Technische Hochschule
Brandenburg
University of
Applied Sciences
Fachbereich
Informatik und Medien

Modellierung patientenorientierter Zielgrößen mit Methoden des Data Mining aus Daten des Behandlungsprozesses beim Mammakarzinom

Benjamin Hoffmann • Masterarbeit • Studiengang Informatik • Fachbereich Informatik und Medien • 17.05.2016

Aufgabenstellung

Ziel dieser Arbeit ist die Erstellung gültiger, transparenter, prädiktiver Modelle zur Vorhersage patientenorientierter Zielgrößen (wie z.B. dem Überleben von Brustkrebspatientinnen) aus den Daten des Tumorzentrums Land Brandenburg e.V. (TZBB). Zugehörige Aufgabenstellungen sind u.a. (a) eine deskriptive und explorative Datenanalyse (EA), (b) die Bestimmung relevanter Merkmale, (c) die Definition von (neuen) Merkmalen und patientenorientierter Zielgrößen, (d) die Modellbildung und -evaluierung und (e) die geeignete patientenorientierte Visualisierung von Zusammenhängen.

Konzept

Sowohl die EA, bei der insb. die Datenstruktur, Verteilungen und Datenqualität näher betrachtet wird, als auch die Modellierung mittels Entscheidungsbäumen und Regelmengen (inkl. Datenvorbereitung und Evaluation) wird in Python implementiert.

Für die Modellevaluation wird die Datenmenge in zwei Mengen aufgeteilt: die 75%-Trainingsmenge und die 25%-Testmenge. Auf der Trainingsmenge wird eine fünffach wiederholte zehnfache Kreuzvalidierung durchgeführt und deren Gütemaße aufgezeichnet. Anhand dieser Messwerte wird eine Menge an Setups ausgewählt, für die das 75%-Modell auf die 25%-Testmenge angewandt wird. Diese Ergebnisse stellt die geschätzte Performanz des mit allen Datensätzen erzeugten 100%-Modells dar (siehe Abb. 1).

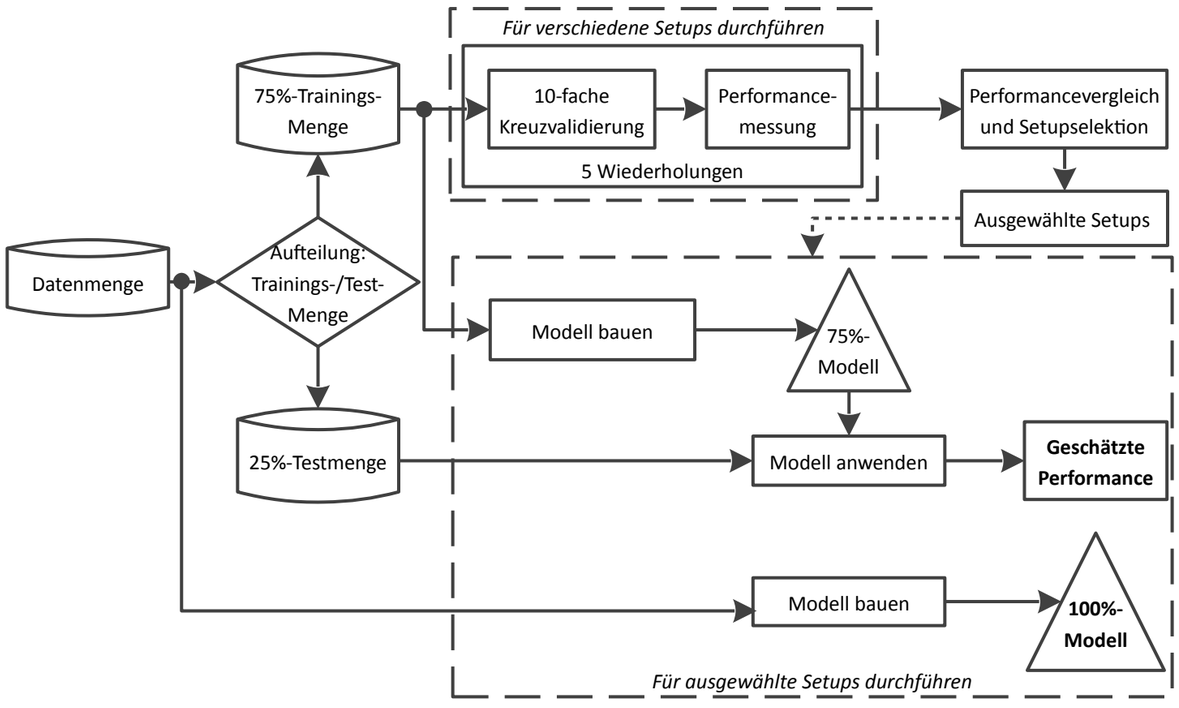

Abb. 1: Evaluationsprozess der Data-Mining-Setups

Explorative Analyse

Die TZBB-Daten enthalten 38.002 Datensätze/Tumore mit 619 Merkmalen v.a. zu Diagnose, Patient und Therapieverlauf. Die meisten Tumore wurden zwischen 1993 u. 2015 diagnostiziert. Informationen zum Todesdatum und zur Todesursache sind ebenso enthalten.

Einige Auffälligkeiten, wie z.B. Widersprüche zwischen Freitextfeldern und zusammengefassten Merkmalen, seltsame Merkmalsausprägungen oder scheinbare duplizierte Spalten, wurden entdeckt. Bei der Betrachtung der Merkmalsentwicklung fielen erwartungsgemäß Merkmale auf, die erst ab einem gewissen Zeitraum vermehrt aufgezeichnet wurden. Insgesamt ergab die Analyse, dass sich die Daten für die Modellierung mithilfe von DM-Techniken eignen.

Datenvorbereitung

Es wurden drei Zielgrößen definiert und evaluiert: (a) 5-Jahres-Gesamtüberleben (GÜ), (b) krankheitsspezifisches 5-J.-Ü. (KSÜ), krankheitsfreies 5-Jahres-Überleben (KFÜ).

Insgesamt dienen 86 Merkmale zur Diagnose, zum Patienten und zur OP bzw. Chemotherapie (inkl. eigens abgeleiteter Merkmale) als Eingabe für den Lernalgorithmus. Datensätze von Männern wurden ebenso ausgeschlossen wie Einträge mit unerklärlichen Datenfehlern und zensierte Einträge (mit einer Nachbeobachtungsdauer von weniger als fünf J. und keinem Eintreten des Ereignisses). Das Klassenverhältnis ist mit einen Anteil an überlebenden Patientinnen von 88,7 %, 80 % bzw. 82 % sehr unausgeglichen.

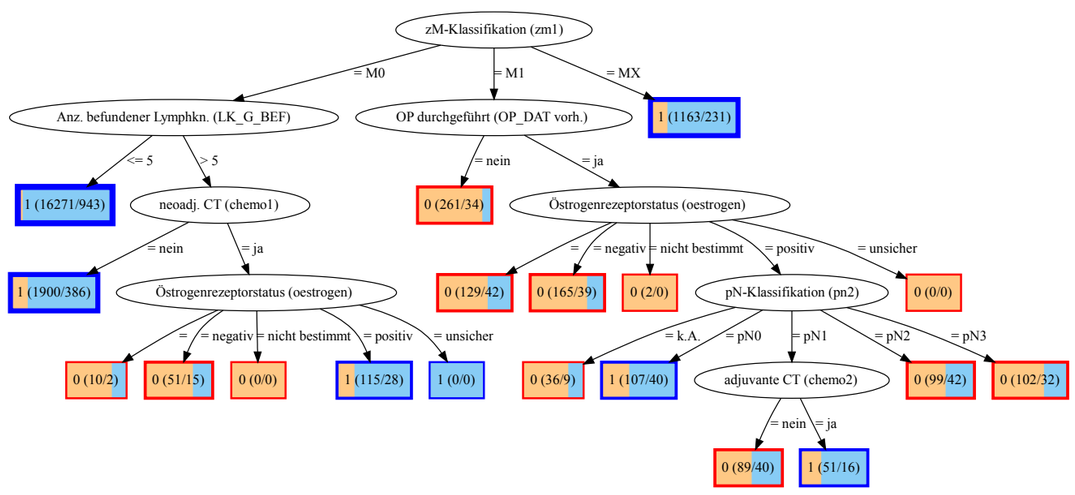

Abb. 2: Finales 100%-Modell für das krankheitsspezifische 5-Jahres-Überleben. In den Blättern verdeutlicht der Farbverlauf das Klassenverhältnis. Blätter der Klasse „0" (überlebt nicht) bzw. „1" (überlebt) werden rot bzw. blau umrandet. Die erste Zahl in der Klammer bezeichnet die Anzahl der Instanzen, die das Blatt erreichen; die zweite Zahl steht für die Anzahl falsch klassifizierter Instanzen.

Ergebnisse

Ein transparenter Entscheidungsbaum für das KSÜ erreicht eine Erfolgsrate von 90,35 % (AUC: 0,67, Sensitivität: 98,6 %, Spezifität: 25,7 %, siehe Abb. 2) auf der Testmenge. Ein durch Anpassung der Fehlklassifizierungskosten gelernter Entscheidungsbaum erreicht eine AUC von 0,794 (Erfolgsrate: 76,39 %) mit einer deutlich höheren Spezifität von 72,6 % (Sensitivität: 76,9 %). Die ausgewählten Modelle erscheinen aus Ärztesicht plausibel. Ähnliche Gütemaße wurden auch von vergleichbaren Arbeiten (vgl. [1, 2]) auf einer anderen Datenmenge berichtet.

Fazit und Ausblick

Es wurden drei patientenorientierte Zielgrößen definiert und für sie transparente DM-Modelle erfolgreich erzeugt, evaluiert und mit einem Arzt diskutiert. Die Visualisierungen und prädiktiven Modelle bilden sowohl für Patienten als auch für Ärzte die Grundlage, für ihren Fall eine Überlebensprognose zu treffen. Einige interessante Patientengruppen, Assoziationen und Modelle wurden identifiziert, die in nachfolgenden Arbeiten näher untersucht werden können.

Quellen

[1] Ya-Qin, Liu; Cheng, Wang; Lu, Zhang: „*Decision Tree Based Predictive Models for Breast Cancer Surivability on Imbalanced Data*", 3rd International Conference on Bioinformatics and Biomedical Engineering (S. 1–4), 2009

[2] Wang, Kung-Jeng; Makond, Bunjira; Wang, Kung-Min: „*An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data*", BMC Medical Informatics and Decision Making 13 (S. 1–14), 2013

Gutachter: Dipl.-Inform. Ingo Boersch • Technische Hochschule Brandenburg • Prof. Dr.-Ing. Sven Buchholz • Technische Hochschule Brandenburg