Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Informatik

Analyse von Visitentexten mittels maschinellem Lernen zur Optimierung der Leitlinienadhärenz auf einer Intensivstation

Bachelorarbeit

zur Erlangung des akademischen Grades
Bachelor of Science (B. Sc.)

eingereicht von: Martin Hoffmann
geboren am: 06.09.1998
geboren in: Berlin

Gutachter: Prof. Dr. med. Dr. rer. nat. Felix Balzer
Dr. med. Fridtjof Schiefenhövel

eingereicht am: ..... verteidigt am: .....

Abstract

Hier kommt ein toller Abstract

Inhaltsverzeichnis

1 Einführung

1.1 Datenerfassung auf Intensivstationen

blabla.

1.1.1 medizinische Scores

Was sind medizinische Scores? Es handelt sich stets um **eindimensionale metrische Skalen**!

1.1.2 Ziel der Arbeit

Das Ziel der vorliegenden Arbeit ist es, mit Hilfe von maschinellem Lernen ein statistisches Modell zu entwickeln, um anhand von Freitexten medizinische Scores möglichst akkurat vorherzusagen.

1.2 Maschinelles Lernen

Der Begriff Maschinelles Lernen bezeichnet einen modernen Ansatz in der Arbeit an künstlicher Intelligenz. (cite).

hat sich im Zeitalter von Big Data und leistungsstarken Rechnern zu einem der Hauptforschungspunkte der Informatik entwickelt (cite).

(et al) definiert den Vorgang maschinellen Lernens folgendermaßen:

> 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.

Im Allgemeinen werden also Algorithmen, die aus großen Datensätzen „lernen“ und damit Vorhersagen über unbekannte Daten machen, als maschinelles Lernen bezeichnet.

1.2.1 Supervised Learning

Neben vielen weiteren Methoden stellt Supervised Learning einen zentralen Ansatz im Gebiet des maschinelles Lernen dar.

1.2.2 Regression vs Klassifizierung

Probleme aus dem Bereich des supervised machine learnings lassen sich im Allgemeinen in eine von zwei Kategorien einordnen:

Klassifizierung bezeichnet den Prozess, bei dem ein Datensatz einer oder mehreren Klassen aus einer endlichen Liste möglicher Klassen zugeordnet wird.

Dieser Ansatz findet beispielsweise bei der automatischen Kategorisierung von E-Mails (Spam oder nicht Spam) oder bei der Erkennung von handschriftlichen Texten (welches Symbol aus einem gegebenen Alphabet ist dargestellt?) Anwendung (CITE). Da es sich bei den betrachteten medizinischen Scores um diskrete, ganzzahlige Werte aus einem endlichen Wertebereich handelt, liegt auch hier die Anwendung eines Klassifizierungs-Verfahrens nahe.

Würde man aber die verschiedenen möglichen Werte eines Scores als separate und voneinander unabhängige Klassen betrachten, so ginge eine wichtige Information über deren Anordnung verloren. Im mathematischen Sinne stellen alle hier betrachteten medizinische Scores Totalordnungen dar. Sie erfüllen also die Anforderungen der Reflexivität, Antisymmetrie, Transitivität und Totalität. Bezeichne M die Menge aller möglichen Werte eines beliebigen medizinischen Scores. Folglich gilt für alle $a, b, c \in M$:

$$
\begin{aligned}
a \leq a & \qquad \text{(Reflexivität)} \\
a \leq b \wedge b \leq a \;\Rightarrow\; a = b & \qquad \text{(Antisymmetrie)} \\
a \leq b \wedge b \leq c \;\Rightarrow\; a \leq c & \qquad \text{(Transitivität)} \\
a \leq b \vee b \leq a & \qquad \text{(Totalität)}
\end{aligned}
$$

Damit lassen sich die verschiedenen Scores vergleichen und in ein Verhältnis setzen.

So ist ein RASS-Wert[1] von -4 (Tief sediert) beispielsweise deutlich näher an -3 (mäßig sediert) als an +1 (unruhig). Bei gängigen Verfahren zur Klassifizierung ginge diese Information verloren, da bei Kenngrößen zur Bewertung solcher Modelle nur betrachtet werden kann, ob ein gegebener Eingabetext der richtigen Kategorie (dem richtigen Score) zugeordnet wurde oder nicht. (Da viele der Scores bei der Vergabe zumindest teilweise auch auf dem persönlichen Ermessen des behandelnden Arztes/der behandelnden Ärztin beruht wäre selbst für menschliche Experten eine genaue Zuordnung eines Textes zu einer Punktzahl problematisch.)

Bei der vorliegenden Arbeit habe ich mich demnach dafür entschieden, die Vergabe von Scores anhand von Eingabetexten als klassisches Regressionsproblem zu betrachten, und die Ausgaben der Modelle im Zweifelsfall auf den nächstmöglichen ganzzahligen Wert zu runden. Dieser Ansatz fand auch bei früheren Arbeiten, die sich mit ähnlichen Fragestellungen befassten, Anwendung (CITE). Die Abweichung des vorhergesagten Werts eines Modells von dem nächst möglichen Wert stellt somit sogar einen rudimentären Ansatz zur Bewertung der Konfidenz bei der Vergabe einzelner Werte dar.

[1] Richmond Agitation-Sedation Scale

2 Übersicht über die Daten

2.1 Datenerfassung an der Charité

information über copra (fridtjof ausquetschen). bla bla bla

Die daten wurden vor dem Export pseudonymisiert. Nach dem Export lagen die Daten in Form mehrerer Datendateien vor. `patienten.csv` enthält Metainformationen über die betrachteten Patienten. Separat gibt `delir.csv` für jeden Patienten an, ob für diesen während seines Aufenthalts ein Delir diagnostiziert wurde.

2.2 Übersicht über die vorliegenden Daten

Der mir für diese Arbeit zur Verfügung stehende Datensatz enthält Informationen über insgesamt 1357 Patienten-Aufenthalte auf den Intensivstationen der Charité Berlin, die allesamt im Zeitraum von Juni 2019 bis Juni 2020 stattfanden. Jeder Aufenthalt wird über eine numerische, inkrementell aufsteigende Nummer[1] eindeutig identifiziert. Hierbei gilt es zu beachten, dass ein Patient bzw. eine Patientin, der/die im Laufe seiner/ihrer Behandlung mehrere Male auf eine Intensivstation verlegt wird, für jeden separaten Aufenthalt eine neue Identifikationsnummer erhält. Für jeden Patient/für jede Patientin liegen Informationen über das Geschlecht, BMI[2], das Alter zum Zeitpunkt der Aufnahme und ob während des Aufenthalts ein Delir[3] diagnostiziert wurde, vor. Weiterhin ist die Dauer des Aufenthalts auf der Intensivstation vermerkt, sowie ob der Patient/die Patientin während seines/ihres Aufenthalts verstorben ist. Die Mortalität während des Aufenthalts lag bei etwa 17% ($n = 225$). Die beobachteten Aufenthalte verliefen über Zeiträume von wenigen Stunden bis zu mehreren Monaten. Die mittlere

[1] in den Datensätzen als n_ID bezeichnet
[2] Body-Mass-Index
[3] F05.* gemäß ICD-10

Aufenthaltsdauer während des Beobachtungszeitraums liegt bei etwa 7,2 Tagen, der Median beträgt 3,2 Tage. Fast die Hälfte aller Aufenthalte endete also nach weniger als drei Tagen. Nur etwas mehr als ein Viertel der Patienten ($n = 375$) wurden für eine Woche oder länger auf der Intensivstation behandelt.

Für jeden Patient/jede Patientin werden während der Dauer seines/ihres Aufenthalts medizinische Scores erfasst sowie Visitentexte geschrieben. Die Visitentexte werden digital erfasst und liegen im Unicode-Zeichensatz vor, folgen allerdings im Allgemeinen keinem einheitlichen Muster. Es handelt sich also um Freitexte, und es liegt im Ermessen der behandelnden Ärzte bzw. Pflegekräfte, einen aussagekräftigen Text zu formulieren. Ebenso können sich Faktoren wie Zeitdruck oder Stress auf Umfang und Genauigkeit der eingetragenen Texte auswirken(CITE).

Bei den erfassten Scores handelt es sich um medizinische Bewertungssysteme, bei denen die Verfassung des betrachteten Patienten anhand klar definierter Regeln anhand einer Punktzahl bewertet wird(CITE). Häufig beziehen sich die Scores dabei nur auf einen kleinen Teil der Gesamtverfassung des Patienten, beispielsweise auf den Grad der Sedierung oder auf Schmerzempfinden, sofern dieses nicht vom Patienten selber mitgeteilt werden kann. Tabelle 2.1 gibt einen Übersicht über alle Scores und Freitexte, die in dem gegebenen Datensatz erfasst wurden. Bei der VarID handelt es sich um eine Zahl, mit der jede Art von auf der Intensivstation erfasstem Wert bzw. eingetragenem Text repräsentiert und eindeutig identifiziert wird.[4]

Es folgt eine detailliertere Beschreibung derjenigen Werte, die für den Inhalt dieser Arbeit besonders hohe Relevanz haben:

Glasgow Coma Scale bla bla lba

2.2.1 Exemplarische Vorstellung eines Patienten

Aufenthalt des Patienten hier detailliert beschreiben, und seinen Scatterplot einfügen (aber noch ohne Pfeile).

[4]Fridtjof fragen ob das stimmt

VarID	Name	Wertebereich
20512769	Glasgow Coma Scale (GCS)	$3 \leq v \leq ?$
20512801	Behavior Pain Scale (BPS)	$3 \leq v \leq 12$
20512802	Delirium Detection Score (DDS)	$3 \leq v \leq 35$*
22085815	Visite_ZNS	Freitext
22085820	Visite_Oberarzt	Freitext
22085836	Visite_Pflege	Freitext
22085897	Ramsay Sedation Scale	$1 \leq v \leq 6$
22085911	NRS/VAS (Visual Analogue Scale)	$0 \leq v \leq 10$
22086067	Vigilanz	Freitext*
22086158	Richmond Agitation Sedation Scale (RASS)	$-5 \leq v \leq 4$
22086169	CAM-ICU	$v \in \{\text{neg.}, \text{pos.}, \text{unmögl.}\}$
22086170	BPS-Bewertung	Freitext*
22086172	NRS/VAS Bedingungen	Freitext*

Tabelle 2.1: Übersicht aller erfassten Scores und Freitexte

2.2.2 Generierung von Schlüssel-Wert-Paaren

Eine besondere Herausforderung stellte dar, dass die verschiedenen Werte und Texte zu unterschiedlichen Zeiten und unabhängig voneinadner eingetragen werden. Die Informationen über Zeitpunkt und Art der Eintragung sowie der eigentliche Wert liegen jeweils als Tripel in Form mehrerer Log-Dateien vor. Bei den Visitentexten entspricht der eingetragene Text dem Wert der Eintragung.

Um ein Machine-Learning Modell aus dem Bereich des supervised learnings zu trainieren ist eine hohe Anzahl von Trainingspaaren notwendig. Jedes Wertepaar enthält einen Text sowie einen medizinischen Score, der den in dem Text angegebenen Informationen über die Verfasssung des Patient/die Patientin möglichst genau entspricht. Zusammen bilden diese Paare die Grundlage der Modelle, selbstständig noch nicht gesehene Texte bewerten zu können. Aufgrund der zeitlichen Unterschiede zwischen den Eintragungen von Texten und Scores erwies sich allerdings ebenjene Zuordnung zueinander als nicht trivial.

2.2.3 Genauigkeit der erfassten Daten

Hier beschreiben, dass viele Texte nicht wirklich den Werten entsprechen. Das mache es schwerer, Modelle zu trainieren, und muss berücksichtigt werden. Gründe:

1. Zeitdruck, Stress bei Ärzten, kann vorkommen dass sie einfach Wert vom letzten mal kopieren
2. Werte können sich innerhalb von Minuten ändern (z.B. RASS von 0 auf -5)
3. Weitere Gründe?

Diese Grüunde sind aber nicht weiter Beobachtungsgegenstand der vorleigenden Arbeit. Dennoch muss das bei der Konzeption, Entwicklung und Bewertung beachtet werden, weil sie ohne weitere Maßnahmen möglicherweise eine obere Schranke für die Performance der Modelle darstellen.

3 Vorgehensweise

4 Fazit

4.1 Auswertung der Ergebnisse

...

4.2 Bewertung der Leitlinienadhärenz auf den Intensivstationen der Charité

...

4.3 Blick in die Zukunft

...

Literaturverzeichnis

Selbständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und noch nicht für andere Prüfungen eingereicht habe. Sämtliche Quellen einschließlich Internetquellen, die unverändert oder abgewandelt wiedergegeben werden, insbesondere Quellen für Texte, Grafiken, Tabellen und Bilder, sind als solche kenntlich gemacht. Mir ist bekannt, dass bei Verstößen gegen diese Grundsätze ein Verfahren wegen Täuschungsversuchs bzw. Täuschung eingeleitet wird.

Berlin, den 11. Juni 2020