Datengestützte Regelgenerierung für die Alarmgebung im Online-Monitoring von Intensivpatienten

Dissertation

zur Erlangung des Grades
einer Doktorin der Naturwissenschaften
der Technischen Universität Dortmund

Der Fakultät Statistik
der Technischen Universität Dortmund

vorgelegt von
Wiebke Sieben

Dortmund 2008

1. Gutachter: Prof. Dr. Ursula Gather

2. Gutachter: JProf. Dr. Uwe Ligges

Tag der mündlichen Prüfung: 15. Januar 2009

Inhaltsverzeichnis

Kapitel 1

Motivation

Die Überwachung von Patienten auf der Intensivstation erfolgt in der Regel mit Hilfe von Patientenmonitoren, die − auf verschiedene Weise − Vitalparameter analysieren. Zurzeit gebräuchliche Geräte prüfen hauptsächlich, ob sich die gemessenen Vitalparameter innerhalb vom Pflegepersonal eingestellter Grenzen befinden. Ist dies nicht der Fall, so wird ein Alarm ausgelöst. Solche Alarmsysteme werden als Schwellwert-Alarmsysteme bezeichnet. Trotz einiger algorithmischer Fortschritte verursachen Schwellwert-Alarmsysteme eine große Zahl von Fehlalarmen (O'Carroll (1986), Lawless (1994), Koski et al. (1990), Tsien und Fackler (1997), Chambrin et al. (1999), Kuhls (2008)), aufgrund derer das Pflegepersonal den Alarmen weniger Bedeutung zumisst, erst spät auf sie reagiert oder die Alarme für einige Vitalparameter sogar vollständig deaktiviert. Dieser Umgang mit den Alarmen eines Systems, das eine hohe Fehlalarmrate aufweist, ist eine natürliche menschliche Reaktion, die in psychologischen Studien beobachtet und untersucht wurde (Edworthy und Hellier (2005)).

Die Entwicklung eines intelligenten Alarmsystems und seine Integration in Patientenmonitore soll das zuverlässige Erkennen und Alarmieren kritischer Gesundheitszustände bei einer gleichzeitigen Reduktion von Fehlalarmen ermöglichen. Das Herzstück eines solchen intelligenten Alarmsystems muss in verbesserten Alarmregeln bestehen, die anstelle der alten Alarmregeln eingesetzt werden. Analog zur Vorgehensweise von Ärzten und Pflegekräften, die bei der Bewertung des Gesundheitszustandes eines Patienten eine Vielzahl von Vitalparametern *in ihrer Kombination* berücksichtigen, sollte ein kritischer Zustand nicht allein dadurch charakterisiert werden, dass der Messwert eines Vitalparameters außerhalb von univariaten Schwellwerten liegt, sondern durch eine Kombination von Eigenschaften und Werten verschiedener Variablen.

Neue Alarmregeln aus medizinischen Lehrbüchern oder Experteninterviews abzuleiten und in Alarmsystemen zu verwenden, ist bisher nicht gelungen und erscheint auch weiterhin äußerst schwierig: „Ideal-Werte“ sind von Patient zu Patient und Situation zu Situation verschieden und die Wirkmechanismen im menschlichen Körper sind weder im gesunden noch im kranken Fall weit genug verstanden, um sie in statistische Modelle für die Alarmgebung zu übertragen. Aus diesem Grund werden in dieser Arbeit Möglichkeiten untersucht, die benötigten multivariaten Alarmregeln aus Daten mittels Klassifikationsverfahren zu generieren.

Die dazu notwendigen Daten sind in einer klinischen Studie des Teilprojekts C4 des Sonderforschungsbereichs (SFB) 475 am Universitätsklinikum Regensburg erhoben worden. Es liegen Zeitreihen sekundengenauer Aufzeichnungen von Messwerten verschiedener Vitalparameter von Patienten auf der Intensivstation der Klinik und Poliklinik für Innere Medizin I vor. Zusätzlich stehen die durch das verwendete Monitoringsystem erzeugten Informationen zu den ausgelösten Alarme und die vom Pflegepersonal eingestellten Schwellwerte zur Verfügung. Die Alarme wurden im Rahmen der klinischen Studie durch Ärzte annotiert, d.h. es wurde die klinische sowie die technische Validität der Alarme bewertet. Der Datensatz weist eine Vielzahl fehlender – weil nicht überwachter oder messbarer – Werte auf. Die Variablen sind diskret, kategoriell so wie metrisch skaliert.

Neben den aus der Datensituation begründeten Anforderungen an die zu verwendenden statistischen Verfahren ergeben sich eine Reihe von Anforderungen aus der Praxis. In Kapitel 2 werden vollständige Prozeduren zur Alarmgebung im intensivmedizinischen Monitoring vorgestellt und Teillösungen diskutiert. Die Güte von Alarmsystemen wird häufig durch ihre Sensitivität und Spezifität beschrieben. Unter Sensitivität wird in diesem Zusammenhang der Anteil richtig erkannter alarmrelevanter Situationen und unter Spezifität der Anteil richtig erkannter nicht alarmrelevanter Situationen verstanden. Die Definitionen von alarmrelevanten und nicht alarmrelevanten Situationen sind jedoch nicht allgemein festgelegt, so dass sie sich in vielen Studien unterscheiden, wodurch ein direkter Vergleich erschwert wird. Die Ansätze zur Alarmgebung werden dem gegenwärtigen Stand der Technik gegenübergestellt und Gründe, die die Umsetzbarkeit gewährleisten bzw. verhindern, ermittelt. Hieraus resultiert ein Anforderungskatalog an Alarmsysteme für die Überwachung von Intensivpatienten. Aus den Anforderungen an Alarmsysteme werden Anforderungen an die zu verwendenden statistischen Verfahren

abgeleitet. In Kapitel 3 werden die erhobenen Daten deskriptiv analysiert. Die Häufigkeit der Alarme und die Bewertung der Situationen, in denen sie ausgelöst wurden, geben Aufschluss über die Defizite eines herkömmlichen Alarmsystems. Die Beschreibung der Datenvorverarbeitung folgt in Kapitel 4. Die Datenvorverarbeitung gilt als wichtiger Schritt in der Klassifikation, da durch sie häufig eine Verbesserung der Klassifikationsergebnisse erzielt werden kann. Die zur Klassifikation verwendete Methodik sowie deren Anpassung zur Lösung der vorliegenden Klassifikationsaufgabe werden in Kapitel 5 beschrieben. Da Fehlklassifikationen von alarmrelevanten Situationen schwerwiegende Folgen haben können, wird ein Verfahren entwickelt, das Alarmregeln mit wählbar hoher Sensitivität erzeugt. Die Eignung dieses Verfahrens wird in Kapitel 6 gezeigt. Hier wird zudem die Möglichkeit diskutiert, aus den beobachteten Zeitreihen abgeleitete charakterisierende Größen zur Verbesserung der Klassifikation mit einzubeziehen. Diese Charakteristika sollen, ähnlich wie in der klinischen Praxis von Ärzten üblich, den gesundheitlichen Verlauf kurz vor einem Alarm als Information in die Bewertung der Situation einbringen. Die Allgemeingültigkeit der erzeugten Alarmregeln wird anhand einer nach Patienten stratifizierten Stichprobe überprüft und diskutiert. Die Arbeit schließt mit einer Zusammenfassung der wichtigsten Ergebnisse in Kapitel 7.

Kapitel 2

Alarmsysteme in der Intensivmedizin

Die Analyse medizinischer Zeitreihen mit dem Ziel, eine Verbesserung in den Bereichen Diagnose, Prognose oder Therapie zu ermöglichen, ist Forschungsgegenstand unterschiedlicher Disziplinen. Die Eignung verschiedenster Verfahren der künstlichen Intelligenz wird in zahlreichen Arbeiten geprüft und aufgezeigt, allerdings ohne dass diese sich in der klinischen Praxis durchsetzen konnten (Hanson und Marshall (2001)). Einen Überblick über die Ansätze seitens der Informatik mit Schwerpunkt auf dem zeitlichen Aspekt der Datenanalyse bietet Augusto (2005). McIntosh (2002) stellt die historische Entwicklung des Monitorings auf Intensivstationen dar und benennt die Mustererkennung als zukünftiges Entwicklungsfeld. Eine ausführliche Zusammenstellung statistischer Ansätze und anderer Alarm-Algorithmen speziell für das intensivmedizinische Monitoring findet sich in Imhoff und Kuhls (2006).

Neben Lösungen zu Teilproblemen des Patientenmonitorings wie Ausreißer- oder Mustererkennung sind auch zahlreiche neue vollständige Alarmsysteme zur Überwachung in der Literatur vorgeschlagen worden. Nachfolgend werden beispielhaft einige Ansätze dargestellt.

2.1 Beispiele vorgeschlagener Alarmsysteme

Koski et al. (1991) und Sukuvaara et al. (1993) entwickeln ein dreistufiges System zur Überwachung des Patientenzustands auf Intensivstationen. Es besteht aus der Datenvorverarbeitung, der Entdeckung pathologischer Muster und der Feststellung kritischer Gesundheitszustände. Zunächst werden durch die Anwendung eines Medianfilters Ar-

tefakte entfernt. Danach werden aus KQ-Regressionen in gleitenden Fenstern der Trend und seine „Verlässlichkeit“ über die Steigung der resultierenden Gerade und die Fehlerquadratsumme geschätzt. Ist die Fehlerquadratsumme zu groß, so wird der Trend als unbekannt eingestuft. Bei kleiner Fehlerquadratsumme wird aufgrund der Steigung eine Einstufung in steigend, fallend oder stabil vorgenommen. Die Messwerte werden so in eine symbolische Darstellung überführt. Die quantitativen Messwerte und die qualitativen Trendsymbole werden mit heuristischen Wenn-Dann-Regeln einer Wissensbasis verglichen, die bestimmte pathologische Zustände beschreiben. Aus den gefundenen Übereinstimmungen wird ein „Alarmscore“ berechnet und gegebenenfalls ein Alarm ausgelöst. Die in der Wissensbasis verwendeten Regeln beruhen zum einen auf Lehrbuchwissen und zum anderen auf Erfahrungswissen zweier Ärzte. In einer Studie erzielt dieses System bezüglich der dort verwendeten Definition von alarmrelevanten Situationen eine Sensitivität von 100% und eine Spezifität von etwa 70% (Koski et al. (1994)). Die Autoren schließen ihre Forschungen mit der Feststellung ab, dass weitere teure und sehr zeitaufwendige Anpassungen in der Wissensbasis und deren Evaluation notwendig seien, um eine Integration in Monitoring-Geräte zu ermöglichen.

Das System TrenDx (Haimowitz et al. (1995)) zielt ebenfalls auf eine Reduktion der Fehlalarmrate ab. Es beruht auf einer Sammlung von multivariaten „Trend-Templates“, die zeitveränderliche Muster wichtiger klinischer Ereignisse in multivariaten Variablen beschreiben. TrenDx liefert eine Rangfolge möglicher Diagnosehypothesen anhand der durchschnittlichen Abweichung zwischen Trend-Template und Daten. Allerdings werden die Untersuchungen des Systems an nur einem Patienten durchgeführt, so dass über die generelle Eignung für den Einsatz auf der Intensivstation keine Aussage möglich ist.

Für die Überwachung von Patienten während Operationen schlagen Becker et al. (1997) ein Alarmsystem vor, das auf einer Wissensbasis von Fuzzyregeln beruht. Die Messwerte von verschiedenen Vitalparametern werden im Schritt „Fuzzification“ zu sprachlichen Aussagen abstrahiert, die zusätzlich mit dem Grad ihres Zutreffens belegt werden. Die dafür nötigen Grenzen werden aus Experteninterviews abgeleitet. In einer klinischen Studie, bei der die Einschätzung des Patientenzustands durch das Alarmsystem mit der Einschätzung durch einen Arzt verglichen wird, wird eine Sensitivität von etwa 96% und eine Spezifität von etwa 98% erreicht.

Schoenberg et al. (1999) schlagen ebenfalls ein mehrstufiges System zur Alarmgebung auf der Intensivstation vor. Zunächst definiert der Nutzer physiologische Trends als Grundlage der Alarmgebung. Während der Überwachung eines Patienten werden zu jedem Zeitpunkt Scores auf Basis der gefundenen Trends berechnet. Liegt der Score über einem Schwellwert, so wird ein Alarm ausgelöst. Dieses System erreicht eine Sensitivität von 82% bei Anwendung auf die Vitalparameter Herzfrequenz, systolischer und diastolischer Blutdruck und Sauerstoffsättigung von sechs Patienten einer Intensivstation über fünf Tage. Die Autoren umgehen in ihrem Vorschlag das Problem der Findung allgemein gültiger Definitionen kritischer Trends, indem sie diese Aufgabe auf das Pflegepersonal übertragen.

Krol und Reich (2000) entwickeln ein System zur Erkennung der zwei Patientenzustände *light anesthesia* und *unstable blood pressure* während Operationen, das durch rechnerische Ableitung von Regeln aus Annotationen erstellt wurde. Mit diesem System wird in einer klinischen Studie eine Sensitivität von 96% und eine Spezifität von 91% erreicht. Da es nur zwei bestimmte kritische Patientenzustände erkennen kann, ist es als Ersatz für die übliche Schwellwert-Alarmgebung nicht geeignet und kann lediglich bei paralleler Verwendung zusätzliche Alarme generieren.

Tsien (2000b) stellt fest, dass es im Allgemeinen schwierig und aufwendig ist, Definitionen kritischer Gesundheitszustände aus Experteninterviews und Lehrbüchern zu extrahieren, da die unterliegenden medizinischen Zusammenhänge bekannt sein müssen, um sie als Regeln formulieren zu können. Sie schlägt stattdessen die „Event Discovery Pipeline“ vor. Diese besteht aus der Benennung des interessierenden klinischen Ereignisses, der Sammlung annotierter Daten, der annotierten Datenvorverarbeitung und der Entwicklung eines Modells zur Entdeckung des Ereignisses. In einer Studie werden Daten aufgezeichnet und bezüglich des interessierenden Ereignisses „wahrer Alarm“ untersucht und annotiert. Die Datenvorverarbeitung umfasst die Berechnung gleitender Mittelwerte und Mediane mit verschiedenen Fensterbreiten und die Berechnung der Steigung von fensterweise angewandten Kleinste-Quadrate-Regressionen. Auf die erhobenen Daten und die durch die Datenvorverarbeitung erzeugten Charakteristika werden verschiedene maschinelle Lernverfahren angewendet. Aufgrund einer unzureichenden zugrunde liegenden Datenbasis betrachtet die Autorin ihre Arbeit nur als Anwendungsbeispiel für die Event Discovery Pipeline (Tsien et al. (2000), Tsien (2000a)).

2.2 Stand der Technik

Die beschriebenen Alarmsysteme sind bislang nicht in die Praxis überführt worden. Grundsätzlich ist die Diskrepanz zwischen Vorschlägen in der Literatur und dem Stand der Technik in der Praxis groß.

Der Stand der Technik auf dem Gebiet des Patientenmonitorings in der Intensivmedizin wird nachfolgend am Beispiel der Infinity Delta® Serie der Firma Dräger beschrieben (Dräger Medical Systems Inc. (2003)). Mit einem Patientenmonitor dieser Serie kann eine Vielzahl von Vitalparametern überwacht werden, etwa Herzfrequenz, Atemfrequenz, invasiver und nichtinvasiver Blutdruck, Temperatur und Sauerstoffsättigung des Blutes. Aber auch die Überwachung anderer Kenngrößen für den Gesundheitszustand ist möglich, so werden z.B. mit Hilfe eines Arrhythmie-Indikators Bradykardie-Ereignissen erkannt und alarmiert.

Die Regeln, nach denen ein Alarm ausgelöst wird, unterscheiden sich je nach Vitalparameter. Invasiv und nichtinvasiv gemessene Blutdrücke, Sauerstoffsättigung, Herzfrequenz, Atemfrequenz und Bluttemperatur werden mit Hilfe von Schwellwerten überwacht. Nach Überschreiten des Schwellwerts wird eine Verzögerungszeitspanne einstellbarer Länge abgewartet. Liegen die Messwerte während dieser Zeit stets außerhalb der Schwellwertgrenzen, so wird ein Alarm ausgelöst. Die Messwerte der Vitalparameter, die mit entsprechenden Alarmregeln verglichen werden, sind dabei in der Regel bereits vorverarbeitet. Zum Beispiel wird die Herzfrequenz aus den R-R-Intervallen der letzten zehn Sekunden als getrimmter Mittelwert berechnet.

Für andere Variablen existieren bereits verfeinerte Alarmregeln, beispielsweise in der Arrhythmie-Erkennung. In einer Lernphase von 30 bis 40 Sekunden wird ein Referenzmuster für den Patienten erstellt, das anschließend zur Beurteilung der gemessenen Herzaktivität herangezogen wird. Je nach Übereinstimmung oder Typ der Abweichung von der Referenz wird ein Alarm zusammen mit einer Klassifizierung der beobachteten Schlagfolge ausgegeben. Beispiele für mögliche Ereignisse sind unter anderem „Asystolie“, „Sinus Bradykardie“ und „Artefakt“.

Außerdem erkennt das Monitoringsystem technische Probleme, die es durch einen Alarm anzeigt. Die Alarmmeldung weist in einem solchen Fall zum Beispiel auf fehlende Messwerte, ein zu schwaches Signal oder einen fehlerhaft platzierten Sensor hin.

Die Hersteller von Patientenmonitoren geben ihre verfeinerten Alarm-Algorithmen und ihre zukünftig geplanten Neuerungen aus Wettbewerbsgründen nicht bekannt. Einen Einblick in die neusten Entwicklungen ermöglichen jedoch Patente auf dem Gebiet des Patientenmonitorings. Unter anderem werden dort adaptive Anpassungen der Alarmgrenzen und die Betrachtung des Produkts aus Zeit und Ausmaß der Überschreitung (Mannheimer (2007)) beschrieben. In einem anderen System (Mannheimer und Li (2007)) werden kurz hintereinander auftretende Alarme des gleichen Vitalparameters unterdrückt, sobald auf einen der Alarme durch das Pflegepersonal reagiert wurde. Wird eine gewisse Zeitspanne lang nicht auf den Alarm reagiert, so wird der Alarmton verändert, um eine gesteigerte Dringlichkeit anzuzeigen. Hutchinson (2003) sowie Hutchinson und Schluter (2004) entwickeln ein Monitoringsystem, das unter Nutzung eines mathematischen Modells physiologischer Abläufe abnormale Zustände erkennt und Alarme produziert. Alternativ dazu schlagen Tivig und Hebler (2006) ein Monitoringsystem vor, in dem Regeln zur Erkennung bestimmter klinischer Ereignisse hinterlegt sind und auch vom Nutzer eingegeben werden können. Die Regeln beinhalten Verknüpfungen von Bedingungen an die Messwerte verschiedener Vitalparameter. Sie können aus den Daten oder aus medizinischem Wissen abgeleitet werden. Es wird ein Alarm ausgelöst, sobald eine vom Nutzer festgelegte Anzahl von Bedingungen für ein Ereignis erfüllt ist. Die Beschreibungen in Patenten beschränken sich jedoch auf Andeutungen zur Architektur des Systems, konkrete Algorithmen bleiben unzugänglich.

2.3 Anforderungen an Alarmsysteme zur Überwachung von Intensivpatienten

Die zu beobachtende Diskrepanz zwischen kommerziell umgesetzten und vorgeschlagenen Monitoringsystemen wirft die generelle Frage nach Bedingungen für die Umsetzbarkeit auf. Die bisher entwickelten Systeme scheinen nicht den Anforderungen der Praxis oder Zulassungsanforderungen zu entsprechen. Um bei der Entwicklung eines neuen Alarmsystems Bedingungen für die Umsetzbarkeit berücksichtigen zu können, wird eine Anforderungsanalyse durchgeführt.

2.3.1 Essentielle Anforderungen

Die offensichtlichste Anforderung ist die sichere und zuverlässige Erkennung kritischer Gesundheitszustände. Gleichzeitig sollen unbedenkliche Gesundheitszustände ebenso sicher und zuverlässig erkannt werden. Die Alarmierung muss zeitnah zum kritischen Gesundheitszustand erfolgen, d.h. Algorithmen, die zu einer zu großen *Verzögerung* führen sind für die Anwendung im Monitoring ungeeignet.

Adaptive Verfahren müssen schon sehr kurze Zeit nach Beginn der Messung zuverlässige Alarme produzieren: Wird der „Normalwert“ für jeden Patienten individuell gelernt, kann für die Lernphase nur eine sehr kurze Zeit eingeräumt werden, während der nach anderen Regeln alarmiert wird. Bei dem Ansatz, den Normalzustand eines Patienten zu Beginn der Messung zu erlernen, ist zu beachten, dass Patienten in der Regel in einem schlechten Gesundheitszustand auf der Intensivstation aufgenommen werden und sich daher der gelernte Normalzustand deutlich vom Optimalzustand unterscheiden kann. Der Optimalzustand eines Patienten kann sich außerdem je nach aktuellem Krankheitsbild mit der Zeit verändern.

Die mangelnde *Signalqualität* macht es erforderlich, dass ein Alarmsystem auf das Auftreten fehlender Werte angemessen reagieren kann. Einzelne fehlende Werte sollten eine Beurteilung der klinischen Situation nicht verhindern. Viele aufeinander folgende fehlende Werte deuten auf ein technisches Problem hin, das die Überwachung stört, und sollten einen technischen Alarm auslösen. Da die Messwerte außerdem verrauscht und mit Ausreißern verschmutzt sein können, ist eine optimale Signalextraktion wichtig. Darunter kann zum Beispiel die einfachste Kurve mit guter Anpassung an die beobachteten Messwerte verstanden werden.

Aus psychologischen Gründen ist die *Transparenz für den Anwender* ein wichtiges Kriterium bei der Kaufentscheidung. Die Funktionsweise des Alarmsystems muss nachvollziehbar und verständlich sein. Nur so können die Alarme vom Pflegepersonal richtig interpretiert werden. Es wird vermutet, dass weniger entscheidend ist, ob die Algorithmen dem Anwender zur Verfügung stehen und verständlich sind, als vielmehr, ob dem Anwender das Prinzip der Funktionsweise vermittelt werden kann. Damit die Alarmgebung stabil und im klinischen Alltag überzeugend ist, müssen die Alarmregeln auf klaren und akzeptierten Definitionen von zu alarmierenden Zuständen und Ereignissen

beruhen. Wie in den Beispielen für vorgeschlagene Alarmsysteme beschrieben, können die Regeln entweder vom Anwender festgelegt und eingegeben werden, aus Lehrbüchern und Experteninterviews gewonnen und in Wissensbasen zusammengefasst werden oder aus entsprechend annotierten Daten abgeleitet werden.

Für die *Zulassung* des Alarmsystems ist entscheidend, dass die Regeln der Alarmgebung allgemeine Gültigkeit haben und bei keiner Patientengruppe zu einer systematischen Fehleinschätzung des Gesundheitszustandes führen können. Die Gültigkeit der Regeln sicherzustellen, ist vermutlich die schwierigste Aufgabe bei der Konstruktion eines neuen Alarmsystems. Sowohl mit wissensbasierten als auch datengestützten Ansätzen ist die allgemeine Gültigkeit nur schwer zu erreichen. Wissensbasierte Ansätze eignen sich nur für eng abgesteckte Überwachungsaufgaben wie die Arrhythmie-Erkennung. Eine Erweiterung auf alle auf der Intensivstation möglichen Diagnosen ist sehr aufwendig. Sie erfordert im Allgemeinen die Aufstellung umfassender Regeln durch Experten. Diese Leistung stattdessen vom Anwender erbringen zu lassen, senkt die Benutzerfreundlichkeit erheblich, da dadurch der Bedienaufwand gesteigert wird. Die datengestützte Alarmregelgenerierung verlangt keine ausformulierten Alarmregeln von Experten oder Anwendern sondern eine breite und repräsentative, annotierte Datenbasis. Eine geeignete Datenbasis kann nur mit hohen Kosten und großem Aufwand erstellt werden, da jede relevante Beobachtung in den Daten durch Experten beurteilt und annotiert werden muss. Bisherige Versuche, wie zum Beispiel von Tsien (2000b), datengestützt gültige Regeln zu erzeugen, scheinen mangels einer adäquaten Datenbasis gescheitert zu sein. Darüber hinaus ist die Evaluation neuer Alarmregeln häufig mit Schwierigkeiten verbunden. Vergleicht ein Experte ein altes und ein neues System in einer Studie am Patientenbett, so wird meist ein System früher einen Alarm auslösen als das andere. Wird in einer kritischen Situation auf einen Alarm des alten Systems hin eine Behandlung eingeleitet, normalisieren sich die Messwerte des Patienten möglicherweise, bevor das neue System ebenfalls alarmiert hat. Ob das neue System die Situation noch rechtzeitig erkannt hätte, kann dann nicht beurteilt und die Sensitivität des neuen Systems nicht bestimmt werden. Alternativ können zwei Alarmsysteme anhand annotierter Daten verglichen werden. Aber auch in diesem Fall lösen die beiden Systeme möglicherweise nicht in exakt der gleichen Sekunde einen Alarm aus. Um die Sensitivität bestimmen zu können, müssen (willkürliche) Regeln dazu aufgestellt werden, wann zwei Alarme sich auf die gleiche kritische Situation beziehen und bis wann ein Alarm als

„noch rechtzeitig gegeben“ gelten kann. Eine ausführliche Diskussion der Möglichkeiten eines Vergleichs verschiedener Alarmsysteme findet sich in Kuhls (2008).

2.3.2 Lösungsstrategie

Die Schwierigkeit, die Sensitivität eines neuen Systems festzustellen, kann umgangen werden, indem die Alarme des alten Systems durch das neue System validiert werden. So wird vom neuen System in exakt der gleichen Sekunde alarmiert oder der Alarm des alten Systems unterdrückt. Auf diese Weise wird ein direkter Vergleich möglich. Die Regeln des neuen Systems könnten in einem Gerät implementiert sein, dass die Daten (Messwerte, Alarme, usw.) des Monitors am Patientenbett als Eingangsdaten erhält und auf dieser Grundlage die Entscheidung trifft, ob ein Alarm tatsächlich gegeben werden soll. Das neue System kann keine zusätzlichen Alarme produzieren und der Anteil zu Recht unterdrückter und zu Recht gegebener Alarme ist leicht und objektiv festzustellen. Alarme des Standard-Monitoringsystems müssen dabei nicht vollständig unterdrückt werden. Es ist ebenso denkbar, die Alarme in verschiedenen Dringlichkeitsstufen auszulösen und Alarme, die nach den neuen Alarmregeln nicht benötigt werden, nur visuell am Patientenbett und in der Zentrale anzuzeigen. Die Vorgehensweise, Alarme durch neue Alarmregeln zu validieren, birgt den zusätzlichen Vorteil, dass kein komplettes Monitoringgerät für die Nutzung auf der Intensivstation erprobt und zugelassen werden muss. Mit dieser Architektur ist die Möglichkeit, die Eignung des neuen Alarmsystems zu überprüfen, gesichert.

In dieser Arbeit wird der Ansatz gewählt, die neuen Alarmregeln, die zur Validierung der Alarme des Standard-Monitoringgeräts verwendet werden, aus Daten abzuleiten und nicht Expertenwissen in einer Wissensbasis zu verankern. Bei der Alarmregelgenerierung wird das Prinzip der Klassifikation bzw. des „überwachten Lernens“ verfolgt. Dieser Ansatz führt zu einer Verzögerung von nur Sekundenbruchteilen im Vergleich zur herkömmlichen Alarmgebung und erfordert, da er nicht adaptiv ist, keine Lernphase zu Beginn der Überwachung eines Patienten. Bei der Wahl des Klassifikationsverfahrens ist die Fähigkeit zum Umgang mit fehlenden Werten sicherzustellen und aus psychologischen Gründen die Vermittelbarkeit der Funktionsweise zu berücksichtigen. Entscheidend für die Zulassung eines neuen Alarmsystems, das auf datengestützt gene-

rierten Alarmregeln beruht, ist die allgemeine Gültigkeit der erzeugten Regeln. Diese sicherzustellen, ist nur mit einer sehr breiten Datenbasis möglich. Ob die hier generierten Regeln allgemein gültig sind, hängt von der Repräsentativität der Daten ab und wird in Kapitel 6 überprüft.

Viele der aufgezeichneten Alarme sind durch das Pflegepersonal, z.B. durch Pflegemaßnahmen oder Medikation, herbeigeführt. Ob ein Alarm durch eine „natürliche“ Veränderung des Patientenzustands oder als Folge einer solchen Manipulation ausgelöst wird, ist ebenfalls in den Annotationen festgehalten. Für die Sicherheit des Patienten ist die zuverlässige Erkennung nicht herbeigeführter alarmrelevanter Situationen maßgeblich. In solchen Situationen verändert sich der Gesundheitszustand des Patienten auf natürliche Weise in eine kritische Richtung. Situationen hingegen, in denen ein Alarm herbeigeführt wurde, müssen für die Sicherheit des Patienten nicht notwendigerweise richtig klassifiziert werden, da das Pflegepersonal selbst für die Überwachung des Gesundheitszustandes ausgebildet ist und mit einer Veränderung der Messwerte als Konsequenz der eigenen Handlung rechnet. Allerdings ist die Anwesenheit einer Pflegekraft im Raum oder eine manipulierende Handlung nicht für den Patientenmonitor erkennbar. Ist die Alarmgebung in solchen Situationen nicht zuverlässig, so wird das System immer noch eine als hoch *wahrgenommene Fehlerquote* haben. Daher ist die Justierung der Sensitivität eines neuen Alarmsystems bezüglich aller, auch der herbeigeführten, Alarme als sinnvoll anzusehen, da es für die Sicherheit des Patienten indirekt über die Beeinflussung des Pflegepersonals durch die Fehlerquote von Bedeutung ist.

2.3.3 Weitere wünschenswerte Eigenschaften und Ideen für zukünftige Ansätze

Neben den oben beschriebenen essentiellen Anforderungen, die in dieser Arbeit in der Lösungsstrategie berücksichtigt werden, gibt es eine Reihe weiterer wünschenswerter Eigenschaften intensivmedizinischer Alarmsysteme. Für weiterführende Untersuchungen kann diese Zusammenstellung als Ideensammlung dienen.

Zu den wünschenswerten Eigenschaften zukünftiger Alarmsysteme gehört die *Früherkennung* kritischer Situationen. Eine Möglichkeit, frühzeitig auf kritische Veränderungen im Gesundheitszustand hinzuweisen, besteht beispielsweise in der Trendbestimmung. In Projekt C4 des SFB 475 sind Signalextraktionsverfahren entwickelt worden

(z.B. Gather et al. (2006), Borowski et al. (2008), Lanius und Gather (2007)), die auch zur Vorhersage der Vitalparameterwerte in den nächsten Sekunden und Minuten genutzt werden können. Ein Vergleich der Vorhersage mit den Schwellwerten des Standard-Monitoringsystems könnte einen Hinweis auf die Dauer bis zur Überschreitung geben. Allerdings produziert ein solches Verfahren zusätzliche Alarme, die vermutlich erst nach gelungener Reduktion der Fehlalarme des übrigen Alarmsystems in der klinischen Praxis akzeptiert würden.

Wird zusätzlich die Variabilität der Vitalparameter-Messwerte geschätzt und in die Alarmgebung einbezogen, kann möglicherweise eine *Priorisierung* der Alarme erreicht werden. Die Variabilität weist zum einen auf Schwankungen im Gesundheitszustand hin, kann zum anderen aber auch als Indikator für die Verlässlichkeit des zuvor bestimmten Trends aufgefasst werden. Ein Alarm, der bei langsam steigendem Trend und großer Variabilität auf Grund einer oberen Grenzverletzung ausgelöst wird, könnte mit einer geringeren Dringlichkeit ausgelöst werden als ein Alarm bei steil ansteigendem Trend und geringer Variabilität. Weit komplizierter, aber auch aussagekräftiger, wäre eine Priorisierung zum Beispiel nach Organsystem und nach Behandlungsmöglichkeit des zugrunde liegenden Problems oder angepasst an den medizinischen Kontext und die vorliegende Diagnose. Neben umfangreichen Daten und medizinischem Wissen zur Erstellung eines solchen Systems zur priorisierten Alarmgebung ist die Eingabe der Diagnose in das System am Patientenbett erforderlich.

Zukünftige Alarmsysteme könnten neben ihrer Alarmfunktion die *Findung der Diagnose* unterstützen (Decision Support System). Dieses Ziel kann unter anderem durch wissensbasierte Diagnosevorschläge erreicht werden, die aus so genannten „clinical guidelines“ abgeleitet werden. Alternativ kann das Pflegepersonal in der Diagnosestellung unterstützt werden, indem der Situation angepasst relevante Informationen dargestellt werden, die die Beurteilung des Gesundheitszustandes sowie die Beurteilung des ausgelösten Alarms im klinischen Kontext erleichtern. Dazu ist eine sinnvolle Komprimierung der Informationsflut notwendig, die beispielsweise durch eine transparente und reproduzierbare Dimensionsreduktion erreicht werden kann.

Über die medizinischen Funktionen hinaus sind auch *technische Verbesserungen* aktueller Alarmsysteme denkbar und lohnend. Zusätzlich zu den Patientenmonitoren produzieren auch andere Geräte Alarme auf der Intensivstation. Eine Zusammenlegung

aller alarmgebender Geräte ist wünschenswert, da so doppelte Alarmsignale vermieden werden können. Darüber hinaus könnte eine zuverlässigere Erkennung technischer Probleme den Arbeitsablauf der intensivmedizinischen Pflege erleichtern und würde dadurch eine sicherere Überwachung der Patienten gewährleisten.

Kapitel 3

Alarme in der Intensivmedizin

3.1 Datengewinnung

Die Grundlage für eine datengestützte Alarmregelgenerierung bilden sorgfältig annotierte Aufzeichnungen von Standard-Patientenmonitoren. Um die generierten Alarmregeln erfolgreich in der Praxis einsetzen zu können, ist ein repräsentativer Datensatz gesundheitskritischer Situationen notwendig, die an einer repräsentativen Auswahl von Patienten beobachtet wurden.

In einer klinischen Studie des Teilprojekts C4 des SFB 475 sind am Universitätsklinikum Regensburg in Zusammenarbeit mit Dräger Medical Systems Monitoring-Daten erhoben worden, die unter anderem die ausgelösten Alarme zusammen mit den Vitalparameter-Messwerten der Patienten umfassen. Zu diesen Daten wurden klinische Annotationen erstellt, die die Bewertung der Alarmsituationen durch Ärzte beinhalten. Hauptziel der Studie ist die klinische Validierung neuer Online-Verfahren zur Alarmgebung im Intensivmonitoring. Die Verwendung der erhobenen Daten zur Alarmregelgenerierung mittels Klassifikationsverfahren stellt eine interessante Alternative zu den in der Studie untersuchten Online-Verfahren dar.

Im Zeitraum von Januar 2006 bis Mai 2007 wurden an 85 Fällen eine Vielzahl von Vitalparametern, Alarmen und Monitoreinstellungen aufgezeichnet und durch Ärzte klinisch bewertet. Es liegen ca. 1250 Beobachtungsstunden vor, in denen 7000 Alarme eines Infinity Monitoring Systems® einer Beurteilung unterzogen wurden. Die Beurteilung wurde mit Hilfe einer eigens für diesen Zweck entwickelten Annotationsmaske erfasst (Kuhls et al. (2006)). Diese Beurteilung erforderte einen erheblichen Zeitauf-

wand und unterscheidet den vorliegenden Datensatz von vielen anderen. Durch sie wird die Anwendung statistischer Klassifikationsverfahren zur Erzeugung neuer Alarmregeln erst ermöglicht.

Abbildung 3.1 zeigt beispielhaft einen Auszug aus den erhobenen Daten. Es werden die Messwerte des systolischen und mittleren arteriellen Blutdrucks, der Herzfrequenz und der Sauerstoffsättigung eines Patienten im Verlauf einer Stunde dargestellt. Graue Abschnitte markieren Zeiträume, in denen die Alarmfunktion des Monitoringsystems deaktiviert war, z.B. weil dem Patienten Blut abgenommen wurde und eine Blutabnahme durch die Druckveränderungen immer einen unnötigen Alarm auslöst. Die Alarmfunktion wird deaktiviert, indem das Pflegepersonal auf eine „Alle-Alarme-Aus“-Taste am Monitor drückt, und bleibt für drei Minuten ausgeschaltet, falls die Deaktivierung nicht durch erneutes Drücken aufgehoben wird. Waagerechte Linien zeigen die eingestellten Schwellwerte an. Senkrechte Linien markieren die Alarme des Standard-Monitoringsystems. Jeder Alarm zeigt eine möglicherweise kritische Situation des Patienten an. Ob die Situation tatsächlich kritisch ist und der Alarm zu Recht gegeben wurde, ist in den Annotationen festgehalten.

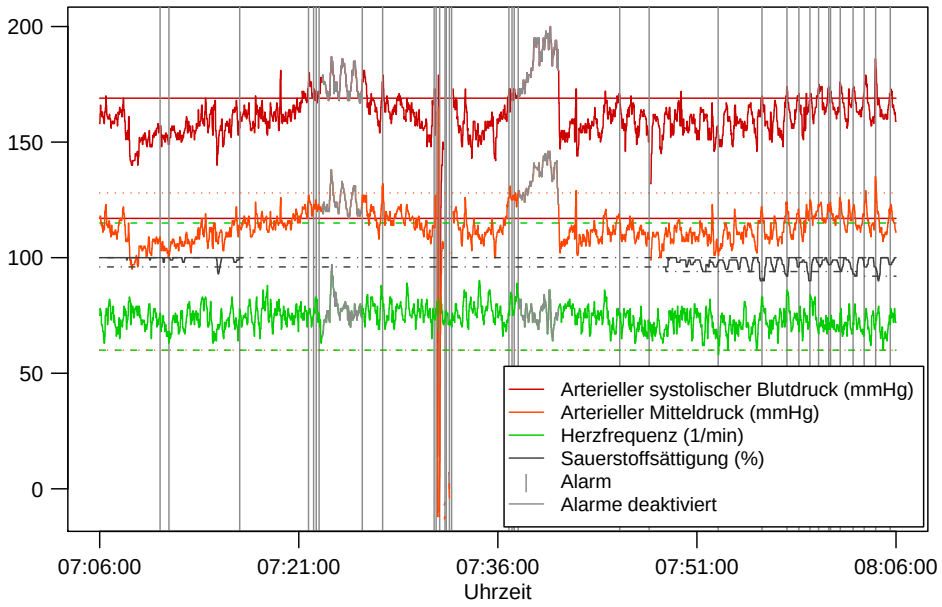

Abbildung 3.1: Intensivmedizinische Zeitreihen (Beispiel)

Die aufgezeichneten Alarme werden sowohl bezüglich ihrer technischen Validität als „technisch richtig“ und „technisch falsch“ als auch bezüglich ihrer klinischen Validität als „alarmrelevant“, „hinweisend“ oder „nicht alarmrelevant“ bewertet. Grundlage der Bewertung sind die Zeitreihen der gemessenen Vitalparameter bis zu 30 Minuten vor und nach dem zu bewertenden Alarm und die eingestellten Schwellwerte, die EKG-Echtzeitkurven und eine Videoaufzeichnung der Situation. Die Bewertung der Situationen hängt dabei in gewissem Maß auch von der persönlichen Erfahrung und Einschätzung des annotierenden Arztes ab und beruht häufig nicht allein auf Fakten, die objektiv und eindeutig festgestellt werden können. Grundlage der Annotation in der vorliegenden Studie bilden folgende Definitionen. Ein Alarm wird als *technisch richtig* annotiert, wenn er aufgrund von korrekt gemessenen Vitalparameterwerten ausgelöst wird oder korrekt auf ein technisches Problem hinweist. Die zu einem Alarm gehörende klinische Situation wird als *alarmrelevant* bewertet, wenn auf den Alarm eine klinische Entscheidung oder eine therapeutische Handlung als Reaktion des Pflegepersonals erfolgt. *Hinweisende Alarme* haben eine klinische Bedeutung, die aber nicht unmittelbar in eine Handlung oder Entscheidung mündet. *Nicht alarmrelevante Situationen* sind solche, in denen ein Alarm gänzlich ohne jede klinische Bedeutung gegeben wurde. Einen wesentlichen Einfluss auf die Beurteilung nimmt die Information, ob ein Alarm durch das Pflegepersonal herbeigeführt wurde (*Manipulation*) oder auf eine durch die Krankheit des Patienten verursachte Veränderung im Gesundheitszustand zurückzuführen ist. Diese Information wird in den Annotationen ebenfalls vermerkt. Obwohl Zeitreihen der gemessenen Vitalparameter über viele Stunden vorliegen, bezieht sich die Annotation nur auf die Sekunde, in der der Alarm ausgelöst wurde.

Da in der Studie nur Situationen bewertet werden, in denen ein Alarm ausgelöst wurde, können keine Situationen entdeckt werden, in denen ein Alarm notwendig gewesen wäre, aber nicht gegeben wurde. Die Daten werden also nicht hinsichtlich kritischer Situationen ohne Alarm analysiert und annotiert. Das Studiendesign beruht demnach auf der Annahme, dass das verwendete Monitoringsystem mit einer Sensitivität von 100% jede alarmrelevante Situation erfasst. Diese Annahme ist diskussionswürdig, bedeutet aber in jedem Fall, dass nur Verbesserungen bezüglich der Spezifität bzw. Fehlalarmreduktion aus den Daten abgeleitet werden können, da für Verbesserungen der Sensitivität nicht die notwendigen Informationen in den Daten enthalten sind.

3.2 Deskriptive Analyse

Von den insgesamt 246 vom Monitoringsystem erfassten Variablen weisen viele während der meisten Alarme fehlende Werte auf. Diese Variablen enthalten daher kaum für die Klassifikation verwertbare Informationen. Daher wird in dieser Arbeit nur ein Teil der Variablen verwendet. Nachfolgend werden die ausgewählten Variablen vorgestellt und die ausgelösten Alarme deskriptiv analysiert. Dazu wird zunächst die Häufigkeit von Alarmen der einzelnen Variablen insgesamt ermittelt und im Anschluss der Bezug zu technischer und klinischer Bewertung hergestellt. Die Betrachtung der klinischen Bewertung der Alarmsituationen, die nicht herbeigeführt wurden, zeigt, wieviele der Alarme der einzelnen Variablen durch ein neues Alarmsystem zuverlässig erkannt werden müssen, um die Sicherheit der Patienten zu gewährleisten.

Der Datensatz, der zur Erstellung der Alarmregeln verwendet wird, enthält nur einen Teil der zur Verfügung stehenden Variablen. Zu den 70 Variablen, die bei mindestens 250 der 7000 ausgelösten Alarme tatsächlich aufgezeichnet wurden und in den Datensatz aufgenommen werden, gehören:

- arterieller systolischer (ART_S), diastolischer (ART_D) und Mitteldruck (ART_M)
- nicht invasiv gemessener systolischer (NBP_S), diastolischer (NBP_D) und Mitteldruck (NBP_M)
- zentralvenöser Druck (CVP)
- Herzfrequenz (HR) und Puls (PLS)
- Sauerstoffsättigung (SpO2) und Atemfrequenz (RESP)
- ein Arrhythmie-Indikator (ARR)
- Temperatur (Ta).

Neben diesen Variablen werden aus dem EKG abgeleitete Kennzahlen sowie die eingestellten Schwellwerte in den Datensatz aufgenommen. Das verwendete Monitoringsystem nimmt selbst eine Einordnung der Alarme in lebensbedrohliche (LT), ernste

(SER) und hinweisend-technische (ADV) Alarme vor und liefert somit eine Einschätzung des „Alarmgrads“. Wegen des Hinweises auf ein lebensbedrohliches Ereignis und der geringen Häufigkeit der nur zehn LT-Alarme werden diese Alarme aus dem Datensatz entfernt und vom zu entwickelnden zukünftigen Alarmsystem uneingeschränkt ausgegeben. Die Unterscheidung in SER- und ADV-Alarme ist für das Verständnis der Annotation bezüglich technischer Validität von Bedeutung: ADV-Alarme sollen auf ein technisches Problem hinweisen, während SER-Alarme einen kritischen Gesundheitszustand anzeigen sollen. In den Annotationen ist festgehalten, ob das Monitoringsystem diese Einteilung korrekt vornimmt. Der Alarmgrad wird ebenso in den Datensatz aufgenommen wie die Variablen, die den Auslösegrund angeben. Darüber hinaus stehen viele weitere, teils technische, teils patientenbezogene Variablen zur Verfügung, die ebenfalls in den Datensatz aufgenommen werden. Eine vollständige Liste aller verwendeter Variablen findet sich im Anhang.

3.2.1 Technische Validität

Einen Überblick über die Variablen, die am häufigsten einen Alarm auslösen, gibt Abbildung 3.2. Ein großer Teil (87%) der Alarme wird durch Schwellwert-überwachte Variablen ausgelöst (z.B. arterieller systolischer und mittlerer Blutdruck, Sauerstoffsättigung und Herzfrequenz). Nicht jeder dieser Alarme muss dabei auch ein Schwellwertalarm sein: Es werden für diese Variablen auch ADV-Alarme ausgelöst, die auf ein technisches Problem, wie beispielsweise fehlende Messwerte durch einen defekten oder falsch platzierten Sensor, hinweisen können.

Etwa 40% der 7000 Alarme werden als technisch falsch beurteilt (vgl. Abb. 3.3). Das bedeutet, dass ein Alarm aus einem falschen Grund ausgelöst wird. Es sagt nichts darüber aus, ob der Alarm aus klinischer Sicht notwendig ist. Herzfrequenzalarme sind größtenteils technisch richtig, Alarme für die Atemfrequenz hingegen sind größtenteils technisch falsch. Alarme des arteriellen systolischen Blutdrucks sind etwa genauso häufig technisch richtig wie technisch falsch. Die technische Zuverlässigkeit der ausgelösten Alarme unterscheidet sich also deutlich je nach Variable. Daher sind für manche Variablen Maßnahmen, die die technische Validität erhöhen sollen, gänzlich überflüssig. Für andere ist das verwendete Konzept der Alarmgebung aufgrund der geringen Zuver-

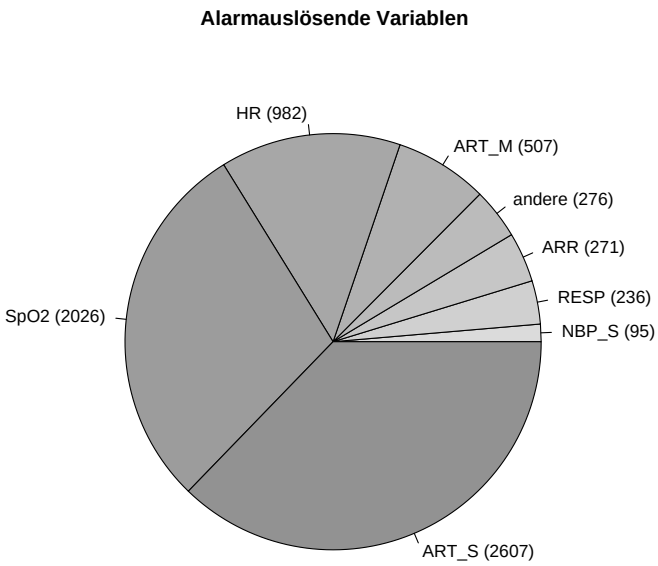

Abbildung 3.2: Häufigkeit der Alarmauslösung durch verschiedene Variablen

lässigkeit insgesamt in Frage zu stellen. Der Fokus dieser Arbeit liegt jedoch nicht auf der technischen Validität, da bei Verbesserungen in der Alarmgebung die Auslösung der Alarme *zum richtigen Zeitpunkt* an erster Stelle steht und erst in zweiter Linie die Auslösung aus dem richtigen Grund angestrebt werden kann.

3.2.2 Klinische Validität

Die klinische Validität ist für die sichere Überwachung der Patienten von größerer Bedeutung als die technische Validität. Sie gibt an, ob das Pflegepersonal aus ärztlicher Sicht durch einen Alarm auf eine Situation aufmerksam gemacht werden muss. Auch technisch falsche Alarme können alarmrelevant sein. Zum Beispiel wird eine Situation als alarmrelevant bewertet, wenn ein Schwellwertalarm aufgrund eines falschen Messwertes ausgelöst wird, der Grund für die falsche Messung aber behoben werden muss. Andererseits können technisch richtige Alarme auch nicht alarmrelevant sein.

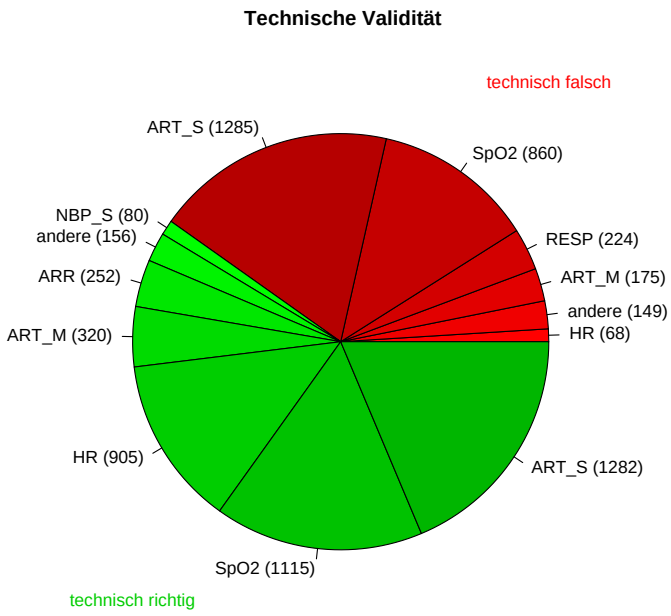

Abbildung 3.3: Technische Validität der ausgelösten Alarme

Dies ist zum Beispiel der Fall, wenn ein Schwellwertalarm durch einen korrekt gemessenen Wert ausgelöst wird, der Alarm aber in der Situation nicht benötigt wird, weil keine Handlung oder Entscheidung erforderlich ist. Die bei der Annotation vorgenommene Einstufung von Alarmen als hinweisend wird in dieser Arbeit nicht benötigt. Hinweisende Alarme werden hier daher zu den nicht alarmrelevanten Situationen gezählt, da sie keine Reaktion beim Pflegepersonal auslösen müssen.

Die klinische Bewertung der Alarmsituationen ist in Abbildung 3.4 dargestellt. Aufgrund des arteriellen systolischen Blutdrucks werden die meisten Fehlalarme ausgelöst. Sauerstoffsättigung, Herzfrequenz und mittlerer arterieller Butdruck verursachen ebenfalls viele Fehlalarme. Insgesamt wird deutlich, dass die ausgelösten Alarme mehrheitlich aus ärztlicher Sicht nicht notwendig sind und das Pflegepersonal als Fehlalarme belästigen. Von allen Alarmen sind nur 15% als alarmrelevant bewertet worden. Fasst man alle übrigen Alarme als Fehlalarme auf, so liegt die Fehlalarmrate bei 85%. Dieses Ergebnis entspricht den Ergebnissen anderer Studien zur Häufigkeit von Fehlalarmen in der Intensivmedizin, die zwischen 68% und 90% Fehlalarme verzeichnen (O'Carroll

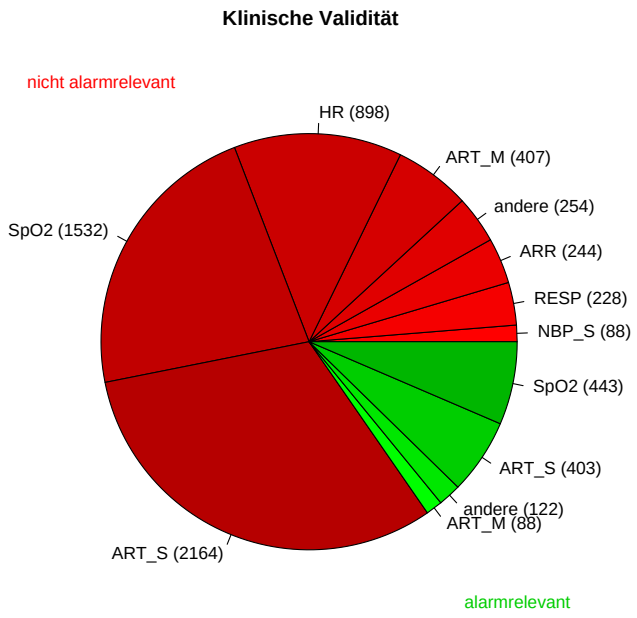

Abbildung 3.4: Klinische Validität der ausgelösten Alarme

(1986), Koski et al. (1990), Lawless (1994), Tsien und Fackler (1997), Chambrin et al. (1999), Chambrin (2001)). Diese Studien, sowie auch die hier festgestellte Fehlalarmrate, weisen auf einen großen Bedarf an Verbesserungen im intensivmedizinischen Monitoring hin.

3.2.3 Klinische Validität und Manipulation

Alarme, die durch Therapie- oder Pflegemaßnahmen herbeigeführt wurden, hätten vermieden werden können, wenn die Alarmfunktion während der Alarm-auslösenden Tätigkeit deaktiviert worden wäre. Abbildung 3.5 zeigt die klinische Validität der nicht durch Manipulation herbeigeführten Alarme und den Anteil der durch Manipulation ausgelösten Alarme. Etwa 40 % aller Alarme sind durch Handlungen des Pflegepersonals verursacht worden. Es ist nicht notwendig sondern wirkt störend, Veränderungen im Gesundheitszustand oder Artefakte durch einen Alarm anzuzeigen, die das Pflege-

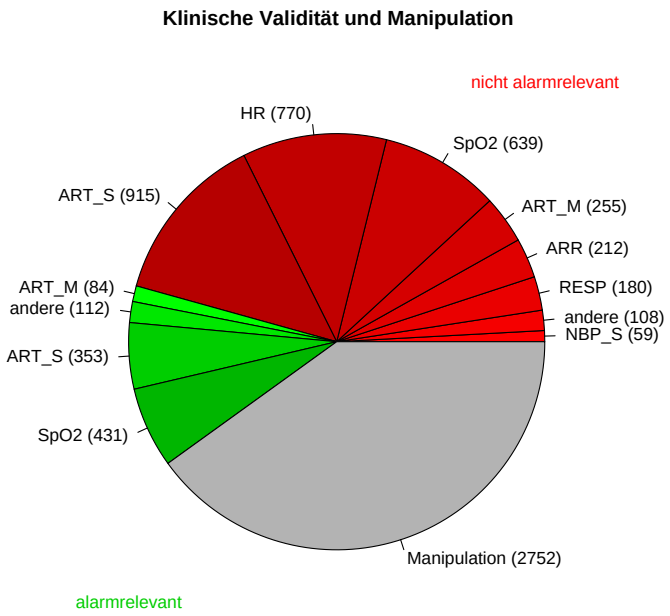

Abbildung 3.5: Klinische Validität nicht durch Manipulation herbeigeführter Alarme

personal selbst ausgelöst hat. Eine Verbesserung der Benutzerfreundlichkeit, die die Deaktivierung der Alarmfunktion erleichtert, bietet also eine etwa ebenso große Möglichkeit, unnötige Alarme zu reduzieren wie eine Verbesserung der Alarm-Algorithmen. Von allen Alarmen, die nicht herbeigeführt wurden, sondern auf eine „natürliche“ Veränderung in den physiologischen Werten des Patienten zurückgehen, sind 24% als alarmrelevant bewertet worden. Es wird somit bei den nicht herbeigeführten Alarmen eine Fehlalarmrate von 76% erreicht. Herbeigeführte Alarme werden mehrheitlich als nicht alarmrelevant eingestuft.

Wie schon zuvor diskutiert, kann es sinnvoll sein, Alarme, die durch Manipulation herbeigeführt wurden, aus der Bestimmung der Sensitivität und Fehlalarmrate eines Alarmsystems auszuschließen oder auch sie mit einzubeziehen (vgl. Kap. 2.3.2). Im Folgenden werden daher meist beide Vorgehensweisen in den Untersuchungen berücksichtigt.

Kapitel 4

Datenvorverarbeitung

Die Datenvorverarbeitung geht der Klassifikation in vielen Anwendungen voraus. Durch Vorverarbeitung der Daten können Charakteristika (Features) extrahiert werden, die eine deutlich bessere Trennbarkeit oder Unterscheidbarkeit der zu unterscheidenden Gruppen ermöglichen. Auf diese Weise können die Klassifikationsergebnisse zum Teil erheblich verbessert werden.

Die Datenvorverarbeitung umfasst hier unter anderem die Ersetzung fehlender Werte und die Erzeugung von Dummy-Variablen, die ersetzte Werte anzeigen. Darüber hinaus werden Regressions-basierte Charakteristika erzeugt, die den gesundheitlichen Verlauf kurz vor einem Alarm wiedergeben. Die Auswahl der Charakteristika orientiert sich an der Vorgehensweise von Ärzten bei der Beurteilung, ob eine Situation kritisch ist.

Wie schon zur deskriptiven Analyse der Daten werden die Datenvorverarbeitung und die später folgende Alarmregelgenerierung mit der Statistik-Software R (R Development Core Team (2008)) durchgefürt.

4.1 Ersetzung fehlender Werte

Einerseits ist für die Verwendbarkeit der später genutzten Verfahren die Ersetzung fehlender Werte Bedingung. Andererseits kann das Fehlen eines Wertes für die Klassifikation von Alarmsituationen eine relevante Information darstellen und zum Beispiel auf einen Defekt der Messgeräte hinweisen. Durch Dummy-Variablen, die ersetzte Werte anzeigen, wird sichergestellt, dass diese Information bei Bedarf mit einbezogen werden kann.

Bei allen Variablen außer Herzfrequenz und Puls werden die fehlenden Werte durch den Median bzw. den Modalwert der entsprechenden Variablen im „Lerndatensatz“ (vgl. Kap. 5) ersetzt. Die Herzfrequenz wird über das EKG und der Puls über einen Clip am Finger des Patienten gemessen. Da beide Variablen die Herzaktivität charakterisieren, werden die fehlenden Werte dieser Variablen mittels einer Kleinste-Quadrate-Regression (KQ) ersetzt. Auch bei den Blutdrücken liegt die Ersetzung fehlender Werte über eine Regression nahe, jedoch treten fehlende Werte hier meist gleichzeitig auf, so dass die Ersetzung durch den Median vorgenommen wird. In der Online-Anwendung werden fehlende Werte auf gleiche Weise durch die aus den Lerndaten gewonnenen Werte ersetzt.

4.2 Charakteristika der zeitlichen gesundheitlichen Veränderung

Ärzte auf der Intensivstation betrachten bei der Einschätzung einer klinischen Situation die Entwicklung des Gesundheitszustandes eines Patienten im zeitlichen Verlauf. Daher ist zu erwarten, dass Alarmregeln, die Charakteristika mit einbeziehen, die die zeitliche physiologische Veränderung wiedergeben, eine höhere Fehlalarmreduktion ermöglichen. Ähnlich wie in Tsien (2000b) werden in dieser Arbeit charakterisierende Größen, wie zum Beispiel die Steigung einer KQ-Regressionsgerade oder die Standardabweichung in gleitenden Fenstern unterschiedlicher Länge berechnet. Zusätzlich dazu werden robuste lineare Regressionsmethoden zur Erzeugung von Charakteristika zum Einsatz gebracht. Ein klassischer Ansatz in der Datenvorverarbeitung von Signalen, wie sie hier vorliegen, ist die Wavelet-Anpassung. Bei dieser Form der Regression können die Wavelet-Koeffizienten als Charakteristika genutzt werden. Neben den genannten univariaten sind auch multivariate Charakteristika denkbar. Die Erzeugung multivariater Charakteristika, die die Entwicklung mehrerer Vitalparameter zusammenfassen, entspricht einer Dimensionsreduktion. Lanius (2004) untersucht verschiedene klassische dimensionsreduzierende Verfahren sowie deren dynamische Erweiterungen auf ihrer Eignung für Daten aus der Intensivmedizin. Eine globale Modellierung intensivmedizinischer Zeitreihen erweist sich in den Untersuchungen als äußerst schwierig und nur lokale Anpassungen liefern sinnvolle Ergebnisse (Lanius (2004), S. 78). Charakteristika, die durch lokale Anpassungen gewonnen werden, die Vitalparameter auf

immer verschiedene Weise zusammenfassen, sind inhaltlich nur schwer zu interpretieren und nicht wie die im Folgenden vorgestellten univariaten Ansätze aus der ärztlichen Praxis zu motivieren. Daher wird in dieser Arbeit auf das Finden geeigneter multivariater Charakteristika verzichtet. Dies erscheint dennoch als Gegenstand zukünftiger Forschungsvorhaben mit dem Fokus auf höchsten Sensitivitäten und Fehlalarmreduktionen − auch zu Lasten der Interpretierbarkeit − sinnvoll.

4.2.1 Charakteristika aus linearer Regression

Sei $(y_t)_{t\in\mathbb{Z}}$, $y_t \in \mathbb{R}$ eine Zeitreihe von Messwerten eines Vitalparameters, wie zum Beispiel ART_S, ART_M, ART_D, RESP, HR, PLS, SpO2 und Ta. Verwende für die Berechnung einer charakterisierenden Größe zum Zeitpunkt t die Beobachtungen y_{t-m+1},..., y_t der Zeitreihe in einem Fenster der Länge m. Unterstellt man in diesem Fenster als Arbeitsmodell für den zugrunde liegenden Prozess einen lokal linearen Zusammenhang

$$Y_{t-m+i} = \mu_t - (m-i)\beta_t + \varepsilon_t, \qquad i = 1, ..., m$$

mit Fehler ε_t, so können das unterliegende Signal μ_t und Steigung β_t mit Hilfe von Regressionsverfahren unter anderem durch die Standardmethode der KQ-Regression oder robust durch eine Repeated-Median-Regression (RM) geschätzt werden.

Der Repeated-Median-Filter hat sich als geeignet erwiesen, Ausreißer und Rauschen aus physiologischen Zeitreihen zu entfernen (Davies et al. (2004), Gather et al. (2006)). In einem gleitenden Zeitfenster wird dabei folgende robuste Regression durchgeführt (vgl. Siegel (1982)):

$$\begin{aligned}
\hat{\beta}_t^{RM} &= \text{Median}_{\,i=1,\ldots,m} \left\{ \text{Median}_{\,i\neq j} \quad \frac{y_{t-m+i} - y_{t-m+j}}{i-j} \right\} \\
\hat{\mu}_t^{RM} &= \text{Median}_{\,i=1,\ldots,m} \left\{ y_{t-m+i} + (m-i)\hat{\beta}_t^{RM} \right\}.
\end{aligned}$$

Die Fensterbreite m ist dabei zu wählen. Sie beeinflusst die Glattheit und die Verzerrung des extrahierten Signals. Ist die schnelle Erkennung einer plötzlich auftretenden gesundheitlichen Veränderung wichtig, so sollte das Zeitfenster klein gewählt werden. Das extrahierte Signal bleibt dann eng an den Beobachtungen. Ist der Gesundheitszustand gleichbleibend, dann sind größere Fensterbreiten vorzuziehen, da sie ein glatteres

Signal erzeugen. Steile Trends und Sprünge können so allerdings weniger schnell erkannt werden. Alternativ zu einer festen Fensterbreite kann diese auch adaptiv den Daten angepasst werden. Ein Test auf Anpassungsgüte basierend auf den Vorzeichen der Residuen im Zeitfenster dient dabei als Entscheidungskriterium, ob die aktuelle Fenstergröße verändert werden muss (Schettlinger et al. (2008)).

Als Charakteristika werden die Online-Schätzungen des Vitalparameterwertes am Ende des Zeitfensters und die Steigungen der geschätzten Geraden (KQ, RM und $RM_{adaptiv}$) gewählt, da sie Aufschluss über den linearen Trend in den letzten m Sekunden geben. Zu jedem Alarmzeitpunkt t beschreiben

- $\hat{\beta}_t^{KQ;m}$: die Steigung der geschätzten Gerade nach KQ-Regression im Fenster der Länge m
- $\hat{\beta}_t^{RM;m}$: die Steigung der geschätzten Gerade nach RM-Regression im Fenster der Länge m
- $\hat{\beta}_t^{RM_{adaptiv}}$: die Steigung der geschätzten Gerade nach $RM_{adaptiv}$-Regression im Fenster variabler Länge mit minimaler Fensterbreite 60 und maximaler Fensterbreite 121
- $\hat{\mu}_t^{KQ;m}$: das geschätzte Signal am Ende des Fensters der Länge m nach KQ-Regression
- $\hat{\mu}_t^{RM;m}$: das geschätzte Signal am Ende des Fensters der Länge m nach RM-Regression
- $\hat{\mu}_t^{RM_{adaptiv}}$: das geschätzte Signal am Ende des Fensters variabler Länge nach $RM_{adaptiv}$-Regression mit minimaler Fensterbreite 60 und maximaler Fensterbreite 121

den linearen Trend kurz vor einem Alarm. Neben der so quantifizierten Information über den Gesundheitszustand kann auch die Variabilität in gleitenden Fenstern als charakterisierend herangezogen werden. Aufschlussreich bezüglich der Entwicklung des Gesundheitszustandes ist sowohl die Variabilität der Messwerte im Zeitfenster als auch die Variabilität der Residuen. Die Standardabweichung als Maß für die Variabilität

ist nicht robust. Rousseeuw und Croux (1993) schlagen als robuste Alternative den Skalenschätzer Qn vor:

$$Qn = 2,2219 \cdot \{|y_{t-m+i} - y_{t-m+j}|; i = 1, ..., m,\ i < j\}_{(k)},$$

mit $k = \frac{\binom{\frac{m}{2}+1}{2}}{4}$.

In Kombination mit den verschiedenen Regressionsmethoden werden zu jedem Alarmzeitpunkt t folgende, die Variabilität charakterisierende Größen berechnet:

- $\hat{\sigma}_t^m$: die Standardabweichung aller Beobachtungen im Fenster der Länge m
- $\hat{\sigma}_t^{KQ;\,Res;\,m}$: die Standardabweichung der Residuen nach KQ-Regression im Fenster der Länge m
- Qn_t^m: der Qn-Schätzer aller Beobachtungen im Fenster der Länge m
- $Qn_t^{RM;\,Res;\,m}$: der Qn-Schätzer der Residuen nach RM-Regression im Fenster mit fester Länge m
- $Qn_t^{RM_{adaptiv}}$: der Qn-Schätzer aller Beobachtungen im Fenster mit variabler Länge gewählt analog zur $RM_{adaptiv}$-Regression.

Die vorgeschlagenen Charakteristika beschreiben also die Veränderungen im Gesundheitszustand eines Patienten, indem sie angeben, ob die Messwerte steigen, fallen oder gleich bleiben, wie stark sie (um eine lineare Regressionsgerade) streuen und welche geschätzten Werte die betrachteten Vitalparameter ohne Rauschen bzw. Ausreißer annehmen würden. Die Berechnungen der RM-basierten Charakteristika werden mit dem R-Package `robfilter` (Fried und Schettlinger (2008)) durchgeführt.

4.2.2 Charakteristika aus Wavelets

In der Signalverarbeitung werden häufig Wavelets genutzt, um das Signal von Rauschen zu befreien. Für eine anschauliche Einführung in die Wavelet-Analyse siehe Vidakovic und Müller (1994) oder Hastie et al. (2001) und für eine tiefergehende und detaillierte Betrachtung Wickerhauser (1996) oder Mallat (1999).

Das Signal $Y_{t-m+1}, ..., Y_t$ im Fenster der Länge $m = 2^n$ wird als gewichtete Summe von Funktionen von x modelliert. Zur Anwendung soll hier das Haar-Wavelet kommen. Die Skalierungsfunktion $\phi(x) = \mathbb{1}_{(0 \leq x < 1)}$ generiert die Haar-Wavelet-Basis. Mit dem Haar-Mutter-Wavelet $\psi(x) = \phi(2x) - \phi(2x-1)$ bilden die Funktionen $\psi_{jk}(x) = b \cdot \psi(2^j x - k)$, b Konstante, $j, k \in \mathbb{Z}$, eine Basis des Raums aller quadratisch integrierbarer Funktionen. Betrachte zum Vektor $\mathbf{y}$ die Funktion $f(x) = \sum_{k=0}^{2^n-1} y_{(t-2^n+1)+k} \cdot \mathbb{1}_{(k2^{-n} \leq x < (k+1)2^{-n})}$ auf $[0, 1)$, dann lautet die Wavelet-Zerlegung von f:

$$f(x) = c_{0,0}\phi(x) + \sum_{j=0}^{n-1} \sum_{k=0}^{2^j-1} d_{j,k}\psi_{jk}(x),$$

$c_{0,0}$ und $d_{j,k}$ werden Wavelet-Koeffizienten genannt. Für eine sparsame Darstellung wird in der Signalverarbeitung üblicherweise ein gewisser Anteil der Koeffizienten $d_{j,k}$ gleich Null gesetzt, wenn sie eine zu wählende Schranke unterschreiten. Aber auch Koeffizienten ab einer gewissen Resolution j gleich Null zu setzen, kann sinnvoll sein.

Nachfolgendes Beispiel soll zeigen, in welcher Weise die Wavelet-Koeffizienten Aufschluss über die gesundheitliche Entwicklung eines Patienten geben können. Betrachte einen Vektor $\mathbf{y} = (y_{100}, ..., y_{107})$, der die Herzfrequenz eines Patienten in den 7 Sekunden vor einem Alarm und in der Sekunde der Alarmauslösung in Sekunde 107 der Aufzeichnung enthält. Die Wavelet-Koeffizienten lassen sich aus der Gleichung

$$\begin{pmatrix} y_{100} \\ y_{101} \\ y_{102} \\ y_{103} \\ y_{104} \\ y_{105} \\ y_{106} \\ y_{107} \end{pmatrix} = \begin{pmatrix} 1 & 1 & \sqrt{2} & 0 & 2 & 0 & 0 & 0 \\ 1 & 1 & \sqrt{2} & 0 & -2 & 0 & 0 & 0 \\ 1 & 1 & -\sqrt{2} & 0 & 0 & 2 & 0 & 0 \\ 1 & 1 & -\sqrt{2} & 0 & 0 & -2 & 0 & 0 \\ 1 & -1 & 0 & \sqrt{2} & 0 & 0 & 2 & 0 \\ 1 & -1 & 0 & \sqrt{2} & 0 & 0 & -2 & 0 \\ 1 & -1 & 0 & -\sqrt{2} & 0 & 0 & 0 & 2 \\ 1 & -1 & 0 & -\sqrt{2} & 0 & 0 & 0 & -2 \end{pmatrix} \cdot \begin{pmatrix} c_{0,0} \\ d_{0,0} \\ d_{1,0} \\ d_{1,1} \\ d_{2,0} \\ d_{2,1} \\ d_{2,2} \\ d_{2,3} \end{pmatrix}$$

$\Leftrightarrow$

$$\frac{1}{8}\begin{pmatrix} 1 & 1 & \sqrt{2} & 0 & 2 & 0 & 0 & 0 \\ 1 & 1 & \sqrt{2} & 0 & -2 & 0 & 0 & 0 \\ 1 & 1 & -\sqrt{2} & 0 & 0 & 2 & 0 & 0 \\ 1 & 1 & -\sqrt{2} & 0 & 0 & -2 & 0 & 0 \\ 1 & -1 & 0 & \sqrt{2} & 0 & 0 & 2 & 0 \\ 1 & -1 & 0 & \sqrt{2} & 0 & 0 & -2 & 0 \\ 1 & -1 & 0 & -\sqrt{2} & 0 & 0 & 0 & 2 \\ 1 & -1 & 0 & -\sqrt{2} & 0 & 0 & 0 & -2 \end{pmatrix}^T \begin{pmatrix} y_{100} \\ y_{101} \\ y_{102} \\ y_{103} \\ y_{104} \\ y_{105} \\ y_{106} \\ y_{107} \end{pmatrix} = \begin{pmatrix} c_{0,0} \\ d_{0,0} \\ d_{1,0} \\ d_{1,1} \\ d_{2,0} \\ d_{2,1} \\ d_{2,2} \\ d_{2,3} \end{pmatrix}$$

bestimmen. Der Koeffizient $c_{0,0}$ ist laut oben stehender Gleichung der Mittelwert des Signals im betrachteten Ausschnitt:

$$c_{0,0} = \frac{1}{8}\sum_{i=100}^{107} y_i.$$

Die Abweichung des Mittelwerts in der ersten bzw. der zweiten Hälfte des Fensters von diesem Gesamtmittel wird durch den Koeffizienten d_{00} beschrieben, da

$$d_{0,0} = \frac{1}{8}\left(\sum_{i=100}^{103} y_i - \sum_{i=104}^{107} y_i\right).$$

Rekonstruiert man das Signal bis zu dieser Resolution, so erhält man für $i = 100, ..., 103$

$$\hat{y}_i = c_{0,0} + d_{0,0} = \frac{1}{8}\sum_{i=100}^{107} y_i + \left(\frac{1}{8}\sum_{i=100}^{103} y_i - \frac{1}{8}\sum_{i=104}^{107} y_i\right) = \frac{1}{4}\sum_{i=100}^{103} y_i,$$

den Mittelwert des Signals in der ersten Hälfte des Fensters, und für $i = 104, ..., 107$

$$\hat{y}_i = c_{0,0} - d_{0,0} = \frac{1}{8}\sum_{i=100}^{107} y_i - \left(\frac{1}{8}\sum_{i=100}^{103} y_i - \frac{1}{8}\sum_{i=104}^{107} y_i\right) = \frac{1}{4}\sum_{i=104}^{107} y_i,$$

den Mittelwert des Signals in der zweitem Hälfte des Fensters. Für alle höheren Resolutionslevel $j = 2, ..., n$ setzt sich dieses Prinzip fort. Der Wavelet-Koeffizient gibt also die Abweichung des Mittelwerts des Signals in einem Ausschnitt des Fensters zum Mittelwert im nächst größeren Ausschnitt des Fensters wieder.

Die Berechnung der Wavelet-Koeffizienten erfolgt mit dem R-Package `Wavethresh` (Nason et al. (2006)). Es werden alle Koeffizienten einer Wavelet-Zerlegung der Messwerte der letzten $2^8 - 1$ Sekunden und der Sekunde der Alarmauslösung berechnet. Jedoch werden nur die Koeffizienten bis Resolutionslevel fünf als Charakteristika verwendet.

Auf diese Weise geht der grobe Verlauf des Signals in die Erstellung von Alarmregeln ein, ohne dass der Umfang der Daten unverhältnismäßig anwächst. Betrachtet man nämlich wie hier $512 = 2^8$ Sekunden, so erzeugt man, falls alle Koeffizienten einbezogen werden sollen, 513 neue Variablen pro Vitalparameter, für den diese Zerlegung durchgeführt wird. Bei Verwendung der Koeffizienten bis Resolutionslevel 5 entstehen nur 65 neue Variablen pro Vitalparameter.

Die beschriebenen Charakteristika werden in Kapitel 6.2 zusammen mit den Daten aus der klinischen Studie genutzt, um mit Hilfe von Klassifikationsverfahren Alarmregeln zu generieren. Dort wird untersucht, ob sie, wie vermutet, das Klassifkationsergebnis verbessern und welche der Variablen auf die Klassifikation großen Einfluss nehmen.

Kapitel 5

Klassifikation

Bei Betrachtung einer Menge Π von Objekten (Gesamtpopulation), die in disjunkte Teilpopulationen Π_1, ...,Π_K zerfällt, bedeutet die Klassifikation eines Objekts $\pi \in \Pi$ seine Zuordnung zu einer dieser Populationen. Die Zuordnung erfolgt anhand einer Realisation $\mathbf{x}$ eines Zufallsvektors $\mathbf{X}$ von p Merkmalen, die an dem Objekt beobachtet wird. Eine Klassifikation wird angestrebt, da die Ausprägung g der (zufälligen) Populationszugehörigkeit G jedes Objekts in der Regel nur mit erheblichem Aufwand zu erheben ist.

Klassifikation entspricht einer Unterteilung des Beobachtungsraums $\mathscr{X}$ von $\mathbf{X}$ in Regionen $B_1, ..., B_K$, so dass Objekte, deren zugehörige Realisation $\mathbf{x}$ in Region B_i fällt, der Population Π_i zugeordnet werden. Ziel ist das Auffinden dieser Regionen unter gewissen Optimalitätskriterien. In der Regel werden die Regionen anhand einer „Lernstichprobe“ bestimmt, in der zusätzlich zu den Realisationen von $\mathbf{X}$ verschiedener Objekte auch deren Populationszugehörigkeit g bekannt ist. Mit diesen Regionen ist es dann möglich Objekte, deren Populationszugehörigkeit unbekannt ist, einer der Populationen zuzuordnen. Die Güte der Klassifikation wird durch die Anwendung auf eine „Teststichprobe“ festgestellt, in der wie in der Lernstichprobe die Realisationen zusammen mit ihrer Populationszugehörigkeit gegeben sind.

5.1 Bayes'sches Modell

Ausgehend von der Annahme, dass die Merkmale $\mathbf{X}$ eines Objekts in den Populationen verschiedenen Verteilungen unterliegen und die Populationszugehörigkeit G eines

Objekts zufällig ist, kann diese Situation als zweistufiges Zufallsexperiment aufgefasst werden. Wir betrachten dazu den Maßraum $(\Pi, \mathfrak{A}, P)$ bestehend aus der Menge Π aller Situationen, in denen ein Alarm ausgelöst wird, einer geeigneten σ-Algebra $\mathfrak{A}$ über Π und einem Wahrscheinlichkeitsmaß P auf $\mathfrak{A}$.

Zufallsexperiment 1: Populationszugehörigkeit eines Objekts

Im hier betrachteten Kontext des intensivmedizinischen Monitorings ist ein zu klassifizierendes Objekt π eine Situation, in der das Standard-Monitoringgerät am Patientenbett einen Alarm auslöst. Jede dieser Alarmsituationen π gehört entweder der Population Π_0 der alarmrelevanten Situationen oder der Population Π_1 der nicht alarmrelevanten Situationen an. Die Vereinigung der $K = 2$ disjunkten Teilpopulationen Π_0 und Π_1 bildet die Gesamtpopulation Π.
Die Populationszugehörigkeit $G(\pi)$ eines Objekts sei eine Zufallsvariable, die Werte aus $\{0, 1\}$ annimmt:

$$G : \Pi \to \{0, 1\}$$

mit den a-priori-Wahrscheinlichkeiten $P_G(i) = P(G^{-1}(i)) = p_i$ $(i = 0, 1)$. Dabei gilt für alle Situationen π aus Π_i: $G(\pi) = i$ mit der Interpretation, dass die Zufallsvariable G den Wert 0 annimmt, wenn die Situation π alarmrelevant ist, und den Wert 1, wenn die Situation π nicht alarmrelevant ist. In der Praxis sind weder G noch P_G bekannt.

Zufallsexperiment 2: Merkmale eines Objekts aus Population Π_i

Die Merkmale $\mathbf{X}$ einer zu klassifizierenden Alarmsituation $\pi \in \Pi_i$ sind in den vorliegenden Daten sowohl stetig als auch kategoriell und binär. Sie bestehen zum einen aus q sekündlich gemessenen stetigen Vitalparametern, wie z.B. der Herzfrequenz oder dem arteriellen systolischen Blutdruck. Neben stetigen sind auch r kategorielle Merkmale, wie z.B. der Alarmgrad mit Ausprägungen „ADV“, „SER“ und „LT“ und s binäre Merkmale, wie z.B. die fehlende Werte anzeigenden Dummy-Variablen, vertreten.
Bezeichne mit $\mathbf{X}(\pi)$ den Zufallsvektor der Merkmale eines Objekts, der jede Situation

$\pi \in \Pi$ in den Beobachtungsraum $\mathscr{X}$ abbildet:

$$\mathbf{X} : \Pi \to \mathbb{R}^q \times \{0,1\}^s \times \{\text{ADV, SER, LT}\} \times ... =: \mathscr{X},$$

mit Verteilung $P_{\mathbf{X}}$ auf einer geeigneten σ-Algebra $\mathfrak{C}$ über $\mathscr{X}$. Eine Grundannahme in der Klassifikation ist, dass sich die Verteilung der Merkmale $\mathbf{X}$ in den Populationen unterscheidet, d.h. $\exists\, B : P_{\mathbf{X}|G=0}(B) \neq P_{\mathbf{X}|G=1}(B)$. Es wird also angenommen, dass sich die Verteilungen der Merkmale von alarmrelevanten Situationen und nicht alarmrelevanten Situationen unterscheiden. Auch diese Verteilungen sind in der Praxis nicht bekannt. Beobachtbar sind allein die Ausprägungen des Zufallsvektors $\mathbf{X}$. Von maßgeblichem Interesse für die Erstellung von Klassifikationsregeln sind die bedingten Verteilungen $P_{X|G=i}$, $i = 0, 1$, mit zugehörigen mehrdimensionalen Dichten f_i bezüglich eines Maßes μ (z.B. eines geeigneten Produktmaßes aus Lebesgue- und Zählmaßen).

Eine binäre *Klassifikationsregel* ist hier eine Entscheidungsregel $\delta : \mathscr{X} \to \{0,1\}$, die jede Beobachtung $\mathbf{x}$ einer Population zuordnet:

$$\delta(\mathbf{x}) = \begin{cases} 0 & : \quad \mathbf{x} \in B_0 \\ 1 & : \quad \mathbf{x} \in B_1 \end{cases}, \quad B_0 \dot{\cup} B_1 = \mathscr{X}.$$

Dies entspricht einer Partition des Beobachtungsraums $\mathscr{X}$ in B_0 und B_1, so dass Beobachtungen aus B_i der Population Π_i zugeordnet werden. Man folgert also, dass die Situation π alarmrelevant ist für $\mathbf{X}(\pi) \in B_0$, und dass sie nicht alarmrelevant ist für $\mathbf{X}(\pi) \in B_1$. Die Regionen B_0 und B_1 können mit verschiedenen Methoden aufgefunden werden. Nachfolgend wird der häufig verwendete Bayes-Ansatz vorgestellt.

Bei der Bestimmung der Regionen B_0 und B_1 können Kosten $c_{j|i}$ für eine Fehlklassifikation eines Objekts aus Population Π_i als zu Π_j gehörend berücksichtigt werden. Die Verlustfunktion $V(i,j) = c_{i|j}$, $i, j = 0, 1$, gibt für $V(1,0) = c_{1|0}$ die Kosten an, die entstehen, wenn eine alarmrelevante Situation als nicht alarmrelevant klassifiziert wird. $V(0,1) = c_{0|1}$ sind die Kosten für eine Fehlklassifikation einer nicht alarmrelevanten Situation als alarmrelevant. Der erwartete Verlust bzw. das Bayes-Risiko einer

Klassifikationsregel $\delta(\mathbf{x})$ ist

$$R(\delta) = \pi_1 c_{0|1} \int_{B_0} f_1(\mathbf{x}) d\mathbf{x} + \pi_0 c_{1|0} \int_{B_1} f_0(\mathbf{x}) d\mathbf{x}.$$

Das Bayes-Risiko wird von der Bayes-Regel

$$\delta^\star(\mathbf{x}) = \underset{i,j \in \{0,1\}, i \neq j}{\arg\min} \quad c_{j|i} p_i f_i(\mathbf{x}) \tag{5.1}$$

minimiert. Die Lösung $\delta^\star$ des Minimierungsproblems ist eindeutig, falls $P(\pi_1 c_{0|1} f_1(\mathbf{x}) = \pi_0 c_{1|0} f_0(\mathbf{x}))$ gleich Null ist (vgl. Anderson (2003), Kap. 6.7). Daraus folgt:

$$B_0 = \{\mathbf{x} \in \mathscr{X} : \pi_1 c_{0|1} f_1(\mathbf{x}) < \pi_0 c_{1|0} f_0(\mathbf{x})\}$$

und

$$B_1 = \{\mathbf{x} \in \mathscr{X} : \pi_0 c_{1|0} f_0(\mathbf{x}) < \pi_1 c_{0|1} f_1(\mathbf{x})\}$$

bilden die gesuchte Partition des Beobachtungsraums, die das Bayes-Risiko minimiert. Äquivalent dazu lassen sich B_0 und B_1 durch die a-posteriori-Verteilungen beschreiben:

$$B_0 = \{\mathbf{x} \in \mathscr{X} : c_{0|1} P(G = 1|\mathbf{X} = \mathbf{x}) < c_{1|0} P(G = 0|\mathbf{X} = \mathbf{x})\}$$

und

$$B_1 = \{\mathbf{x} \in \mathscr{X} : c_{1|0} P(G = 0|\mathbf{X} = \mathbf{x}) < c_{0|1} P(G = 1|\mathbf{X} = \mathbf{x})\}.$$

Die a-priori-Verteilungen, die bedingten Dichten sowie die a-posteriori-Verteilungen und Fehlklassifikationskosten sind in der Regel unbekannt. Es können Verteilungsannahmen getroffen werden und Fehlklassifikationskosten (idealerweise aus dem Sachproblem begründet) gewählt werden. Die bekanntesten statistischen Klassifikationsverfahren basierend auf Verteilungsannahmen sind die lineare bzw. quadratische Diskriminanzanalyse und die logistische Regression. Ein Überblick über verschiedene Klassifikationsmethoden mit weiterführenden Literaturempfehlungen findet sich z.B. in Hastie et al. (2001), McLachlan (1992) oder Anderson (2003).

Für intensivmedizinische Daten sind sinnvolle Verteilungsannahmen schwer zu treffen. Besonders die Annahme der Normalverteilung ist in der Regel nicht gerechtfertigt. Stattdessen kann eine geeignete Partition des Beobachtungsraums auf nichtparametrische Weise zum Beispiel mit Hilfe von Entscheidungsbäumen gefunden werden. Bei diesem Vorgehen gelangt man über die Konstruktion eines Entscheidungsbaums zur Klassifikationsregel δ.

5.2 Beurteilung der Güte von Klassifikationsregeln

Die Güte der Klassifikationsregel wird häufig mit Hilfe des Anteils richtig klassifizierter Beobachtungen der Teststichprobe bewertet und als ihre Klassifikationsgüte bezeichnet. Aber auch differenziertere Betrachtungen sind möglich und im Fall der Alarmklassifikation auch notwendig. Bei der Klassifikation von Alarmsituationen wird eine Klassifikationsentscheidung als *richtig positiv* gewertet, wenn eine alarmrelevante Situation als solche erkannt und ein Alarm ausgelöst wird. Wird sie fälschlicherweise nicht erkannt und kein Alarm ausgelöst, wird die Entscheidung *falsch negativ* genannt. Lautet die Klassifikationsentscheidung in einer nicht alarmrelevanten Situation richtigerweise keinen Alarm auszulösen, dann wird die Entscheidung als *richtig negativ* bezeichnet. In einer solchen Situation einen Alarm auszulösen, entspricht einer *falsch positiven* Entscheidung. Diese Bezeichnungen sind schematisch dargestellt in Tabelle 5.1.

Neben der Klassifikationsgüte werden auch andere Größen, wie beispielsweise die Fehlklassifikationsrate (= 1 − Klassifikationsgüte), Sensitivität, Spezifität, der positive prädiktive und negative prädiktive Wert oder das so genannte F1-Maß zur Beurteilung herangezogen. Die Wahl der geeigneten Größe hängt von der Aufgabenstellung und Zielsetzung ab, zu deren Zweck klassifiziert wird. Die erwähnten Gütemaße sind wie folgt definiert:

- $\text{Klassifikationsgüte} = \dfrac{\text{richtig positive} + \text{richtig negative}}{\text{alle}}$
- $\text{Fehlklassifikationsrate} = \dfrac{\text{falsch positive} + \text{falsch negative}}{\text{alle}}$
- $\text{Sensitivität} = \dfrac{\text{richtig positive}}{\text{richtig positive} + \text{falsch negative}}$
- $\text{Spezifität} = \dfrac{\text{richtig negative}}{\text{richtig negative} + \text{falsch positive}}$
- $\text{positiver prädiktiver Wert} = \dfrac{\text{richtig positive}}{\text{richtig positive} + \text{falsch positive}}$
- $\text{negativer prädiktiver Wert} = \dfrac{\text{richtig negative}}{\text{richtig negative} + \text{falsch negative}}$
- $\text{F1-Maß} = \dfrac{2 \cdot \text{positiver prädiktiver Wert} \cdot \text{Sensitivität}}{\text{positiver prädiktiver Wert} + \text{Sensitivität}}$.

Hier wird zur Beurteilung der Güte der durch Klassifikation gewonnenen Alarmregeln zum einen die Sensitivität verwendet, da sie den Anteil der alarmrelevanten Situationen

angibt, auf die zu Recht durch einen Alarm hingewiesen wird. Um einen Vergleich der Fehlalarmhäufigkeiten des neuen und des alten Systems zu ermöglichen, wird zum anderen die Fehlalarmreduktion betrachtet:

$$\text{Fehlalarmreduktion} = \frac{\text{richtig negative (neues Alarmsystem)}}{\text{falsch positive (altes System)}}.$$

Diese beiden Größen geben an, wie häufig ein Patient durch einen ausbleibenden Alarm gefährdet wird und um wie viel weniger das Pflegepersonal durch unnötige Alarme belästigt wird.

Situation	Alarm wird gegeben	Alarm wird nicht gegeben
alarmrelevant	richtig positiv	falsch negativ
nicht alarmrelevant	falsch positiv	richtig negativ

Tabelle 5.1: Bezeichnung richtig bzw. falsch positiver und negativer Ereignisse

5.3 Entscheidungsbäume

5.3.1 Definition

Ein *Entscheidungsbaum* ist ein gerichteter, azyklischer, endlicher Graph, d.h. ein Tupel (E, D) mit einer nichtleeren Menge E von Knoten $A \in E$, A Element der σ-Algebra $\mathfrak{C}$ über $\mathscr{X}$, und einer nichtleeren Menge D von gerichteten Kanten $K = (A_1, A_2)$, $A_1, A_2 \in E$ (vgl. z.B. Edwards (2000)). Bei der Kante $K = (A_1, A_2)$ wird Knoten A_1 Elternknoten und Knoten A_2 Kindknoten genannt. Ein endlicher Graph besitzt eine endliche Anzahl Knoten und Kanten ($|E|, |D| < \infty$). Ein Graph heißt azyklisch, falls er nur Kantenfolgen $K_1, ..., K_l \in D$ mit $K_i = (A_{i-1}, A_i)$, $i = 1, ..., l$ enthält, deren Anfangspunkt A_0 und Endpunkt A_l verschieden sind.

Jeder Knoten eines Entscheidungsbaums ist hier eine Teilmenge von $\mathscr{X}$, genauer eine Element von $\mathfrak{C}$. $\mathscr{X}$ selbst ist ebenfalls ein Knoten, der sogenannte Wurzelknoten, $\emptyset$ ist im Allgemeinen kein Knoten. In den Wurzelknoten gehen keine Kanten ein, d.h. $D \supset \{(A_1, A_2) : A_1, A_2 \in E, A_2 \neq \mathscr{X}\}$. Alle übrigen Knoten des Entscheidungsbaums haben genau eine eingehende Kante, d.h. $(A_1, A), (A_2, A) \in E \Rightarrow A_1 = A_2$. Von jedem

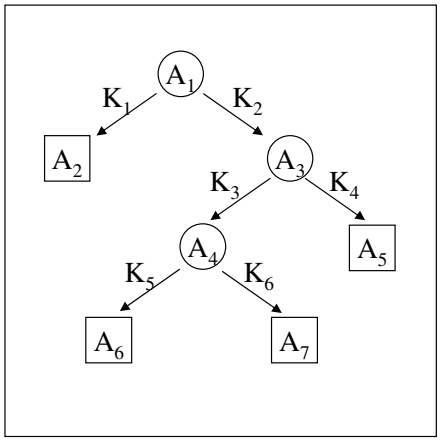

Abbildung 5.1: Entscheidungsbaum

Knoten gehen keine oder mindestens zwei Kanten aus. Ein solcher *Split* mit mindestens zwei Kindknoten bildet eine Partition des Elternknotens, d.h. zum Beispiel für einen binären Split, dass $(A, A_1), (A, A_2) \in E \rightarrow A_1 \dot{\cup} A_2 = A$. Knoten ohne ausgehende Kanten heißen Blattknoten. Abbildung 5.1 veranschaulicht beispielhaft einen binären Entscheidungsbaum. Er besteht aus sieben Knoten $A_1, ..., A_7$ und sechs gerichteten Kanten $K_1, ..., K_6$. Knoten A_1 ist der Wurzelknoten des Entscheidungsbaums. Knoten A_3 ist der Elternknoten von seinen beiden Kindknoten A_4 und A_5, die Vereinigung der beiden disjunkten Knoten A_4 und A_5 ist also gerade gleich Knoten A_3. Die Knoten A_2, A_5, A_6 und A_7 sind die Blattknoten des Entscheidungsbaums.

Entscheidungsbäume können mit Hilfe rekursiver Partitionierung konstruiert werden. Über die Anteile der Populationen in den Blattknoten des konstruierten Entscheidungsbaums werden die Regionen B_0 und B_1 der gesuchten Klassifikationsregel δ definiert.

5.3.2 Rekursive Partitionierung

Rekursive Partitionierung beinhaltet die schrittweise Zerlegung des $(q + r + s)$-dimensionalen Beobachtungsraums $\mathscr{X} = \mathscr{X}_1 \times ... \times \mathscr{X}_{(q+r+s)}$, in disjunkte Teilmengen anhand einer Stichprobe $\{(\mathbf{x}_1, g_1), ..., (\mathbf{x}_n, g_n)\}$. Im ersten Schritt bei der Konstruktion eines binären Entscheidungsbaums wird $\mathscr{X}$ in die Kindknoten A_1 und A_2 zerlegt, wobei

$\mathscr{X} = A_1 \dot{\cup} A_2$ und

$$\begin{aligned} A_1 &= \mathscr{X}_1 \times ... \times \quad \tilde{A} \quad \times ... \times \mathscr{X}_{(q+r+s)} \quad \text{und} \\ A_2 &= \mathscr{X}_1 \times ... \times \quad \mathscr{X}_i \setminus \tilde{A} \quad \times ... \times \mathscr{X}_{(q+r+s)}, \end{aligned}$$

mit z.B. $\tilde{A} = \{x : x \in \mathscr{X}_i, x \leq \tilde{a} \in \mathbb{R}\}$ im Fall $\mathscr{X}_i = \mathbb{R}$. Die Dimension i und die Menge $\tilde{A}$ werden dabei so gewählt, dass eine Funktion der Beobachtungen aus der Stichprobe, die die Vermischung der Populationen der entstehenden Kindknoten misst, maximal wird. Diese Vermischung wird als Unreinheit bezeichnet. Die Zerlegung, genannt *Split*, erfolgt nach der so genannten *Splitting-Regel*, die die Funktion festlegt, mit der die Zunahme an Reinheit gemessen wird. Die resultierenden Teilmengen werden wiederum in gleicher Weise geteilt, bis eine *Stop-Splitting-Regel* greift. Auf diese Weise entsteht ein binärer Entscheidungsbaum. Die Klassifikationsregel δ wird dann durch folgende Festlegung der Regionen B_0 und B_1 konstruiert:

$$\begin{aligned} B_0 &= \bigcup \{A_j : |\{g_i = 0, \mathbf{x}_i \in A_j\}| >= |\{g_i = 1, \mathbf{x}_i \in A_j\}|, A_j \text{ Blattknoten}\} \\ B_1 &= \bigcup \{A_j : |\{g_i = 0, \mathbf{x}_i \in A_j\}| < |\{g_i = 1, \mathbf{x}_i \in A_j\}|, A_j \text{ Blattknoten}\}. \end{aligned}$$

Die Vereinigung aller Blattknoten, in denen die Mehrheit der Objekte Population Π_0 angehört, definiert also Region B_0, und die Vereinigung aller übrigen Blattknoten definiert Region B_1.

Verschiedene Algorithmen stehen zur Durchführung einer rekursiven Partitionierung zur Verfügung.

5.3.3 Algorithmen

Die drei bekanntesten Algorithmen zur rekursiven Partitionierung sind

- „Chi Square Automatic Interaction Detection“ (CHAID) (Kass (1980)),
- „Classification and Regression Trees“ (CART) (Breiman et al. (1984)),
- „C4.5“ (Quinlan (1993)).

Der CHAID Algorithmus wurde in erster Linie für kategoriale Variablen entworfen. Stetige Variablen werden klassiert. Die Splitting-Regel basiert auf einem χ^2-Test und kann zu einer Aufteilung eines Knotens in mehr als zwei Kindknoten führen. Das Splitting wird beendet, sobald die p-Werte zu groß sind.

Im Gegensatz dazu können die Algorithmen CART and C4.5 eine rekursive Partitionierung sowohl für kategoriale als auch stetige Variablen vornehmen, ohne dass die stetigen Variablen klassiert werden. Während der CART Algorithmus binäre Bäume erzeugt, teilt der C4.5 Algorithmus beim Splitting auf einer kategorialen Variablen den Knoten in je einen Kindknoten pro Kategorie. Im Fall stetiger und binärer Variablen, wie sie im intensivmedizinischen Monitoring hauptsächlich zu beobachten sind, unterscheiden sich die beiden Algorithmen nicht wesentlich. Da CART auf einen allgemeineren Ansatz zur Messung der Reinheit von Knoten aufbaut (siehe unten) als der ausschließlich auf der Entropie basierende C4.5-Algorithmus, wird im Folgenden der CART Algorithmus detaillierter beschrieben und später zur Generierung von Alarmregeln verwendet.

CART

Im Fall eines Zwei-Populationen-Problems sei $a_{k|t} = |\{\mathbf{x}_i : \mathbf{x}_i \in A_t, g_i = k, i = 1, ..., n\}|/|\{A_t\}|$, $k = 0, 1$, der Anteil an Objekten aus Population Π_k in Knoten A_t mit $a_{0|t}, a_{1|t} \geq 0$ und $a_{0|t} + a_{1|t} = 1$. Zur Messung der (Un-)Reinheit eines Knotens A_t kann eine Funktion $\phi_{A_t} : [0, 1] \to \mathbb{R}$, $A_t \subseteq \mathscr{X}$ herangezogen werden, die folgende Eigenschaften hat:

ϕ hat ihr einziges Maximum bei $a_{0|t} = 0,5$,

ϕ hat ihre Minima bei $a_{0|t} = 0$ und $a_{0|t} = 1$,

ϕ ist symmetrisch.

Beispiele (vgl. Abb. 5.2) für solche, die Unreinheit messende Funktionen ϕ sind

- Gini Index: $\phi(a_{0|t}) = 2a_{0|t}(1 - a_{0|t})$
- Entropie: $\phi(a_{0|t}) = -a_{0|t}log(a_{0|t}) - (1 - a_{0|t})log(1 - a_{0|t})$.

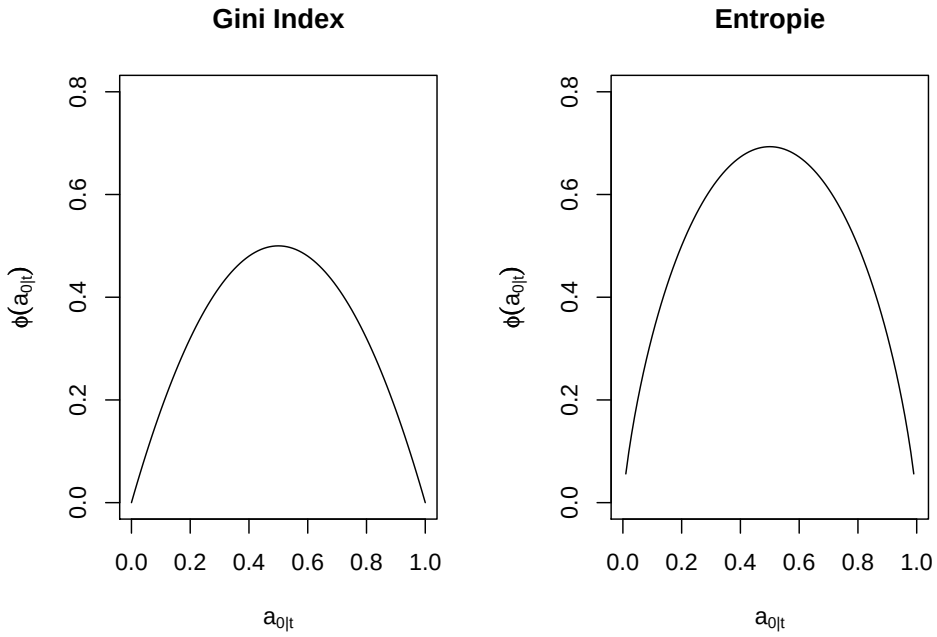

Abbildung 5.2: Gini Index und Entropie

Ein Split s teilt einen Knoten A_t des Entscheidungsbaums wie zuvor beschrieben in einen linken Kindknoten A_l und einen rechten Kindknoten A_r. Die Abnahme an Unreinheit eines Knotens durch einen Split s wird definiert als

$$\psi(s,t) = \phi_{A_t}(a_{0|t}) - \frac{|A_l|}{|A_t|}\phi_{A_l}(a_{0|l}) - \frac{|A_r|}{|A_t|}\phi_{A_r}(a_{0|r}).$$

Die Splitting-Regel wählt den Split $s^\star$ für einen Knoten A_t, der zur größten Abnahme an Unreinheit bzw. zur größten Zunahme an Reinheit führt, d.h.

$$s^\star = \arg\max_s \quad \psi(s,t).$$

Splits werden im Allgemeinen nur auf einer Variablen vorgenommen: Sei beispielsweise X_1 eine stetige Variable mit Wertebereich $\mathbb{R}$, so könnte der beste Split $s^\star$ derjenige sein, der die Objekte mit $X_1 < 3$ von den Objekten mit $X_1 \geq 3$ trennt. Es ist aber auch möglich, Splits auf (zufälligen) Kombinationen von Variablen zuzulassen.

Stop-Splitting-Regeln beinhalten in der Regel eine minimale Knotengröße oder eine mindestens durch einen Split zu erzeugende Abnahme an Unreinheit.

5.3.4 Eigenschaften und Beispiele

Entscheidungsbäume liefern Klassifikationsregeln, die der menschlichen Entscheidungsbildung sehr ähnlich sind. Sie lassen sich als Abfolge von „Ja-Nein-Fragen“ darstellen und münden je nach Abfolge von „Ja“ und „Nein“ in einer Entscheidung.

Abbildung 5.3 zeigt ein Beispiel für einen Entscheidungsbaum zur Unterscheidung alarmrelevanter und nicht alarmrelevanter Situationen in der Intensivmedizin. Eine Situation wird hier beispielsweise als alarmrelevant eingestuft, falls die Atemfrequenz niedriger ist als 54,4 Atemzüge pro Minute, der untere Schwellwert für den arteriellen systolischen Blutdruck von den Pflegern auf unter 97 mmHg eingestellt wurde, die Sauerstoffsättigung mindestens 97,5% beträgt und gleichzeitig ein Puls von 123 Schläge pro Minute oder höher gemessen wird. Diese Einstufung als alarmrelevant erfolgt unabhängig davon, für welche Variable und aus welchem Grund der Alarm ausgelöst wurde. Die anschauliche Darstellbarkeit auch komplizierterer Regeln unterscheidet Entscheidungsbäume von vielen anderen Klassifikationsverfahren, wie zum Beispiel Support Vector Machines oder Neuronalen Netzen. Sie ermöglicht eine inhaltliche Interpretation und kann Sachzusammenhänge offenlegen.

Ein Nachteil von Entscheidungsbäumen ist ihre Instabilität (Breiman et al. (1984), S.156). Bei einer Änderung der Aufteilung in Lern- und Teststichprobe kann ein völlig anderer Entscheidungsbaum entstehen (Abb. 5.4). Auch bei diesem neuen Entscheidungsbaum ist eine medizinische Interpretation möglich − allerdings eine möglicherweise stark abweichende. Die Tatsache, dass mit einer anderen Lernstichprobe bereits ein anderer Entscheidungsbaum konstruiert wurde, lässt daher an der medizinischen Bedeutung der Regeln zweifeln.

Die dargestellten Entscheidungsbäume sind nach ihrer Konstruktion zurückgeschnitten worden („Pruning“, vgl. Breiman et al. (1984)), um eine bei Entscheidungsbäumen häufig auftretende Überanpassung an die Lerndaten zu vermeiden. Überanpassung bedeutet hier, dass die erzeugte Partition auf den Lerndaten so fein ist, dass nicht die allgemeine Struktur sondern die Struktur genau dieser Lerndaten wiedergegeben wird. Typischerweise wird ein hoher Anteil der Beobachtungen aus dem Lerndatensatz durch einen überangepassten Entscheidungsbaum richtig klassifiziert, aber nur ein deutlich geringerer Anteil der Beobachtungen aus dem Testdatensatz.

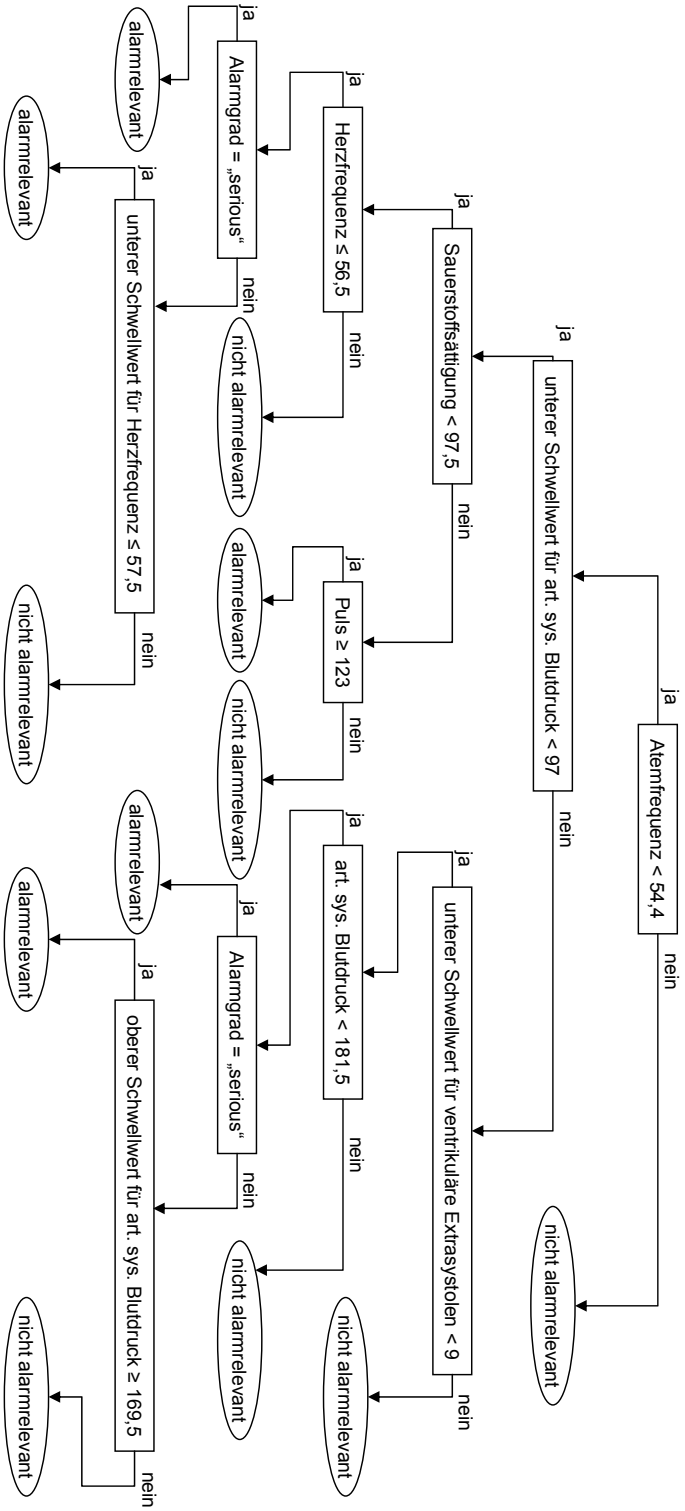

Abbildung 5.3: Entscheidungsbaum für die intensivmedizinische Alarmgebung

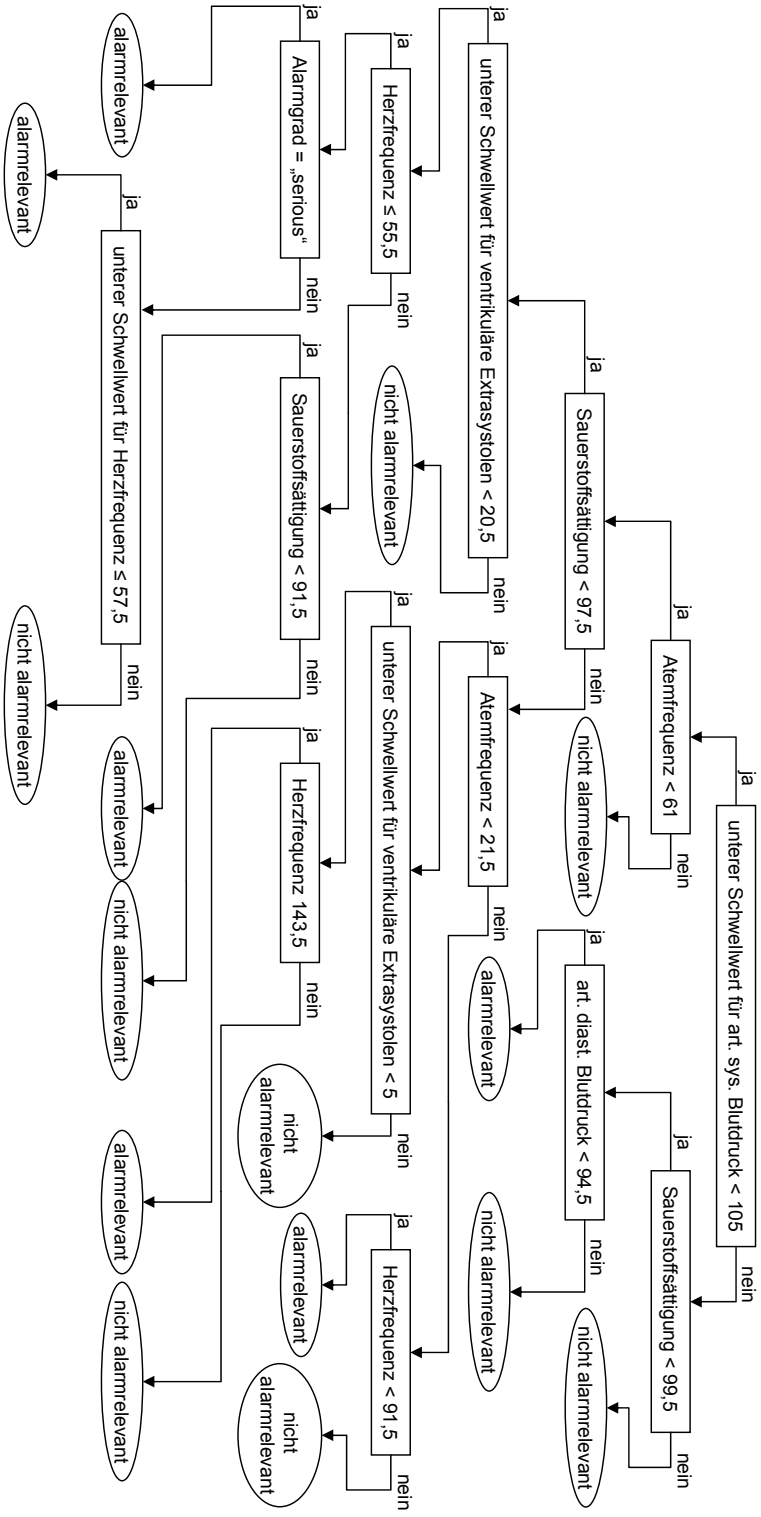

Abbildung 5.4: Entscheidungsbaum für die intensivmedizinische Alarmgebung

Um diesem Effekt entgegenzuwirken, wird die Partition durch Pruning künstlich vergröbert, um so die Anteile der richtig klassifizierten Beobachtungen in den Lern- und Testdaten einander anzunähern. Darüber hinaus sind Entscheidungsbäume anderen Verfahren wie Support Vektor Machines oder gebaggten (Breiman (1996)) oder geboosteten (Schapire (2003)) Verfahren in ihrer Klassifikationsgüte oft unterlegen (Breiman (2001)).

Aufgrund der beiden genannten Probleme von Bäumen, der Instabilität und der geringen Klassifikationsgüte, werden im Folgenden statt einzelner Bäume Ensembles von Bäumen verwendet, die als eine Lösung für diese Probleme bekannt sind (Breiman (2001)). Der Verlust der Darstellbarkeit der resultierenden Regeln kann dabei in Kauf genommen werden, da das Ziel der Klassifikation von Alarmsituationen in dieser Arbeit nicht die Einsicht in medizinische Zusammenhänge sondern eine möglichst zuverlässige Alarmgebung ist. Ensembles von anderen Klassifikationsregeln sind ebenfalls denkbar, werden aber hier nicht näher betrachtet, da sich die Vermittelbarkeit von Entscheidungsbäumen auf das Ensemble überträgt: Die *Konstruktion* eines Ensembles von Bäumen lässt sich leicht und anschaulich erklären, da Bäume leicht und anschaulich zu erklären sind. Es wird ein kleiner Einblick in die „Black Box“ ermöglicht, wodurch, so wird vermutet, die Akzeptanz der Anwender positiv beeinflusst wird.

5.4 Wälder

5.4.1 Definition

Ein Wald im Sinne eines *Random Forest* (Breiman (2001)) ist ein *Ensemble* von Entscheidungsbäumen mit folgender Arbeitsweise: Jeder Entscheidungsbaum wird auf einer unabhängig und zufällig gewählten Bootstrap-Stichprobe $\{(\mathbf{x}_{b_1}, g_{b_1}), ..., (\mathbf{x}_{b_m}, g_{b_m})\}$ der Lernstichprobe $\{(\mathbf{x}_{j_1}, g_{j_1}), ..., (\mathbf{x}_{j_m}, g_{j_m})\} \subset \{(\mathbf{x}_1, g_1) ,..., (\mathbf{x}_n, g_n)\}$, $m \leq n$, konstruiert. Die Dimensionen, in denen ein möglicher Split erfolgt, werden auf eine feste Anzahl in jeder der Rekursionen zufällig ausgewählter Kandidaten beschränkt. Ein Objekt wird von jedem einzelnen Entscheidungsbaum klassifiziert und der Population mit den Stimmen der meisten Bäume zugeordnet.

Bezeichne $\delta_j(\mathbf{x})$ den j-ten von N Entscheidungsbäumen des Waldes,

$$\delta_j(\mathbf{x}) = \begin{cases} 0 & : \quad \mathbf{x} \in B_0^j \\ 1 & : \quad \mathbf{x} \in B_1^j \end{cases}, \quad j = 1, ..., N,$$

dann ist der Wald $\delta_{Wald}(\mathbf{x})$ definiert als

$$\delta_{Wald}(\mathbf{x}) = \begin{cases} 0 & : \quad \sum_{j=1}^{N} \delta_j(\mathbf{x}) < N/2 \\ 1 & : \quad \sum_{j=1}^{N} \delta_j(\mathbf{x}) \geq N/2 \end{cases}.$$

5.4.2 Algorithmen

Der von Breiman und Cutler (for) entwickelte Fortran-Code zur Konstruktion eines Random Forest steht zur nicht-kommerziellen Nutzung der Allgemeinheit zur Verfügung. In der Statistik-Software R ist dieser Code im `randomForest`-Package (Liaw und Wiener (2002)) implementiert. Die Entscheidungsbäume der so erzeugten Wälder werden nach dem Prinzip des CART-Algorithmus konstruiert. Andere Algorithmen zur Konstruktion von Entscheidungsbäumen zu verwenden, ist ebenso möglich.

5.4.3 Eigenschaften

Bagging bezeichnet eine Methode, die auf der Anwendung eines Algorithmus (Klassifikation, Regression, etc.) auf Bootstrap-Stichproben der Lernstichprobe und der Aggregation der Ergebnisse basiert. In diesem Sinne sind Wälder Ensembles von „gebaggten“ und randomisierten CART-Bäumen. In empirischen Studien konnte ihre Überlegenheit bezüglich Fehlklassifikation und Stabilität gegenüber einzelnen Bäumen in vielen Fällen gezeigt werden (Breiman (2001)). Generelle theoretische Erkenntnisse über die Gründe hierfür liegen bislang noch nicht in umfangreichem Maß vor. Erste theoretische Untersuchungen des Bagging führen z.B. Buja und Stuetzle (2006) und Bühlmann und Yu (2002) durch.

Buja und Stuetzle (2006) untersuchen den Effekt von Bagging auf U-Statistiken. Sie zeigen, dass Stichprobenziehungen mit Zurücklegen (u.a. Bootstrap) und Stichprobenziehungen ohne Zurücklegen zu den gleichen gebaggten U-Statistiken führen, falls der

Anteil der Stichprobe mit Zurücklegen a_m an der Lernstichprobe zum Anteil der Stichprobe ohne Zurücklegen a_o in folgendem Verhältnis steht: $a_m = a_o/(1 - a_o)$. Das bedeutet, dass eine klassische Bootstrap-Stichprobe zur gleichen U-Statistik führt wie eine Stichprobe ohne Zurücklegen vom halben Umfang (*Half-Sampling*), wodurch der Rechenaufwand reduziert wird. Darüber hinaus zeigen sie, dass unter gewissen Bedingungen die Varianz, der Bias und auch beide zugleich durch Bagging reduziert werden, worin generell der Grund für die Überlegenheit von gebaggten Methoden bzgl. Vorhersagefehlern liegen kann. Obwohl CART-Bäume und U-Statistiken verschieden sind, zeigt sich in Simulationen, dass Bagging mit Zurücklegen zu ähnlichen Ergebnissen bzgl. Bias, Varianz und MSE führt wie Bagging ohne Zurücklegen mit entsprechendem Stichprobenumfang. Die Autoren vermuten daher, dass der von ihnen gefundene Zusammenhang nicht nur für U-Statistiken gilt sondern möglicherweise universell für Bagging. Diese Vermutung ist für die nachfolgend in dieser Arbeit durchgeführten Simulationen bei der Wahl der Stichprobenziehung von Bedeutung.

Bühlmann und Yu (2002) formalisieren den Aspekt der Instabilität und leiten daraus Aussagen über die Varianz-reduzierende Wirkung von Bagging ab. Auch diese Autoren beschäftigen sich mit der Variante des Bagging, bei der ohne Zurücklegen gezogen wird, da sie leichter zu analysieren ist. Sie stellen in Simulationen ebenfalls fest, dass das weniger rechenintensive Ziehen ohne Zurücklegen von halbem Stichprobenumfang nur geringfügig schlechtere Klassifikationsergebnisse als klassisches Bootstrapping liefert. Breiman (2001) zeigt selbst, dass die erwartete Fehlklassifikationsrate von Wäldern nach oben beschränkt ist, und der Wald mit wachsender Anzahl an Bäumen nicht zur Überanpassung neigt.

Wälder nach Breiman sind für die Klassifikation von Alarmsituationen nur bedingt geeignet. Sie liefern zwar eine insgesamt geringe Fehlklassifikationsrate, ein Kontrollmechanismus, mit dem die Fehlklassifikation alarmrelevanter Situationen weitgehend vermieden werden kann, existiert jedoch nicht. Nachfolgend wird zunächst der Zusammenhang zwischen der Klassifikation von Alarmsituationen und statistischen Tests beleuchtet. Auf dieser Grundlage wird daraufhin eine Modifikation für Wälder vorgeschlagen, die es erlaubt, die Fehlklassifikationsrate alarmrelevanter Situationen zu kontrollieren.

5.5 Klassifikation aus der Sichtweise statistischer Tests

5.5.1 Übereinstimmungen der zugrunde liegenden Situationen

In der Intensivmedizin besteht ein Ungleichgewicht der Konsequenzen bei der Fehlklassifikation von alarmrelevanten und nicht alarmrelevanten Situationen. Wird eine alarmrelevante Situation fälschlicherweise als nicht alarmrelevant eingestuft und deshalb kein Alarm ausgelöst, so kann dies schwerwiegende Konsequenzen für die Gesundheit des Patienten bedeuten. Wird hingegen eine nicht alarmrelevante Situation fälschlicherweise als alarmrelevant eingestuft und ein unnötiger Alarm ausgelöst, bedeutet dies lediglich eine Störung oder Belästigung für das Personal. Aus diesem Grund muss in der Klassifikation von Alarmsituationen die Wahrscheinlichkeit, eine alarmrelevante Situation falsch zu klassifizieren, kontrolliert werden. Unter dieser Bedingung kann die Wahrscheinlichkeit, eine nicht alarmrelevante Situation falsch zu klassifizieren, reduziert werden.

Es ist hier also nicht, wie im Bayes-Ansatz üblich, die Wahrscheinlichkeit irgendeines Klassifikationsfehlers von Interesse. In der Regel wird aus klassischer Bayes-Sicht heraus die Klassifikationsregel so gewählt, dass der erwartete Verlust bei üblicherweise gleich großen Fehlklassifikationskosten minimiert wird. Möchte man im Bayes-Ansatz die besondere Situation im Patienten-Monitoring berücksichtigen und die Fehlklassifikation für eine Klasse möglichst vermeiden, so können Fehlklassifikationen mit unterschiedlich hohen Kosten belegt werden. Eine falsch klassifizierte alarmrelevante Situation müsste in diesem Fall mit erheblich höheren Kosten verbunden sein. Wie hoch genau die Kosten sein müssen, ist allerdings nur schwer zu entscheiden. Inhaltlich lassen sich sinnvolle Kosten nicht ableiten, daher müssen bei diesem Ansatz Klassifkationsregeln mit verschiedenen Verlustfunktionen generiert werden, bis eine akzeptable Regel gefunden ist.

Alternativ kann eine Grenze für die Fehklassifikationswahrscheinlichkeit alarmrelevanter Situationen vorgegeben und unter dieser Bedingung die beste Klassifikationsregel gewählt werden. Dieses Vorgehen entspricht der Idee, die statistischen Tests nach der Neyman-Pearson-Theorie zugrunde liegt.

5.5.2 Neyman-Pearson Lemma

In der Klassifikation ist eine Klassifikationsregel ein Test, bei dem die beiden Hypothesen den beiden Populationen entsprechen. Das bisher beschriebene Vorgehen zur Konstruktion einer Klassifikationsregel, die das Bayes-Risiko minimiert, entspricht einem Bayes-Test.

Betrachte nun die einfachen Hypothesen

$$H_0 : f = f_0 \qquad \text{vs.} \qquad H_1 : f = f_1,$$

$f_i = f(\mathbf{x}, \theta_i)$, $i = 0, 1$, θ_0 und θ_1 bekannt und $X_1, ..., X_n$ eine Zufallsstichprobe aus f_0 oder f_1. Dann besagt das Neyman-Pearson (NP) Lemma, dass jeder Test der Form

$$\delta_{NP}(\mathbf{x}) = \begin{cases} 0 & : \quad L(\mathbf{x}, f_0, f_1) < q \\ \gamma & : \quad L(\mathbf{x}, f_0, f_1) = q \\ 1 & : \quad L(\mathbf{x}, f_0, f_1) > q \end{cases},$$

mit $\gamma \in [0, 1]$ und dem Likelihood-Quotienten $L(\mathbf{x}, f_0, f_1) = \frac{f_1(\mathbf{x})}{f_0(\mathbf{x})}$ bester Test zum Niveau α ist, d.h. für den gilt $P\left(\delta_{NP}(\mathbf{X}) = 1 | H_0 \text{ wahr}\right) \leq \alpha$ (vgl. z.B. Mood et al. (1974)).

Die Konstruktion der Regel δ gemäß dem NP-Lemma ermöglicht die Kontrolle des Fehlers erster Art bei Minimierung des Fehlers zweiter Art. Scott und Nowak (2005) untersuchen die Eigenschaften von Klassifikationsregeln, die ohne Kenntnis von f_0 und f_1 allein auf Datenbasis nach dem NP-Prinzip erstellt werden, hinsichtlich PAC („probably approximately correct“) Schranken, Orakelungleichungen, Konvergenzraten und Konsistenz.

5.6 NP-Wälder: Übertragung des NP-Prinzips auf Wälder

5.6.1 Vorüberlegungen

In der hier betrachteten Anwendung liegen keine Informationen über die Verteilungen f_0 und f_1 vor sondern lediglich die Daten der Stichprobe. Dennoch kann die Idee des Neyman-Pearson Lemmas als Konstruktionsheuristik für einen Wald verwendet werden.

Dieser Wald ist jedoch nicht notwendigerweise der beste Test (d.h. Wald) zu einem festen Niveau α.

In Analogie zum NP-Lemma wird die Nullhypothese getestet, dass eine klinische Situation alarmrelevant ist, gegen die Alternative, dass die Situation nicht alarmrelevant ist. Ist eine Situation π alarmrelevant, so unterliegt der beobachtete zugehörige Vektor $\mathbf{x}$ der Dichte f_0 oder andernfalls der Dichte f_1, wobei beide Dichten unbekannt sind. Die Hypothesen lauten also:

$$H_0 : f = f_0 \qquad \text{vs.} \qquad H_1 : f = f_1 \,.$$

Die Testentscheidung zu diesem Testproblem wird mit folgender Funktion getroffen:

$$\delta_{NP}(\mathbf{x}) = \begin{cases} 0 & : \quad L(\mathbf{x}, f_0, f_1) < q \\ 1 & : \quad L(\mathbf{x}, f_0, f_1) \geq q \end{cases} .$$

Für $q = \frac{\pi_0}{1-\pi_0}$ handelt es sich um einen Bayes-Test, d.h. den Test, der das erwartete Risiko bei gleich hohen Fehlklassifikationskosten minimiert. Da im Fall der Klassifikation von Alarmsituationen nicht das erwartete Risiko zu minimieren ist, sondern die Wahrscheinlichkeit, eine alarmrelevante Situation falsch zu klassifizieren, kontrolliert werden muss, ist der kritische Wert q gesucht, so dass der Fehler erster Art für diesen Test kleiner oder gleich α ist.

Die Dichten f_0 und f_1 sind unbekannt und werden während der Konstruktion von Wäldern auch nicht geschätzt. Sie können nicht zur Berechnung der Teststatistik und für die Testentscheidung genutzt werden. Daher ist eine alternative Teststatistik eines durch monotone Transformationen gewonnenen äquivalenten Tests zu finden. Beruht diese alternative Teststatistik auf Größen, die im Verlauf der Erstellung von Wäldern geschätzt werden, kann sie zur Konstruktion eines geeigneten approximierenden Tests genutzt werden.

Nach dem Bayes-Theorem (vgl. z.B. Wasserman (2004)) gilt:

$$\begin{aligned}
\frac{P(G=1|\mathbf{X}=\mathbf{x})}{P(G=0|\mathbf{X}=\mathbf{x})} &= \frac{P(G=1)f(\mathbf{x}|G=1)}{P(G=0)f(\mathbf{x}|G=0)} \\
&= \frac{(1-\pi_0)}{\pi_0}\frac{f_1(\mathbf{x})}{f_0(\mathbf{x})} \\
&= \frac{(1-\pi_0)}{\pi_0}L(\mathbf{x}, f_0, f_1) \\
\Leftrightarrow \frac{P(G=1|\mathbf{X}=\mathbf{x})}{1-P(G=1|\mathbf{X}=\mathbf{x})} &= \frac{(1-\pi_0)}{\pi_0}L(\mathbf{x}, f_0, f_1) \\
\Leftrightarrow L(\mathbf{x}, f_0, f_1) &= \frac{P(G=1|\mathbf{X}=\mathbf{x})\pi_0}{(1-\pi_0)(1-P(G=1|\mathbf{X}=\mathbf{x}))}.
\end{aligned}$$

Statt des Likelihood-Quotienten kann also $\frac{P(G=1|\mathbf{X}=\mathbf{x})\pi_0}{(1-\pi_0)(1-P(G=1|\mathbf{X}=\mathbf{x}))}$ oder auch $P(G=1|\mathbf{X}=\mathbf{x})$ als Teststatistik eines äquivalenten Tests verwendet werden, da $\frac{P(G=1|\mathbf{X}=\mathbf{x})\pi_0}{(1-\pi_0)(1-P(G=1|\mathbf{X}=\mathbf{x}))}$ monoton steigend in $P(G=1|\mathbf{X}=\mathbf{x})$ für festes π_0 ist.

Die a-posteriori-Wahrscheinlichkeit $P(G=1|\mathbf{X}=\mathbf{x})$ ist ebenfalls unbekannt, wird jedoch während der Konstruktion eines Waldes von jedem Baum geschätzt: Fällt eine Beobachtung $\mathbf{x}$ in einen Blattknoten A_t eines Entscheidungsbaums, so wird wie gehabt der Anteil $a_{0|t}$ von Beobachtungen des Lerndatensatzes aus Population Π_0 und der Anteil $a_{1|t}$ von Beobachtungen des Lerndatensatzes aus Population Π_1 in diesem Knoten ermittelt. Die Beobachtung $\mathbf{x}$ wird durch den Entscheidungsbaum der Population mit dem höheren Anteil zugeordnet. Da jede Beobachtung in nur genau einen Blattknoten fallen kann, bezeichne die entsprechenden Anteile kurz mit a_0 bzw. a_1. Fasse $a_1^{(i)}$ als Schätzung der a-posteriori-Wahrscheinlichkeit $P(G=1|\mathbf{X}=\mathbf{x})$ des i-ten Entscheidungsbaums eines Waldes auf und schreibe die Klassifikationsregel $\delta_i(\mathbf{x})$ des i-ten Entscheidungsbaums als $\delta_i(\mathbf{x}) = \mathbb{1}_{(a_1^{(i)}>0,5)}$. Dann nimmt $\sum_{i=1}^{N}\delta_i(\mathbf{x})$ große Werte an, wenn viele Bäume des Waldes die a-posteriori-Wahrscheinlichkeit $P(G=1|\mathbf{X}=\mathbf{x})$ größer als 0,5 schätzen. Diese Summe wird daher als Teststatistik eines geeigneten approximierenden Tests betrachtet.

5.6.2 Definition

Ein NP-Wald ist eine Entscheidungsregel $\delta_{NP-Wald}(\mathbf{x})$ zu einem gegebenen Signifikanzniveau α und einem kritischen Wert $q^\star$, so dass

$$\delta_{NP-Wald}(\mathbf{x}) = \begin{cases} 0 & : \quad \sum_{i=1}^{N} \delta_i(\mathbf{x}) < q^\star \\ 1 & : \quad \sum_{i=1}^{N} \delta_i(\mathbf{x}) \geq q^\star \end{cases},$$

und $P(\delta_{NP-Wald}(\mathbf{X}) = 1|G = 0) \leq \alpha$. Dieser kritische Wert $q^\star$ ist das $(1-\alpha)$-Quantil der Verteilung der Teststatistik $\sum_{i=1}^{N} \delta_i(\mathbf{X})$ unter H_0, da

$$\begin{aligned} & P(\delta_{NP-Wald}(\mathbf{X}) = 1|G = 0) \leq \alpha \\ \Leftrightarrow \quad & P\left(\sum_{i=1}^{N} \delta_i(\mathbf{X}) > q^\star|G = 0\right) \leq \alpha \\ \Leftrightarrow \quad & P\left(\sum_{i=1}^{N} \delta_i(\mathbf{X}) \leq q^\star|G = 0\right) \geq 1 - \alpha. \end{aligned}$$

Durch die Wahl von α kann die Sensitivität des resultierenden Alarmsystems kontrolliert werden.

5.6.3 Verteilung der Teststatistik

Zur Konstruktion eines Waldes unter Ausnutzung der oben beschriebenen Analogie wird die Verteilung der Teststatistik unter der Nullhypothese benötigt. Da diese Verteilung unbekannt ist, folgen Überlegungen, wie sie geschätzt werden kann.

Sei $d_i = P(\delta_i(\mathbf{X}) = 1|G = 0)$ die Wahrscheinlichkeit, dass Entscheidungsbaum δ_i eine alarmrelevante Beobachtung als nicht alarmrelevant klassifiziert. Wären alle d_i gleich und wären die Bäume unabhängig, d.h. $d_1 = ... = d_n$ und $P(\delta_1(\mathbf{X}) = j_1, ..., \delta_n(\mathbf{X}) = j_n) = \prod_{i=1}^{n} P(\delta_i(\mathbf{X}) = j_i)$, $j_1, ..., j_n \in \{0, 1\}$, so wäre die gesuchte Verteilung der Teststatistik unter der Nullhypothese eine Binomialverteilung.

Die Voraussetzung gleicher d_i ist jedoch in der Regel nicht erfüllt. Unter der Voraussetzung der Unabhängigkeit aber bei verschiedenen d_i lässt sich die Verteilung der

Teststatistik kombinatorisch ermitteln:

$$
\begin{aligned}
P\left(\left(\sum_{i=1}^{N}\delta_i(\mathbf{X})\right)=1|G=0\right) &= \sum_{i=1}^{N} d_i \prod_{j\neq i}(1-d_j) \\
P\left(\left(\sum_{i=1}^{N}\delta_i(\mathbf{X})\right)=2|G=0\right) &= \sum_{i=1}^{N-1}\sum_{j>i}^{N} d_i d_j \prod_{k\neq i,j}(1-d_k) \\
&\text{etc.}
\end{aligned}
$$

Allerdings ist auch die Bedingung der Unabhängigkeit häufig nicht erfüllt. Zudem ist die Berechnung des benötigten Quantils bei Wäldern mit vielen Bäumen mit erheblichem Rechenaufwand verbunden. Aus diesem Grund wird die Verteilung der Teststatistik aus der Lernstichprobe geschätzt. Dazu wird die Lernstichprobe geteilt. Zunächst wird auf der ersten Hälfte der Lernstichprobe der Wald in beschriebener Weise konstruiert. Der Wald wird anschließend auf die als alarmrelevant annotierten Beobachtungen angewendet. Es wird für jede dieser Beobachtungen ermittelt, von wie vielen Bäumen des Waldes sie als nicht alarmrelevant klassifiziert werden. Das $(1-\alpha)$-Quantil dieser Werte ist der gesuchte kritische Wert des NP-Waldes.

Kapitel 6

Alarmregelgenerierung

6.1 Auswahl des Stichprobenverfahrens

Mit Hilfe von NP-Wäldern können zu einer vorgegebenen Sensitivität Alarmregeln zur Klassifikation von Situationen erzeugt werden, in denen der Standard-Patientenmonitor einen Alarm auslöst. Um die Eignung der NP-Wälder für diesen Zweck zu überprüfen, werden alle Alarme des in der klinischen Studie erhobenen Datensatzes herangezogen, dabei werden die im Anhang aufgelisteten Variablen sowie Dummy-Variablen zur Kennzeichnung ersetzter Werte genutzt. Es werden Wälder mit der Zielsetzung konstruiert, 95% bzw. 98% der als alarmrelevant annotierten Alarme zu erkennen. Dazu wird der Datensatz für jeden Wald zufällig in Lern-, Schätz- und Teststichproben gleicher Größe aufgeteilt, wobei die Anteile alarmrelevanter und nicht alarmrelevanter Situationen denen im gesamten Datensatz entsprechen. Da die erzielten Sensitivitäten und Fehlalarmreduktionen von der zufälligen Aufteilung abhängen, wird der Datensatz 1000 mal zufällig in diese Teilstichproben zerlegt und die Prozedur angewendet. Die erreichten Sensitivitäten und Fehlalarmreduktionen auf den 1000 Teststichproben erlauben so Schlüsse auf die generelle Eignung der Prozedur.

Wie zuvor beschrieben, besteht die Vermutung, dass das volle Bootstrapping durch Half-Sampling ersetzt werden kann, ohne nennenswerte Nachteile in der Klassifikation in Kauf nehmen zu müssen. Da bei Half-Sampling nur halb so viele Daten verarbeitet werden wie bei vollem Bootstrap, ist es weniger rechenintensiv. Auch die Zielsensitivität kann die erreichbare Fehlalarmreduktion beeinflussen. Daher werden je für $\alpha = 0,02$ und $\alpha = 0,05$ Wälder erzeugt. Darüber hinaus ist eine Unterscheidung sinnvoll, ob es sich um herbeigeführte oder nicht herbeigeführte Alarme handelt. Ein Alarmsy-

stem sollte bezüglich aller nicht herbeigeführten, also durch „natürliche“ Veränderungen des Gesundheitszustandes verursachte Alarme eine gewisse Sensitivität erreichen. In Situationen, in denen ein Alarm auf Grund einer Manipulation entsteht, ist dies zur Sicherheit der Patienten nicht notwendig. Andererseits wird auch ein unnötiger, herbeigeführter Alarm als Fehlalarm wahrgenommen. Eine hohe Sensitivität und Fehlalarmreduktion, auch bezüglich manipulierter Alarme, ist daher aus psychologischen Gründen wünschenswert.

Um zu entscheiden, welches Verfahren der Stichprobenziehung am geeignetsten ist, werden je 1000 Wälder mit

- vollem Bootstrapping
  - und $\alpha = 0,02$ auf der Grundlage aller Alarme
  - und $\alpha = 0,05$ auf der Grundlage aller Alarme
  - und $\alpha = 0,02$ auf der Grundlage aller nicht herbeigeführten Alarme
  - und $\alpha = 0,05$ auf der Grundlage aller nicht herbeigeführten Alarme
- Half-Sampling
  - und $\alpha = 0,02$ auf der Grundlage aller Alarme
  - und $\alpha = 0,05$ auf der Grundlage aller Alarme
  - und $\alpha = 0,02$ auf der Grundlage aller nicht herbeigeführten Alarme
  - und $\alpha = 0,05$ auf der Grundlage aller nicht herbeigeführten Alarme
- 200 Objekte ohne Zurücklegen (kleines Bootstrapping)
  - und $\alpha = 0,02$ auf der Grundlage aller Alarme
  - und $\alpha = 0,05$ auf der Grundlage aller Alarme
  - und $\alpha = 0,02$ auf der Grundlage aller nicht herbeigeführten Alarme
  - und $\alpha = 0,05$ auf der Grundlage aller nicht herbeigeführten Alarme

konstruiert und die erreichten Sensitivitäten und Fehlalarmreduktionen verglichen.

Jeder der Wälder besteht aus 1000 Bäumen, die auf einer zufälligen Stichprobe nach einer der drei Methoden konstruiert werden. Die Anzahl der zufällig ausgewählten Variablen, auf denen Splits in Betracht gezogen werden, ist entsprechend der allgemeinen Empfehlung und Voreinstellung im R-Package `RandomForest` auf die Wurzel der eingehenden Variablen beschränkt. Die Stop-Splitting-Regel ist hier so gewählt, dass eine minimale Blattknotengröße von fünf Beobachtungen gefordert wird. Die unbekannte Verteilung der Teststatistik wird durch die empirische Verteilungsfunktion geschätzt, indem jede alarmrelevante Situation aus der Schätzstichprobe von jedem Entscheidungsbaum des Waldes klassifiziert wird und die Anzahl der Stimmen für „nicht alarmrelevant“ bestimmt wird. Das empirische $(1 - \alpha)$-Quantil dieser Häufigkeiten ist der kritische Wert q. Die Alarmsituationen der Teststichprobe werden mit Hilfe des so bestimmten kritischen Wertes durch den Wald klassifiziert. Dabei wird die Summe der Stimmen für „nicht alarmrelvant“ mit dem kritischen Wert verglichen: Eine Alarmsituation wird der Population der alarmrelevanten Situationen zugeordnet, falls weniger als q Bäume für „nicht alarmrelevant“ stimmen.

Die erreichten Sensitivitäten und Fehlalarmreduktionen werden in Tabelle 6.1 zusammengefasst. Die Zielsensitivitäten von 95% bzw. 98% werden bei allen untersuchten Stichprobenverfahren im Median nur leicht unter- oder überschritten. Daraus kann geschlossen werden, dass die nach dem Neyman-Pearson-Prinzip modifizierten Wälder zur Konstruktion von Alarmregeln bei Vorgabe einer Zielsensitivität geeignet sind. Es zeigen sich Unterschiede in den im Median erreichten Fehlalarmreduktionen. Wie nach den Ergebnissen von Bühlmann und Yu (2002) zu erwarten, sind die im Median erreichten Fehlalarmreduktionen bei vollem Bootstrap und Half-Sampling sehr ähnlich. Deutlich schlechtere Ergebnisse in Bezug auf Fehlalarmreduktion wird mit Bootstrap-Stichproben der Größe 200 erzielt. Aufgrund der kürzeren Rechenzeiten und der geringen Unterschiede in den Ergebnissen wird in allen weiteren Untersuchungen Half-Sampling als Stichprobenverfahren zur Erzeugung der Wälder verwendet.

Es fällt auf, dass für nicht herbeigeführte Alarme bessere Alarmregeln mit Hilfe der Wälder gefunden werden können als für alle auftretenden Alarme. Könnte also erreicht werden, dass das Pflegepersonal die Alarmfunktion bei Pflegemaßnahmen konsequent deaktiviert, um herbeigeführte Alarme zu vermeiden, ließen sich die verbleibenden Fehlalarme je nach Zielsensitivität im Median um etwa 38% bis 39% bzw. um etwa 55% verringern.

Verfahren	Sensitivität		Fehlalarmreduktion	
	alle Alarme	nicht herbeig. Alarme	alle Alarme	nicht herbeig. Alarme
$\alpha = 0,02$				
volles Bootstrap	97,73	97,85	26,29	38,91
Half-Sampling	97,73	97,85	26,50	39,87
Bootstrap (klein)	98,01	97,85	23,07	35,56
$\alpha = 0,05$				
volles Bootstrap	94,89	94,79	39,65	55,45
Half-Sampling	94,89	94,79	40,94	55,93
Bootstrap (klein)	94,89	95,09	35,32	48,95

Tabelle 6.1: Mediane der erzielten Sensitivitäten und Fehlalarmreduktionen bei unterschiedlichen Stichprobenverfahren

Neben den Medianen der erreichten Sensitivitäten und Fehlalarmreduktionen ist auch ihre Streuung von Interesse. Die Alarmgebung auf Basis der erzeugten Wälder sollte nicht nur im Median sondern generell zu Sensitivitäten nahe der vorgegebenen 95% bzw. 98% führen. Daher folgt eine detaillierte Betrachtung sowie die grafische Darstellung der Klassifikationsergebnisse bei Half-Sampling.

Bei einem Signifikanzniveau von 5% und Half-Sampling werden von den Wäldern auf den Teststichproben bei Betrachtung aller Alarme die erwarteten 95% Sensitivität im arithmetischen Mittel und Median mit 94,89% erreicht (Abb. 6.1). Die Hälfte der Wälder ergeben Alarmregeln mit Sensitivitäten zwischen 93,75% und 96,31%. Die überwiegende Mehrheit der Wälder führt zu einer Sensitivität über 92% (Abb. 6.2). Die Streuung der erreichten Sensitivitäten ist gering: nur wenige Aufteilungen in Lern-, Schätz- und Teststichprobe bewirken eine Alarmgebung mit einer Sensitivität unter 90%. Höhere Sensitivitäten als die erwarteten 95% haben zwar kein höheres Risiko für den Patienten zur Folge, alarmrelevante Situationen nicht zu erkennen, allerdings geht mit ihnen häufig eine geringere Reduktion der Fehlalarme einher (Abb. 6.3).

Daher sollte ein Wald, der als Grundlage für ein neues Alarmsystem dient, einen Kompromiss hinsichtlich der beiden Kriterien Sensitivität und Fehlalarmreduktion bieten. Ein Beispiel für einen solchen Wald kann in den 1000 erzeugten Wäldern gefunden werden. Mit Hilfe dieses Waldes (Tab. 6.2) werden mit einer Sensitivität von etwa 96% alle alarmrelevanten Situationen in der Teststichprobe richtig erkannt, während

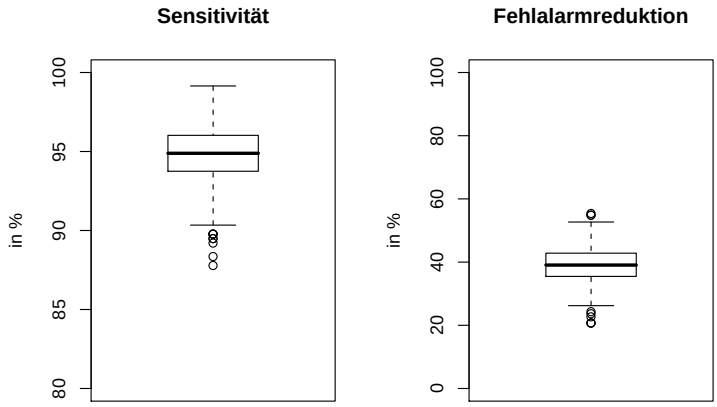

Abbildung 6.1: Sensitivitäten und Fehlalarmreduktionen (Signifikanzniveau 5%, Half-Sampling, alle Alarme)

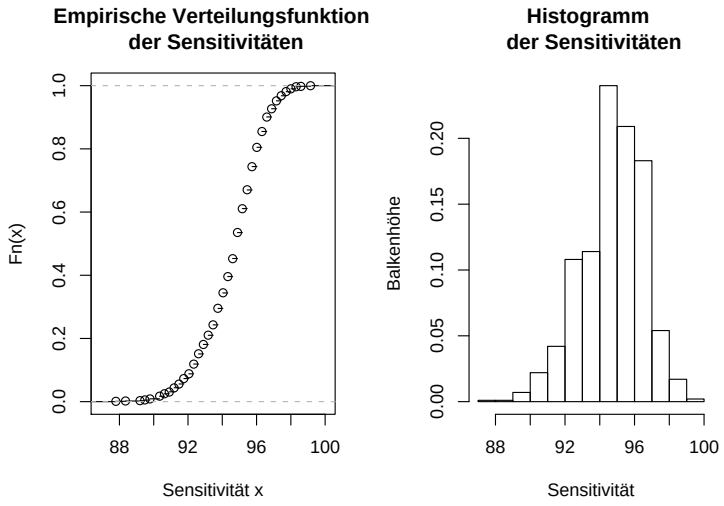

Abbildung 6.2: Empirische Verteilungsfunktion und Histogramm der Sensitivitäten (Signifikanzniveau 5%, Half-Sampling, alle Alarme)

die Fehlalarme um etwa 46% reduziert werden. Der kritische Wert für diesen Wald ist $q = 958,9$, d.h. dass von 1000 Bäumen mindestens 959 für nicht alarmrelevant stimmen müssen, damit eine Situation als nicht alarmrelevant klassifiziert wird.

Das Klassifikationsergebnis ist bei ausschließlicher Betrachtung der nicht herbeigeführ-

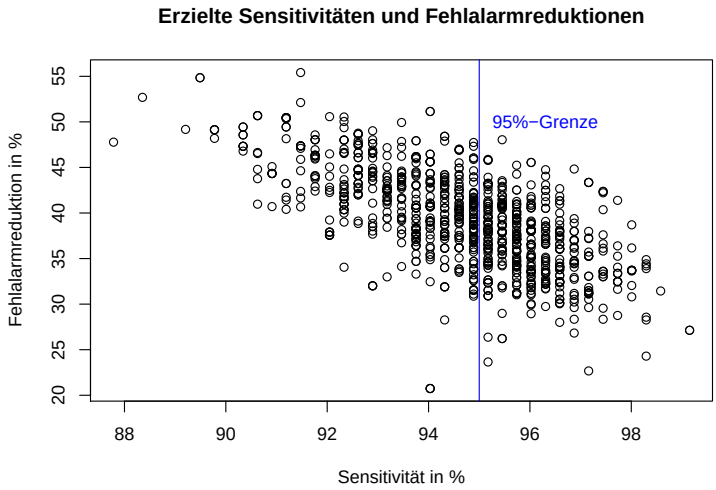

Abbildung 6.3: Zusammenhang der erzielten Sensitivitäten und Fehlalarmreduktionen (Signifikanzniveau 5%, Half-Sampling, alle Alarme)

Situation	klassifiziert als	
	alarmrelevant	nicht alarmrelevant
alarmrelevant	337	15
nicht alarmrelevant	1051	887

Tabelle 6.2: Confusion Matrix eines der besten erzeugten Wälder (Signifikanzniveau 5%, Half-Sampling, alle Alarme)

ten Alarme noch besser. Zum Signifikanzniveau von 5% werden die angestrebten 95% Sensitivität im Mittel mit 94,74% und im Median mit 94,79% erreicht (Abb. 6.4). Der Abstand zwischen dem ersten Quartil (93,56%) und dem dritten Quartil (95,71%) ist mit 2,15 gering. Die erreichte Fehlalarmreduktion ist mit im Mittel 55,76% und Median 55,93% wesentlich höher im Vergleich zum Klassifikationsergebnis für alle Alarme. Ließen sich also die ca. 40% durch Manipulation herbeigeführten Alarme beispielsweise durch Schulungen im Umgang mit dem Patientenmonitor vermeiden, so könnten von den verbleibenden 45% Fehlalarmen im Mittel etwa die Hälfte durch die Verwendung der Alarmregeln eines Waldes vermieden werden. Die größte durch einen Wald erreichte Fehlalarmreduktion liegt in dieser Gruppe bei 67,78%, allerdings bei einer vergleichsweise niedrigen Sensitivität von nur 90,80%. Einer der besten Wälder (Tab. 6.3) in beiden Kriterien klassifiziert 95% aller alarmrelevanten Situationen in der Teststich-

Situation	klassifiziert als	
	alarmrelevant	nicht alarmrelevant
alarmrelevant	310	16
nicht alarmrelevant	397	649

Tabelle 6.3: Confusion Matrix eines der besten erzeugten Wälder (Signifikanzniveau 5%, Half-Sampling, ohne herbeigeführte Alarme)

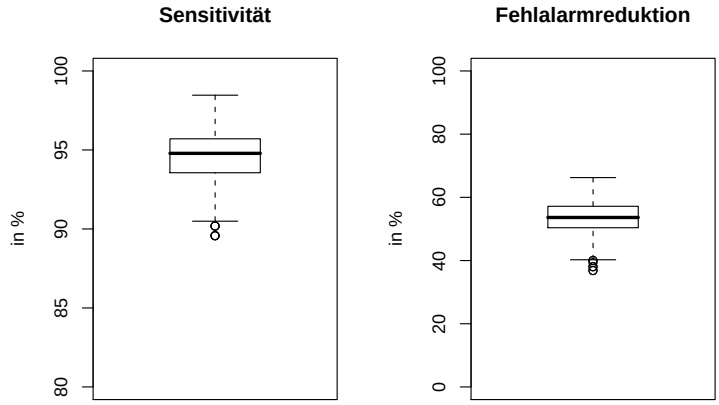

Abbildung 6.4: Sensitivitäten und Fehlalarmreduktionen (Signifikanzniveau 5%, Half-Sampling, ohne herbeigeführte Alarme)

probe richtig, während 62% der Fehlalarme verhindert werden. Für diesen Wald liegt der kritische Wert bei $q = 889,1$. Es müssen also mindestens 890 der 1000 Bäume für nicht alarmrelevant stimmen, um eine Situation als nicht alarmrelevant zu klassifizieren.

Eine Sensitivität von 95% ist für ein Alarmsystem auf der Intensivstation möglicherweise nicht angemessen hoch. Eine höhere Sensitivität führt jedoch zu einer niedrigeren Reduktion der Fehlalarme, wie die Ergebnisse von 1000 Wäldern mit einer angestrebten Sensitivität von 98% zeigen (Tab. 6.1). Der Median und der arithmetische Mittelwert der auf den Teststichproben erreichten Sensitivitäten liegen bei Half-Sampling nah an den angestrebten 98% mit 97,85% bzw. 97,66% (ohne herbeigeführte Alarme) und 97,85% bzw. 97,73% (alle Alarme). Die Mehrheit der Wälder erreicht in beiden Gruppen Sensitivitäten zwischen 95% und 100% (Abb. 6.5, Abb. 6.6) jedoch mit Unterschieden in den erreichten Fehlalarmreduktionen. Die Fehlalarme werden im Median um 26,50%

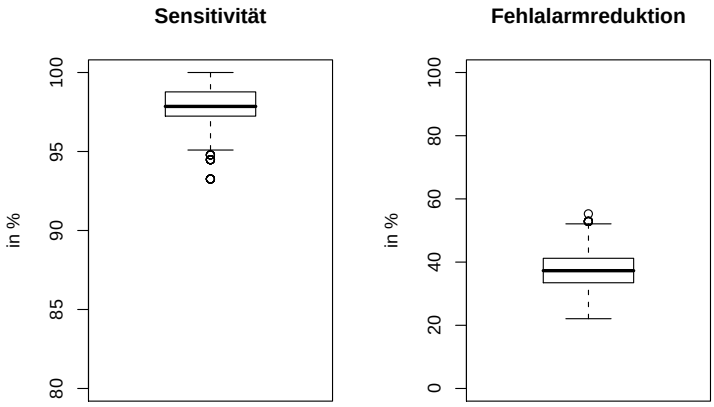

Abbildung 6.5: Sensitivitäten und Fehlalarmreduktionen (Signifikanzniveau 2%, Half-Sampling, ohne herbeigeführte Alarme)

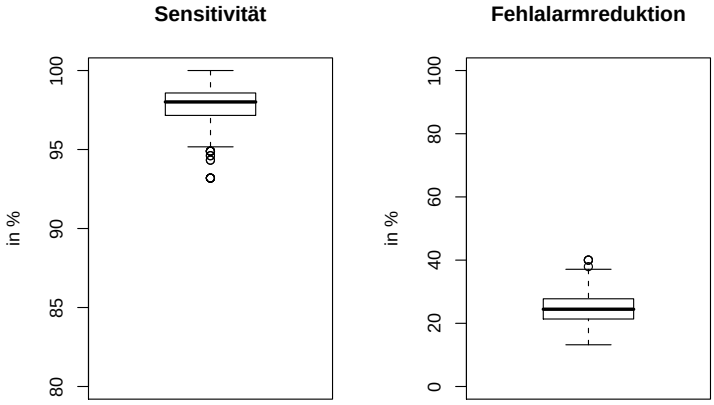

Abbildung 6.6: Sensitivitäten und Fehlalarmreduktionen (Signifikanzniveau 2%, Half-Sampling, alle Alarme)

(alle Alarme) und 39,87% (ohne herbeigeführte Alarme) reduziert. Die Streuung der erreichten Sensitivitäten bleibt auch bei diesem Signifikanzniveau klein. Trotz der hohen Vorgabe von 98% Sensitivität ist eine erkennbare Reduktion der Fehlalarme weiterhin möglich.

Situation	klassifiziert als	
	alarmrelevant	nicht alarmrelevant
alarmrelevant	320	6
nicht alarmrelevant	553	493

Tabelle 6.4: Confusion Matrix eines der besten erzeugten Wälder (Signifikanzniveau 2%, Half-Sampling, ohne herbeigeführte Alarme)

Situation	klassifiziert als	
	alarmrelevant	nicht alarmrelevant
alarmrelevant	345	7
nicht alarmrelevant	1225	713

Tabelle 6.5: Confusion Matrix eines der besten erzeugten Wälder (Signifikanzniveau 2%, Half-Sampling, alle Alarme)

Beispiele für Wälder mit guten Ergebnissen hinsichtlich Sensitivität und Fehlalarmreduktion sind in den Tabellen 6.4 und 6.5 dargestellt. Die Wälder klassifizieren 345 (alle Alarme) bzw. 320 (nicht herbeigeführte Alarme) der 352 bzw. 326 alarmrelevanten Situationen in der Teststichprobe korrekt (Sensitivität ca. 98%) und reduzieren die Fehlalarme von 1938 bzw. 1046 auf 1225 (alle Alarme) bzw. auf 553 (nicht herbeigeführte Alarme) um etwa 37% bzw. 47%. Die kritischen Werte für diese beiden Wälder liegen für alle Alarme bei $q = 972,96$ und für nicht herbeigeführte Alarme bei $q = 947,8$.

Die dargestellten Ergebnisse zeigen, dass mit NP-Wäldern eine Adjustierung der Sensitivität des resultierenden Alarmsystems entsprechend der speziellen Anforderungen des Monitorings von Intensivpatienten möglich ist, und dass gleichzeitig die Fehlalarmrate merklich reduziert werden kann. Sie eignen sich daher für die Klassifikation von intensivmedizinischen Alarmsituationen. Der geringe Unterschied in den Klassifikationsergebnissen bei vollem Bootstrapping und Half-Sampling erlaubt bei weiteren Untersuchungen auf das zeitlich weniger aufwendige Half-Sampling zurückzugreifen. Darüber hinaus ist offensichtlich die Findung geeigneter Alarmregeln für alle Alarme schwieriger als nur für nicht herbeigeführte Alarme.

6.2 Einbeziehung der Charakteristika in die Alarmregelgenerierung

Klassifikationsergebnisse können in der Regel durch Hinzunahme von Charakteristika bzw. Features, die aus den Rohdaten gewonnen werden, verbessert werden. Naheliegend in dieser medizinischen Anwendung ist, der Vorgehensweise von Pflegern und Ärzten am Patientenbett zu folgen: Muss in einer Situation, in der der Patientenmonitor einen Alarm ausgelöst hat, eine Entscheidung darüber getroffen werden, ob dieser Alarm eine Reaktion erfordert oder nicht, so wird in der Regel der Gesundheitsverlauf in der Zeit vor dem Alarm in der Entscheidungsfindung hinzugezogen. Daher werden Charakteristika gewählt, die in einem gewissen Zeitabschnitt vor einem Alarm verschiedene Aspekte des Gesundheitsverlaufs wie die Variabilität, den linearen Trend oder einen entrauschten Schätzwert für einen der Vitalparameter wiedergeben. Aus der Betrachtung dieser Charakteristika in verschieden großen Zeitabschnitten werden plötzliche aber auch schleichende Veränderungen erkennbar. Ist zum Beispiel die Variabilität in den vergangenen Minuten eher gering und in der letzten Minute hoch, so ist das ein Anzeichen für eine Änderung im Gesundheitszustand. Als weitere Indikatoren für solche Veränderungen werden Haar-Wavelet-Koeffizienten der letzten etwa acht Minuten (511 Sekunden) verwendet. Sie stehen im Zusammenhang mit den Differenzen der Mittelwerte eines betrachteten Vitalparameters in Teilstücken dieser Zeitspanne. Da es sich um Informationen handelt, die bislang nicht in die Alarmregelgenerierung eingegangen sind und ihre Relevanz aus dem Sachzusammenhang begründet ist, wird vermutet, dass durch ihre Hinzunahme das Klassifikationsergebnis in Form einer höheren Fehlalarmreduktion weiter verbessert werden kann.

Um diese Vermutung zu überprüfen, werden Wälder auf der Grundlage der auch schon zuvor verwendeten Daten und 1240 zusätzlich erzeugter Charakteristika konstruiert. Es handelt sich dabei um die in Tabelle 6.6 aufgelisteten Charakteristika je für die Variablen ART_S, ART_D, ART_M, HR, PLS, SpO2, Ta und RESP. Die Notation entspricht der in Kapitel 4.

Die Rohdaten, die erzeugten Charakteristika sowie die Dummy-Variablen zur Kennzeichnung fehlender Werte werden in folgenden Gruppen als Eingangsvariablen für NP-Wälder verwendet:

- Rohdaten und Steigungs-Charakteristika

Charakteristika	klassisch	robust
Steigung	$\hat{\beta}_t^{KQ;61}$, $\hat{\beta}_t^{KQ;121}$, $\hat{\beta}_t^{KQ;601}$	$\hat{\beta}_t^{RM;61}$, $\hat{\beta}_t^{RM;121}$, $\hat{\beta}_t^{RM;601}$, $\hat{\beta}_t^{RM_{adaptiv}}$
Variabilität	$\hat{\sigma}_t^{61}$, $\hat{\sigma}_t^{121}$, $\hat{\sigma}_t^{601}$, $\hat{\sigma}_t^{KQ;Res;61}$, $\hat{\sigma}_t^{KQ;Res;121}$, $\hat{\sigma}_t^{KQ;Res;601}$	Qn_t^{61}, Qn_t^{121}, Qn_t^{601}, $Qn_t^{RM;Res;61}$,$Qn_t^{RM;Res;121}$, $Qn_t^{RM;Res;601}$, $Qn_t^{RM_{adaptiv}}$
geschätztes Signal	$\hat{\mu}_t^{KQ;61}$, $\hat{\mu}_t^{KQ;121}$, $\hat{\mu}_t^{KQ;601}$	$\hat{\mu}_t^{RM;61}$, $\hat{\mu}_t^{RM;121}$, $\hat{\mu}_t^{RM;601}$, $\hat{\mu}_t^{RM_{adaptiv}}$
Wavelet-Koeffizienten	$c_{0,0}$, $d_{0,0}$, $d_{1,0}$, $d_{1,1}$, $d_{2,0}$, $d_{2,1}$, $d_{2,2}$, $d_{2,3}$, $d_{3,0}$, $d_{3,1}$, $d_{3,2}$, $d_{3,3}$, $d_{3,4}$, $d_{3,5}$, $d_{3,6}$, $d_{3,7}$, $d_{4,0}$, $d_{4,1}$, $d_{4,2}$, $d_{4,3}$, $d_{4,4}$, $d_{4,5}$, $d_{4,6}$, $d_{4,7}$, $d_{4,8}$, $d_{4,9}$, $d_{4,10}$, $d_{4,11}$, $d_{4,12}$, $d_{4,13}$, $d_{4,14}$, $d_{4,15}$, $d_{5,0}$, $d_{5,1}$, $d_{5,2}$, $d_{5,3}$, $d_{5,4}$, $d_{5,5}$, $d_{5,6}$, $d_{5,7}$, $d_{5,8}$, $d_{5,9}$, $d_{5,10}$, $d_{5,11}$, $d_{5,12}$, $d_{5,13}$, $d_{5,14}$, $d_{5,15}$, $d_{5,16}$, $d_{5,17}$, $d_{5,18}$, $d_{5,19}$, $d_{5,20}$, $d_{5,21}$, $d_{5,22}$, $d_{5,23}$, $d_{5,24}$, $d_{5,25}$, $d_{5,26}$, $d_{5,27}$, $d_{5,28}$, $d_{5,29}$, $d_{5,30}$, $d_{5,31}$, $d_{5,32}$	$c_{0,0}^{RM}$, $d_{0,0}^{RM}$, $d_{1,0}^{RM}$, $d_{1,1}^{RM}$, $d_{2,0}^{RM}$, $d_{2,1}^{RM}$, $d_{2,2}^{RM}$, $d_{2,3}^{RM}$, $d_{3,0}^{RM}$, $d_{3,1}^{RM}$, $d_{3,2}^{RM}$, $d_{3,3}^{RM}$, $d_{3,4}^{RM}$, $d_{3,5}^{RM}$, $d_{3,6}^{RM}$, $d_{3,7}^{RM}$, $d_{4,0}^{RM}$, $d_{4,1}^{RM}$, $d_{4,2}^{RM}$, $d_{4,3}^{RM}$, $d_{4,4}^{RM}$, $d_{4,5}^{RM}$, $d_{4,6}^{RM}$, $d_{4,7}^{RM}$, $d_{4,8}^{RM}$, $d_{4,9}^{RM}$, $d_{4,10}^{RM}$, $d_{4,11}^{RM}$, $d_{4,12}^{RM}$, $d_{4,13}^{RM}$, $d_{4,14}^{RM}$, $d_{4,15}^{RM}$, $d_{5,0}^{RM}$, $d_{5,1}^{RM}$, $d_{5,2}^{RM}$, $d_{5,3}^{RM}$, $d_{5,4}^{RM}$, $d_{5,5}^{RM}$, $d_{5,6}^{RM}$, $d_{5,7}^{RM}$, $d_{5,8}^{RM}$, $d_{5,9}^{RM}$, $d_{5,10}^{RM}$, $d_{5,11}^{RM}$, $d_{5,12}^{RM}$, $d_{5,13}^{RM}$, $d_{5,14}^{RM}$, $d_{5,15}^{RM}$, $d_{5,16}^{RM}$, $d_{5,17}^{RM}$, $d_{5,18}^{RM}$, $d_{5,19}^{RM}$, $d_{5,20}^{RM}$, $d_{5,21}^{RM}$, $d_{5,22}^{RM}$, $d_{5,23}^{RM}$, $d_{5,24}^{RM}$, $d_{5,25}^{RM}$, $d_{5,26}^{RM}$, $d_{5,27}^{RM}$, $d_{5,28}^{RM}$, $d_{5,29}^{RM}$, $d_{5,30}^{RM}$, $d_{5,31}^{RM}$, $d_{5,32}^{RM}$

Tabelle 6.6: Charakteristika des gesundheitlichen Verlaufs

- Rohdaten und Variabilitäts-Charakteristika
- Rohdaten und Schätzwerte des unterliegenden Signals
- Rohdaten und Wavelet-Koeffizienten
- Rohdaten und robuste Charakteristika
- Rohdaten und klassische Charakteristika
- Rohdaten und alle Charakteristika zugleich.

Es werden unter Verwendung von Half-Sampling Regeln für alle Alarme mit einer Sensitivität von 98% angestrebt, da bisher in diesem Fall die Fehlalarmreduktion nicht so hoch ist wie bei angestrebten 95% oder für nicht herbeigeführte Alarme und somit größere Verbesserungen möglich erscheinen. Für jede der Gruppen werden 100

Wälder mit jeweils 1000 Bäumen erzeugt und die erreichten Sensitivitäten und Fehlalarmreduktionen bestimmt. Sie können zur Beurteilung der Eignung der betrachteten Charakteristika herangezogen werden. Wie zuvor wird die vorgegebene Sensitivität gut eingehalten, die Fehlalarmreduktionen unterscheiden sich jedoch deutlich, sowohl von den bisherigen Ergebnissen als auch innerhalb der Gruppen (Tab. 6.7).

Gruppe	Fehlalarmreduktion
Steigungs-Charakteristika	22,45%
Variabilitäts-Charakteristika	26,01%
Schätzwerte des unterliegenden Signals	24,95%
Wavelet-Koeffizienten	10,24%
klassische Charakteristika	15,51%
robuste Charakteristika	15,27%
alle Charakteristika zugleich	13,16%

Tabelle 6.7: Fehlalarmreduktionen bei Verwendung von Rohdaten, Charakteristika und Dummy-Variablen, alle Alarme, $\alpha = 0,02$

Bisher konnten die Fehlalarme mit Half-Sampling für alle, auch herbeigeführte Alarme im Median um 26,50% reduziert werden. Die Hinzunahme zusätzlicher charakterisierender Merkmale hat nicht wie erwartet die Fehlalarmreduktion weiter erhöht sondern zum Teil erheblich verringert. Lediglich die Variabilitäts-Charakteristika und die Schätzwerte für das unterliegende Signal führen mit im Median 26,01% bzw. 24,95% zu vergleichbar hohen Fehlalarmreduktionen. Diese Ergebnisse werfen die Frage nach dem Einfluss („Importance“) der eingehenden Variablen auf das Klassifikationsergebnis auf.

6.2.1 Importance der Variablen in Wäldern

Bei einem Wald handelt es sich, wie bei vielen anderen Verfahren des maschinellen Lernens auch, um ein so genanntes Black-Box-Verfahren: Auf welche Weise die Variablen in das Modell eingehen, das zur Vorhersage bzw. Klassifikation genutzt wird, bleibt dem Anwender verborgen. Solche Verfahren sind auf das Ziel ausgerichtet, möglichst gute Vorhersagen zu treffen, und nicht (Sach-)Zusammenhänge offen zu legen. Dennoch gibt es Ansätze, den Einfluss (*Importance*) der Variablen in Wäldern zu messen.

Zwei Importance-Maße sind im R-Package `RandomForest` implementiert. Zum einen wird pro Variable und Baum die Zunahme an Reinheit bezüglich des Gini-Kriteriums (vgl. Kap. 5.3.3) aufsummiert. Aus dem arithmetischen Mittel über alle Bäume des Waldes ergibt sich die Wichtigkeit der betrachteten Variablen. Zum anderen wird auf Permutationen der Daten basierend auf den Einfluss einer Variablen geschlossen. In der klassischen Konstruktionsweise eines Waldes wird für jeden Baum eine Bootstrap-Stichprobe gezogen. Die Beobachtungen, die nicht in der Stichprobe enthalten sind, werden oob-Beobachtungen („out of bag“) genannt. Um die Wichtigkeit einer Variablen zu messen, werden ihre Realisationen in allen oob-Beobachtungen permutiert und durch den entsprechenden Baum klassifiziert. Die Differenz aus den Anteilen richtig klassifizierter permutierter und nicht permutierter oob-Beobachtungen wird für jeden Baum bestimmt. Der Durchschnitt dieser Differenzen über alle Bäume im Wald misst die Importance der betrachteten Variablen.

Strobl et al. (2008) weisen darauf hin, dass die Permutation der Beobachtungen einer Variablen bei Beibehaltung der Beobachtungen aller übrigen Variablen und der Klassenzugehörigkeit nicht nur zu großen Importance-Werten führt, wenn die Variable tatsächlich einen großen Einfluss auf das Klassifikationsergebnis ausübt. Auch Abhängigkeiten zwischen der betrachteten und den übrigen Variablen können beim Permutations-basierten Importance-Maß zu großen Werten führen. Variablen können also nach diesem Kriterium wichtiger erscheinen als sie tatsächlich sind. Als Alternative schlagen die Autoren eine aufwendige bedingte Permutation vor, die diesen Effekt mildern kann, aber nicht vollständig vermeidet.

Auch das Gini-basierte Importance-Maß kann nur als Anhaltspunkt bei der Beurteilung des Einflusses der eingehenden Variablen betrachtet werden. Bei der Wahl der besten Splits in jedem Baum des Waldes nach dem Gini-Kriterium werden Variablen, die mehr Werte annehmen, gegenüber solchen mit weniger Werten bevorzugt (Breiman et al. (1984)). Daher kann auch der Durchschnitt des Gini-Zuwachses Variablen für das Klassifikationsergebnis wichtiger erscheinen lassen als gerechtfertigt. Bestehen Abhängigkeiten zwischen Variablen können die Gini-Zuwächse dieser Variablen ihre Wichtigkeit geringer erscheinen lassen. Die Bevorzugung von manchen Variablen in der Konstruktion von Wäldern wirkt sich auch auf das Permutations-basierte Importance-Maß aus. Aus den Nachteilen beider Maße ist schwer festzustellen, welches in der vorliegenden Anwendung geeigneter ist. Hier wird das Gini-Importance-Maß gewählt, anhand

dessen die Variablen, die zur Klassifikation von Alarmsituationen verwendet werden sollen, bestimmt werden. Das Permutations-basierte Importance-Maß könnte ebenso gut verwendet werden.

6.2.2 Auswahl der geeignetsten Variablen

Die Hinzunahme von Charakteristika hat bislang die Klassifikationsergebnisse nicht verbessern können. Der Grund hierfür liegt vermutlich darin, dass viele der Variablen nur wenig für die Klassifikation nützliche Information tragen. In einem solchen Fall kann es bei Bestimmung des besten Splits innerhalb der Konstruktion der Bäume zu einer Auswahl zwischen ausschließlich schlecht geeigneten Kandidaten kommen, wodurch die resultierende Klassifikationsregel negativ beeinflusst wird. Daher sollen zunächst die Variablen ermittelt werden, die die Klassifikation stören.

Dazu werden für die 100 Wälder jeweils in den oben genannten Gruppen (Tab. 6.7) die maximal erreichten Importance-Werte pro Variable ermittelt. Die kleinsten Maxima haben die Dummy-Variablen zur Kennzeichnung ersetzter Werte. Sie tragen also am wenigsten zur Unterscheidung von alarmrelevanten und nicht alarmrelevanten Situationen bei und werden daher aus dem Datensatz entfernt. Mit dem so verkleinerten Datensatz werden erneut in ausgewählten Gruppen je 100 Wälder konstruiert.

Tabelle 6.8 enthält die Mediane der erreichten Fehlalarmreduktionen bei Verwendung der Rohdaten und Charakteristika, aber nicht der Dummy-Variablen. In den betrachteten Gruppen kann die Fehlalarmreduktion durch Weglassen der Dummy-Variablen gesteigert werden. Besonders die Wavelet-Koeffizienten führen nun zu Alarmregeln mit deutlich weniger Fehlalarmen (21,81%) als zuvor (10,24%). Allerdings kann die Fehlalarmreduktion durch Hinzunahme von Charakteristika nicht gesteigert werden.

Um zu überprüfen, ob die hier betrachteten Charakteristika auch bei einer angestrebten Sensitivität von 95% und auch für die nicht durch Manipulation herbeigeführten Alarme nicht zu einer besseren Alarmklassifikation beitragen, werden mit diesen Vorgaben und für die in Tabelle 6.8 aufgeführten Gruppen jeweils 100 Wälder erzeugt. Die Auswahl der Wavelet-Koeffizienten wird dabei zuvor entsprechend der Gini-Importance reduziert.

Gruppe	Fehlalarm-reduktion	Steigerung in Prozentpunkten
Rohdaten ohne Dummy-Variablen und ohne zusätzliche Charakteristika	**27.86%**	**1,36**
Steigungs-Charakteristika	22,65%	0,20
Variabilitäts-Charakteristika	27,50%	1,49
Schätzwerte des unterliegenden Signals	26,26%	1,31
Wavelet-Koeffizienten	21,81%	11,57

Tabelle 6.8: Fehlalarmreduktionen bei Verwendung von Rohdaten, Charakteristika ohne Dummy-Variablen, alle Alarme, $\alpha = 0,02$

Die erreichten Fehlalarmreduktionen (Tab. 6.9, S. 77) zeigen, dass die betrachteten Charakteristika, die den gesundheitlichen Verlauf der Patienten beschreiben, das Klassifikationsergebnis nicht weiter verbessern können. Sie bestätigen allerdings auch, dass bei alleiniger Verwendung der Rohdaten wesentlich verbesserte Alarmregeln konstruiert werden können. Nachfolgend wird daher überprüft, ob diese Alarmregeln allgemein gültig sind und auch für andere Patienten verwendet werden können.

6.3 Überprüfung der Generalisierbarkeit

Die Generalisierbarkeit ist grundlegende Voraussetzung für die Anwendung der generierten Alarmregeln in der Praxis. Generalisierbarkeit bedeutet, dass auch Patienten, deren Vitalparameterwerte und ihr Monitoring nicht zur Konstruktion der Regeln verwendet wurden, zuverlässig mit den neuen Regeln überwacht werden können.

Bisher wurden die Alarme zufällig in Lern-, Schätz- und Teststichproben aufgeteilt. Unter der Annahme, dass der Datensatz repräsentativ ist bezüglich der Auswahl der enthaltenen „Patiententypen“, ist dieses Vorgehen gerechtfertigt. Um zu überprüfen, ob die Annahme gerechtfertigt ist und tatsächlich die Daten der Patienten in der Studie ausreichen, um auf den Gesundheitszustand anderer Patienten zu schließen, wird die Konstruktion der Wälder mit *stratifizierten* Lern-, Schätz- und Teststichproben wiederholt.

Bei der Aufteilung wird nun gemäß der „Leave-one-out-Methode“ jeweils ein Patient nicht in die Konstruktion der Alarmregeln einbezogen sondern nur für die Feststel-

lung deren Güte herangezogen. Es wird für jeden Patienten ein Wald erzeugt und die Sensitivität und Fehlalarmreduktion ermittelt. Bei allgemein gültigen Regeln sollten die erzielten Sensitivitäten und Fehlalarmreduktionen mit denen vergleichbar sein, die ohne stratifizierte Stichproben erzeugt wurden. Dann hätte es keine Auswirkung auf die Sicherheit eines Patienten, ob seine Daten zur Konstruktion der Alarmregeln, mit denen er überwacht wird, verwendet wurden.

Es treten bei 67 Fällen alarmrelevant annotierte Situationen auf, die für jeden Fall sowohl teils herbeigeführt als auch teils nicht herbeigeführt wurden. Bei 85 Fällen wurden nicht alarmrelevant oder hinweisend annotierte Situationen beobachtet, die bei 7 Fällen vollständig durch Manipulationen ausgelöst wurden.

Bei Betrachtung aller, auch der herbeigeführten Alarme wird bei angestrebten Sensitivitäten von 95% und 98% für die meisten Fälle eine Sensitivität von 100% erreicht (Abb. 6.7, Abb. 6.8). Es treten aber auch sehr niedrige Sensitivitäten von bis zu 50% auf. Die erreichten Fehlalarmreduktionen sind dabei häufig gering. Sie liegen im Median bei 3,23% (95% Sensitivität) bzw. 0% (98% Sensitivität). Für einige Patienten erscheint die erreichte Reduktion der Fehlalarme hoch. Allerdings ist dabei die zum Teil geringe Anzahl an nicht alarmrelevanten Situationen zu beachten. Zum Beispiel wird im Extremfall der Fehlalarmreduktion um 100% die einzige beobachtete und als nicht alarmrelevant annotierte Situation des Falls korrekt erkannt. Dies muss genauso bei sehr niedrigen Sensitivitäten bedacht werden. Der Extremfall einer Sensitivität von nur 50% wird für einen Fall erreicht, bei dem eine der zwei als alarmrelevant annotierten Situationen nicht erkannt wird. Ähnliche Ergebnisse sind für die nicht herbeigeführten Alarme zu beobachten. Die erreichte Sensitivität ist für fast alle Fälle 100% (Abb. 6.9, Abb. 6.10). Im Median und sogar im dritten Quartil kann bei beiden angestrebten Sensitivitäten keine Fehlalarmreduktion (0%) erreicht werden.

Insgesamt bleiben die erreichten Fehlalarmreduktionen bei Stichprobenaufteilung unter Berücksichtigung der Fallnummer deutlich hinter denen bei rein zufälliger Stichprobenaufteilung zurück. Abbildung 6.11 zeigt den Zusammenhang zwischen erreichten Sensitivitäten und Fehlalarmreduktionen für alle bzw. nur die nicht herbeigeführten Alarme und mit Zielsensitivitäten von 95% und 98%. Zur besseren Unterscheidbarkeit sind die Werte in x-Richtung leicht versetzt dargestellt. Die senkrechten grauen Linien markieren hier die Zielsensitivitäten. Die waagerechten grauen Linien zeigen die klein-

ste erreichte Fehlalarmreduktion bei rein zufälliger Aufteilung der Stichproben an. Nur für wenige Fälle liegt die erreichte Fehlalarmreduktion in der Nähe oder oberhalb dieser Minima. Es werden mehrheitlich deutlich schlechtere Fehlalarmreduktionen erreicht.

Darüber hinaus kann auf Grundlage der vorliegenden Daten nicht festgestellt werden, ob dies auch für Patienten auf anderweitig spezialisierten Intensivstationen gilt. Möglicherweise sind zum einen manche Krankheitsbilder, die auf anderen Intensivstationen zu beobachten sind, in den Studiendaten nicht enthalten. Zum anderen beziehen sich die generierten Alarmregeln auf die Alarme und Einstellungen des Standardmonitors, die sich auf verschieden ausgerichteten Intensivstationen unterscheiden können.

Unter der Voraussetzung, dass die Daten repräsentativ bezüglich der auftretenden Patiententypen sind, ist das verwendete Verfahren für eine datengestützte Alarmregelgenerierung geeignet. Die großen Unterschiede in den Klassifikationsergebnissen deuten jedoch darauf hin, dass die vorliegenden Daten nicht ausreichen, um allgemein gültige Alarmregeln zu generieren. Neue Patienten können zwar (meist) sicher überwacht werden, eine Reduktion der Fehlalarme ist in der Mehrzahl der Fälle jedoch nicht möglich. Die hohen Fehlalarmreduktionen auf rein zufällig aufgeteilten Daten lassen weitere Forschungsvorhaben, insbesondere eine Weiterführung und Ausweitung der klinischen Studie auch auf Intensivstationen anderen Typs lohnenswert erscheinen. Darüber hinaus könnten NP-Wälder auch in anderen Anwendungsgebieten, in denen die Konsequenzen einer Fehlklassifikation unerwünscht und ausreichend Daten beschaffbar sind, zum Beispiel bei Spam-Filtern für E-Mail-Dienste oder bei der Entdeckung krimineller Handlungen (Fraud Detection), eingesetzt werden.

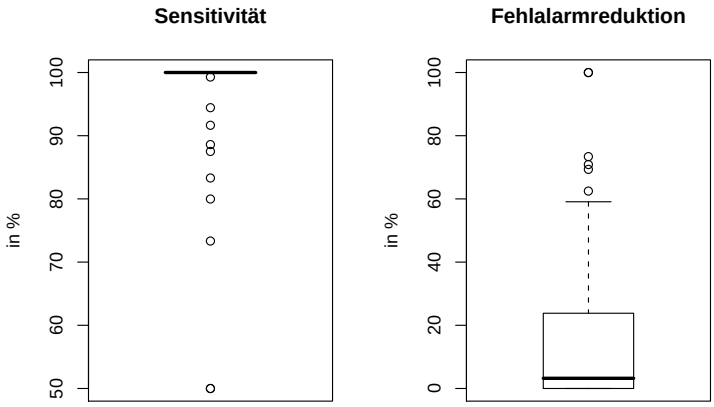

Abbildung 6.7: Sensitivitäten und Fehlalarmreduktionen (Signifikanzniveau 5%, Half-Sampling, alle Alarme, Leave-one-out)

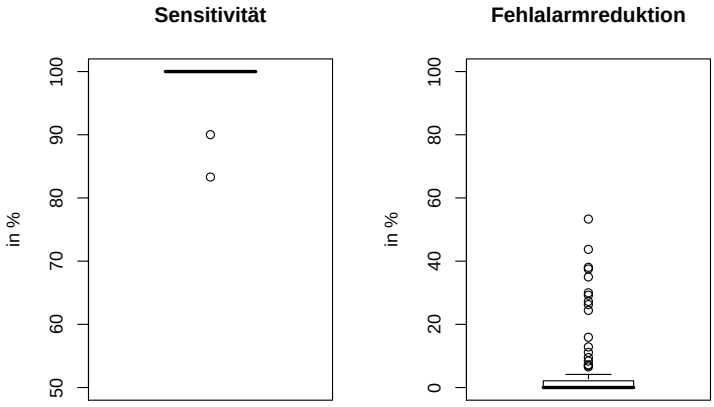

Abbildung 6.8: Sensitivitäten und Fehlalarmreduktionen (Signifikanzniveau 2%, Half-Sampling, alle Alarme, Leave-one-out)

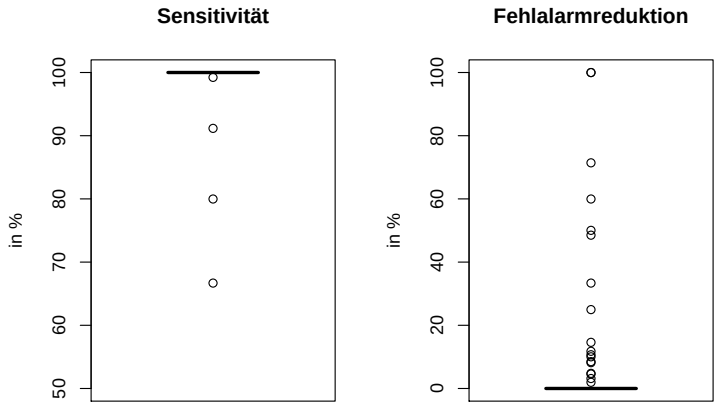

Abbildung 6.9: Sensitivitäten und Fehlalarmreduktionen (Signifikanzniveau 5%, Half-Sampling, nicht herbeigeführte Alarme, Leave-one-out)

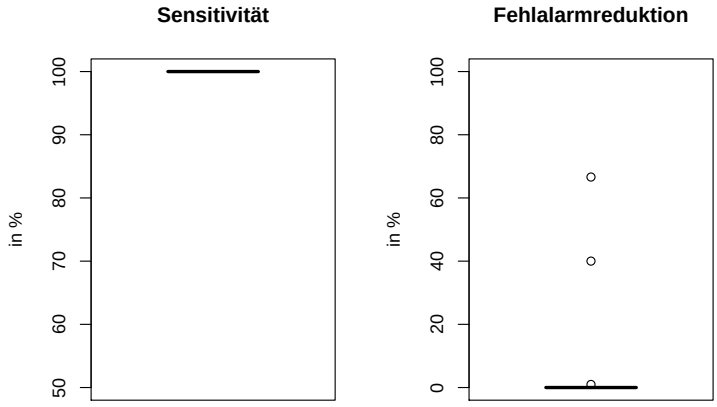

Abbildung 6.10: Sensitivitäten und Fehlalarmreduktionen (Signifikanzniveau 2%, Half-Sampling, nicht herbeigeführte Alarme, Leave-one-out)

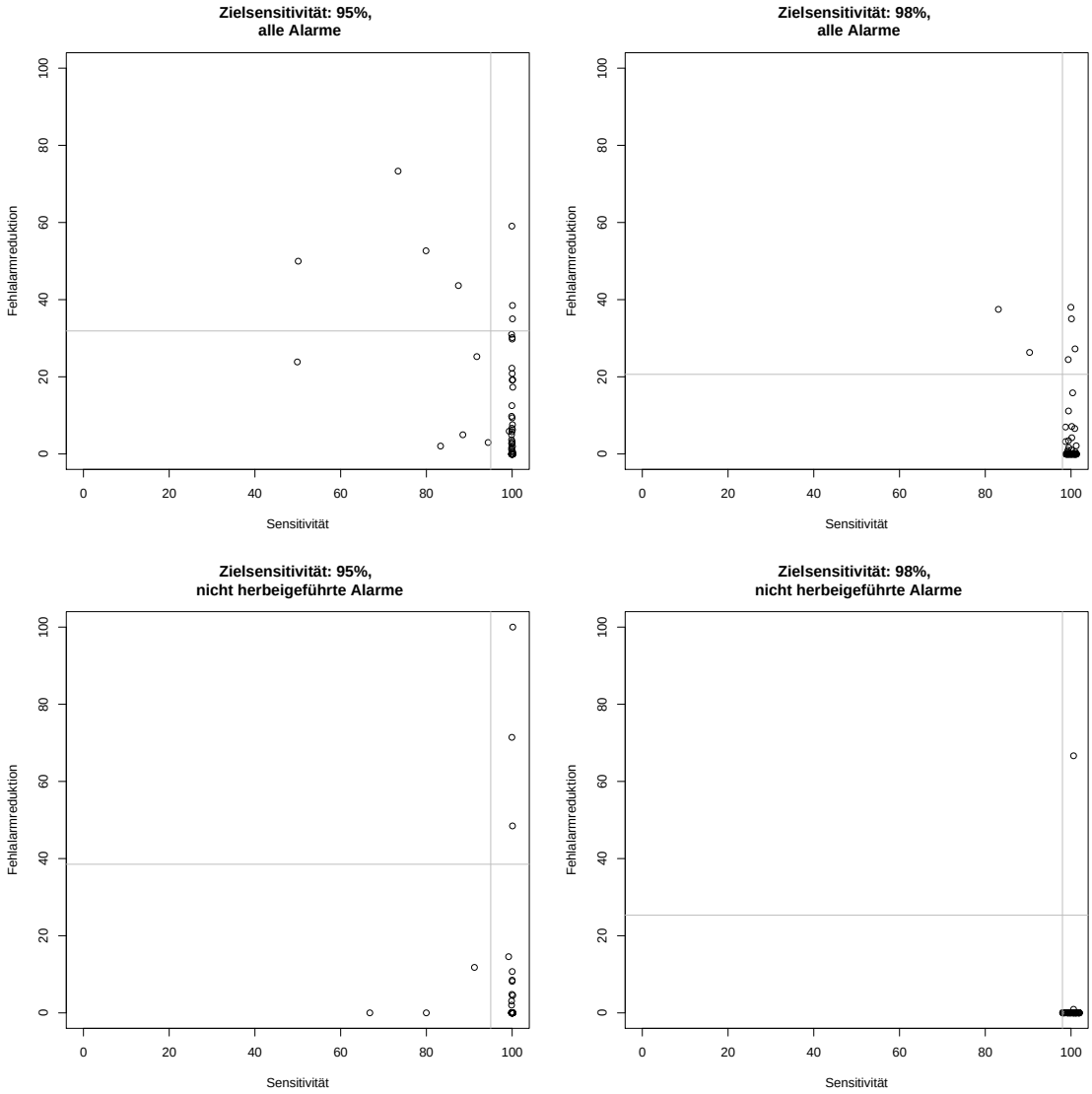

Abbildung 6.11: Sensitivitäten und Fehlalarmreduktionen im Vergleich

Gruppe	Fehlalarmreduktion, alle Alarme, $\alpha = 0,05$
Rohdaten ohne Dummy-Variablen und ohne zusätzliche Charakteristika	**43,21%**
Steigungs-Charakteristika	33,90%
Variabilitäts-Charakteristika	39,86%
Schätzwerte des unterliegenden Signals	38,73%
Wavelet-Koeffizienten	35,71%
Gruppe	Fehlalarmreduktion, nicht herbeigeführte Alarme, $\alpha = 0,05$
Rohdaten ohne Dummy-Variablen und ohne zusätzliche Charakteristika	**55,40%**
Steigungs-Charakteristika	47,28%
Variabilitäts-Charakteristika	50,86%
Schätzwerte des unterliegenden Signals	49,76%
Wavelet-Koeffizienten	48,61%
Gruppe	Fehlalarmreduktion, nicht herbeigeführte Alarme, $\alpha = 0,02$
Rohdaten ohne Dummy-Variablen und ohne zusätzliche Charakteristika	**42,07%**
Steigungs-Charakteristika	35,09%
Variabilitäts-Charakteristika	38,34%
Schätzwerte des unterliegenden Signals	39,10%
Wavelet-Koeffizienten	33,37%

Tabelle 6.9: Fehlalarmreduktionen bei Verwendung von Rohdaten, Charakteristika ohne Dummy-Variablen

Kapitel 7

Zusammenfassung

In dieser Arbeit wird die Eignung datengestützt generierter Alarmregeln für die Überwachung von Patienten auf der Intensivstation untersucht. Hierfür wird ein bekanntes Klassifikationsverfahren modifiziert, so dass es Alarmregeln mit einer wählbar hohen Sensitivität erzeugt. Dieses neue Verfahren wird offline an Patientenmonitoring-Daten getestet.

Die Reduktion von Fehlalarmen im Patientenmonitoring auf der Intensivstation ist aufgrund von Fehlalarmraten von bis zu 90% seit etwa 1980 Gegenstand der Forschung verschiedener Disziplinen. Die wissenschaftlichen Fortschritte haben jedoch bislang keinen Eingang in die klinische Praxis gefunden. Eine einfache Schwellwertüberwachung der Vitalparameter ist gegenwärtiger Stand der Technik auf dem Gebiet der Monitoring-Geräte. Daher wird eine Anforderungsanalyse durchgeführt, um Bedingungen für die Überführbarkeit neuer Verfahren in die Praxis zu ermitteln. Die offensichtlichste Anforderung ist die schnelle und sichere Erkennung Gesundheits-kritischer, aber ebenso auch Gesundheits-unbedenklicher Situationen. Des Weiteren darf eine Lernphase von adaptiven Verfahren nur kurz andauern und muss immer zu zuverlässigen Regeln führen. Neue Verfahren müssen angemessen mit fehlenden Werten und mit Signal-Verschmutzungen wie Rauschen oder Ausreißern umgehen können. Aus psychologischen Gründen ist die Transparenz der Funktionsweise eines neuen Alarmsystems für den Anwender von Bedeutung. Um für den Einsatz auf der Intensivstation zugelassen zu werden, muss ein neues Alarmsystem evaluiert werden können. Besonders, wenn das Standard-Monitoringsystem und ein neues Alarmsystem nicht zum selben Zeitpunkt Alarme auslösen, stellt dies eine besondere Schwierigkeit dar. Gemäß der festgestellten Anforderungen wird ein Ansatz zur Alarmregelgenerierung gewählt, der mit sehr geringem Zeitverzug die Alarme des Standard-Patientenmonitors validiert. Es werden also neue Alarmregeln erzeugt, nach denen ein Alarm des Patientenmonitors unterdrückt

oder ausgegeben wird. Bei ausreichend großem und annotiertem Datensatz können für diesen Zweck allgemein gültige Alarmregeln mit Hilfe von Klassifikationsverfahren aus den Daten abgeleitet werden.

Die Daten, die in dieser Arbeit zur Alarmregelgenerierung herangezogen werden, wurden in einer klinischen Studie des Teilprojekts C4 des Sonderforschungsbereichs 475 am Universitätsklinikum Regensburg erhoben. Es liegen sekundengenaue Aufzeichnungen von Vitalparametern, die ausgelösten Alarme, technische Informationen vom Monitoringgerät und klinische Annotationen vor. In den Annotationen ist durch eine Ärztin festgehalten, ob ein Alarm technisch richtig gegeben wurde, ob er klinisch relevant ist und ob er durch eine Manipulation des Pflegepersonals herbeigeführt wurde. Bei Betrachtung aller Alarme liegt die Fehlalarmrate bei 85%, bezogen auf Alarme, die nicht durch das Pflegepersonal herbeigeführt wurden, liegt sie bei 76%. Dabei werden die Alarme − sowohl klinisch relevante als auch Fehlalarme − in erster Linie durch Schwellwert-überwachte Variablen ausgelöst. In der Überwachung dieser Vitalparameter sind demnach die Defizite von Standard-Patientenmonitoren zu finden.

Die Datenvorverarbeitung gilt als wichtiger Schritt, der der Anwendung von Klassifikationsverfahren vorangeht, und in vielen Anwendungen das Klassifikationsergebnis verbessern kann. Entsprechend der Vorgehensweise von Ärzten am Patientenbett, wenn sie den Gesundheitszustand eines Patienten einschätzen, soll der Verlauf des Gesundheitszustands kurz vor einem Alarm in die Klassifikation einbezogen werden. Aus diesem Grund werden aus den beobachteten Daten Charakteristika konstruiert, die den gesundheitlichen Verlauf eines Patienten kurz vor Auslösung eines Alarms wiedergeben. Dies geschieht zum einen mit Hilfe lokaler linearer Regression. Es werden der Trend in den Vitalparameter-Messwerten und ihre Variabilität sowohl auf robuste als auch klassische Weise bestimmt. Zusätzlich wird, wie in der Signalverarbeitung häufig, eine Waveletzerlegung durchgeführt. Die Wavelet-Koeffizienten bei Verwendung des Haar-Wavelets stehen in Beziehung zu arithmetischen Mittelwerten in Teilstücken des betrachteten Zeitraums vor einem Alarm. Sie eignen sich daher als Charakteristika und spiegeln auf alternative Weise zu den lokalen linearen Regressionen die gesundheitliche Veränderung wider.

Alarmregeln für die Überwachung von Intensivpatienten müssen eine sehr hohe Sensitivität aufweisen, da Fehlklassifikationen von alarmrelevanten Situationen schwerwie-

gende Folgen haben können. Fehlkassifikationen von nicht alarmrelevanten Situationen gefährden die Gesundheit der Patienten nicht, sondern bedeuten eine Störung der Pflegekräfte in ihrer Arbeit. Bei Verwendung bekannter Klassifikationsverfahren kann diesem Ungleichgewicht mit stark unterschiedlichen Fehlklassifikationskosten begegnet werden. Allerdings ist die Wahl der Kosten nicht aus dem Sachzusammenhang zu begründen und nur durch viele Versuche adäquat zu treffen. Dieser Ansatz eignet sich nicht, um Alarmregeln mit einer gewissen Zielsensitivität zu generieren. Aus diesem Grund wird das häufig genutzte Klassifikationsverfahren „Random Forest“ entsprechend der Anforderungen in der Klassifikation von Alarmen modifiziert. Analog zur Konstruktion statistischer Tests nach dem Neyman-Pearson Lemma ist die Konstruktion von Wäldern zu einer vorgegebenen Sensitivität möglich.

Die Eignung dieses Verfahrens wird für Zielsensitivitäten von 95% und 98% sowohl für alle Alarme als auch für nicht herbeigeführte Alarme gezeigt. Dazu werden je 1000 Wälder auf zufälligen Teilstichproben der Studiendaten erzeugt. Die nicht verwendeten Daten werden als Teststichproben zur Bewertung der Eignung des Verfahrens herangezogen. Die vorgegebenen Sensitivitäten werden in den betrachteten Fällen im arithmetischen Mittel und Median bei geringer Variabilität in den Ergebnissen erreicht. Gleichzeitig können die Fehlalarme um bis zu 55% im Median reduziert werden. Die geringste erreichte Fehlalarmreduktion ist bei einer Vorgabe von 98% Zielsensitivität und Verwendung aller, auch der herbeigeführten Alarme zu beobachten. Selbst in diesem Fall beträgt die Reduktion der Fehlalarme noch etwa 27%. Charakteristika des gesundheitlichen Verlaufs können diese Ergebnisse nicht weiter steigern. Die Allgemeingültigkeit der erzeugten Alarmregeln wird anhand einer nach Patienten stratifizierten Stichprobe überprüft. In diesem Fall bleiben die Fehlalarmreduktionen deutlich hinter den bisherigen Ergebnissen zurück. Dabei wird bei den meisten Patienten eine Sensitivität von 100% erreicht. Dies deutet darauf hin, dass die vorliegenden Daten nicht ausreichen, um allgemein gültige Regeln zu generieren.

Das im Rahmen dieser Arbeit entwickelte Verfahren kann bei Erhebung weiterer Daten zur Alarmregelgenerierung für die Patientenüberwachung genutzt werden. Dabei erscheint auch eine Ausweitung der annotierten Datensammlung auf weitere Typen von Intensivstationen sinnvoll. Solche Daten können auch zur Entwicklung von Verfahren zur Früherkennung von kritischen Situationen, zur priorisierten Alarmgebung entsprechend der Dringlichkeit einer Intervention oder zur Diagnose genutzt werden.

In Anwendungen, in denen die Konsequenzen von Fehlklassifikationen bei einer von zwei Populationen äußerst unerwünscht sind und Fehlklassifikationskosten nicht aus dem Sachzusammenhang zu begründen sind, ist der vorgestellte Ansatz zur Lösung des Klassifikationsproblems geeignet und leicht übertragbar. Dabei können bei Beibehaltung des Bagging anstatt der randomisierten CART-Bäume auch andere Klassifizierer verwendet werden. Das in dieser Arbeit entwickelte Konzept, die Klassifikations-Entscheidung eines gebaggten Ensembles von Klassifizierern analog zu einem Neyman-Pearson-Test zu bilden, ist daher in vielen Anwendungsgebieten von Interesse.

Symbolverzeichnis

A	Knoten, $A \in E$	40
$a_{k\|t}$	Anteil an Objekten aus Population Π_k in Knoten A_t	43
$\mathfrak{A}$	σ-Algebra über Π	36
α	Signifikanzniveau	55
B_0, B_1	Partition von $\mathscr{X}$, Interpretation: Situation π ist alarmrelevant für $\mathbf{X}(\pi) \in B_0$ und nicht alarmrelevant für $\mathbf{X}(\pi) \in B_1$	37
$\hat{\beta}_t^{KQ;\,m}$	die Steigung der geschätzten Gerade nach KQ-Regression im Fenster der Länge m	30
$\hat{\beta}_t^{RM;\,m}$	die Steigung der geschätzten Gerade nach RM-Regression im Fenster der Länge m	30
$\hat{\beta}_t^{RM_{adaptiv}}$	die Steigung der geschätzten Gerade nach $RM_{adaptiv}$-Regression im Fenster variabler Länge mit minimaler Fensterbreite 60 und maximaler Fensterbreite 121	30
$c_{j\|i}$	Fehlklassifikationskosten	37
$c_{0,0}$, $d_{j,k}$	Wavelet-Koeffizienten	32
$\mathfrak{C}$	σ-Algebra über $\mathscr{X}$	37
D	Menge von gerichteten Kanten	40
d_i	d_i=$P(\delta_i(\mathbf{X}) = 1\|G = 0)$ Wahrscheinlichkeit, dass Entscheidungsbaum δ_i eine alarmrelevante Beobachtung als nicht alarmrelevant klassifiziert	55
$\delta(\mathbf{x})$	Entscheidungs- oder Klassifikationsregel	37
$\delta^\star(\mathbf{x})$	Bayes-Regel	38
$\delta_j(\mathbf{x})$	Entscheidungsregel des j-ten von N Entscheidungsbäumen eines Waldes	49
$\delta_{Wald}(\mathbf{x})$	Entscheidungsregel eines Waldes	49
$\delta_{NP-Wald}(\mathbf{x})$	Entscheidungsfunktion eines NP-Waldes	55

E	Menge von Knoten	40
f_i	mehrdimensionale Dichten zu den bedingten Verteilungen $P_{X\|G=i}$, $i = 0, 1$	37
$G(\pi)$	Zufallsvariable, Populationszugehörigkeit der Alarmsituation π, Interpretation: $G(\pi) = 0$ wenn Situation alarmrelevant, $G(\pi) = 1$ wenn Situation nicht alarmrelevant	36
K	gerichtete Kante, $K = (A_1, A_2)$, $A_1, A_2 \in E$	40
$L(\mathbf{x}, f_0, f_1)$	Likelihood-Quotient	52
$\hat{\mu}_t^{KQ;m}$	das geschätzte Signal am Ende des Fensters der Länge m nach KQ-Regression	30
$\hat{\mu}_t^{RM;m}$	das geschätzte Signal am Ende des Fensters der Länge m nach RM-Regression	30
$\hat{\mu}_t^{RM_{adaptiv}}$	das geschätzte Signal am Ende des Fensters variabler Länge nach $RM_{adaptiv}$-Regression	30
P	Wahrscheinlichkeitsmaß auf $\mathfrak{A}$	36
$P_{\mathbf{X}}$	Verteilung von $\mathbf{X}$ auf $\mathfrak{C}$	37
p_i	a-priori-Wahrscheinlichkeiten, $P_G(i) = P(G^{-1}(i)) = p_i$, $i = 0, 1$	36
Π	Gesamtpopulation, Menge aller Alarmsituationen	35
Π_0	Population aller alarmrelevanten Situationen	36
Π_1	Population aller nicht alarmrelevanten Situationen	36
π	Alarmsituation, d.h. Situation, in der das Standard-Monitoring-Gerät einen Alarm auslöst, zu klassifizierendes Objekt	36
Qn	robuster Skalenschätzer	31
Qn_t^m	der Qn-Schätzer aller Beobachtungen im Fenster der Länge m	31
$Qn_t^{RM;Res;m}$	der Qn-Schätzer der Residuen nach RM-Regression im Fenster mit fester Länge m	31
$Qn_t^{RM_{adaptiv}}$	der Qn-Schätzer aller Beobachtungen im Fenster mit variabler Länge gewählt analog zur $RM_{adaptiv}$-Regression	31
$R(\delta)$	Bayes-Risiko einer Entscheidungsregel δ	38

$s^\star$	Split, der zur größten Abnahme an Unreinheit führt	44
$\hat{\sigma}_t^m$	die Standardabweichung aller Beobachtungen im Fenster der Länge m	31
$\hat{\sigma}_t^{KQ;\,Res;\,m}$	die Standardabweichung der Residuen nach KQ-Regression im Fenster der Länge m	31
$V(i,j)$	Verlustfunktion	37
ϕ_{A_t}	Funktion zur Messung der (Un-)Reinheit eines Knotens	43
$\mathbf{X}(\pi)$	Zufallsvektor, charakterisierende Merkmale einer zu klassifizierenden Alarmsituation $\pi \in \Pi$	36
$\mathscr{X}$	Beobachtungsraum von $\mathbf{X}$	37
$\psi(s,t)$	Funktion zur Messung der Abnahme an Unreinheit eines Knotens	44

Literaturverzeichnis

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. Hoboken, NJ: Wiley-Interscience.

Augusto, J. (2005). Temporal reasonig for decision support in medicine. *Artificial Intelligence in Medicine 33*, 1–24.

Becker, K., B. Thull, H. Käsmacher-Leidinger, J. Stemmer, G. Rau, G. Kalff, und H. Zimmermann (1997). Design and validation of an intelligent patient monitoring and alarm system based on a Fuzzy logic process model. *Artificial Intelligence in Medicine 11*, 33–53.

Borowski, M., K. Schettlinger, und U. Gather (2008). Multivariate real time signal extraction by a robust adaptive regression filter. *Communications in Statistics – Simulation and Computation* (erscheint).

Breiman, L. (1996). Bagging predictors. *Machine Learning 24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning 45*(1), 5–32.

Breiman, L., J. H. Friedman, R. A. Olshen, und C. J. Stone (1984). *Classification and Regression Trees*. Statistics/Probability Series. Belmont, California, U.S.A.: Wadsworth Publishing Company.

Bühlmann, P. und B. Yu (2002). Analyzing bagging. *Annals of Statistics 30*(4), 927–961.

Buja, A. und W. Stuetzle (2006). Observations on bagging. *Statistica Sinica 16*, 323–351.

Chambrin, M., P. Ravaux, D. Calvelo-Aros, A. Jaborska, C. Chopin, und B. Boniface (1999). Multicentric study of monitoring alarms in the adult intensive care unit (ICU): A descriptive analysis. *Intensive Care Medicine 25*, 1360–1366.

Chambrin, M.-C. (2001). Alarms in the intensive care unit: How can the number of false alarms be reduced? *Critical Care 5*, 184–188.

Davies, P., R. Fried, und U. Gather (2004). Robust signal extraction for on-line monitoring data. *Journal of Statistical Planning and Inference, Special Issue: Contemporary Data Analysis: Theory and Methods in Honor of John W. Tukey 122*, 65–78.

Dräger Medical Systems Inc. (2003). *Serie Infinity Delta, Gebrauchsanweisung*. Danvers.

Edwards, D. (2000). *Introduction to Graphical Modelling* (2nd ed.). New York: Springer-Verlag.

Edworthy, J. und E. Hellier (2005). Fewer but better auditory alarms will improve patient safety. *Quality and Safety in Health Care 14*, 212–215.

Fried, R. und K. Schettlinger (2008). robfilter: Robust time series analysis. http://cran.r-project.org/web/packages/robfilter/index.html.

Gather, U., K. Schettlinger, und R. Fried (2006). Online signal extraction by robust linear regression. *Computational Statistics 21*(1), 33–51.

Haimowitz, I., P. Le, und I. Kohane (1995). Clinical monitoring using regression-based trend templates. *Artificial Intelligence in Medicine 7*, 473–496.

Hanson, C. und B. Marshall (2001). Artificial intelligence applications in the intensive care unit. *Critical Care Medicine 29*(2), 427–435.

Hastie, T., R. Tibshirani, und J. Friedman (2001). *The Elements of Statistical Learning*. New York: Springer.

Hutchinson, G. (2003). System for and method of establishing monitoring alarm limits based on physiological variables. GE Medical Systems Information Technologies, Inc. (Anmelder), US Patent, US 6,585645 B2.

Hutchinson, G. und P. Schluter (2004). Method and apparatus for monitoring using a mathematical model. GE Medical Systems Information Technologies, Inc. (Anmelder), US Patent, US 2004/0236188 A1.

Imhoff, M. und S. Kuhls (2006). Alarm algorithms in critical care monitoring. *Anesthesia & Analgesia 102*(5), 1525–1537.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorial data. *Journal of Applied Statistics 29*(2), 119–127.

Koski, E., A. Mäkivirta, T. Sukuvaara, und A. Kari (1990). Frequency and reliability of alarms in the monitoring of cardiac postoperative patients. *International Journal of Clinical Monitoring and Computing 7*, 129–133.

Koski, E., A. Mäkivirta, T. Sukuvaara, und A. Kari (1991). Development of an expert system for haemodynamic monitoring: computerized symbolization of on-line monitoring data. *Journal of Clinical Monitoring and Computing 8*(4), 289–293.

Koski, E. M. J., T. Sukuvaara, und A. Mäkivirta A.and Kari (1994). A knowledge-based alarm system for monitoring cardiac operated patients − assessment of clinical performance. *International Journal of Clinical Monitoring and Computing 11*, 79–83.

Krol, M. und D. Reich (2000). Development of a decision support system to assist anesthesiologists in operating room. *Journal of Medical Systems 24*(3), 141–146.

Kuhls, S. (2008). *Planung, Durchführung und Analyse einer klinischen Studie zur Bewertung und zum Vergleich von Alarmsystemen in der Intensivmedizin.* Dissertation, Technische Universität Dortmund.

Kuhls, S., S. Siebig, F. Stößel, und M. Imhoff (2006). Entwicklung einer Eingabemaske füur die Erfassung klinischer Annotationen. Technical Report 15, SFB 475, Universität Dortmund.

Lanius, V. (2004). *Statistische Extraktion relevanter Information aus multivariaten Online-Monitoring-Daten der Intensivmedizin.* Dissertation, Fachbereich Statistik, Universität Dortmund.

Lanius, V. und U. Gather (2007). Robust online signal extraction from multivariate time series. Technical Report 38, SFB 475, Technische Universität Dortmund.

Lawless, S. T. (1994). Crying wolf: False alarms in a pediatric intensive care unit. *Critical Care Medicine 22*(6), 981–985.

Liaw, A. und M. Wiener (2002). Classification and regression by randomForest. *R News 2*(3), 18–22.

Mallat, S. G. (1999). *A wavelet tour of signal processing*. San Diego: Academic Press.

Mannheimer, P. (2007). Nuisance alarm reduction in a physiological monitor. Nellcor Puritan Bennett Inc. (Anmelder), US Patent, US 2007/0032714 A1.

Mannheimer, P. und L. Li (2007). Patient monitoring alarm escalation system and method. Nellcor Puritan Bennett Inc. (Anmelder), US Patent, US 2007/0106426 A1.

McIntosh, N. (2002). Intensive care monitoring: past, present and future. *Clinical Medicine 2*(4), 349–355.

McLachlan, G. J. (1992). *Discriminant analysis and statistical pattern recognition*. New York: Wiley-Interscience.

Mood, A., F. Graybill, und D. Boes (1974). *Introduction to the Theory of Statistics*. New York: McGraw-Hill.

Nason, G., A. Kovac, und M. Maechler (2006). wavethresh: Software to perform wavelet statistics and transforms. http://cran.r-project.org/web/packages/wavethresh/index.html.

O'Carroll, T. (1986). Survey of alarms in an intensive therapy unit. *Anaesthesia 41*, 742–744.

Quinlan, J. R. (1993). *C4.5 : programs for machine learning*. San Mateo, Calif.: Morgan Kaufmann.

R Development Core Team (2008). R: A Language and Environment for Statistical Computing. http://www.R-project.org. ISBN 3-900051-07-0.

Rousseeuw, P. J. und C. Croux (1993). Alternatives to the median absolute deviation. *Journal of the American Statistical Association 88*(424), 1273–1283.

Schapire, R. E. (2003). The boosting approach to machine learning: an overview. In D. Denison, M. Hansen, C. Holmes, B. Mallick, und B. Yu (Eds.), *Nonlinear Estimation and Classification*, Volume 171 of *Lecture Notes in Statistics*, pp. 149–172.

Schettlinger, K., R. Fried, und U. Gather (2008). Real time signal processing by adaptive repeated median filters. *International Journal of Adaptive Control and Signal Processing* (erscheint).

Schoenberg, R., D. Sands, und C. Safran (1999). Making ICU alarms meaningful: a comparison of traditional vs. trend-based algorithms. In *Proceedings of American Medical Informatics Association Fall Symposium*, pp. 370–383.

Scott, C. und R. Nowak (2005). A Neyman-Pearson approach to statistical learning. *IEEE Transactions on Information Theory 51*(11), 3806 – 3819.

Siegel, A. (1982). Robust regression using repeated medians. *Biometrika 69*, 242–244.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, und A. Zeileis (2008). Conditional variable importance for random forests. *BMC Bioinformatics 9*(307).

Sukuvaara, T., M. Koksi, A. Mäkivirta, und A. Kari (1993). A knowledge-based alarm system for monitoring cardiac operated patients − technical construction and evaluation. *International Journal of Clinical Monitoring and Computing 10*, 117–126.

Tivig, G. und S. Hebler (2006). Medical monitoring method and system. Koninklijke Philips Electronics N.V. (Anmelder), Internationales Patent, WO 2006/067725 A2.

Tsien, C. (2000a). Event discovery in medical time-series data. In *Proceedings of the American Medical Informatics Association Annual Fall Symposium 2000*, pp. 858–862.

Tsien, C. (2000b). *TrendFinder: Automated detection of alarmable trends*. Ph. D. thesis, Massachusetts Institute of Technology.

Tsien, C. und J. Fackler (1997). Poor prognosis for existing monitors in the intensive care unit. *Critical Care Medicine 25*, 614–619.

Tsien, C., I. Kohane, und N. McIntosh (2000). Multiple signal integration by decision tree induction to detect artifacts in the neonatal intensive care unit. *Artificial Intelligence in Medicine 19*(3), 189–202.

Vidakovic, B. und P. Müller (1994). Wavelets for kids - a tutorial introduction. Discussion Paper Series 94-13, Institute of Statistics and Decision Science, Georgia Institute of Technology.

Wasserman, L. (2004). *All of Statistics : A Concise Course in Statistical Inference*. New York: Springer.

Wickerhauser, M. V. (1996). *Adaptive Wavelet-Analysis*. Braunschweig, Wiesbaden: Vieweg.

Anhang: Verwendete Variablen

Monitoring-Variable	Beschreibung
ARR	Arrhythmie-Indikator
ARR_LOW	untere Alarmgrenze für ARR
ARR_HIGH	obere Alarmgrenze für ARR
ART_D	arterieller diastolischer Blutdruck
ART_M	arterieller mittlerer Blutdruck
ART_M_LOW	untere Alarmgrenze für ART_M
ART_M_HIGH	obere Alarmgrenze für ART_M
ART_M_ALMATTR	Alarminformation für ART_M
ART_S	arterieller sytolischer Blutdruck
ART_S_LOW	untere Alarmgrenze für ART_S
ART_S_HIGH	obere Alarmgrenze für ART_S
ART_S_ALMATTR	Alarminformation für ART_S
BEWERTUNG	Annotation: technische und klinische Validität
CONDSTR	Alarmgrund (z.B. untere Grenzwertverletzung)
CVP	zentralvenöser Druck
etCO2	CO_2-Gehalt im Atemweg am Ende des Atemzyklus
GRADE	Alarmgrad (SER, ADV, LT)
HR	Herzfrequenz
HR_LOW	untere Alarmgrenze für HR
HR_HIGH	obere Alarmgrenze für HR
HR_ALMATTR	Alarminformation für HR
iCO2	CO_2-Gehalt im Atemweg während der Inspiration
iO2	O_2-Gehalt im Atemweg während der Inspiration
MAP	mittlerer Atemwegsdruck
MVe	expiriertes Minutenvolumen

Monitoring-Variable	Beschreibung
MVi	inspiriertes Minutenvolumen
NBP_D	nicht invasiver diastolischer Blutdruck
NBP_M	nicht invasiver mittlerer Blutdruck
NBP_M_LOW	untere Alarmgrenze für NBP_M
NBP_M_HIGH	obere Alarmgrenze für NBP_M
NBP_S	nicht invasiver systolischer Blutdruck
NBP_S_LOW	untere Alarmgrenze für NBP_S
NBP_S_HIGH	obere Alarmgrenze für NBP_S
P2a_D	automatisch bezeichnetes Drucksignal
P2a_M	automatisch bezeichnetes Drucksignal
P2a_S	automatisch bezeichnetes Drucksignal
PA_D	pulmonalarterieller diastolische Druck
PA_M	pulmonalarterieller mittlerer Druck
PA_M_LOW	untere Alarmgrenze für PA_M
PA_M_HIGH	obere Alarmgrenze für PA_M
PA_M_ALMATTR	Alarminformation für PA_M
PA_S	pulmonalarterieller systolischer Druck
Pause	Pausendruck
PEEP	endexspiratorischer Spitzendruck
PIDSTR	auslösende Avriable
PIP	inspiratorischer Spitzendruck
PLS	Puls
PVC_min	Vorzeitige ventrikuläre Kontraktionen pro Minute
PVC_min_HIGH	obere Alarmgrenze für PVC_min
PVC_min_ALMATTR	Alarminformation für PVC_min
RESP	Atemfrequenz
RESP_LOW	untere Alarmgrenze für RESP
RESP_HIGH	obere Alarmgrenze für RESP
RESP_ALMATTR	Alarminformation für RESP
RRc	Atemfrequenz
RRv	Atemfrequenz

Monitoring-Variable	Beschreibung
SpO2	Sauerstoffsättigung
SpO2_LOW	untere Alarmgrenze für SpO2
SpO2_HIGH	obere Alarmgrenze für SpO2
SpO2_ALMATTR	Alarminformation für SpO2
STI STII STIII stavr STaVF STaVL STV	Ausgaben der ST-Analyse (EKG)
Ta	Temperatur
TVe	expiriertes Tidalvolumen
TVi	inspiriertes Tidalvolumen
InspT	Inspirationszeit