2 Methoden

Grundlage dieser Arbeit sind recht umfangreiche Analysen. Um einen Überblick über das Vorgehen zu geben, erfolgt zuerst eine kurze Darstellung der Auswertungsschritte. Die Verwendung der Methoden ist dann in einzelnen Abschnitten ausführlicher beschrieben.

Verwendete Daten

- Simulierte Daten

  Für diese Arbeit wurden Daten simuliert, mit denen die Leistungsfähigkeit künstlicher neuronaler Netze im Vergleich zu klassischen statistischen Methoden demonstriert werden sollte. Trotzdem sollten die Daten so aufgebaut sein, daß möglicherweise auch derartige empirischen Daten vorkommen könnten.
- Daten aus dem Zentralregister Malignes Melanom
- Daten aus dem Zentralregister Malignes Melanom unter Verwendung von Fällen mit fehlenden Angaben

Analyseschritte

Bei allen Auswertungen wurde einheitlich nach dem in Tabelle 5 aufgeführten Muster vorgegangen. Dazu wurden die Daten im Verhältnis 1:2 per Zufallsgenerator in jeweils zwei unabhängige Dateien geteilt. An der größeren Datei (Trainingsdaten) wurden die statistischen Modelle bzw. die künstlichen neuronalen Modelle entwickelt und trainiert. An der kleineren Datei (Testdaten) wurden die Modelle angewendet zur Parameterschätzung. Die geschätzten Parameter wurden bei den Trainings- und bei den Testdaten mit den beobachteten Zahlen verglichen.

Tabelle 5: Arbeitsschritte bei der Analyse der Daten

Deskriptive Analyse der Daten
Kaplan-Meier-Kurven
Log-Rank-Test
CART-Analyse (Rpart)
Cox-PH-Analyse
Künstliche neuronale Netze
ROC-Analyse für die geschätzten Parameter

2.1 Beschreibung der verwendeten Daten

2.1.1 Simulierte Daten

Zur Simulation der Daten wurden Zufallszahlen generiert, die zum Teil verknüpft in die auszuwertende Datei übernommen wurden. Dabei wurde Wert darauf gelegt, eine Datei zu generieren, bei der einfache lineare Modelle nicht die optimale Beschreibung liefern können. Außerdem sollten die Daten einen hohen Anteil zensierter Beobachtungen enthalten, um Ähnlichkeit zu den gleichfalls analysierten Registerdaten herzustellen. Das Programm zur Generierung der Daten ist im Anhang abgedruckt. Bei der Simulation der Daten mußte darauf geachtet werden, daß die generierten Daten nicht die Modellannahmen des Cox-PH-Modells verletzen.

2.1.2 Registerdaten

Die verwendeten Daten stammen aus dem Zentralregister Malignes Melanom der Deutschen Dermatologischen Gesellschaft (DDG). Für die Analyse wurden Daten von Patienten ausgewählt, bei denen ein Melanom erstmalig im Zeitraum zwischen 1976 und 1996 aufgetreten ist. Damit bestand für alle Fälle die Chance für eine mindestens dreijährige Beobachtungszeit.

Die Erhebung der Daten erfolgte überwiegend prospektiv. Um eine hohe Datenqualität sicherzustellen, wurden nur die Daten aus den Zentren einbezogen, bei denen einer hohe Vollständigkeit bei der Verlaufsdokumentation vorlag.

Aus der Literatur sind eine Vielzahl prognostischer Faktoren bekannt. Für diese Analyse wurden Alter, Geschlecht, histologischer Typ, Invasionslevel nach Clark, Lokalisation des Primärtumors, Tumordicke nach Breslow, Regression und Ulzeration untersucht.

Üblicherweise erfolgt die Auswertung derartiger Daten mit Methoden der Überlebensanalyse. Für diese Methoden ist eine Zeitangabe über die Dauer zwischen Erstdiagnose des Melanoms und Ende des Untersuchungszeitraums nötig. Die Wahl der Zeitpunkts ‚Erstdiagnose' kann zu erheblichen Fehlern bei einzelnen Fällen führen. Verschiedene Autoren haben gezeigt, daß ein Melanom über viele Jahre bestehen kann, bevor es entdeckt und behandelt wird. In bisherigen Auswertungen wurde der Zeitpunkt der ersten Operation des Melanoms als Startpunkt der Beobachtungszeit mit dem Argument verwendet, daß der Patient zu diesem Zeitpunkt wieder klinisch tumorfrei ist.

Für die Bewertung der Ergebnisse der Modelle ist die Überprüfung an unabhängigen Daten nötig. Für das Training künstlicher neuronaler Netze wird generell ein Trainings-, ein Valdierungs- und ein Test-Set gefordert. Daher wurden die vorhandenen Daten in drei gleich große Gruppen geteilt. Die Zuordnung der Patienten zu den Gruppen erfolgte zufällig.

2.2 Durchführung der Analysen

Die Cox-Regressionsanalyse wurde entsprechend den Vorschlägen von Harell et al. [75] vorgenommen.

Vorbereitung

- Definition der Zielvariablen
- Auswahl potentieller Prädiktoren (Umgang mit fehlenden Angaben)
- Auswahl plausibler Interaktionen

Datenreduktion

Um die Genauigkeit eines Modells zu erhöhen, muß eine kleine Zahl von Faktoren (genauer: Spalten in der Designmatrix) verwendet werden. Die Anzahl der Faktoren und die Komplexität des Modells muß sich nach Fallzahl der Beobachtungen richten. Als grobe Daumenregel gilt, daß die Zahl der untersuchten Faktoren nicht m/10 überschreiten darf, wobei m die Zahl der unzensierten Ereignisse darstellt.

Überprüfung der Modellannahmen

- Linearitätsannahme
- Additivitätsannahme
- Verteilungsannahme

Quantifizierung der Vorhersagegenauigkeit

Für die Quantifizierung der Vorhersagegenauigkeit gibt es verschiedene Gründe

- Quantifizierung des Wertes eines Prädiktors oder eines Modells für Screening oder für die Identifizierung von Individuen mit hohem Risiko
- Überprüfung eines gegebenen Modells auf Overfitting oder nichtausreichende Güte (falsche Modellspezifikation, Nichtberücksichtigung wichtiger Prädiktoren, ungenügende Modellanpassung)
- Vergleich von Methoden oder Modellen

2.3 Kaplan-Meier-Schätzung und Log-Rank-Test

Zur Durchführung der Kaplan-Meier-Schätzung, zur Darstellung der Kaplan-Meier-Überlebenskurven und für die Berechnungen des Log-Rank-Tests wurden die bei S-Plus 4.0 mitgelieferten Funktionen `surv` und `survdiff` genutzt.

2.4 CART-Analyse für zensierte Daten

Die für die Konstruktion von Klassifikations- und Regressionsbäumen verwendete Bibliothek Rpart [160] von T. Therneau basiert auf den Arbeiten zu Klassifikations- und Regressionsbäumen von Breiman, Friedman, Olshen, und Stone [19].

2.5 Multifaktorielle proportionale Hazard-Regressionsanalyse

Die in S-Plus verwendete Funktion zur Berechnung des proportionalen Hazards-Modell [38] beruht auf der Formulierung des Problems als Zählprozess [5, 161]. Harrell hat diese usprünglich von Therneau entwickelte Funktion an die Besonderheiten der Design-Bibliothek angepaßt. In dieser Arbeit wurde die in der Design-Bibliothek enthaltene Funktion `coxph` für die Duchführung der Berechnungen des Cox-PH-Modells verwendet.

2.6 Neuronale Netze für zensierte Verlaufsdaten

Zur Verwendung neuronaler Netze gibt es bisher wenige brauchbare Ansätze [98, 132]. Der nach bisherigen Erkenntnissen methodisch beste Vorschlag stammt von Faraggi & Simon [59]. Bei diesem Ansatz werden alle Modellbedingungen des Cox-PH-Modells beibehalten. Im Unterschied zum statistischen Ansatz erfolgt die

Anpassung der Parameter nicht über eine Maximum-Likelihood-Schätzung, sondern unter Verwendung eines vorwärts gerichteten künstlichen neuronalen Netzes mit einer verdeckten Schicht. Für S-Plus lag dieser Ansatz in Form einer Bibliothek von R. Ripley vor [134].

Letztendlich läßt sich das verwendete Modell als Regressionsgleichung darstellen. Eine Erweiterung diese Ansatzes stellt die *Projection Pursuit Regression* dar [66].

2.7 Generierung neuer Werte bei fehlenden Angaben

Fehlende Angaben bei multivariaten Untersuchungen führen dazu, daß jeweils eine komplette Zeile aus dem Datensatz gelöscht werden muß. Um diesen Informationsverlust zu vermeiden, können die fehlenden Angaben geschätzt werden. Eine einfache Form der Schätzung besteht in der Verwendung des Medians bei stetigen Variablen und des häufigsten Wertes bei kategorialen Daten. Wenn die Variablen, die fehlende Angaben enthalten, mit anderen Variablen korreliert sind, wird die Schätzung genauer, wenn die anderen Variablen berücksichtigt werden.

Die grundsätzliche Aufgabe bei dieser Form der Vervollständigung der Daten besteht also darin, ein multivariates (Regressions-)Modell zu finden, das in der Lage ist, fehlende Angaben in einer Datenzeile durch Schätzungen aufgrund von Verteilungsannahmen und vorhandenen Daten anderer Variablen durchzuführen. Diskutiert werden verschiedene Ansätze und deren Bedeutung bei Schaefer [145], Rubin [140] und Vach [164].

Die Generierung neuer Werte bei fehlenden Angaben wurde unter Verwendung der Funktion *transcan* in der S-Plus-Bibliothek *Hmisc* (F. Harrell) durchgeführt, die den gleichzeitigen Umgang mit stetigen und mit kategorialen Daten erlaubt.

In einem vorbereitenden Schritt wurden in der Datenmatrix die Fälle nach Mustern der fehlenden Angaben sortiert, stetige Variablen wurden automatisch zentriert und skaliert.

Zur Ergänzung der fehlenden Angaben wurde eine nichtlineare additive Funktion verwendet [140]. Diese Funktion transformiert stetige und kategoriale Daten so, daß sie eine maximale Korrelation mit der besten Linearkombination der anderen Variablen haben. Dabei wurden stetige Variablen unter der Verwendung kubischer Splines transformiert, kategoriale Variablen wurden dummy-codiert. Wenn bei einer Variablen Werte fehlten, wurden diese ergänzt unter Verwendung einer multiplen Regression (Kleinste-Quadrate-Schätzung). Um eine Überanpassung zu vermeiden, wurde ein Shrinkage-Verfahren verwendet [166].

2.8 Vergleich der Klassifikationsgüte der Modelle

Der Vergleich der Klassifikationsgüte der Modelle stellt ein schwieriges Problem dar. Zwar gibt es für jedes einzelne Modell auch Parameter zur Einschätzung der Prognosegüte, jedoch sind diese Parameter nicht direkt vergleichbar.

Zur Bewertung der Ergebnisse wurden zwei verschiedene Ansätze gewählt. Beim ersten Ansatz werden die ROC-Kurven als Maß für die Sensitivität (Wahrscheinlichkeit für die korrekte Vorhersage des Tods), Spezifität (Wahrscheinlichkeit für die korrekte Vorhersage des Überlebens) und Genauigkeit (Anteil korrekter Vorhersagen insgesamt in Prozent) ausgewählt. Diese Parameter werden für vier Zeitpunkte angegeben und geben damit Hinweise für die praktische Anwendbarkeit der Prognoseschätzungen.

ROC steht für *Receiver Operating Characteristic.* Der Name ist historisch bedingt und liefert keine Beschreibung des Verfahrens. ROC-Kurven dienen dazu, für ein stetiges Merkmal Dichotomisierungseigenschaften darzustellen. Dazu werden für jeden Wert s des Merkmals auf der Abszisse die Rate der falsch-positiven und auf der Ordinate 1 - Rate der falsch negativen Eingruppierungen bei Verwendung des Wertes s die Punkte

eingetragen. Mit diesen Punkten kann die ROC-Kurve beschrieben werden. Der Flächeninhalt unter der ROC-Kurve c ist eng verwandt mit dem Konkordanzmaß Somers's D: D = 2(c-0,5).

Im zweiten Ansatz wird das von Harrell vorgeschlagene Konkordanzmaß c verwendet [80]. Der Werte c wird berechnet unter Verwendung aller möglichen Paarungen von Patienten. Für jede Paarung werden die Prognoseschätzungen als konkordant mit dem Ergebnis angesehen, wenn die Fälle mit günstiger eingeschätzter Prognose auch tatsächlich eine längere Verweildauer aufweisen. Wenn bei beiden Fällen eines Paares die Verweildauern zensiert sind oder wenn bei einem Fall das Zielereignis eingetreten ist und der zweite Fall eine kürzere Verweildauer aufweist als der erste Fall, wird die Paarung nicht berücksichtigt. Der c-Index ist das Verhältnis von konkordanten Paaren und allen Paarungen, für die Verweildauern bestimmt werden können. Wenn c den Wert 0,5 hat, besitzt das zugrundeliegende Modell keinen Wert für die Vorhersage. Der Wert c kann als ROC-Statistik für zensierte Daten verstanden werden. Zur Berechnung wurde die Funktion *rcorr.cens*[2] in der Bibliothek *Hmisc* verwendet.

2.9 Software für die Analysen

Die Aufbereitung der Daten erfolgte unter *SPSS*® 9.0 für Windows. Die Analysen wurden durchgeführt unter Verwendung des Statistikpakets *S-Plus*® 4.0 für Windows® der Firma MathSoft.

Zur Generierung neuer Werte bei fehlenden Angaben wurde testweise die S-Plus-Bibliothek MIX (Multiple imputation functions for mixed continuous and categorical data) Version 1/98 verwendet[145]. Wegen Problemen mit dieser Software wurde dann aber die *transcan*-Funktion der Bibliothek *Hmisc* verwendet.

Die Cox-PH-Analysen wurden vorgenommen unter Verwendung der S-Plus-Bibliotheken *Hmisc* und *Design* (F. Harrell). Die Berechnung und Validierung der Klassifikations- und Regressionsbäume (CART) erfolgte unter Verwendung der Bibliothek *RPart* (T. Therneau). Für die Analysen zensierter Daten mit neuronalen Netzen wurde die Bibliothek *SurvNN* (Ripley) verwendet.

Vorbereitende Analysen für diese Arbeit wurden mit einer eigenen Anpassung des *Stuttgarter Neuronale Netze Simulators* (*SNNS*, A. Zell) unter Linux vorgenommen.

2.9.1 Statistikprogramme für Prognoseschätzungen

Es gibt einen großen Markt für Statistiksoftware. Entsprechend sind mittlerweile auch zahlreiche Auswertungssysteme vorhanden. Eine marktbeherrschende Stellung nimmt dabei SAS (SAS Institute) ein, da es nach wie vor das einzige System ist, das für die Auswertung von Zulassungsstudien durch die FDA (Food and Drug Administration, zuständig u.a. für die Zulassung von Medikamenten) in den Vereinigten Staaten anerkannt ist.

2.9.2 Statistikpaket S-Plus

Die meisten Berechnungen in dieser Arbeit wurden unter Verwendung des Statistikpakets *S-Plus 4.0* unter Windows NT durchgeführt.

Das an den Bell Laboraties entwickelte Auswertungssystem S ist in S-Plus integriert. Außerdem bietet es ein flexibles und mächtiges Auswertungssystem für viele statistische Fragestellungen und verfügt über eines der derzeit besten Systeme zur Erstellung wissenschaftlicher Grafiken.

---

[2] Die Funktion rcorr.cens() ist fehlerhaft, weshalb die geschätzten Parameter der Modelle negiert übergeben werden mußten

Unter anderem bietet dieses System eine große Bibliothek statistischer Werkzeuge, die aufeinander abgestimmt sind. Dazu gehören u.a. Werkzeuge für lineare und nichtlineare statistische Modelle. Zusätzlich verfügt das System über die Implementierung der Sprache S über eine objektorientierte Programmiersprache, die leicht erweitert werden kann und für die eine große Anzahl zusätzlicher Module existiert. Ein großer Teil der aktuellen anwendungsorientierten statistischen Forschung basiert auf der Verwendung von S. Für rechentechnisch besonders intensive Probleme können Objektmodule aus Fortran oder C in S-Plus-Funktionen eingebunden werden.

Zu den Nachteilen des Systems gehören die relativ aufwendige Einarbeitung, wobei Grundlagenwissen in der objektorientierten Programmierung erforderlich ist, und einige Instabilitäten, die zu gelegentlichen Abstürzen der Software führen. Bedauerlicherweise ist dieses hervorragende System teuer und dadurch für den Einsatz in der Ausbildung nur in besonderen Fällen geeignet.

2.9.3 Statistikpaket R

Eine weitere Implementierung der Sprache S wurde beim Statistikpaket R verwirklicht[1]. Bei der ursprünglichen Entwicklung waren Ross Ihaka und Robert Gentleman von der University of Auckland (Neuseeland) die Initiatoren. Seit Mitte 1997 wird dieses System von einer internationalen Gruppe (R Core Team) koordiniert.

Zum Erlernen der Sprache S oder zur Verwendung ausgewählter Prozeduren kann dieses Paket problemlos verwendet werden. Es erreicht nicht die Benutzerfreundlichkeit von S-Plus, kann aber im Bereich von universitärerer Forschung und Lehre frei benutzt werden.

2.9.4 Simulationen mit dem Stuttgarter Neuronale Netze Simulator (SNNS)

Für erste Experimente zum Vergleich neuronaler Netze mit Regressionsmodellen wurde das Programm SNNS (Stuttgarter Neuronale Netze Simulator) verwendet. Dieses umfangreiche Paket, daß über eine Vielzahl von vorbereiteten neuronalen Netzen verfügt, ist für wissenschaftliche Analysen und Ausbildungszwecke frei verfügbar.

Bei diesem Programm handelt es sich um einen Simulator, der am Institut für Parallele und Verteilte Höchstleistungsrechner (IPVH) der Universität Stuttgart entwickelt wurde. Im wesentlichen besteht SNNS aus einem Modul für die Simulation, einer grafischen Oberfläche zur Visualisierung und Veränderung generierter Netze und einem Compiler zur Generierung neuronaler Netze aus einer speziell entwickelten Hochsprache zur Beschreibung der Netze.

Laut Dokumentation sind im SNNS u.a. folgende Netzwerkmodelle enthalten:

- Backpropagation
- Quickprop
- Counterpropagation
- Backpercolation
- Rprop
- Cascade-Correlation
- Time-Delay-Netze
- ART-1 und ART-2 (Adaptive Resonanztheorie)

- Backpropagation und Quickpropagation through time
- Selbstorganisierende Karten

Das System ist für verschiedene Rechnersysteme verfügbar, unter anderem für Workstations von Hewlett Packard, IBM und Sun, für Windows und in einer Linux-Version. Im Rahmen dieser Dissertation wurde die Linux-Version unter X-Windows eingesetzt.