wortes.
  Eine genauere Anpassung dieser Verteilung liefert die Mandelbrotverteilung nach [41]:

  $$TF_i = \frac{c}{(k+r)^{\phi}} \tag{7.2}$$

Wobei die Parameter c, k und ϕ einer besseren Anpassung dienen. Diese Verteilung wird oft auch als *generalisiertes Zipfsches Gesetz* bezeichnet.
Das Modell bezieht sich auf die Anwendbarkeit der SVM. Die SVM wählt genau die Hyperebene, die die Beispiele bzgl. der Klassenzugehörigkeiten am besten (mit maximalem Abstand) trennt. Jedes Beispiel hat zur Hyperebene mindestens einen Abstand von δ. Wie in Kapitel 4.4.3 bereits erwähnt, wird δ als Rand (*margin*) der Hyperebene bezeichnet.
Joachims [34] zeigte, dass eine Separation der Beispiele mit einem großen Rand einhergeht. Die Kombination eines großen Rands mit einem geringen Trainingsfehler führt zu einer hohen Generalisierungsfähigkeit. Joachims gibt in [34] eine Schranke für den erwarteten Fehler an:

Theorem 7.1 (Schranke für den erwarteten Fehler)
Die erwartete Fehlerrate $\varepsilon(Err^n(h_{SVM}))$ einer SVM mit weichem Rand basierend auf n Trainingsbeispielen mit $c \leq \kappa(\vec{x_i}, \vec{x_j}) \leq c + R^2$ für eine Konstante c ist beschränkt durch

$$C \geq \frac{1}{\rho R^2} : \varepsilon(Err^n(h_{SVM})) \leq \frac{\rho\varepsilon\left(\frac{R^2}{\delta_2}\right) + C\rho R^2 \varepsilon\left(\sum_{i=1}^{n+1} \xi_i\right)}{n+1} \tag{7.3}$$

$$C < \frac{1}{\rho R^2} : \varepsilon(Err^n(h_{SVM})) \leq \frac{\rho\varepsilon\left(\frac{R^2}{\delta_2}\right) + \rho\left(CR^2+1\right)\varepsilon\left(\sum_{i=1}^{n+1}\xi_i\right)}{n+1} \tag{7.4}$$

δ bezeichnet den Rand, R die Länge der Dokumentenvektoren und ξ den Trainingsfehler.

Definition 7.3 (Homogene TCat-Konzepte)
Das TCat-Konzept

$$TCat([p_1 : n_1 : f_1], ..., [p_s : n_s : f_s])$$

beschreibt eine binäre Klassifikationsaufgabe mit s disjunkten Mengen von Merkmalen. Die i-te Menge enthält f_i Merkmale. Jedes positive Beispiel enthält p_i Merkmale aus der jeweiligen Menge und jedes negative Beispiel enthält n_i Merkmale. Das gleiche Merkmal kann mehrmals in einem Dokument vorkommen.

Joachims zeigt in [34] eine untere Schranke für den Rand δ der trennenden Hyperebene:

Lemma 7.2 (Untere Schranke für den Rand von TCat-Konzepten)
Für $TCat([p_1 : n_1 : f_1], ..., [p_s : n_s : f_s])$-Konzepte existiert immer eine Hyperebene durch den Ursprung und die Hyperebene hat einen Rand beschränkt durch

$$\delta^2 \geq \frac{ac - b^2}{a + 2b + c} \tag{7.5}$$

$$mit\ a = \sum_{i=1}^{s} \frac{p_i^2}{f_i}, \qquad b = \sum_{i=1}^{s} \frac{p_i n_i}{f_i} \quad und \quad c = \sum_{i=1}^{s} \frac{n_i^2}{f_i}$$

Die Separierbarkeit der Beispiele impliziert, dass der Trainingsfehler null ist. Um Theorem 7.1 anwenden zu können, wird eine Schranke für die maximale Euklidische Länge R der Dokumentenvektoren benötigt. Die Euklidische Länge des Dokumentenvektors eines Dokuments mit l Worten kann nicht größer sein als l. Diese Schranke ist für reale Dokumente nicht genau genug. Unter Verwendung des Zipfschen Gesetzes kann eine präzisere Schranke für R angegeben werden. Angenommen, die Termfrequenzen in jedem Dokument folgen dem generalisiertem Zipfschen Gesetz, dies impliziert nicht, dass ein bestimmtes Wort mit einer gewissen Häufigkeit in jedem Dokument vertreten ist. Das Zipfsche Gesetz besagt nur, dass das r-häufigste Wort mit einer gewissen Häufigkeit in einem Dokument vertreten ist. Das r-häufigste Wort kann somit in unterschiedlichen Dokumenten verschieden sein. Das folgende Lemma verbindet die Länge der Dokumentenvektoren mit dem Zipfschen Gesetz.

Lemma 7.3 (Euklidische Länge der Dokumentenvektoren)
Wenn die gerankten Termfrequenzen TF_i in einem Dokument mit l Termen dem generalisiertem Zipfschen Gesetz folgen

$$TF_i = \frac{c}{(k+r)^{\phi}} \tag{7.6}$$

dann ist die quadratische euklidische Länge des Dokumentenvektors $\vec{x}$ der Termfrequenzen begrenzt durch

$$\|\vec{x}\| \leq \sqrt{\sum_{r=1}^{d} \left(\frac{c}{(k+r)^{\phi}} \right)^2} \quad \textit{wobei für } d \textit{ gilt} \quad \sum_{r=1}^{d} \frac{c}{(k+r)^{\phi}} = l \tag{7.7}$$

Die Tatsache, dass die Termfrequenzen dem Zipfschen Gesetz folgen, hat einen starken Einfluss auf die Lernbarkeit von Textklassifikationsaufgaben. Das Zipfsche Gesetz beinhaltet, dass sich die meisten Worte nicht oft wiederholen und die Anzahl der unterschiedlichen Terme d hoch ist. Würde das Zipfsche Gesetz nicht gelten, könnte ein einzelnes Wort l-mal vorkommen und der Dokumentenvektor hätte eine Euklidische Länge von l. Durch das Zipfsche Gesetz erhält man vergleichsweise kurze Dokumentenvektoren und dadurch bedingt einen niedrigen Wert für R^2 in der Schranke für die erwartete Generalisierungsperformanz.

Durch Kombination von Lemma 7.2 und Lemma 7.3 erhält man folgendes Theorem für die Lernbarkeit von TCat-Konzepten:

Theorem 7.4 (Lernbarkeit von TCat-Konzepten)
Für TCat$([p_1 : n_1 : f_1], ..., [p_s : n_s : f_s])$-Konzepte und Dokumente mit l Termen, deren Verteilung dem generalisierten Zipf Gesetz $TF_i = \frac{c}{(k+r)^{\phi}}$ entsprechen, ist der erwartete Generalisierungsfehler einer SVM nach dem Training auf n Beispielen beschränkt durch

7.3.2. Übertragung des TCat-Modells auf Neuroblastom-Daten

Es wurde untersucht, ob die Eigenschaften der Textklassifikation auch für die vorliegenden Neuroblastom-Daten gelten und die Daten sich als TCat-Modell modellieren lassen. Mit diesem Modell könnte dann der Generalisierungsfehler der SVM auf den Daten abgeschätzt und eine Einschätzung für die Schwierigkeit der Lernaufgabe gegeben werden.
Um auf den vorliegenden Daten ein TCat-Konzept erstellen zu können, wird zunächst überprüft, ob die Eigenschaften von Texten auch für die vorliegenden Neuroblastom-Daten gelten. Bei der Textklassifikation wird ein Dokument als

$$\varepsilon(Err^n(h_{SVM})) \leq \rho\frac{R^2}{n+1}\frac{a+2b+c}{ac-b^2} \quad \text{mit} \quad a = \sum_{i=1}^{s}\frac{p_i^2}{f_i}, \; b = \sum_{i=1}^{s}\frac{p_i n_i}{f_i}, \; c = \sum_{i=1}^{s}\frac{n_i^2}{f_i}, \; R^2 = \sum_{r=1}^{d}\left(\frac{c}{(k+r)^\phi}\right)^2$$

wenn nicht $\forall_{i=1}^{s} : p_i = ni$ ist. d wird so gewählt, dass $\sum_{r=1}^{d}\frac{c}{(k+r)^\phi} = l$. Für unverzerrte SVMs ist $\rho = 1$ und für verzerrte SVMs ist $\rho = 2$.

Beispiel betrachtet. Für die vorliegenden Daten kann ein Patienten als ein „Dokument“ aufgefasst werden. Ein Exon (Merkmal) entspricht dann einem „Wort“ in einem Dokument. Das „Wörterbuch“ für die Exon-Daten besteht aus allen verschiedenen Exons (184.985 verschiedene „Wörter“).
Zusammengefasst:

- Exon = Wort
- Patient = Dokument
- verschiedene Exons = Wörterbuch

Bei der Textklassifikation werden Dokumente als Vektoren aufgefasst. Jeder Vektor hat die Dimension des zugrunde liegenden Wörterbuchs. Ein Vektoreintrag entspricht dem Vorkommen des entsprechenden Wortes (Termfrequenz) im Text (siehe 7.4). Für die vorliegenden Daten kann der Expressionswert als Termfrequenz angesehen werden. Hat ein Patient beispielsweise für ein Exon einen Expressionswert von 5, würde dies im übertragenden Sinn einem Wortvorkommen im Dokument von 5 entsprechen.
Die gegebenen Daten zeichnen sich durch einen hochdimensionalen Merkmalsraum (184.985 verschiedene Merkmale) aus. Ein heterogener Gebrauch der Merkmale kann vermutet werden, da - wie bereits erläutert - verschiedene Klassifikatoren auf unterschiedlichen Merkmalen ähnlich gute Ergebnisse erbringen können. Die Eigenschaft der Spärlichkeit ist für die Daten nicht gegeben. Jeder Patient enthält für jedes Exon einen Expressionswert (der Dokumentennvektor ist somit vollständig besetzt).
Um eine Spärlichkeit der Vektoren zu erreichen, wurde ein Filter für die Expressionswerte definiert. Alle Werte, die einen vorgebenden Schwellwert unterschreiten, wurden auf null gesetzt. Somit wurden einzelne Exons je nach Ausprägung für einzelne Patienten „aus“ geschaltet.
In einem ersten Versuch wurden die Daten zunächst auf das Intervall $[0, 1]$ nor-

malisiert. Anschließend wurde der Filter in 0,1er-Schritten von 0,0 auf 1,0 erhöht, um zu sehen, wie sich die Spärlichkeit der Vektoren verändert. Problematisch bei dieser Vorgehensweise ist die anschließende Interpretation:

- Ein Expressionswert eines Exons wird als Anzahl des Wortvorkommens interpretiert, somit würde ein Wort (Exon) bei dieser Transformation maximal einmal pro Patient (Dokument) vorkommen.
- Die Daten nicht zu transformieren und die ursprünglichen Expressionswerte (Wertebereich 1-14) als Anzahl des Wortvorkommens zu interpretieren, stellt sich ebenfalls als problematisch dar. Nicht für jedes Exons erstrecken sich die Expressionswerte in einem Wertebereich von 1 bis 14, so dass bei einer Filterung (zur Erzwingung einer Spärlichkeit) in 1ner Schritten ab einem bestimmten Wert für ein Exon alle Expressionswerte der Patienten auf Null gesetzt wurden. So wurde zwar eine Dimensionsreduktion erreicht, aber die Spärlichkeit der Vektoren nur gering herabgesetzt.
- Um die erwähnten Probleme zu vermeiden, wurden die Daten in ein Intervall von [0,10] transformiert und eine Filterung in 1ner Schritten von 1 bis 10 vorgenommen. Im Diagramm 7.5 ist die Anzahl der „angeschalteten“ Exons pro Schwelle dargestellt.

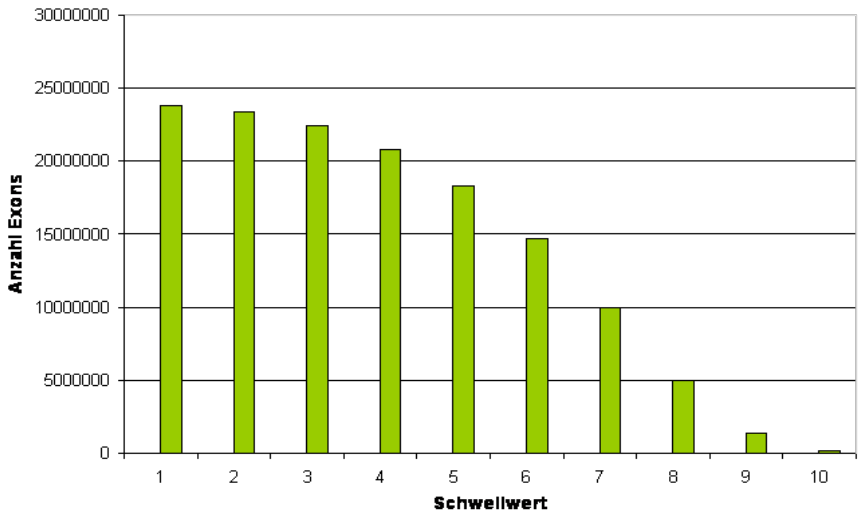


Abbildung 7.5.: Spärlichkeit

Bei 184.985 Exons und 131 Patienten können maximal 24233025 ($184.985 \cdot 131$) Exons „angeschaltet“ sein. Bei einem Schwellwert von 5 sind noch ca. 75% der Exons „angeschaltet“, bei einem Schwellwert von 7 noch ca. 40%. Jedes Exon

kann durch die Skalierung maximal eine Ausprägung von 10 aufweisen. Die Werte enthalten Nachkommastellen; zur weiteren Verarbeitung wurden die Werte gerundet.
Die durchschnittliche Euklidische Länge der Dokumentenvektoren (Patientenvbektoren) pro Schwelle ist im Diagramm 7.6 abgebildet.

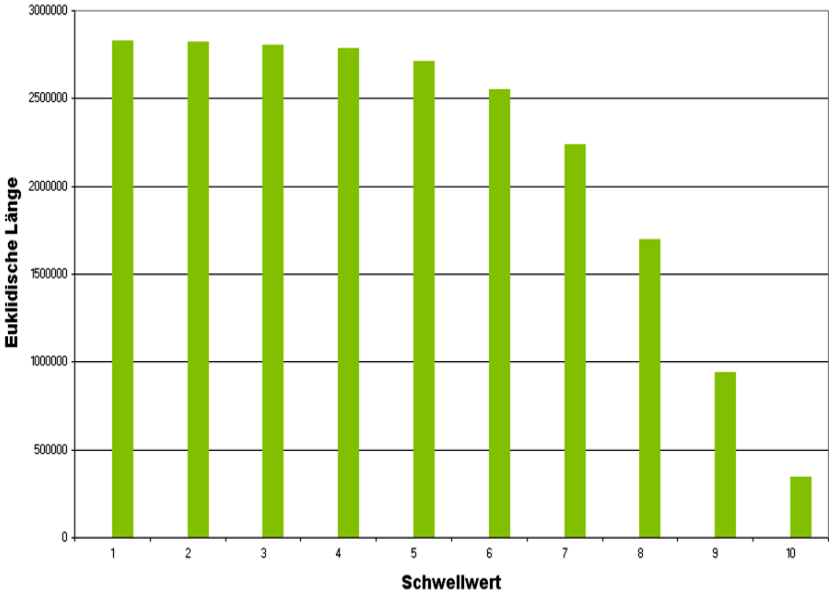


Abbildung 7.6.: durchschnittliche Euklidische Länge

Die weiteren Schritte wurden für alle Schwellwerte getestet. Exemplarisch werden hier die Ergebnisse für einen Schwellwert von 7 erläutert. Um die Daten als TCat-Modell modellieren zu können, ist eine Einteilung der Merkmale anhand ihrer Termfrequenzen in hoch- ,mittel- und niedrigfrequente Merkmale nötig. Die Termfrequenz eines Merkmals X_i berechnet sich als Summe der Termfrequenzen dieses Merkmals bei allen Patienten unabhängig von der Klassenzugehörigkeit:

$$TF(X_i) = \sum_{j=1}^{131} TF(X_i, C_j) \tag{7.8}$$

Anhand dieser berechneten Termfrequenzen werden die Merkmale absteigend sortiert und jedes Merkmal erhält eine Rangzahl. Die Rangzahl 1 bekommt das Merkmal mit der höchsten Termfrequenz. Dieses Merkmal ist Exon 3.453.742 mit einer Termfrequenz von 1226. Den letzten Rang teilen sich 1.394 Exons mit einer Termfrequenz von 10.
Die Einteilung in hoch-, mittel- und niedrigfrequente Merkmale ist wie folgt:

- hochfrequente Merkmale haben eine Termfrequenz zwischen 1.051-1.226

- mittelfrequente Merkmale haben eine Termfrequenz zwischen 525-1.050
- niedrigfrequente Merkmale haben eine Termfrequenz zwischen 0-524

Um die Elemente der drei Klassen in positive, negative und irrelevante Merkmale aufzuteilen, wird die „Odds Ratio“ (Quotenverhältnis) verwendet. Diese ist wie folgt definiert:

Definition 7.4 (Odds Ratio)
Für ein Ereignis A berechnet sich die Odds aus der Wahrscheinlichkeit, dass das Ereignis eintritt und der Wahrscheinlichkeit, dass es nicht eintritt (Gegenwahrscheinlichkeit):

$$\frac{P(A)}{1 - P(A)} \tag{7.9}$$

dabei ist P die Wahrscheinlichkeit, dass das Ereignis A eintritt.

Für zwei Ereignisse A_1 und A_2 mit den Wahrscheinlichkeiten $P(A_1)$ und $P(A_2)$ beschreibt die Odds Ratio das Verhältnis der Odds von Ereignis A_1 und Ereignis A_2:

$$OddsRatio = \frac{\frac{P(A_1)}{1 - P(A_2)}}{\frac{P(A_1)}{1 - P(A_2)}} \tag{7.10}$$

Für die Anwendung der Odds Ratio auf die vorliegenden Daten bedeutet dies:
A_1 ist das Ereignis, dass bei einem Patientenprofil der Klasse EFS das entsprechende Exon exprimiert ist. A_2 steht für das Ereignis, dass bei einem Patientenprofil der Klasse Event das Exon exprimiert ist. Die Odds Ratio für ein Exon berechnet sich aus den Odds des Exons für die Klasse EFS und die Klasse Event:

$$\frac{P(A_1)}{1 - P(A_1)} = \frac{\|n \in N \mid \text{Klasse} = \text{EFS und } TF(X_i) > 0\|}{\|n \in N \mid \text{Klasse} = \text{EFS und } TF(X_i) = 0\|} \tag{7.11}$$

$$\frac{P(A_2)}{1 - P(A_2)} = \frac{\|n \in N \mid \text{Klasse} = \text{Event und } TF(X_i) > 0\|}{\|n \in N \mid \text{Klasse} = \text{Event und } TF(X_i) = 0\|} \tag{7.12}$$

wobei N die Anzahl aller Beispiele (131) ist.
Die Odds Ratio kann nicht berechnet werden, wenn eine Null in der Berechnung vorkommt (Division durch Null). Um die Odds Ratio dennoch berechnen zu können, wurde bei Vorkommen einer Null eine Stetigkeitskorrektur von 0,5 verwendet.

Die Zuteilung der Indikatoren erfolgt in dieser Arbeit wie folgt:

- positive Indikatoren: Odds Ratio > 2

- negative Indikatoren: Odds Ratio $< 0{,}5$
- irrelavante Indikatoren: Odds Ratio $\geq 0{,}5$ und ≤ 2

Durch die Berechnung der Odds Ratio können die Merkmale den Gruppen als positive, negative und irrelevante Indikatoren wie folgt zugeteilt werden:

- Gruppe hochfrequenter Merkmale: 1.633 positive Indikatoren und 85 negative Indikatoren
- Gruppe mittelfrequenter Merkmale: 11.527 positive Indikatoren und 7.626 negative Indikatoren
- Gruppe niedrigfrequenter Merkmale: 17.168 positive Indikatoren und 26.871 negative Indikatoren
- 120.075 Merkmale wurden als irrelevante Indikatoren bewertet

Die gesamte Dokumentenlänge (Anzahl aller Exon-Vorkommen) liegt bei 81.033.215. Ein durchschnittlicher Patient (Dokument) enthält ca. 618.574 Exons (Wörter), wovon ca. 76.443 unterschiedliche Exons (Wörter) sind. Bei einem vollständig besetzten Vektor könnte ein Patient maximal 1.849.850 Exons (Wörter) enthalten (ein Exon kann durch die Normalisierung maximal 10 mal pro Patient enthalten sein).
Für einen durchschnittlichen Patienten aus der Klasse EFS (positive Klasse) stammen 2,3% der Exons aus der Klasse der hochfrequenten, positiven Indikatoren, während bei einem durchschnittlich negativen Dokument (Patient mit Klassenzugehörigkeit Event) etwa 2,1% der Exons aus Klasse der hochfrequenten positiven Indikatoren stammen. Die weiteren relativen Vorkommenshäufigkeiten der Exons sind Tabelle 7.4 zu entnehmen.

	hochfrequent		**mittelfrequent**		**niedrigfrequent**		
	1633 pos.	**85 neg.**	**11527 pos.**	**7626 neg.**	**17168 pos.**	**26871 neg.**	**120075 Rest**
EFS	2,3%	0,1 %	12,2%	5,9%	6,8%	5,1%	67,6%
Event	2,1%	0,1%	8,7%	7,9%	2,9%	10,9%	67,4%

Tabelle 7.4.: Zusammensetzung durchschnittlich positiver Patienten (EFS) und negativer Patient(Event)

Aus Tabelle 7.4 lässt sich durch Anwendung der Prozentzahlen auf die durchschnittliche Anzahl von Exons ein TCat-Konzept erstellen:

```
TCat ( [14103 : 13010 : 1633] ,   [688 : 710 : 85]                 # hochfrequent
       [75734 : 53348 : 11527] , [36468 : 48971 : 7626]            # mittelfrequent
       [42114 : 18310 : 17168] , [31576 : 67121 : 26871]           # niedrigfrequent
       [417892 : 411047 : 120075]                                  # Rest
     )
```

Nach Theorem 1.4 ist der erwartete Generalisierungsfehler einer SVM nach dem Training auf n Beispielen beschränkt durch:

$$\frac{R^2}{n+1}\frac{a+2b+c}{ac-b^2}$$

Die Werte für a, b und c können aus dem TCat-Konzept berechnet werden:

- $a = \sum_{i=1}^{7} \frac{p_i^2}{f_i} = 2394125{,}70$
- $b = \sum_{i=1}^{7} \frac{p_i n_i}{f_i} = 2278209{,}20$
- $c = \sum_{i=1}^{7} \frac{n_i^2}{f_i} = 2307032{,}80$

R^2 ist die maximale Euklidische Länge eines Merkmalsvektors in den Trainingsdaten. R^2 lässt sich abschätzen, wenn die Termfrequenzen dem generalisierten Zipfschen Gesetz folgen. Nach Theorem 1.4 ist:

$$R^2 = \sum_{r=1}^{d} \left(\frac{c}{(k+r)^{\phi}} \right)^2$$

In Grafik 7.7 ist die Anpassung der Mandelbrotverteilung (grüne Linie) dargestellt. Die Termfrequenz ist gegen den Rang aufgetragen (rote Linie). Trifft man die Annahme, dass alle Patienten (Dokumente) hinreichend homogen sind, gilt das generalisierte Zipfsche Gesetz auch für jeden einzelnen Patienten und somit ergibt sich für R^2:

$$\sum_{i=1}^{76443} TF_i = 2474596{,}04 \geq R^2$$

In Theorem 1.4 eingesetzt ergibt sich, dass der erwartete Generalisierungsfehler nach dem Training auf 131 Beispielen bei unter 52% liegt.
Der Versuch, anstelle der L2-Norm[2] (Euklidische Norm) die L1-Norm[3] zur Berechnung von R^2 zu verwenden, erbrachte eine bessere Abschätzung des Generalisierungsfehlers (unter 10%). Diese Abschätzung muss jedoch als zu optimistisch

[2] Wurzel aus der Summe aller quadrierten Werte.
[3] Summe aller Koeffizienten-Beträge ohne Diagonalelemente.

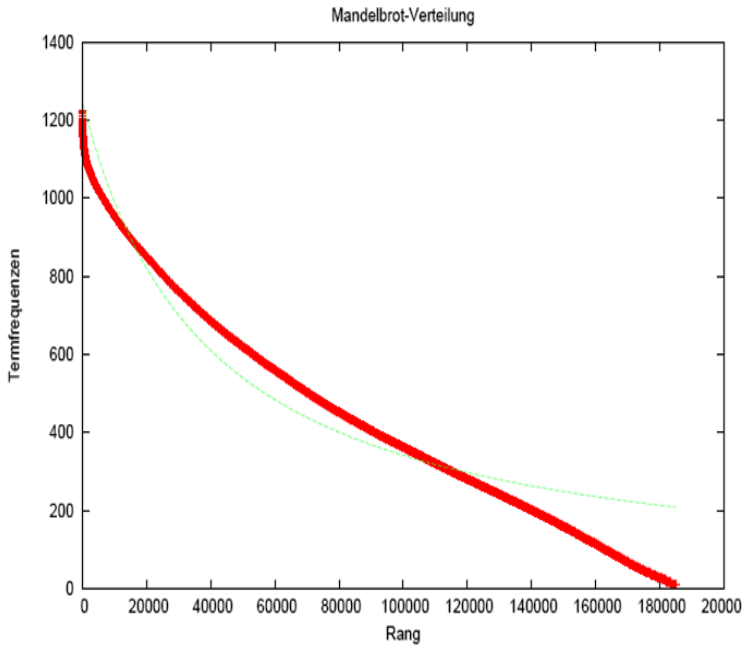


Abbildung 7.7.: Mandelbrotverteilung

angesehen werden und kann durch die Experimente nicht bestätigt werden. Die SVM erzielte auf den transformierten Daten (TF-Repräsentation) eine Accuracy von 77,12% (+/- 6.78). Auch eine Transforamtion in eine TFIDF-Repräsentation brachte keine besseren Lernergebnisse.

Aus dem T-Cat-Konzept lässt sich ablesen, dass das Verhältnis von p_i und n_i für die einzelnen Häufigkeitsklassen relativ gleich verteilt ist (siehe hierzu auch Tabelle 7.4). Dies spiegelt die Schwierigkeit der Lernaufgabe wider. Durch den berechneten Vorhersagefehler von über 50% wird die Schwierigkeit bestätigt. Eine Zuordnung der Patienten anhand der Anzahl ihrer Merkmale aus den ermittelten Häufigkeitsklassen ist somit nicht gegegeben.
Die Interpretation der Expressionswerte als Worte und der damit verbundenen Übertragung des vorliegenden Problems auf die Textklassifikation muss sehr kritisch betrachtet werden. Problematisch zeigte sich die Überführung der vollständigen Merkmalsvektoren in spärliche Vektoren. Durch die Filterung von Expressionswerten, die eine vorgegebene Schwelle nicht erreichen, werden nur noch Exons betrachtet, die sehr hochexprimiert sind (im Übertragenden Sinne nur noch Wörter berücksichtigt, die häufig vorkommen). Dieser Ansatz entspricht eigentlich nicht der Textklassifikation, in der sehr häufig vorkommende Wörter als Stoppwörter betrachtet und nicht berücksichtigt werden. Auch die Interpretation der Ausprägungen der Expressionswerte als Häufigkeitsvorkommen zu betrachten, ist

problematisch bei der Interpretation der erstellten TCat-Konzepte.

8. Zusammenfassung und Ausblick

In diesem Kapitel werden die Ergebnisse dieser Arbeit zusammengefasst und anschließend ein Ausblick auf mögliche weitere Untersuchungen gegeben.

8.1. Zusammenfassung

Im Rahmen der vorliegenden Arbeit wurden Daten von Patienten, die an der Krebsart „Neuroblastom“ erkrankt sind, anhand von Verfahren und Methoden aus dem Bereich des maschinellen Lernens und der Statistik untersucht. Bei diesen zu untersuchten Daten, die mit Hilfe der Micorarray-Technologie gewonnen wurden, handelt es sich um sog. *Exon-Expressionswerte* (Exons sind Untereinheiten von Genen), während in bisher aus der (Bio)Informatik bekannten Untersuchungen *Gen-Expressionswerte* verarbeitet wurden.

Ziel der Untersuchungen war die Beantwortung der beiden Fragestellungen

- Kann ein (guter) Klassifikator zur Vorhersage des Krankheitsverlaufes gefunden werden?

- Können signifikante Exons ermittelt werden, die biologische Rückschlüsse auf die Überlebensprognose der Patienten zulassen und die für die weitere Neuroblastom-Forschung von Bedeutung sind?

Zum Verständnis dieser Aufgabenstellung war zunächst eine Auseinandersetzung mit den Eigenschaften der Krebserkrankung und den Verfahren zur Gewinnung von Exon-Expressionswerten notwendig (siehe Kap. 2 und 3). Danach wurden für die Untersuchungen passende Lernverfahren (siehe Kap. 4) und Methoden zur Merkmalsauswahl (siehe Kap. 5) aus dem Bereich des maschinellen Lernens und der Statistik recherchiert. Für die Experimente sollte die Lernumgebung *RapidMiner* verwendet werden; für die nicht in RapidMiner implementierten Methoden wurden entsprechende Operatoren realisiert.
Im Rahmen unterschiedlicher Experimentreihen wurden diese Verfahren und Methoden auf die Neuroblastom-Daten angewendet.

Experimente ohne Merkmalsauswahl

Zunächst wurden Untersuchungen zur Normalisierung durchgeführt, aus denen hervorging, dass auf eine Normalisierung der Daten verzichtet werden kann. Als Ergebnis der anschließenden Experimente zur Klassifikation ist festzuhalten, dass ohne eine Merkmalsauswahl die Stützvektormethode (SVM) mit ca. 84% die besten Ergebnisse liefert.

Experimente mit Merkmalsauswahl

Durch den Einsatz von Ensembles zur Merkmalsauswahl wurden bessere Resultate erlangt als mit einer einfachen Merkmalsauswahl. Dabei konnte interessanterweise die SVM ihr gutes Ergebnis ohne eine Merkmalsauswahl nicht übertreffen. Für die o.g. **erste Fragestellung** bedeutet dies, dass eine Klassifikation der Patienten anhand ihrer Expressionswerte prinzipiell möglich ist; die erreichte Vorhersagegenauigkeit kann jedoch aus biologischer Sicht (noch) nicht als zufriedenstellend bewertet werden. Wird ausschließlich ein guter Klassifikator gesucht und ist die dabei zu erhebende Anzahl von Merkmalen pro Patient von untergeordnetem Interesse, sollte die SVM auf alle Merkmale angewendet werden. Liegt das Interesse dagegen an einer kleineren Menge von Merkmalen, so erzielen die Verfahren Naive Bayes und kNN auf einer ausgewählten Merkmalsmenge die besten Ergebnisse.
Bei den Methoden zur Merkmalsauswahl erwiesen sich die Methoden *Relief* und *SAM* als eine gute Wahl. Mit den von diesen Methoden bewerteten ausgewählten Merkmalen konnten im Vergleich zu den anderen Auswahlmethoden bessere Ergebnisse erzielt werden. Auch bei den Experimenten zur Robustheit lieferten diese beiden Methoden die besten Ergebnisse.
Erwartungsgemäß führte eine zu kleine Anzahl von Merkmalen zu schlechteren Lernergebnissen, so dass die in der Einleitung formulierte Hypothese - nicht ein bestimmtes Exon determiniert die Klassifikation, sondern das Zusammenspiel einer Anzahl von Exons ist für die Zugehörigkeit zu einer Klasse verantwortlich - durch die Experimente insoweit bestätigt wurde, dass durch eine höhere Anzahl von Exons (Merkmalen) ein besseres Lernergebnis erreicht werden kann.

Die Beantwortung der o.g. **zweiten Fragestellung** erwies sich als deutlich schwieriger. Sie kann gleichgesetzt werden mit der Suche nach einer robusten Merkmalsauswahl. Von einer robusten Auswahl eines Merkmals wird gesprochen, wenn dieses Merkmal von mehreren Methoden hoch bewertet wird oder in variierenden Datensätzen immer oder häufig enthalten ist. Durch eine robuste Auswahl wird die Vermutung, dass eine genetische Grundlage für die Ausprägung dieses Merkmals zwischen den Patientengruppen vorliegt, bestärkt.
Für keine der Auswahlmethoden konnte bei einer einfachen Auswahl eine aus-

reichend hohe Robustheit erreicht werden. Dagegen stellte sich heraus, dass für eine relativ robuste Auswahl von Merkmalen hinsichtlich der Fragestellung zur Entdeckung signifikanter Exons das Ensemble und als Bewertungsfunktion SAM oder Relief verwendet werden sollte. SAM zeigte bei der Auswahl von Merkmalen mit dem Ensemble eine Robustheit von 0,651%. Unter den anhand von SAM ausgewählten Merkmalen fanden sich mehrere bereits aus biologichen Untersuchungen bekannte relevante Gene. Daher liegt die Vermutung nahe, dass noch weitere interessante Exons innerhalb dieser Menge vorhanden sind, die Einfluss auf die Überlebensprognose haben könnten. Auf die Empfehlung, diese Exons biologisch weiter zu untersuchen, liegt noch keine biologische Beurteilung vor.

In Kapitel 7 dokumentiert, wurde im Zusammenhang mit der Relevanz bestimmter Merkmale den interessanten Aspekten nachgegangen, ob die durch die Merkmalsauswahl selektierten Merkmalsmengen Übereinstimmungen aufweisen und ob die Merkmale innerhalb der Mengen korreliert sind. Hinsichtlich der Übereinstimmungen der Merkmalsmengen wiesen die verschiedenen Mengen gemeinsame Merkmale auf. Allen Mengen war gemein, dass sie in 4 Exons, die zu einem Gen gehören, übereinstimmen. Innerhalb der einzelnen Mengen konnten Exons bzw. deren zugehörige Gene ermittelt werden, die bereits von biologischer Seite als relevant für die Neuroblastomforschung eingeschätzt wurden und somit in Verbindung zur Prognose der Patienten gebracht werden können. So wurden beispielsweise Exons des Gens „PLXNA4“ von allen Auswahlmethoden als signifikant bewerten. Dieses Gen ist biologisch bekannt und wird meistens mit einer guten Prognose der Erkrankung assoziiert.
Die Untersuchung der Merkmalskorrelationen führte zu dem Ergebnis, dass innerhalb der Merkmalsmengen je nach Auswahlmethode mehr oder weniger korrelierte Merkmale enthalten sind. Es zeigte sich aber, dass diese korrelierten Merkmale für das Lernen in ihrer Kombination keinen Nutzen bringen (dieses Ergebnis muss als sehr optimistisch angesehen werden, da nur eine geringe Anzahl manuell selektierter Merkmale untersucht wurde). Auch aus biologischer Sicht sind die Korrelationen dieser Merkmale eher uninteressant, da die korrelierten Merkmale immer einem gemeinsamen Gen angehören. Aus dieser Erkenntnis könnte sich jedoch ein weiterer interessanter Ansatzpunkt zur Analyse der Daten ergeben (siehe Ausblick).

Weitergehende Untersuchungen

Vergleiche der unterschiedlichen **Patientenprofile** führten zu dem Ergebnis, dass die Patienten einer Klasse in ihren Expressionswerten (teilweise) sehr heterogen sind. Hieraus lässt sich ableiten, dass eine Klassifikation der Patienten anhand ihrer Expressionswerte nicht ohne weiteres möglich ist, was sich in den erzielten

Lernergebnissen widerspiegelt.

Abschließend wurde untersucht, ob sich die vorliegenden Daten anhand eines bekannten statistischen Lernmodells (**TCat-Modell**) modellieren lassen und ob die statistischen Eigenschaften der Textklassifikation auch für die vorliegenden Daten gelten. Hierbei erwies sich die Darstellung der vorliegenden Neuroblastom-Daten als TCat-Konzept als problematisch. Bei einer gelungenen Modellierung könnte man folgern, dass es wie bei Texten nur wenig unwichtige Merkmale gibt und daher die Klassifikationsgüte durch eine Reduktion der Merkmalsmenge sinkt. Ob sich diese Aussage auf die Neuroblastom-Daten übertragen lässt, kann nicht eindeutig beantwortet werden. Dafür sprechen würde allerdings, dass die SVM auf allen Merkmalen bessere Ergebnisse erlangt als auf einer ausgewählten Merkmalsmenge.

8.2. Ausblick

Für die Überführung der Bilddaten in Expressionswerte (Low-Level-Analyse, siehe Kap. 3.3.1) existieren zahlreiche Methoden. Da die vorliegenden Daten ausschließlich mit dem RMA-Verfahren (siehe Kap. 3.3.2) vorverarbeitet wurden, dürfte es hier interessant sein, alternative Techniken zur Low-Level-Analyse für die Daten einzusetzen und die dann erreichten Lernergebnisse miteinander zu vergleichen.

Alle Patientendaten, die in dieser Arbeit verwendet wurden, wurden mit dem selben Arraytyp (Affymetrix GeneChip®) gewonnen. In weiteren Untersuchungen sollten Daten von unterschiedlichen Arraytypen einbezogen und dann untersucht werden. Hierzu könnten die Methoden zur Merkmalsauswahl eingesetzt werden, um Übereinstimmungen zwischen den selektierten Mengen verschiedener Arraytypen zu ermitteln. Dies könnte ggf. zu interessanten Entdeckungen signifikanter Exons/Gene führen.

Mit biologisch-/genetischem Hintergrundwissen könnte eine Vorauswahl von zu betrachtenden Exons getroffen werden. So könnten z.B. nur Exons von Genen betrachtet werden, von denen eine biologische Relevanz bekannt ist. Ein anderer Aspekt wäre, Exons bekannter Gene zu untersuchen und zu überprüfen, ob Zusammenhänge (z.B. über die Korrelation) mit anderen Exons bestehen, um so neue Erkenntnisse zu gewinnen.

In Kapitel 7.1.2 wurden Korrelationen innerhalb ausgewählter Mengen betrachtet. Hierbei zeigte sich, dass korrelierte Merkmale zu Gruppen zusammengefasst werden und diese Merkmale innerhalb dieser Gruppen gemeinsamen Genen zuge-

ordnet werden konnten. Da diese Analyse nicht auf allen Merkmalen durchgeführt wurde, ist eine Analyse für alle Merkmale ein interessanter Ansatzpunkt. Hierbei könnte dann der Frage nachgegangen werden, ob es für das „Lernen“ einen Informationsgewinn bringt, die Exon-Ebene zu betrachten.

A. Anhang

A.1. Übereinstimmungen der Merkmalsmengen

Die nachfolgenden Tabellen enthalten

- in der ersten Spalte die entsprechende Exon-Id (probeset-id)
- in der zweiten Spalte die Genzugehörigkeits-Id (transcript-id) für das jeweilige Exon
- in der dritten Spalte die entsprechenden Gen-Informationen (Namen, Beschreibung des Genprodukts und Lokalisationsort)

Die Zuordnung der Exon-Ids mit den entsprechenden Informationen erfolgte mit der Software R [57]. Exons, die einem gemeinsamen Gen zugeordnet werden können, sind farbig hinterlegt.

Folgende Übereinstimmungen werden aufgelistet:

- SAM, Welch-Test und t-Statistik (Tabelle A.1)
- SVM-Gewichtung und Relief (Tabelle A.2)
- SAM, Welch-Test, t-Statistik, Relief und SVM-Gewichtung (Tabelle A.3)

A.2. Anhand von SAM ausgewählte Merkmalsmenge

In der Tabelle A.4 sind die 100 auf Basis von SAM ausgewählten Merkmale des Ensembles dargestellt.

probeset-id	transcript-cluster-id	gene-assignment
2327561	2327542	NR-003109 // TRSPAP1 // tRNA selenocysteine associated protein 1 // 1p35.3 // 54952
2361778	2361761	NM-001007792 // NTRK1 // neurotrophic tyrosine
2394521	2394478	NM-015557 // CHD5 // chromodomain helicase DNA binding protein 5 // 1p36.31 // 26038
2394537	2394478	NM-015557 // CHD5 // chromodomain helicase DNA binding protein 5 // 1p36.31 // 26038
2834753	2834743	NM-000024 // ADRB2 // adrenergic
2834754	2834743	NM-000024 // ADRB2 // adrenergic
2870999	2870964	NM-022140 // EPB41L4A // erythrocyte membrane protein band 4.1 like 4A // 5q22.2 // 64097
2871006	2870964	NM-022140 // EPB41L4A // erythrocyte membrane protein band 4.1 like 4A // 5q22.2 // 64097
2871011	2870964	NM-022140 // EPB41L4A // erythrocyte membrane protein band 4.1 like 4A // 5q22.2 // 64097
2871049	2870964	NM-022140 // EPB41L4A // erythrocyte membrane protein band 4.1 like 4A // 5q22.2 // 64097
2975744	2975741	NM-003980 // MAP7 // microtubule-associated protein 7 // 6q23.3 // 9053
3025352	3025291	NM-144648 // LRGUK // leucine-rich repeats and guanylate kinase domain containing // 7q33 // 136332
3073483	3073267	NM-020911 // PLXNA4 // plexin A4 // 7q32.3 // 91584
3073484	3073267	NM-020911 // PLXNA4 // plexin A4 // 7q32.3 // 91584
3073485	3073267	NM-020911 // PLXNA4 // plexin A4 // 7q32.3 // 91584
3727828	3727787	NM-153228 // ANKFN1 // ankyrin-repeat and fibronectin type III domain containing 1 // 17q22 // 162282
3727830	3727787	NM-153228 // ANKFN1 // ankyrin-repeat and fibronectin type III domain containing 1 // 17q22 // 162282

Tabelle A.1.: Übereinstimmungen SAM, Welch-Test und t-Statistik

probeset-id	transcript-cluster-id	gene-assignment
2736160	2736060	NM-001510 // GRID2 // glutamate receptor
3020378	3020343	NM-000245 // MET // met proto-oncogene (hepatocyte growth factor receptor) // 7q31 // 4233
3073472	3073267	NM-020911 // PLXNA4 // plexin A4 // 7q32.3 // 91584
3073483	3073267	NM-020911 // PLXNA4 // plexin A4 // 7q32.3 // 91584
3073484	3073267	NM-020911 // PLXNA4 // plexin A4 // 7q32.3 // 91584
3073485	3073267	NM-020911 // PLXNA4 // plexin A4 // 7q32.3 // 91584
3073486	3073267	NM-020911 // PLXNA4 // plexin A4 // 7q32.3 // 91584
3797414	3797295	NM-173464 // L3MBTL4 // l(3)mbt-like 4 (Drosophila) // 18p11.31 // 91133

Tabelle A.2.: Übereinstimmungen SVM und Relief

probeset-id	transcript-cluster-id	gene-assignment
2736160	2736060	NM-001510 // GRID2 // glutamate receptor
3073483	3073267	NM-020911 // PLXNA4 // plexin A4 // 7q32.3 // 91584
3073484	3073267	NM-020911 // PLXNA4 // plexin A4 // 7q32.3 // 91584
3073485	3073267	NM-020911 // PLXNA4 // plexin A4 // 7q32.3 // 91584

Tabelle A.3.: Übereinstimmungen aller Merkmalsmengen

probeset-id	transcript-cluster-id	gene-assignment
2323000	2322957	NM-018125 // ARHGEF10L // Rho guanine nucleotide exchange factor (GEF) 10-like // 1p36.13 // 55160
2327561	2327542	NR-003109 // TRSPAP1 // tRNA selenocysteine associated protein 1 // 1p35.3 // 54952
2357195	2357193	AK000992 // GenBank // Homo sapiens cDNA FLJ10130 fis, clone HEMBA1003035. // chr1 // 100
2361761	2361761	NM-001007792 // NTRK1 // neurotrophic tyrosine kinase
2361771	2361761	NM-001007792 // NTRK1 // neurotrophic tyrosine kinase
2361776	2361761	NM-001007792 // NTRK1 // neurotrophic tyrosine kinase
2361777	2361761	NM-001007792 // NTRK1 // neurotrophic tyrosine kinase
2361778	2361761	NM-001007792 // NTRK1 // neurotrophic tyrosine kinase
2361779	2361761	NM-001007792 // NTRK1 // neurotrophic tyrosine kinase
2361781	2361761	NM-001007792 // NTRK1 // neurotrophic tyrosine kinase
2361783	2361761	NM-001007792 // NTRK1 // neurotrophic tyrosine kinase
2361784	2361761	NM-001007792 // NTRK1 // neurotrophic tyrosine kinase
2361785	2361761	NM-001007792 // NTRK1 // neurotrophic tyrosine kinase
2375731	2375706	NM-001001396 // ATP2B4 // ATPase
2375766	2375706	NM-001001396 // ATP2B4 // ATPase
2394480	2394478	NM-015557 // CHD5 // chromodomain helicase DNA binding protein 5 // 1p36.31 // 26038
2394482	2394478	NM-015557 // CHD5 // chromodomain helicase DNA binding protein 5 // 1p36.31 // 26038
2394487	2394478	NM-015557 // CHD5 // chromodomain helicase DNA binding protein 5 // 1p36.31 // 26038
2394521	2394478	NM-015557 // CHD5 // chromodomain helicase DNA binding protein 5 // 1p36.31 // 26038
2394530	2394478	NM-015557 // CHD5 // chromodomain helicase DNA binding protein 5 // 1p36.31 // 26038
2394537	2394478	NM-015557 // CHD5 // chromodomain helicase DNA binding protein 5 // 1p36.31 // 26038
2412634	2412624	NM-002867 // RAB3B // RAB3B
2412635	2412624	NM-002867 // RAB3B // RAB3B
2432011	2431886	NM-014644 // PDE4DIP // phosphodiesterase 4D interacting protein (myomegalin) // 1q12 // 9659
2438637	2438612	NM-014215 // INSRR // insulin receptor-related receptor // 1q21-q23 // 3645
2527918	2527895	NM-032726 // PLCD4 // phospholipase C
2560945	2560881	NM-024993 // LRRTM4 // leucine rich repeat transmembrane neuronal 4 // 2p12 // 80059
2560946	2560881	NM-024993 // LRRTM4 // leucine rich repeat transmembrane neuronal 4 // 2p12 // 80059
2585479	2585476	NM-002976 // SCN7A // sodium channel
2585481	2585476	NM-002976 // SCN7A // sodium channel
2585484	2585476	NM-002976 // SCN7A // sodium channel
2585486	2585476	NM-002976 // SCN7A // sodium channel
2585491	2585476	NM-002976 // SCN7A // sodium channel
2585493	2585476	NM-002976 // SCN7A // sodium channel
2585494	2585476	NM-002976 // SCN7A // sodium channel
2585496	2585476	NM-002976 // SCN7A // sodium channel
2585497	2585476	NM-002976 // SCN7A // sodium channel
2585498	2585476	NM-002976 // SCN7A // sodium channel
2585501	2585476	NM-002976 // SCN7A // sodium channel
2585504	2585476	NM-002976 // SCN7A // sodium channel
2585507	2585476	NM-002976 // SCN7A // sodium channel

2585517	2585476	NM-002976 // SCN7A // sodium channel
2585523	2585476	NM-002976 // SCN7A // sodium channel
2592623	2592598	NM-016192 // TMEFF2 // transmembrane protein with EGF-like and two follistatin-like domains 2 // 2q32.3 // 23671
2602777	2602770	NM-139072 // DNER // delta/notch-like EGF repeat containing // 2q36.3 // 92737
2603857	2603844	NM-004826 // ECEL1 // endothelin converting enzyme-like 1 // 2q36-q37 // 9427
2624677	2624639	NM-018398 // CACNA2D3 // calcium channel
2624681	2624639	NM-018398 // CACNA2D3 // calcium channel
2624684	2624639	NM-018398 // CACNA2D3 // calcium channel
2624686	2624639	NM-018398 // CACNA2D3 // calcium channel
2624695	2624639	NM-018398 // CACNA2D3 // calcium channel
2624696	2624639	NM-018398 // CACNA2D3 // calcium channel
2624699	2624639	NM-018398 // CACNA2D3 // calcium channel
2624719	2624639	NM-018398 // CACNA2D3 // calcium channel
2624730	2624639	NM-018398 // CACNA2D3 // calcium channel
2624740	2624639	NM-018398 // CACNA2D3 // calcium channel
2624743	2624639	NM-018398 // CACNA2D3 // calcium channel
2624745	2624639	NM-018398 // CACNA2D3 // calcium channel
2624773	2624639	NM-018398 // CACNA2D3 // calcium channel
2736160	2736060	NM-001510 // GRID2 // glutamate receptor
2771343	2771342	NM-004439 // EPHA5 // EPH receptor A5 // 4q13.1 // 2044
2771345	2771342	NM-004439 // EPHA5 // EPH receptor A5 // 4q13.1 // 2044
2797067	2796995	NM-021069 // SORBS2 // sorbin and SH3 domain containing 2 // 4q35.1 // 8470
2834753	2834743	NM-000024 // ADRB2 // adrenergic
2834754	2834743	NM-000024 // ADRB2 // adrenergic
2870999	2870964	NM-022140 // EPB41L4A // erythrocyte membrane protein band 4.1 like 4A // 5q22.2 // 64097
2871006	2870964	NM-022140 // EPB41L4A // erythrocyte membrane protein band 4.1 like 4A // 5q22.2 // 64097
2871011	2870964	NM-022140 // EPB41L4A // erythrocyte membrane protein band 4.1 like 4A // 5q22.2 // 64097
2871027	2870964	NM-022140 // EPB41L4A // erythrocyte membrane protein band 4.1 like 4A // 5q22.2 // 64097
2871029	2870964	NM-022140 // EPB41L4A // erythrocyte membrane protein band 4.1 like 4A // 5q22.2 // 64097
2871030	2870964	NM-022140 // EPB41L4A // erythrocyte membrane protein band 4.1 like 4A // 5q22.2 // 64097
2871032	2870964	NM-022140 // EPB41L4A // erythrocyte membrane protein band 4.1 like 4A // 5q22.2 // 64097
2871046	2870964	NM-022140 // EPB41L4A // erythrocyte membrane protein band 4.1 like 4A // 5q22.2 // 64097
2871049	2870964	NM-022140 // EPB41L4A // erythrocyte membrane protein band 4.1 like 4A // 5q22.2 // 64097
2909962	2909948	NM-003221 // TFAP2B // transcription factor AP-2 beta (activating enhancer binding protein 2 beta) // 6p12 // 7021
2909967	2909948	NM-003221 // TFAP2B // transcription factor AP-2 beta (activating enhancer binding protein 2 beta) // 6p12 // 7021
2926347	2926323	NM-004100 // EYA4 // eyes absent homolog 4 (Drosophila) // 6q23 // 2070
2926362	2926323	NM-004100 // EYA4 // eyes absent homolog 4 (Drosophila) // 6q23 // 2070
2926365	2926323	NM-004100 // EYA4 // eyes absent homolog 4 (Drosophila) // 6q23 // 2070
2926372	2926323	NM-004100 // EYA4 // eyes absent homolog 4 (Drosophila) // 6q23 // 2070
2926374	2926323	NM-004100 // EYA4 // eyes absent homolog 4 (Drosophila) // 6q23 // 2070
2926380	2926323	NM-004100 // EYA4 // eyes absent homolog 4 (Drosophila) // 6q23 // 2070
2975744	2975741	NM-003980 // MAP7 // microtubule-associated protein 7 // 6q23.3 // 9053
3025352	3025291	NM-144648 // LRGUK // leucine-rich repeats and guanylate kinase domain containing // 7q33 // 136332

3039342	3039247	NM-004080 /// DGKB /// diacylglycerol kinase
3073483	3073267	NM-020911 /// PLXNA4 /// plexin A4 /// 7q32.3 /// 91584
3073484	3073267	NM-020911 /// PLXNA4 /// plexin A4 /// 7q32.3 /// 91584
3073485	3073267	NM-020911 /// PLXNA4 /// plexin A4 /// 7q32.3 /// 91584
3141621	3141589	NM-000880 /// IL7 /// interleukin 7 /// 8q12-q13 /// 3574
3266289	3266279	NM-003054 /// SLC18A2 /// solute carrier family 18 (vesicular monoamine)
3266312	3266279	NM-003054 /// SLC18A2 /// solute carrier family 18 (vesicular monoamine)
3266314	3266279	NM-003054 /// SLC18A2 /// solute carrier family 18 (vesicular monoamine)
3266318	3266279	NM-003054 /// SLC18A2 /// solute carrier family 18 (vesicular monoamine)
3391664	3391653	NM-000795 /// DRD2 /// dopamine receptor D2 /// 11q23 /// 1813
3727828	3727787	NM-153228 /// ANKFN1 /// ankyrin-repeat and fibronectin type III domain containing 1 /// 17q22 /// 162282
3727830	3727787	NM-153228 /// ANKFN1 /// ankyrin-repeat and fibronectin type III domain containing 1 /// 17q22 /// 162282
3727843	3727787	NM-153228 /// ANKFN1 /// ankyrin-repeat and fibronectin type III domain containing 1 /// 17q22 /// 162282
3727865	3727787	NM-153228 /// ANKFN1 /// ankyrin-repeat and fibronectin type III domain containing 1 /// 17q22 /// 162282
3746605	3746574	NM-000304 /// PMP22 /// peripheral myelin protein 22 /// 17p12-p11.2 /// 5376

Tabelle A.4.: ausgewählte Merkmale SAM

A.3. Patientenvergleiche

A.3.1. Grafiken - SAM

In Abbildung A.1 sind die Abweichungen der Patienten zu allen anderen Patienten innerhalb der Klasse EFS dargestellt. Ein Punkt steht für einen Patienten und die Anzahl der Abweichungen seiner Expressionswerte zu den Expressionswerten der anderen Patienten bei den Vergleichen. Maximal kann ein Patient 8100 Abweichungen innerhalb der Vergleiche zu den anderen Patienten aufweisen.

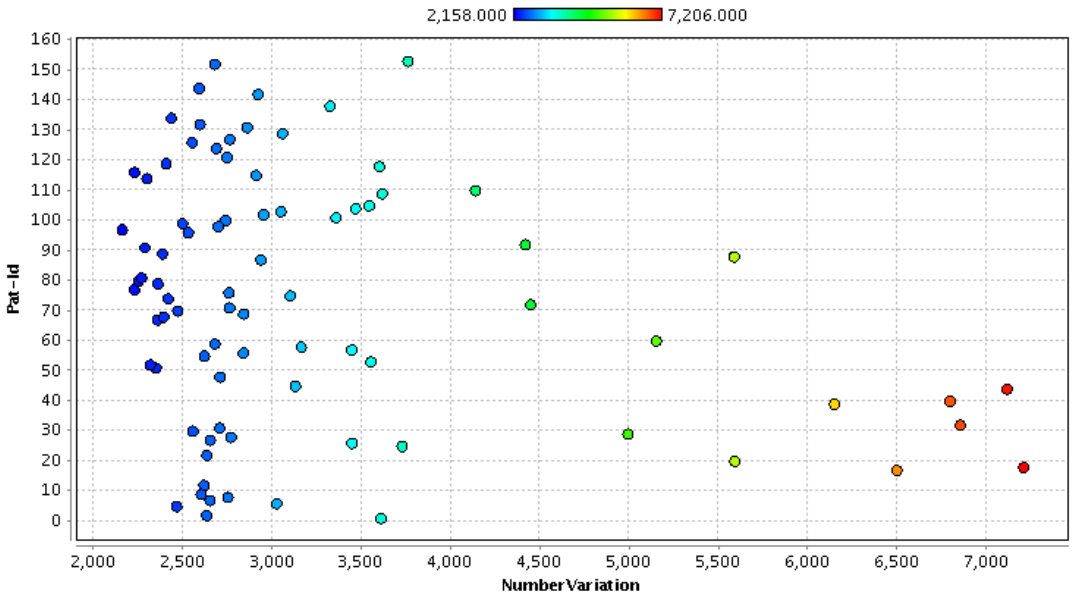


Abbildung A.1.: Anzahl der Expressionswert-Abweichungen einzelner Patienten zu allen anderen Patienten (SAM)

Grafik A.2 visualisiert die Anzahl der Abweichungen pro Exon bei den Vergleichen für die Klasse EFS. Ein Punkt stellt ein Exon und dessen Anzahl an Abweichungen dar.

Gefundene Gruppen auf Basis der SAM-Bewertung

In Tabelle A.5 sind die Exons und ihre Informationen dargestellt, die innerhalb der Klasse EFS „ähnlich“ exprimiert sind.

probeset-id	transcript-cluster-id	gene-assignment
2602777	2602770	NM-139072 // DNER // delta/notch-like EGF repeat containing // 2q36.3 // 92737
2624681	2624639	NM-018398 // CACNA2D3 // calcium channel
2624686	2624639	NM-018398 // CACNA2D3 // calcium channel
2624745	2624639	NM-018398 // CACNA2D3 // calcium channel
2870999	2870964	NM-022140 // EPB41L4A // erythrocyte membrane protein band 4.1 like 4A // 5q22.2 // 64097
2871027	2870964	NM-022140 // EPB41L4A // erythrocyte membrane protein band 4.1 like 4A // 5q22.2 // 64097
3266289	3266279	NM-003054 // SLC18A2 // solute carrier family 18 (vesicular monoamine)

Tabelle A.5.: innerhalb der Klasse EFS ähnlich exprimierte Exons

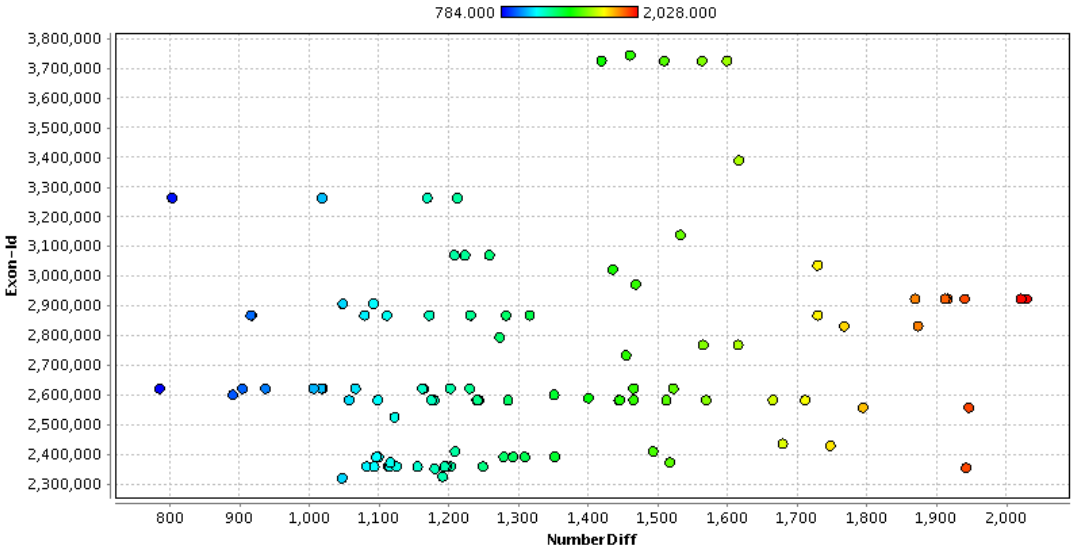


Abbildung A.2.: Anzahl der Abweichungen der Expressionswerte bei Patienten-Vergleichen für einzelne Exons (SAM)

B. Implementierung

Im Folgenden erfolgt eine kurze Beschreibung der Operatoren für RapidMiner, die im Rahmen dieser Arbeit entwickelt wurden. Die einzelnen Operatoren sind zu einem Plugin *Exon_ Analysis* zusammengefügt, wodurch eine leichte Einbindung in RapidMiner ermöglicht wird.
Da die Operatoren für die vorliegenden sehr hochdimensionalen Daten implementiert worden sind, wurde auf die Verwendung einer geeigneten Datenstruktur geachtet, um übermäßig lange Laufzeiten zu vermeiden.

B.1. Operatoren

Exon_SAM

Dieser Operator berechnet für jedes Merkmal einen „SAM-score“ nach der in Kap. 5.3.2 beschriebenen Formel. Über den Parameter *normalize_ weights* kann angegeben werden, ob die berechneten Werte normalisiert werden sollen.

Exon_WelchTest

Der Operator Exon_WelchTest berechnet für jedes Merkmal einen p-Wert auf Basis der in Kapitel 5.3.1 vorgestellten Formel. Der Operator verfügt über zwei Parameter. Über den Parameter *normalize_ weights* kann angegeben werden, ob die berechneten Werte normalisiert werden sollen; der Parameter *Benjamini-Hochberg-Adjustment* führt zur p-Werte Adjustierung (Kontrolle der „False Discovery Rate“).

Exon_t-Statistik

Durch diesen Operator wird für jedes Merkmal ein t-Wert nach der Formel in Kap. 5.3.1 berechnet. Die Möglichkeit zur Normalisierung der Werte ist über den Parameter *normalize_ weights* gegeben.

Exon_EnsemblesCombi

Dieser Operator kann verwendet werden, um die Ergebnisse einzelner Ensembles (Operator: EnsembleFeatureSelection) zu kombinieren. Die zu kombinierenden

Operatoren (EnsembleFeatureSelection) werden als innere Operatoren angehängt. Über den Parameter *normalize_ weights* können die Ergebniswerte normalisiert werden.

Exon_ExpressionsFilter

Bei diesem Operator kann über den Parameter *threshold* ein Schwellwert übergeben werden, so dass alle Attributausprägungen die unterhalb des Schwellwertes liegen auf Null gesetzt werden.

Exon_ExpressionsFilterBinaer

Dieser Operator entspricht dem Operator Exon_ExpressionsFilter. Der Unterschied zwischen beiden Operatoren besteht darin, dass nicht nur alle Attributausprägungen unterhalb des über den Parameter *threshold* eingestellten Schwellwertes auf Null gesetzt werden, sondern die übrigen Attributausprägungen (oberhalb des Schwellwertes) auf 1 gesetzt werden.

Exon_TCatModel

Zur Berechnung eines TCat-Modells kann der Operator Exon_TCatModel benutzt werden. Als Parameter benötigt der Operator dazu eine Pfadangabe, wo die berechneten Merkmalsmengen gespeichert werden sollen. Über den Parameter *indikator* können die Grenzwerte für die Häufigkeitsklassen eingestellt werden. Als Ausgabe speichert der Operator die ermittelten Merkmalsmengen als Textdokumente unter dem eingegebenen Pfad ab. In der GUI werden die Informationen zum Erstellen des TCat-Konzept und selbiges dargestellt. Des Weiteren werden die ermittelten Odds-Werte für jedes einzelne Merkmal ausgegeben.

Exon_Analysis_intra

Zur Untersuchung der Patientenprofile innerhalb einer Klasse wurde der Operator Exon_Analysis_intra implementiert. Hierbei wird für alle Vergleichskombinationen von zwei Patienten (Beispielen) für jedes einzelne Merkmal die Differenz ihrer Ausprägungen berechnet. Über den Parameter *threshold* wird bestimmt, bis zu welcher Differenz zwei Ausprägungen als „gleich“ akzeptiert werden. Als Ausgabe sind die einzelnen Patientenvergleiche und ihre Differenzen für jedes Merkmal, für jedes Merkmal die Anzahl der Abweichungen bei den Vergleichen, für jeden Patienten die Anzahl der Abweichungen seiner Attributausprägungen bei den Vergleichen mit den übrigen Patienten und eine Darstellung der Anzahl der Abweichungen zwischen den Patienten als Matrix zu betrachten. Zu beachten sei, das vor Anwendung des Operators eine Filterung der Patienten zu einer Klasse stattfinden sollte.

Exon_Analysis_extra

Dieser Operator ähnelt dem Operator Exon_Analysis_intra mit dem Unterschied, dass der Operator Exon_Analysis_extra die Patientenprofile zwischen zwei verschiedenen Klassen untersucht. Der Operator berücksichtigt die Klassenzugehörigkeit bei den Vergleichskombinationen von zwei Patienten (Beispielen). Ansonsten zeigt der Operator Exon_Analysis_extra die selben Ausgaben wie der Operator Exon_Analysis_intra, nur muss keine vorherige Filterung erfolgen. In der Matrix werden die Patienten der unterschiedlichen Klassen gegenüber gestellt.

Exon_StatistikExample

Mit diesem Operator lässt sich für jeden Patienten seine Euklidische Länge, die einfache Länge und die Anzahl der Merkmale mit einer Ausprägung größer Null anzeigen. Über den Parameter *threshold* kann bestimmt werden, ob vor der Berechnung Attributausprägungen auf Null gesetzt werden sollen.

Literaturverzeichnis

[1] Affymetrix: *http://www.affymetrix.com*.

[2] Alpaydin, Ethem: *Maschinelles Lernen*. Oldenbourg, 2008.

[3] Bayes, Thomas: *An Essay towards solving a Problem in the Doctrine of Chances*. 1763.

[4] Bellmann, R. E.: *Adaptive Control Processes*. Princeton University Press, 1961.

[5] Bererle, Christoph und Gabriele Kern-Isberner: *Methoden wissensbasierter Systeme: Grundlagen, Algorithmen, Anwendungen*. Viweg+Teubner Verlag, 4. Auflage, 2008.

[6] Berthold, F. und B. Hero: *Neuroblastomstudie NB 97, Studienprotokoll*. 1998.

[7] Berthold, F. und B. Hero: *Neuroblastoma - Current Drug Therapy Recommendations as part of the Total Treatment Approch*. Drugs, 59/6:1261–1277, 2000.

[8] Bioconductor: *http://www.bioconductor.org*.

[9] Biotechnology Online: *http://www.biotechnologyonline.gov.au*.

[10] Blum, A. und P. Langley: *Selection of relevant features and examples in machine learning*. Artifical Intelligence, 97(1-2):245–271, 1997.

[11] Breiman, Leo: *Bagging predictors*. Machine Learning,, 26:123–140, 1996.

[12] Brodeur, G.M.: *Neuroblastoma: Biological insights into a clinical enigma*. Nature Reviews Cancer, 3:203–216, 2003.

[13] Brodeur, G.M. et al.: *Revisions of the international criteria for neuroblastoma diagnosis, staging, and response to treatment*. Journal of Clinical Oncology, 11:1466–1477, 1993.

[14] Brown, M., W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet und T.S. Furey: *Knowlwdge-based analysis of microarray gene expression data by using support vector machines.* Proc. Natl. Acad. Sci., Seiten 262–267, 2000.

[15] Brown, Terence A.: *Genome und Gene.* Spektrum Akademischer Verlag, 2007.

[16] Burges, C.: *A Tutorial on Support Vector Machines for Pattern Recognition.* Kluwer Academic Publishers, 1998.

[17] Chih-Wei, Hsu, Chang Chih-Chung und Lin Chih-Jen: *A Pratical Guide to Support Vector Classification.* Technischer Bericht, Department of Computer Science, National Taiwan University, Taiwan,Taipei 106, 2008.

[18] Dietterich, Thomas G.: *Ensemble Methods in Machine Learning.* Seiten 1–15. Springer-Verlag, 2000.

[19] Dudoit, S., J. Fridlyand und T.P. Speed: *Comparision of discrimination methods for the classification of tumors using gene expression data.* J Am Stat Assoc, 97:77–87, 2002.

[20] Eckey, Hans-Freidrich, Reinhold Kosfeld und Martina Renger: *Multivariate Statistik.* Dr. Th. Gabler Verlag, 2002.

[21] Ein-Dor, Liat, Itai Kela, Gad Getz, David Givol und Eytan Domany: *Outcome signature genes in breast cancer: is there a unique set?* Bioinformatics, 21(2):171–178, 2004.

[22] Elpelt, Bärbel und Joachim Hartung: *Grundkurs Statistik.* Oldenbourg, 3. Auflage, 2004.

[23] Fix, E. und J. Hodges: *Discriminatory analysis: nonparametric discrimination: Consistency properties.* Technischer Bericht 21-49-004(4), USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[24] Fodor, S.P., J.L. Read, M.C. Pirrung, L. Stryer, A.T. Lu und D. Solas: *Light-directed, spatially addressable parallel chemical synthesis.* Sience, 251:767.

[25] Gautier, L., L. Cope, B.M. Bolstad und R.A. Irizarry: *affy-analysis of Affymetrix GeneChip data at the probe level.* Bioinformatics, 20(3):307–315, 2004.

[26] Golub, T., D. Slomin, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Dowing, M. Caligiuri, C. Bloomfield und E. Lander: *Molecular classification of cancer: class discovery and class prdiction by gene eypression monitoring.* Sience, 286:531–537, 1999.

[27] Guyon, I. `http://www.mathworks.com/`.

[28] Guyon, I., J. Weston, S. Barnhill und V. Vapnik: *Gene Selection for Cancer Classification using Support Vector Machines.* Machine Learning, 46(1/3):389–422, 2002.

[29] Hastie, Trevor, Robert Tibshirani und Jerome H. Friedman: *The Elements of Statistical Learning. Data Mining, Inference, and Prediction.* Springer Verlag, 2001.

[30] Hsu, C. W., C. C. Chang und C. J. Lin: *A Practical Guide to Support Vector Classification.* Technischer Bericht, Taipei, 2003.

[31] Irizarry, R.A., B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf und T.P. Speed: *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 4(2):249.

[32] Irizarry, R.A, Z. Wu und H.A. Jaffe: *Comparision of affymetrix genechip expression measure.* Bioinformatics, 22:789–794, 2006.

[33] Joachims, Thorsten: *Text Categorization with Support Vector Machines: Learning with Many Relevant Features.* In: *Proceedings of the European Conference on Machine Learning*, Seiten 137–142. Springer, 1998.

[34] Joachims, Thorsten: *The Maximum-Margin Approach to Learning Text Classifiers Methods, Theory, and Algorithms.* Doktorarbeit, Fachbereich Informatik, Universität Dortmund, 2001.

[35] Kambhampati, Dev: *Protein Microarray Technology.* Wiley-VCH, 1. Auflage, 2004.

[36] Kira, Kenji und Larry A. Rendell: *The Feature Selection Problem: Traditional Methods and a New Algorithm.* AAAI, Seiten 129–134, 1992.

[37] Kira, Kenji und Larry A. Rendell: *A practical approach to feature selection.* ML92: Proceedings of the ninth international workshop on machine learning, Seiten 249–256, 1992.

[38] Koller, D. und M. Sahami: *Toward optimal feature selection.* Technical Report, Stanford InfoLab, Februar 1996.

[39] Kononenko, Igor: *An adaption of Relief for attribute estimation in regression.* European Conference on Machine Learning, Seiten 171–182, 1994.

[40] Kononenko, Igor: *Estimating Attributes: Analysis and Extensions of RELIEF.* 1994.

[41] Mandelbrot, Benoit: *A note on a class of skew distribution functions: Analysis and critique of a paper by H. A. Simon.* Information and Control, 2:90–99, 1959.

[42] Marhöfer, Richard, Andreas Rohwer und P.M. Selzer: *Angewandte Bioinformatik: Eine Einführung.* Springer, 1. Auflage, 2003.

[43] Matthias Rudolf, Wiltrud Kuhlisch: *Biostatistik: Eine Einführung für Biowissenschaftler.* Pearson Verlag, 1. Auflage, 2008.

[44] Mercer, John: *Functions of positive and negative type and their connection with the theory of integral equations.* Philos. Trans. Roy. Soc. London, Seiten 415–446, 1909.

[45] Mierswa, Ingo, Michael Wurst, Ralf Klinkenberg, Martin Scholz und Timm Euler: *YALE: Rapid Prototyping for Complex Data Mining Tasks.* In: Eliassi-Rad, Tina, Lyle H. Ungar, Mirk Craven und Dimirios Gunopulos (Herausgeber): *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)*, Seiten 935–940. ACM Press, 2006.

[46] Mitchell, T.: *Machine Learning.* Mcgraw-Hill, 1997.

[47] Morik, Katharina: *Stützvektormethode.* Folien zur Vorlesung: Wissensentdeckung in Datenbanken, Universität Dortmund Fakultät Informatik Lehrstuhl für Künstliche Intelligenz, Mai 2009.

[48] Müller, Hans-Joachim und Thomas Röder: *Der Experimentator: Microarrays.* Spektrum Akademischer Verlag, 1. Auflage, 2004.

[49] On-Line Biology Book: *http://www.emc.maricopa.edu/faculty/farabee/BIOBK/BioBookTOC.html*, 2009.

[50] Saeys, Y., I. Inza und Larranaga P.: *A review of feature selection techniques in bioinformatics.* Bioinformatics, 13(19):2507–2517, 2007.

[51] Saeys, Yvan, Thomas Abeel und Yves Van de Peer: *Robust Feature Selection Using Ensemble Feature Selection Techniques.* In: *ECML/PKDD (2)*, Band 5212 der Reihe *Lecture Notes in Computer Science*, Seiten 313–325. Springer, 2008.

[52] Schapier, Robert E.: *The Strength of Weak Learnability.* Machine Learning„ Bd. 5, Nr. 2:197–227, 1990.

[53] Schilling, F.H., C. Six, F. Berthold, R. Erttmann, N. Fehse, B. Hero, G. Klein, J. Sander, K. Schwarz, J. Treuner, U. Zorn und J. Michaelis: *Neuroblastoma screening at one year of age.* N Engl J Med, 346:1047–1053, 2002.

[54] Simon, Herbert A.: *Why Should Machines Learn?* In: Michalski, Ryszard S., T. W. Mitchel und Tom M. Mitchell (Herausgeber): *Machine Learning: An Artificial Intelligence Approach*, Seiten 25–37. Springer, 1984.

[55] Simon, Richard: *Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n).* SIGKDD Explor. Newsl., 5(2):31–36, 2003.

[56] Singer, M. und P. Berg: *Gene und Genome.* Spektrum Akademischer Verlag, 1992.

[57] The R Project for Statistical Computing: *http://www.r-project.org/*.

[58] Thorell1, Kaisa, Annika Bergman, Helena Caren, Staffan Nilsson, Per Kogner, Tommy Martinsson und Frida Abel: *Verification of genes differentially expressed in neuroblastoma tumours: a study of potential tumour suppressor genes.* BMC Medical Genomics, 2:53, 2009.

[59] Tusher, V.G., R. Tibshirani und G. Chu: *Significance analysis of microarrays applied to the ionizing radiation response.* Statistics, Genetics, 98(9):5116–5121, 2001. The National Academy of Siences.

[60] Vapnik, Vladimir: *Estimation of Dependencies Based on Empirical Data.* Springer, 1982.

[61] Vapnik, Vladimir: *The Nature of Statistical Learning Theory (Information Science and Statistics).* Springer, 1999.

[62] Watson, J.D. und F.H.C. Crick: *A Structure for Deoxyribose Nucleic Acid.* Nature, 171:737.

[63] Welch, B.L.: *The generalization of students problem when several population variances are involved.* Biometrika, 34:28–35, 1947.

[64] Wikimedia Foundation, Inc.: *Maschinelles Lernen.* `http://de.wikipedia.org/wiki/Maschinelles_Lernen`, Juli 2009.

[65] Wikipedia: *http://de.wikipedia.org/wiki/Protein*, 2009.

[66] William S. Klug, Michael R. Cummings, Charlotte A. Spencer: *Genetik*. Pearson Studium, 8. aktualisierte Auflage, 2007.

[67] Witten, Ian H. und Eibe Frank: *Data Mining. Praktische Werkzeuge und Techniken für das maschinelle Lernen*. Hanser Verlag, 2001.

[68] Wittgenstein, Ludwig: *Philosophical Investigations*. Blackwell, 2 Auflage, 1967.

[69] W.K. Purves, D. Sadava, G.H. Orians H.C. Heller: *Biologie*. Elsevier, Spektrum Akad. Verl, 7. Auflage, 2006.

[70] Wrobel, M., K. Morik und T. Joachims: *Maschinelles Lernen und Data Mining*. In: Schneeberger, J. (Herausgeber): *Handbuch der Künstlichen Intelligenz*, Seiten 517–597. Oldenbourg, 2000.

[71] Ye, Kai, K. Anton Feenstra, Jaap Heringa, Adrian P. IJzerman und Elena Marchiori: *Multi-RELIEF: a method to recognize specificity determining residues from multiple sequence alignments using a Machine-Learning approach for feature weighting*. Bioinformatics, 24(1):18–25, 2008.

[72] Zipf, Georg Kinsley: *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press Inc, 1949.