Visualisierung kontinuierlicher, multimodaler Schmerz Scors am Beispiel akustischer Signale

Masterarbeit

Franz Anders
HTWK Leipzig

Januar 2017

Abstract

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

2 Grundlagen der medizinischen Schrei-Forschung

2.1 Schmerz Scores

Bei erwachsenen Menschen wird der Schmerzgrad typischerweise durch eine Selbsteinschätzung des Patienten unter der Leitung gezielter Fragen des Arztes vorgenommen. Bei Kindern unter 3 Jahren ist diese Selbsteinschätzung nicht möglich. Schmerz drückt sich in Veränderungen des psychologischen, körperlichen und biochemischen Verhaltens des Säuglings aus. Die für den Arzt am leichtesten feststellbaren Verhaltensänderungen sind von außen wahrnehmbaren Merkmale, wie zum Beispiel ein Verkrampfen des Gesichtsausdrucks, erhöhte Körperbewegungen oder lang anhaltendes Weinen. Um eine weitesgehend objektive Schmerzfeststellung zu ermöglichen, wurden sogenannte *Pain-Scores* entwickelt, die durch ein Punktesystem den insgesamten Schmerzgrad des Babies quantifizieren.[19] Es existieren *eindimensionale* Pain-Scores, die den Schmerz nur Aufgrund der Beobachtung eines Merkmals beurteilen, so wie beispielsweise die reine Beurteilung des Gesichtsausdruckes. *Mehrdimensionale* (auch *multimodale*) Pain-Scores beziehen mehrere Faktoren in das Scoring mit ein.[1]. Tabelle 2.1 zeigt das Scoring-System „Neonatal Infant Pain Scale“(NIPS) als Beispiel für eine multimodale Pain-Score. Der Säugling wird anhand der aufgeführten Kategorien bewertet und alle vergebenen Punkte aufsummiert. Ein insgesamter Wert von > 3 zeigt Schmerz an, ein Wert von > 4 großen Schmerz.[9]

Tabelle 2.1: NIPS-Scoring

NIPS	**0 points**	**1 point**	**2 points**
Facial Expr.	Relaxed	Contracted	-
Cry	Absent	Mumbling	Vigorous
Breathing	Relaxed	Different than basal	-
Arms	Relaxed	flexed/stretched	-
Legs	Relaxed	flexed/stretched	-
Alertness	Sleeping	uncomfortable	-

In den meisten mehrdimensionalen Scoring-Systeme werden die Schreigeräusche mit einbezogen. Tabelle 2.2 zeigt eine Übersicht über eine ausgewählte Menge an multimodalen Pain-Scores. Alle Pain-Scores sind für Kleinkinder bis 3 Jahren gedacht. In der Übersicht wird nicht wiedergegeben, welche weiteren Merkmale jeweils in das Scoring mit einbezogen werden, oder welche Insgesamtpunktzahlen auf welche Schmerzintensität hinweisen. Es soll an dieser Stelle nur verdeutlicht werden, welche unterschiedlichen Ansätze zur Bewertung des Schreiens aus medizinischer Sicht im Zusammenhang mit Pain-Scores existieren. Folgende Beobachtungen lassen sich aus der Übersicht ziehen:

1. Die zu beobachtenden Eigenschaften des Weinens werden mit subjektiv behafteten Werten charakterisiert. Beispielsweise wird im N-PASS-System ist ein Schmerz-Schrei

als „High-pitched or silent-continuous crying“ beschrieben. Es wird nicht fest definiert, was als „crying“ gilt oder welche Tonhöhe als „high-pitched“ ist. Auch die Erstquellen geben keine festen Definitionen.

2. Es gibt verschiedene Ansätze zur Bewertung des Weinens. Bei CRIE ist die Tonhöhe, bei BIIP die Länge und bei COMFORT die Art des Weinens entscheidend.

3. Die Beschreibungen sind kurz und prägnant gehalten, der Arzt hat in keinem der Modelle auf mehr als drei Parameter des Schreiens zu achten. Die Begründung liegt darin, dass bei allen Modellen a.) das Schreien nur eines von mehreren Faktoren ist, und b.) Die Schmerzbestimmung in einem vorgegebenen Zeitrahmen durchführbare sein muss.

System	**P.**	**Description**
FLACC[25]	0	No cry (awake or asleep)
	1	Moans or whimpers; occasional complaint
	2	Crying steadily, screams or sobs, frequent complaints
N-PASS[20]	-2	No cry with painful stimul
	-1	Moans or cries minimally with painful stimuli
	0	Appropiate Crying
	1	Irritable or Crying at Intervals. Consolable
	2	High-pitched or silent-continuous crying. Not consolable
BIIP[7]	0	No Crying
	1	Crying <2 minutes
	2	Crying >2 minutes
	3	Shrill Crying >2 minutes
CRIES[3]	0	If no cry or cry which is not high pitched
	1	If cry high pitched but baby is easily consoled
	2	If cry is high pitched and baby is inconsolable
COVERS[12]	0	No Cry
	1	High-Pitched or visibly crying
	2	Inconsolable or difficult to soothe
PAT[10]	0	No Cry
	1	Cry
DAN[4]	0	Moans Briefly
	1	Intermittent Crying
	2	Long-Lasting Crying, Continuous howl
COMFORT[17]	0	No crying
	1	Sobbing or gasping
	2	Moaning

	3	Crying
	4	Screaming
MBPS[18]	0	Laughing or giggling
	1	Not Crying
	2	Moaning quiet vocalizing gentle or whimpering cry
	3	Full lunged cry or sobbing
	4	Full lunged cry more than baseline cry

Tabelle 2.2: Übersicht über Pain-Scores

2.2 Schmerz-Schrei aus medizinischer Sicht

Die Frage ist: Woher kommen diese unterschiedlichen Bewertungen des Weinens in Tabelle 2.2? Gibt es eine Pain-Score, die aus wissenchaftlicher Sicht „recht hat“? Dieser Fragestellung unterliegen unterliegen zwei grundlegendere Fragen: 1.) Ist es überhaupt möglich, anhand der akustischen Eigenschaften den Grund für den Schrei abzuleiten, also beispielsweise Hunger, Einsamkeit oder Schmerz? Anders formuliert: Gibt es überhaupt so etwas wie einen Schmerz-Schrei? 2.) Ist es möglich, anhand der akustischen Eigenschaften den Schweregrad des Unwohlseins abzuleiten (also beispielsweise den Grad des Schrei-Versursachenden Schmerzes)?

Die Annahme, dass es möglich ist, aus dem Schreien den Grund abzuleiten, wird als „Cry-Types Hypothesis“ bezeichnet. Die berühmtesten Befürworter dieser Hypothese ist eine skandinavische Forschungsgruppe, auch bezeichnet als „Scandinavian Cry-Group“, die diese Idee in dem Buch „Infant Crying: Theoretical and Research Perspectives“[2] publik machte. Die Annahme ist, dass die verschiedenen Ursachen *Hunger, Freude, Schmerz, Geburt und Anderes* klare Unterschiede hinsichtlich ihrer akustischen Merkmale aufweisen, welche an einem Spektogramm ablesbar seien. Entsprechende Beispiele werden in dem Buch gegeben. Nur einige Jahre Später zeigte Müller et al [6] in einem Paper, dass bei leichter Veränderung der Bedingungen der Experimente die Unterscheidung nicht möglich ist. Die Gegenhypothese ist, dass Weinen „nichts als undifferenziertes Rauschen“ sei. 50 Jahre später liegt kein anerkannter Beweis für die eine oder andere Hypothese vor. Es gibt nur starke Hinweise dafür, dass die Plötzlichkeit des Eintretens des Schreigrundes hörbar ist. Ein plötzliches Ereignis, wie ein Nadelstich oder ein lautes Geräuch, führen auch zu einem plötzlich beginnenden Schreien. Ein langsam einretendes Ereignis, wie ein langsam immer stärker werdender physischer Schmerz oder langsam eintretender Hunger führen auch zu einem langsam eintretenden Weinen. Da keine Einigung herrscht, wird empfohlen, den Grund aus dem Kontext abzuleiten.[24]

Die Zweite Frage nach der Ableitung der Stärke des Unwohlseins aus den akustischen Eigenschaften des Geschreis wird in der Fachsprache unter dem Begriff *Cry as a graded Signal* subsumiert. Je „stärker“ das Weinen, desto höher das Unwohlsein (*Level of Distress (LoD)*) des Säuglings. Tatsächlich bemessen wird dabei der von dem Beobachter vermutete Grad des Unwohlsein des Babies, und nicht der tatsächliche Grad, da dieser ohne die Möglichkeit der direkten Befragung des Kindes nie mit absoluter Sicherheit bestimmt werden kann. Dieser vermutete LoD wird entweder durch das subjektive Empfinden der Beobachter oder durch Pain-Scores festgestellt. Ein hohes Level of Distress hat vor allem

eine schnelle Reaktion der Aufsichtspersonen zur Beruhigung des Babies zur Folge, womit dem Geschrei eine Art Alarm-Funktion zukommt. Es gibt starke Hinweise darauf, dass das Level of Distress anhand objektiv messbarer Eigenschaften des Audiosignals bestimmt werden kann. So herrscht beispielsweise weitesgehend Einigung darüber, dass ein „lang“ anhaltendes Geschrei auf einen hohen Level of Distress hinweist. Insofern aus dem Kontext des Schreiens Schmerz als wahrscheinlichste Ursache eingegrenzt werden kann, kann aus einem hohen Level of Distress ein hoher Schmerz abgeleitet werden. [24] und [23]

Es herrscht wiederum keine Einigung darüber, welche akustischen Eigenschaften im Detail ein hohes Level of Distress anzeigen. Carlo V Bellieni et al [4] haben festgestellt, dass bei sehr hohem Schmerz in Bezug auf die DAN-Scala (siehe Tabelle 2.2) die Tonhöhe des Geschreis steigt. Qiaobing Xie et al [23] haben festgestellt, dass häufiges und „verzerrtes“ Schreien (ohne feststellbares Grundfrequenz, da der Ton stimmlos erzeugt wird) auf einen hohen Level of Distress hinweist.[24] Diese Uneinigkeit hat wahrscheinlich zu den verschiedenen Bewertungen in den Pain-Scores geführt. 2.2.

2.3 Phyisio-Akustische Modellierung des Weinens

Das Ziel dieses Kapitels ist die Schaffung eines einheitlichen Vokabulares, auf den sich bezogen wird, um das Schreien eines Babys zu beschreiben. Die hier vorgestellten Begriffe stammen sowohl aus dem Buch „A Physioacoustic Model of the Infant Cry “ H Golub und M Corwin [11] als auch aus dem Paper „Rythmic organization of the Sound of Infant Cry “ von Zeskind et al.[21]

Die Lautäußerung eines Neugeborenen, umgangssprachlich auch als „Weinen“ oder „Schreien“ bezeichnet, lässt sich im allgemeinen beschreiben als das „rythmische Wiederholen eines beim ausatmen erzeugen Geräusches, einer kurzen Pause, einem Einatmungs-Geräusch, einer zweiten Pause, und dem erneuten Beginnen des Ausatmungs-Geräusches.“[26].

Das Vokabular, welches insbesondere von H Golub und M Corwin geschaffen wurde, ist sehr umfassend. An dieser Stelle wird eine Auswahl grundlegender Begrifflichkeiten vorgestellt, die in dieser Arbeit gebraucht werden. Sie werden in Abbildung 2.1 veranschaulicht.

Expiration beschreibt den Klang, der bei einem einzelnen, ununterbrochenem Ausatmen mit Aktivierung der Stimmbänder durch das Baby erzeugt wird. [21]. Der von Golub et al [11] verwendete Begriff **Cry-Unit** wird in dieser Arbeit synonym verwendet. Umgangssprachlich ist handelt es sich um einen einzelnen, ununterbrochenen *Schrei*.

Inspiration beschreibt den Klang, der beim Einatmen durch das Baby erzeugt wird.

Burst beshreibt die Einheit von einer Expiration und der darauf folgenden Inspiration. Das heisst, dass die zeitliche Dauer eines Bursts sowohl das Expiration-Geräusch, das Inspiration-Geräusch als auch die beiden Pausen zwischen diesen Geräuschen umfasst. Praktisch ergibt sich das Problem, dass vor allem bei stärkerem Hintergrundrauschen die Inspiration-Geräusche häufig weder hörbar noch auf dem Spektrogramm erkennbar sind. Daher wird die Zeitdauer eines Bursts oder Cry-Unit vom Beginn einer Expiration bis zum Beginn der darauf folgenden Expiration definiert und somit allein von den Expirations auf die Bursts geschlossen. Implizit wird somit eine Inspiration zwischen zwei Expirations angenommen.

Cry die insgesamte klangliche Antwort zu einem spezifischen Stimulus. Eine Gruppe meh-

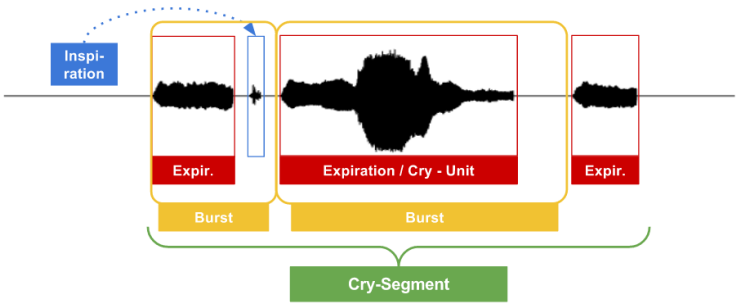

Abbildung 2.1: Veranschaulichung des Grundvokabulars

rerer Cry-Units.[11] In dieser Arbeit wird ein *Cry* als **Cry-Segment** bezeichnet, um Verwechslungen zu vermeiden.

Weiterhin wurden von H Golub und M Corwin [11] Cry-Units in eine der folgenden drei Kategorien eingeführt:

Phonation beschreibt eine Cry-Unit mit einer „vollen Vibration der Stimmbänder“ mit einer Grundfrequenz zwischen 250 und 700 Hz. Entspricht umgangssprachlich einem Weinen mit einem „klaren, hörbaren Ton“.

Hyper-Phonation beschreibt eine Cry-Unit mit einer „falsetto-artigem Vibration der Stimmbänder“ mit einer Grundfrequenz zwischen 1000 und 2000 Hz. Entspricht umgangssprachlich einem Weinen mit einem „sehr hohen, aber klaren, hörbaren Ton“.

Dysphonation beschreibt eine Cry-Unit ohne klar feststellbare Tonhöhe, produziert durch Turbulenzen an den Stimmbändern. Entspricht umgangsprachlichen dem „Brüllen oder Krächzen“.

Eine Cry-Unit gehört dabei mindestens einer dieser Kategorien an, kann aber auch in seinem zeitlichen Verlauf die Kategorie wechseln. H Golub und M Corwin [11] stellen weiterhin eine Reihe an charakteristischen Eigenschaften vor, die in Bezug auf ein Cry-Segment berechnet werden.

Latency-Period beschreibt die Dauer zwischen dem zufügen eines Schmerz-Stimulus und dem beginn des ersten Cry-Bursts des Segmentes

Duration beschreibt die insgesamte Zeitdauer des Cry-Segmentes. Es wird keine genaue Definition gegeben, wodurch Beginn und Ende definiert werden. Das Segment endet dort, wo es „scheint, aufzuhören“.

Maximum-Pitch beschreibt die höchste festgetellte Grunfrequenz des Segmentes.

... und viele weitere, die in [11] nachgelesen werden können, aus Platzgründen an dieser Stelle jedoch nicht vollständig genannt werden.

3 Überwachtest Lernen

Überwachtes Lernen ist ein Wissenschaftsgebiet des *Maschinellen Lernen*. Es existieren verschiedene Definitionen für Maschinelles Lernen. Eine der meist zitierten Definitionen lauteet wie folgt:

> A Computer Program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.[16, S. 2]

Diese weitreichende Definition lässt sich auf verschiedene Wissenschaftsgebiete anwenden. Ein Beispiel ist ein Computer Programm, welches lernt, Dame zu spielen. Angewandt auf die eben genannte Definition lassen folgende Aufgabenbereiche definieren:

- **Task** T: Dame spielen.
- **Performance Measure** P: Prozentsatz gewonnener Spiele gegen Gegner
- **Training Experience** E: Übungsspiele gegen sich selber.

Selbstverständlich könnten P und E in diesem Beispiel beliebig anders gewählt werden. P könnte auch die Freude sein, die der menschliche Gegenspieler beim Spiel erfindet. [16, S. 2 -3]

Die Klasse an Aufgaben, die für diese Arbeit Bedeutung besitzen, ist die des *Überwachten Lernens*. Beim Überwachten exisitert ein *Trainings-Datensatz* mit korrekten Antworten auf die Problem-Fragestellung. Der Algorithmus generalisiert diese Trainings-Beispiele, um auf alle möglichen Datensätze die richtige Lösung zu gewährleisten. [15, S 6]

Ein Beispiel für ein Problem des überwachten Lernens ist die *Erkennung von Handschrift*. Die Aufgabenbereiche werden folgendermaßen identifiziert:

- **Task** T: Erkennung handgeschriebener Worte in Bildern und Zuordnung zu dem tatsächlichen Wort.
- **Performance Measure** P: Prozentsatz korrekt erkannter Wörter
- **Training experience** E: Eine Datenbank mit handgeschriebener Wörter und mit dem tatsächlich geschriebenen Wort. [16, S. 3 - 4]

Dieses Problem gehört zur Unterkategorie der *Klassifzierung*, beschrieben in Kapitel 3.1.

Eine zweite Klasse an Aufgaben des überwachten Lernens, die Bedeutung in dieser Arbeit hat, ist die *Regression*. Ein Beispiel für eine Regressionsaufgabe ist die *Schätzung des Verkaufspreieses einges gebrauchten PKWs*. Folgende Aufgebenbereiche werden identifiziert:

- **Task** T: Schätzung des Marktwertes eines gebrauchten PKWs.
- **Performance Measure** P: Abweichung des geschätzten Wertes zum tatsächlichen Verkaufswert

- **Training experience** E: Eine Datenbank mit gebrauchten PKWs und ihrem tatsächlichen Verkaufswert.

3.1 Klassifizierung

Das Klassifzierungs-Problem wird folgendermaßen modelliert:

Es existieren *Instanzen* x. Jede Instanzen hat eine Reihe an Eigenschaften, bezeichnet als *Features* oder *Attribute* f, wobei jedes Feature einen eigenen Wertebereich, bezeichnet als *Domain* hat. Menge aller möglichen Feautre-Kombinationen wird als *Feature-Raum* X bezeichnet.

$$\begin{aligned} \text{Feature-Raum}: \quad & X = \{\ dom(f_1)\times, \ldots, \times dom(f_n)\ \} \\ \text{Instanz}: \quad & x \in X \end{aligned} \tag{3.1}$$

Außerdem exisistiert eine Menge an *Klassen* $C = \{1, \ldots, k\}$. Die *Klassifizierungs-Funktion*, *Predictor* oder *Classifier* c bestimmt für eine Instanz eine Klasse.

$$\begin{aligned} \text{Classes}: \quad & C = \{1, \ldots, k\} \\ \text{Classifier}: \quad & c : X \mapsto C \end{aligned} \tag{3.2}$$

Es gibt einen Datensatz D mit einer Menge an Instanzen. Für jede der Instanzen ist die zugehörige Klasse bekannt. Ein Paar aus Instanz und Klasse wird als *Example* e bezeichnet. Die einer Instanz x_i zugewiesenen Klasse c_i wird als *Label* beschrieben.

$$\begin{aligned} \text{Datensatz}: \quad & D = \{\langle x_1, c_1\rangle, \ldots, \langle x_1, c_1\rangle\} \\ \text{Example}: \quad & e \in D \end{aligned} \tag{3.3}$$

Die Fehlerfunktion E zählt für einen Datensatz die Menge aller nicht richtig klassifzierten Instanzen

$$\mathrm{E}(D, C) = \operatorname*{count}_{\langle x_i, c_i\rangle \in D} (C(x_i) \neq c_i) \tag{3.4}$$

Das Ziel des Klassfikations-Problems ist es, diejenige Funktion C zu finden, die für einen Test-Datensatz $D_{test} \subseteq D$ die Anzahl falsch klassifzierter Examples minimiert. Nach dem in Kapitel 3 vorgestellten Muster nach T, P und E ergibt sich folgende Aufgabenbeschreibung. [5, S. 8 - 9] [13, S. 14] [15, S. 7 - 10, 18]

- **Task** T: Für einen Test-Datensatz $D_{test} \subseteq D$, finde eine Klassfikations-Funktion c, die die Funktion E minimiert, das heißt: $E(D_{Test}, c) \mapsto min$
- **Performance Measure** P: Die Fehler-Funktion $E(D_{test}, c)$.
- **Training experience** E: Ein Trainings-Datensatz $D_{training} \subseteq D$

Im Zusammenahang mit Klassifikation haben die Klassen C die Eigenschaft, dass die Klassen eine *qualitativen* Charakter, und keinen *quantitativen*. Das heißt, dass die Klassen untereinander keine hierarchische Ordnung besitzen, bei der eine Klasse „besser ist als die andere“. Außerdem handelt es sich um eine diskrete Menge, und keinen kontinuierlichen

Zahlenbereich. [8, S. 127]. Ein besonderer Fall der Klassifikation ist ein sogenannter *binärer Klassfikator*, bei dem es nur zwei Klassen gibt: $C = \{0, 1\}$ (oder $C = \{yes, no\}$). Die Domains der Features können ebenfalls qualitativer natur sein, das heißt einen Wert in einem diskreten, ungeordneten Raum annehmen, oder quantitativer Natur, das heißt, einen Wert in einem kontinuierlichen, geordneten Zahlenraum annehmen. [16, S. 54]

Eine andere Art und Weise, die Aufgabenstellung der Klassifikation zu betrachten, ist die *Generalisation zur Prongnose*. Das heißt, dass die Ableitung der Klassen aus den Instanzen des Datensatzes verallgemeinert wird, um in Zukunft für neue, noch nicht bekannte Instanzen die Klasse vorhersagen (prognostizieren) zu können. [15, S. 6 - 7]

Tabelle 3.1 gibt einen Beispiel-Datensatz für eine Klassifkationsproblem. In diesem Beispiel geht es darum, ob abhängig von der Tageszeit und der Temperatur ein Federball-Match Spaß gemacht hat, oder nicht. Das Problem wird folgendermaßen modelliert:

- Es gibt zwei Features: $f_1 = Temperatur$, mit $dom(f_1) = R$, ein quantitives Feature. $f_2 = Tageszeit$, mit $dom(f_2) = morgens, mittags, abends$, ein qualitatives Feature. Der Feature-Raum ist $X = dom(f_1) \times dom(f_2)$.
- Es gibt zwei Klassen, $C = Ja, Nein$.
- Der Datensatz hat fünf Instanzen, $D = \{x_1, \ldots, x_5\}$. Für jede Instanz ist das Label $c_1, \ldots, c_5$ bekannt, welches besagt, ob das Federball spielen Spaß gemacht hat, oder nicht.

Tabelle 3.1: Beispieldatensatz D für eine Klassifikation

x_i	Temperatur	Tageszeit	c_i = Spaß?
x_1	20	morgens	Ja
x_2	15	abends	Ja
x_3	8	mittags	Nein
x_4	23	mittags	Ja
x_5	10	morgens	Nein

Das Ziel ist, in Zukunft abschätzen zu können, allein durch die Kenntniss der Temperatur und Tageszeit abschätzen zu können, ob das Federball-Spielen Spaß machen wird, damit man von vorneherein keine Matches beginnt, die keine Aussicht auf Spaß haben. An dieser Stelle wählt man einen Algorithmus, der einen Classifier baut, der dieses Problem löst.

Es gibt eine Reihe an Algorithmen, die Classifikatoren nach unterschiedlichen Methoden erstellen. Beispiele sind für *k-NN*, *Support-Vector-Maschienen* oder *künstliche Neuronale Netze*. Eine Klasse an Klassifikations-Algorithmen, die in dieser Arbeit Anwendung finden, sind die sogenannten *Entscheidungsbäume*

3.2 Entscheidungsbäume

Es gibt drei Algorithmen zur Erzeugung von Entscheidungsbäumen, die weitreichende Einsatz finden: *ID3*, *C.45* und *CART*, wobei die letzteren Erweiterungen der grundlegenden Idee des *ID3*-Algorithmus darstellen. Daher wird an dieser Stelle zuerst der *ID3*-Algorithmus vorgestellt.

Es wird zunächst davon ausgegangen, dass alle Features diskret und nicht kontinuierlich sind. Tabelle 3.2 gibt einen Beispieldatensatz, an dessen Beispiel ein Classificator mit Hilfe des ID3 erzeugt wird. Es geht ähnlich dem Beispiel aus Tabelle 3.1 um die Frage, ob Federball-Spielen abhängig von Temperatur und Tageszeit Spaß macht, nur sind in diesem Fall alle Features diskret.

Tabelle 3.2: Beispieldatensatz D für die Kassfikation mit ID3

x_i	Temperatur	Tageszeit	c_i = Spaß?
x_1	warm	Tag	Ja
x_2	kalt	Tag	Ja
x_3	normal	Nacht	Nein
x_4	kalt	Nacht	Nein
x_5	normal	Tag	Ja
x_6	warm	Nacht	Ja

Abbildung 3.1 zeigt einen Klassifikator, den der ID-3 Algorithmus für diesen Datensatz baut. Es handelt sich um einen Entscheidungsbaum. In Jedem Knoten steht ein Feature, welches einen Ast für jeden möglichen Wert dieses Features bildet. In den Blättern stehen die Klassen.[15, S. 134]

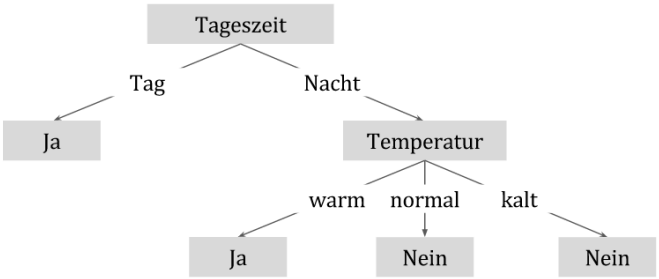

Abbildung 3.1: Entscheidungsbaum, der durch den ID3-Algorithmus für den Datensatz aus Beispiel 3.2 erzeugt wurde.

Der Entscheidungsbaum lässt sich in eine Reihe von `if ... then ...` -Regeln transformieren. Jeder Weg von der Wurzel bis zu einem Blatt ergibt eine Entscheidungsregel, bei der Feauture-Werte der betretenen Kanten konjunktiv Verknüpft werden und die Klasse implizieren. Die Entscheidungsregeln für den Baum aus Abbildung 3.1 sind: [15, S. 134]

- `if` *Tageszeit* = *Tag* `then` *Spaß* = *Ja*
- `if` *Tageszeit* = *Nacht* `and` *Temperatur* = *warm* `then` *Spaß* = *Ja*
- `if` *Tageszeit* = *Nacht* `and` *Temperatur* = *normal* `then` *Spaß* = *Nein*
- `if` *Tageszeit* = *Nacht* `and` *Temperatur* = *kalt* `then` *Spaß* = *Nein*

Der Klassifikator, das heißt der Entscheidungsbaum, wird beim ID3 Algorithmus nach folgenden Muster erstellt: Der Baum wird Top-Down erzeugt, dass heisst beginnend bei der Wurzel bis zu den Blättern. Da in jedem Knoten genau ein Feature aufgespalten wird, wird an der Wurzel die Frage gestellt „*Welches Feature sollte zuerst getestet werden?*“. Um diese Frage zu beantworten, wird jedes Feature einem statistischen Test unterzogen und festzustellen, wie „gut“ es zur Klassifikation der Trainings-Daten beiträgt. Das „beste“

Attribut wird ausgewählt und als Wurzel festgelegt. Nun wird ein Kind für jeden möglichen Wert des Features gebildet. Der Datensatz des Elternknotens wird in disjunkte Teilmengen aufteilt, wobei jedes Kind die Untermenge erhält, die den jeweiligen Feature-Wert bestitzt. Daraufhin beginnt für jedes Kind der Prozess des Auswählen des „besten“ Attributes von vorn. Ein Kind wird dann zu einem Blatt, wenn seine Teilmenge an Daten nur noch aus Instanzen einer Klasse besteht und somit kein weiteres Aufteilen notwendig ist.[16, S. 55]

Das Wort „gut“ wurd ein dieser Beschreibung in Anführungsstrichen geschrieben, da es subjektiv ist und quantifiziert werden muss. Zur Quantifizierung der Information wird die Entropie nach Formel 3.5 als Hilfsmittel definiert. p_i ist die Wahrscheinlichkeit, dass in einem Datensatz D eine Instanz mit der Klasse $i \in C$ angetroffen wird.

$$H(p) = -\sum_{i \in C} p_i \cdot \log_2 p_i \tag{3.5}$$

Die Entropie quantifiziert die *Unreinheit des Datensatzes*. Angenommen, ein Datensatz hat zwei Klassen, $C = \{+, -\}$. Existiert der gesamte Datensatz nur aus einer der beiden Klasse, ist die Entropie $-p_+ \log_2 p_+ - p_- \log_2 p_- = -1log_21 - 0log_20 = 0$. Das heißt, dass die *Unreinheit des Datensatzes* 0 beträgt. Ist die *Unreinheit des Datensatzes* hingegen maximal, das heißt es liegen exakt 50% positive und 50% negative Samples vor, ist die Entropie $-p_+ \log_2 p_+ - p_- \log_2 p_- = -0.5log_{0.5}1 - 0.5log_20.5 = 1$. [15, S. 135]

Es ist das Attribut in einem Knoten zu wählen, welches den höchsten *Informatoinsgewinn* gewährleistet, das heißt, zu einer bestmöglichen *Reinheit* bei der alleinigen Unterteilung des Datensatzes auf Basis dieses Attributs führt. Der informationsgewinn eines Features f für den Datensatz D wird nach Formel 3.6 definiert. v sind alle möglichen Werte dieses Features. $|D|$ beschreibt die Anzahl an Instanzen des Datensatzes. D_v ist die Untermenge an Instanzen, die für das Feature f den Wert v besitzen.[15, S. 136 - 137]

$$\text{Gain}(D, f) = H(D) - \sum_{v \in dom(f)} \frac{|D_v|}{|D|} H(D_v) \tag{3.6}$$

Für das Beispiel aus Tabelle 3.1 ergibt sich für den ersten Test folgende Berechnung des Informationsgewinnes der beiden Features *Temperatur* und *Tageszeit*. Da die Tageszeit den höheren Informationsgewinn gewährleistet, wird dieses Features in der Wurzel gewählt.

$$H(D) = -p_+ \log_2 p_+ - p_- \log_2 p_- = -\frac{4}{6}\log_2(\frac{4}{6}) - \frac{2}{6}\log_2(\frac{2}{6}) = 0.91 \tag{3.7}$$

$$Gain(D, Tageszeit) = 0.91 - \Big(\overbrace{\frac{3}{6} \cdot (-\frac{3}{3}\log_2\frac{3}{3} - -\frac{0}{3}\log_2\frac{0}{3})}^{Tag} \underbrace{\frac{3}{6} \cdot (-\frac{1}{3}\log_2\frac{1}{3} - -\frac{2}{3}\log_2\frac{2}{3})}_{Nacht} \Big) = 0.86 \tag{3.8}$$

$$Gain(D, Temperatur) = 0.91 - \Big(\overbrace{\frac{2}{6} \cdot (-\frac{2}{2}\log_2\frac{2}{2} - -\frac{0}{2}\log_2\frac{0}{2})}^{warm} \overbrace{\frac{2}{6} \cdot (-\frac{1}{2}\log_2\frac{1}{2} - -\frac{1}{2}\log_2\frac{1}{2}}^{normal} \overbrace{\frac{2}{6} \cdot (-\frac{1}{2}\log_2\frac{1}{2} - -\frac{1}{2}\log_2\frac{1}{2})}^{kalt} \Big) = 0.66 \tag{3.9}$$

Algorithmus1 zeigt den Ablauf des ID-3 in Psuedocode.

Algorithm 1 ID3-Algorithmus in Pseudocode

```
1: tree = {}
2: function ID3(D, X, C, f_current)
3:     if ∀e ∈ D : ∃k ∈ C : e.c = k then
4:         return k
5:     else
6:         if isEmpty(X) then
7:             return most common Label in D
8:         else
9:             f_best = max_{f∈X} Gain(D, f)
10:            parentKnot ← f_best
11:            X ← X/f_dom
12:            for v ∈ f_best do
13:                D_f ← ∀e ∈ D : e.f_best = v
14:                CalcSquare(a)
15:                tree = tree ∪ ⟨f_current, f_⟩
16:            end for
17:        end if
18:    end if
19: end function
```

Der ID3-Algorithmus hat folgende **Vorteile**:

- Der Klassifizierer versucht, möglichst kurze Entscheidungsbäume zu bauen, indem Features mit hohem Informationsgewinn bevorzugt werden. Dies ist eine Umsetzung von *Ocam's Razor*: „Bevorzuge die kürzeste Hypothese“
- Der Klassifikator ist für den Menschen verständlich (im Gegensatz zu zum Beispiel Neuronalen Netzen)[16, S. 65]

Der ID3-Algorithmus hat folgende **Nachteile**

3.3 Gütemaße binärer Klassifikatoren

Bei einer binären Klassifikation gibt es vier mögliche Klassifikationsergebnisse. Die beiden Klassen werden im allgemeinen als *Positive* und *Negative* beschrieben, was in diesem

speziellen Fall *Stimme* und *Stille* entspricht. Eine Klassifikation, bei der ein tatsächliches Positive richtig als Positive vorhergesagt wird, spricht man von einem *True Positive* [TP]. Wird hingegen ein tatsächliches Positive fälschlicherweise als Negative vorhergesagt, spricht man von einem *False-Negative* [FN]. Das System wird entsprechend für die Klassifikation tatsächlicher Negatives angewandt und ergibt. *True-Negatives* [TN] und *False-Positives* [FP]. Die *Confusion Matrix* in Abbildung 3.2 gibt eine Übersicht über die vier möglichen Klassifikations-Ergebnisse.

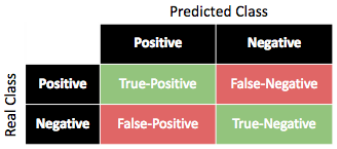

Abbildung 3.2: Confusion-Matrix

Die insgesamte Güte einer Klassifikation wird durch die *Accuracy* nach Formel 3.10 bestimmt. Eine Accuracy von 100% bedeutet, dass *alle* Instanzen richtig klassifiziert werden, eine Accuracy von 50% bedeutet, dass die Hälfte aller instanzen richtig klassifiziert werden, was der Güte einer rein zufälligen Wahl entspricht.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP} \tag{3.10}$$

Die Accuracy beziffert die insgesamte Performance des Klassifikators, gibt jedoch keinen Aufschluss darüber, ob der Klassifikator eher eine Tendenz zur falschen Klassifizierung von Positives oder Negatives hat. Bei einer Datenbank mit der selben Anzahl an Positives und Negatives kann eine Accuracy von 50% beispielsweise dadurch entstehen, dass *alle* Instanzen als Positives markiert werden, also sowohl die Positives richtigerweise als Positives, aber die Negatives fälschlicherweise ebenfalls als Positives. Im Umgedrehten Fall ergibt die Klassifizierung aller Instanzen als Negatives ebenfalls eine Accuracy von 50%. In einem dritten Fall irrt sich die Klassifikator gleich oft bei der Einordnung der Negatives und Positives.Die Maße *Sensitivity* und *Specificity* geben Aufschluss über die Güte der Klassifikation hinsichtlich der Positives und Negatives. Die *Sensitivity*, auch bezeichnet als *True-Positive-Rate*, bemisst den Anteil tatsächlicher Positives, die auch als solche erkannt wurden, nach Formel 3.11. Eine Sensitivity von 100% bedeutet, dass alle Positives durch den Klassifikator erkannt wurden. Die Erkennungsrate der Negatives hat keinen Einfluss auf die Sensitivity. Eine hohe Sensitivity lässt sich somit „einfach“ erzielen, in dem man *alle* Instanzen immer als Positives klassifiziert. Die Specificity nach Formel 3.12 bestimmt analog zur Sensitivity den Anteil der korrekt als Negatives bestimmten Instanzen. Ein Klassifikator, der alle Instanzen als Positives markiert, hat zwar eine Sensitivity von 100%, aber eine Specificity von 0%. Ergeben zwei verschiedene Klassifikationsmodelle sehr ähnliche Accuracies, hilft die Bestimmung der Sensitivity und Specificity bei der Auswahl des für den Anwendugnsfall Adequäteren Klassifikators. So ist beispielsweise bei der Bestimmung von schweren Krankheiten eventuell ein Klassifikator mit höherer Sensitivity wünschbar, um die Wahrscheinlichkeit zu minimieren, dass die entsprechende Krankheit nicht erkannt wird. [14] [22]

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3.11}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3.12}$$

3.4 Klassifikationsalgorithmen

3.4.1 CARD

3.4.2 Andere

4 Zusammenfassung

Literaturverzeichnis

[1] K J S Anand. *Pain in Neonates and Infants*. Elsevier, 2007.

[2] Zachariah Boukydis Barry Lester. *Infant Crying: Theoretical and Research Perspectives*. Springer, 1985.

[3] Judy Bildner. *CRIES Instrument Assessment Tool of Pain in Neonates*. City of Hope Pain, 1997. Online unter `http://prc.coh.org/pdf/CRIES.pdf`.

[4] R Sisto & Giuseppe Buonocore Carlo Bellieni, Franco Bagnoli. Cry features reflect pain intensity in term newborns: An alarm threshold. *Pediatric Research*, 5:142–146, 1. Online unter `https://www.researchgate.net/publication/297827342_Cry_features_reflect_pain_intensity_in_term_newborns_An_alarm_threshold`.

[5] Alin Dobra. Introduction to classification and regression, 2005. Online erhältlich unter: `https://www.cise.ufl.edu/~adobra/datamining/classif-intro.pdf`.

[6] H. Hollien & T Murry E Müller. Perceptual responses to infant crying: identification of cry types. *Journal of Child Language*, 1(1):89–95, 1974. Online unter `https://www.cambridge.org/core/journals/journal-of-child-language/article/perceptual-responses-to-infant-crying-identification-of-cry-types/4F0F8088116FCE381851D8D560697A5F`.

[7] Jan Hamers & Peter Gessler Eva Cignac, Romano Mueller. Pain assessment in the neonate using the Bernese Pain Scale for Neonates. *Early Human Development*, 78(2):125–131, 2004. Online unter `http://www.sciencedirect.com/science/article/pii/S0378378204000337`.

[8] Trevor Hastie Gareth James, Daniela Witten and Robert Tibshirani. *An Introduction to Statistical Learning*. Springer, 2013.

[9] Health Facts For You. *Using Pediatric Pain Scales Neonatal Infant Pain Scale (NIPS)*, 2014. Online unter `https://www.uwhealth.org/healthfacts/parenting/7711.pdf`.

[10] Hodgkinson. Neonatal Pain Assessment Tool , 2012. Online unter `http://www.rch.org.au/uploadedFiles/Main/Content/rchcpg/hospital_clinical_guideline_index/PAT%20score%20update.pdf`.

[11] Michael J Corwin Howard L Golub. A Physioacoustic Model of the Infant Cry. In *Infant Crying - Theoretical and Research Perspectives*, chapter 3, pages 59 – 82. Plenung, 1985.

[12] Donna Geiss Laura Wozniak & Charles Hall Ivan Hand, Lawrence Noble. COVERS Neonatal Pain Scale: Development and Validation. *International Journal of Pediatrics*, 2010, 2010. Online unter `https://www.hindawi.com/journals/ijpedi/2010/496719/`.

[13] Wei-Yin Loh. A Comparative Performance Study of Several Pitch Detection Algorithms. *WIREs Data Mining Knowl Discovery*, 1:14–23, 2011.

[14] Tze-Wey Loong. Understanding sensitivity and specificity with the right side of the brain. *BMJ*, 327(7417), 2003. Online unter https://www.ncbi.nlm.nih.gov/pmc/articles/PMC200804/.

[15] Stephen Marsland. *Machine Learning - An Algorithmic Perspective*. Chapman & Hall / CRC, 2009.

[16] Tom M Mitchell. *Machine Learning*. WCB McGraw-Hill, 1997.

[17] Hans M Koot Dick Tibboel Jan Passchier & Hugo Duivenvoorden Monique van Dijk, Josien de Boer. The reliability and validity of the COMFORT scale as a postoperative pain instrument in 0 to 3-year-old infants. *Pain*, 84(2):367—-377, 2000. Online unter http://www.sciencedirect.com/science/article/pii/S0304395999002390.

[18] Taddio Nulman. A revised measure of acute pain in infants. *J Pain Symptom Manage*, 10:456–463, 1995. Online unter http://geriatricphysio.yolasite.com/resources/Modified%20Behavioral%20Pain%20Scale%20(MBPS)%20in%20infants.pdf.

[19] J L Mathew P J Mathew. Assessment and management of pain in infants. *Postgrad Med J*, 79:438–443, 2003. Online unter http://pmj.bmj.com/content/79/934/438.full.

[20] Steven Creech & Marc Weiss. Patricia Hummel, Mary Puchalski. N-PASS: Neonatal Pain, Agitation and Sedation Scale – Reliability and Validity. *Pediatrics/Neonatology*, 2(6), 2004. Online unter http://www.anestesiarianimazione.com/2004/06c.asp.

[21] Susan Parker-Price & Ronald Barr Philip Zeskind. Rythmic organization of the Sound of Infant Cry. *Dev Psychobiol*, 26(6):321–333, 1993. Online unter https://www.ncbi.nlm.nih.gov/pubmed/8119482.

[22] David M W Powers. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011. Online unter https://dl.dropboxusercontent.com/u/27743223/201101-Evaluation_JMLT_Postprint-Colour.pdf.

[23] R Ward & C Laszlo Qiaobing Xie. Automatic Assessment of Infants' Levels-of-Distress from the Cry Signals. *IEEE Transanctions on Speech and Audio Processing*, 4(4):253–265, 1996. Online unter http://ieeexplore.ieee.org/document/506929/.

[24] Brian Hopkins & James Green Ronald Barr. *Crying as a Sign, a Symptom, and a Signal*. Mac Keith Press, 2000.

[25] J R Shayevitz & Shobha Malviya Sandra Merkel, Terri Voepel-Lewis. The FLACC: A Behavioral Scale for Scoring Postoperative Pain in Young Children. *Pediatric Nursing*, 23(3):293–7, 1996. Online unter https://www.researchgate.net/publication/13998379_The_FLACC_A_Behavioral_Scale_for_Scoring_Postoperative_Pain_in_Young_Children.

[26] P H Wolff. The role of biological rhythms in early psychological development. *Bulletin of the Menninger Clinic*, 31:197–218, 1967.

Appendices

Tabelle .1: Accuracy-Werte der Grenzwertfindung mit REPTree

$S_{Training}$	3 dB				50 dB				50+3 dB			
A_{Test}	3 dB	50 dB	7 dB*	Mean	3 dB	50 dB	7 dB*	Mean	3 dB	50 dB	7 dB*	Mean
Zeit	77.81%	79.02%	86.04%	80,96%	49.33%	94.70%	48.66%	64,23%	77.54%	92.47%	84.38%	84,80%
Freq	82.05%	89.28%	82.71%	84,68%	70.52%	94.37%	55.06%	73,31%	81.75%	91.22%	74.90%	82,62%
Ceps	88.98%	94.72%	92.96%	**92,22%**	86.83%	94.68%	92.83%	**91,45%**	88.98%	94.72%	92.96%	**92,22%**
Corr	80.45%	73.47%	84.89%	79,60%	73.07%	87.14%	77.98%	79,39%	77.90%	84.88%	82.84%	81,87%
Zeit+Freq	82.05%	89.28%	82.71%	84,68%	70.52%	94.37%	55.06%	73,31%	81.75%	91.22%	74.90%	82,62%
Zeit+Ceps	88.98%	94.72%	92.96%	**92,22%**	86.83%	94.68%	92.83%	**91,45%**	88.98%	94.72%	92.96%	**92,22%**
Zeit+Corr	80.45%	73.47%	84.89%	79,60%	49.33%	94.70%	48.66%	64,23%	80.32%	92.35%	88.22%	86,96%
Freq+Ceps	88.98%	94.72%	92.96%	**92,22%**	70.65%	94.75%	55.06%	73,49%	88.98%	94.72%	92.96%	**92,22%**
Freq+Corr	82.05%	89.28%	82.71%	84,68%	70.52%	95.60%	95.60%	87,24%	81.75%	94.42%	74.90%	83,69%

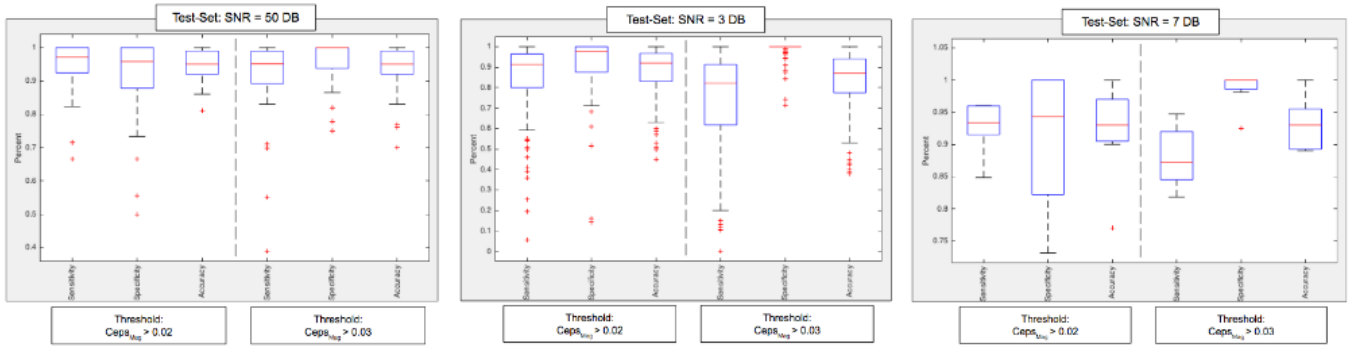

Abbildung .1: Boxplot-Auswertung über Sensitivity, Specificity und Accuracy der beiden VAD-Modelle