Aus dem
CharitéCentrum 4 für Therapieforschung
Institut für Medizinische Informatik
Direktor: Prof. Dr. Thomas Tolxdorff

Habilitationsschrift

Maschinelle Lernverfahren für nieder- und hochdimensionale Probleme: Zusammenführung und Analyse biomedizinischer Daten

zur Erlangung der Lehrbefähigung
für das Fach Medizinische Informatik, Biometrie und Epidemiologie

vorgelegt dem Fakultätsrat der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Dr. rer. physiol. Murat Sariyar

Eingereicht: März 2015

Dekan: Professor Dr. med. Axel R. Pries

1. Gutachter: Professor Dr. biol. hom. Hans-Ulrich Prokosch

2. Gutachter: Professor Dr. rer. nat. Ulrich Sax

“[Modern scientific] theories can necessarily never be more than hypothetical, for facts in themselves are always susceptible of diverse explanations and so never have been and never will be able to guarantee the truth of any theory.”

(Rene Guenon)

Inhaltsverzeichnis:

1. Einleitung

1.1 Methodenbasiert an wissenschaftliche Probleme herangehen

Typischerweise wird in der Wissenschaft von Problemen und ungeklärten Sachverhalten ausgegangen, die einer Lösung bzw. Erklärung bedürfen. Die Wissenschaftlichkeit der Herangehensweise wird dabei durch Theorien und Methoden sichergestellt, welche selbst gewisse Kriterien erfüllen, also zum Beispiel methodisch etabliert sein müssen. Dieser Selbstbezüglichkeit von Methoden/Kriterien – oder der Paradoxie des Anfangsproblems: welche Kriterien/Methoden dürfen gültige Methoden etablieren? – ist die Erkenntnis Thomas Kuhns komplementär, dass es keine puren Fakten gibt (Kuhn 1996). Einstein formulierte in diesem Sinne den Satz: "Erst die Theorie entscheidet darüber, was man beobachten kann“ (in Heisenberg 1971). Das lässt folgende Schlussfolgerung zu: jede Form der Problemortung und –definition erfolgt unter Rückgriff auf vorgegebene kognitive / theoretische Filter.

Die Selbstbezüglichkeit bei der Methodenfindung führt dazu, dass gewisse Methoden a priori und axiomatisch ein explanatorisches Gewicht zugewiesen bekommen, welches ihnen erlaubt, als fundamentales Mittel der objektiven Wissensgenerierung zu dienen. Zu einem solchen fundamentalen Methodenkanon gehören in Einzelwissenschaften neben der formalen Logik zum Beispiel auch solche Erkenntnisgewinnungsverfahren, über die ein fächerübergreifender Konsens hinsichtlich ihrer epistemischen Relevanz existiert. Je mehr grundlegende Methoden es gibt, desto größer wird der Bereich, in welchem stabile Perspektiven und Herangehensweisen an potentielle Probleme existieren. Im Sinne Luhmanns kann man auch sagen, dass sich Eigenwerte ausbilden (vgl. dazu auch (Luhmann 1992; Luhmann 1996)).

Es besteht nun ein Unterschied zwischen Problemen, die erst auf Basis von etablierten Theorien oder Methoden als solche gesehen werden und Problemen, bei denen das nicht der Fall ist. Beispielsweise besteht das Problem des multiplen Testens im Rahmen der Statistik in der Inflationierung des Fehlers erster Art; ohne einen statistischen Rahmen erscheint es „unwahrscheinlich“, dass sich ein Bewusstsein für diese Problematik in dieser Form entwickelt. Das im Bereich des maschinellen Lernens intensiv besprochene Thema des Overfittings erscheint dem

multiplen Testen möglicherweise ähnlich, besteht jedoch aus einer anders gelagerten Problematik: hier wird ein unrichtiges Modell ausgewählt (siehe auch (Bishop 2007)). Für die Frage, was ein methodenbasiertes (also ein sich auf gewisse Methoden beschränkendes) Herangehen an wissenschaftliche Probleme vornehmlich charakterisiert, bedeutet dies, dass nicht primär die wissenschaftliche Qualität berührt wird, sondern Art, Reichweite und die Komplexität von Problemen, die man auf der Basis erkennen und adressieren kann. Als weiteres Beispiel sei ein Thema dieser Schrift betrachtet: Maschinelle Lernverfahren für ein fehlerfreies Zusammenführen von Daten. Es existiert eine Vielzahl von Verfahren für diese Aufgabe: statistische, deterministische, manuelle etc. Sich dieser Aufgabe auf der Basis von maschinellen Lernverfahren zu nähern, bedeutet, es primär als kognitives Klassifikationsproblem zu betrachten. Dies führt dazu, dass Fragestellungen der Form „wie hoch ist die Wahrscheinlichkeit für einen Synonymfehler bei der Zusammenführung von Datenbasis X mit Datenbasis Y?“ nicht adäquat bzw. nur schwer zu beantworten sind. Hier wäre die klassische statistische Herangehensweise geeigneter, welche ein stochastisches Inferenzproblem modelliert. Auf der anderen Seite kann man für das Ziel „Richtiges Klassifizieren von Datensatzpaaren als Homonym oder Synonym“ deshalb an Komplexitätsadäquanz gewinnen, da eine ganze Reihe von Klassifikationsverfahren zur Verfügung steht, die versprechen, eine annähernde Simulation des kognitiven und komplexen Klassifikationsmechanismus abzubilden.

In der Zusammenfassung heißt das alles: eine methodenbasierte Herangehensweise impliziert eine Beschränkung auf bestimmte Probleme, erlaubt für einen Teil dieser aber auch eine höhere Tiefenschärfe als ein generelles und nicht auf eine bestimmte Methodenklasse rekurrierendes Angehen von Problemen. Letztendlich hängt es – erneut selbstbezüglich – von den Problemen ab, ob eine methodenbasierte Perspektive adäquat ist oder eher eine fundamentalere Perspektive eingenommen werden sollte. In dieser Schrift werden Probleme in zwei unterschiedlichen Domänen (Record Linkage und hochdimensionale Daten) mit Hilfe von maschinellen Lernverfahren angegangen. Die Annahme ist, dass diese zu Erkenntnissen und Lösungen führen, zu denen man mit deterministischen und klassisch-statistischen Mitteln nur schwer oder gar nicht gelangen würde.

1.2. Maschinelle Lernverfahren in der biomedizinischen Forschung

Maschinelles Lernen bezeichnet ein Wissensgebiet, das sich mit der Konstruktion, Anwendung und Analyse von Algorithmen beschäftigt, welche generalisiertes Wissen aus Daten zum Zwecke von Vorhersage, Erklärung und Beschreibung ableiten (Mitchell 1997; Bishop 2007). Charakteristisch ist eine auf Trainingsdaten aufbauende Lernphase, in welcher ein allgemeines Modell für den die Daten generierenden Mechanismus erzeugt wird. Auf diese Trainingsphase folgt zumeist eine Anwendung des erzeugten Modells auf unbekannte Daten. Hauptsächliches Ziel ist dabei die Vorhersage neuer oder zukünftiger Ereignisse. Zwei weitere Ziele beziehen sich auf das Erzeugen und Testen von kausalen Modellen/Theorien (Erklärung) sowie auf die Repräsentation (Beschreibung) der vorhandenen und potentiellen Daten in einer kompakten Form (Shmueli 2010). Im Unterschied zum eng verwandten Wissensgebiet Data-Mining, geht es also nicht ausschließlich um das selbstständige Entdecken von neuen Mustern oder Gesetzmäßigkeiten, sondern vor allem auch um eine verbesserte Vorhersage.

In der biomedizinischen Forschung finden häufig stochastische Modelle Anwendung, da diese eine beschreibende Modellierungsperspektive einnehmen, welche eine gewisse Transparenz hinsichtlich der Vorhersagemechanismen impliziert. Dagegen ist ein nicht-formales Beschreiben von algorithmisch formulierten Modellen aus dem Kontext Maschinelles Lernen selbst für einen Fachspezialisten häufig schwierig. Dies gilt vor allem dann, wenn Black-Box-Modelle wie künstliche neuronale Netze genutzt werden. Nichtsdestotrotz gewinnen maschinelle Verfahren in solchen Bereichen an Bedeutung, in denen die Vorhersage und nicht so sehr die Beschreibung im Vordergrund steht (Schrom 2014; Cleophas und Zwinderman 2013). Typische Anwendungen sind Diagnoseunterstützung, Vorhersage von molekularen Strukturen und Phänotypen oder Hypothesengenerierung im Rahmen des ‚Clusterns' in einer Wissensdatenbank (Dua, Acharya, und Dua 2013). Es kommt dabei häufig vor, dass maschinelle Lernverfahren zunächst für komplexe Problemstellungen genutzt werden, für die sich herkömmliche statistische Verfahren als inadäquat erwiesen haben. Bei positiven Resultaten werden sie dann häufig auch in Bereichen eingesetzt, für die zwar schon etablierte Verfahren existieren, in denen aber ein gewisses Optimierungspotential gesehen wird. Optimieren lässt sich dabei in vielen

Fällen Modellflexibilität und die Transparenz hinsichtlich des Tunings der eingesetzten Verfahren (Boulesteix und Schmid 2014).

Es gibt viele Möglichkeiten, maschinelle Lernverfahren zu kategorisieren. Da das Lernen bei maschinellen Lernverfahren das primäre Charakteristikum ist, rekurriert die häufigste Kategorisierung auch darauf: überwachtes Lernen, unüberwachtes Lernen, semi-überwachtes Lernen, bestärkendes Lernen und Tiefenlernen. Bei überwachtem Lernen besitzen die Lerndaten den Wert (Label), den der zu generierende Algorithmus bei neuen Daten vorhersagen soll. In dieser Kategorie gibt es die größte Anzahl an Verfahren; hierzu gehören unter anderem: klassische neuronale Netze, Support Vector Machines, Entscheidungsbaumverfahren, Random Forests, Conditional Random Fields, k-Nearest Neighbor, Diskriminanzanalyse, Boosting, etc. (Mohri, Rostamizadeh und Talwalkar 2012). Bei unüberwachtem Lernen sind die Zielwerte nicht bekannt, so dass ohne Anpassungsmöglichkeiten an die „Wahrheit“ selbstständig interessante Muster gefunden werden müssen. Klassische Vertreter sind: Self-Organizing Maps, Assoziationsregeln und Clusteringalgorithmen. Bei semi-überwachten Verfahren stehen wenige „gelabelte“ und meist viele ungelabelte Daten zur Verfügung; in einem iterativen Lernprozess werden dann immer mehr ungelabelte Daten zu gelabelten (Zhu und Goldberg 2009; Chapelle, Scholkopf und Zien 2010). Bestärkendes Lernen bedeutet, gemäß Markov-Entscheidungsmodell eine Abfolge von Operationen so mit Nutzenwerten zu belegen, dass ein Modell entsteht, welches eine optimale Optionenfolge für beliebige Situationen erlaubt (Wiering, Otterlo und Otterlo 2012). Schließlich geht es im Rahmen des Tiefenlernens um die Transformation von Daten in solch eine Repräsentation, die ein weiteres Prozessieren durch weitere maschinelle Lernverfahren erleichtert (Bell 2014; Deng und Yu 2014). Eine zusammenfassende Übersicht über diese maschinellen Lernverfahren mit beispielhaften Vertretern bietet Abbildung 1.

Darüber hinaus existiert ergänzend dazu der Bereich „Active Learning“, der sich mit optimaler Auswahl von Trainingsdaten beschäftigt (Settles 2009; Muslea, Minton und Knoblock 2006; Tuia et al. 2011). Im Unterschied zu den herkömmlichen Lernverfahren gibt es hier einen vorgelagerten (nicht transformierenden) Schritt, in dem determiniert wird, welche Daten für das Lernen in dem Sinne wertvoll sind, dass

sie in gelabelter Form ein die Domäne gut repräsentierendes Modell produzieren lassen. Dabei möchte man die Anzahl der Trainingsdaten so gering wie möglich halten, da im Allgemeinen ein mit Kosten verbundenes manuelles Labeln der Daten notwendig ist, wenn man aus allen (potentiell) verfügbaren Daten zu einer Auswahl von Trainingsdaten kommen möchte.

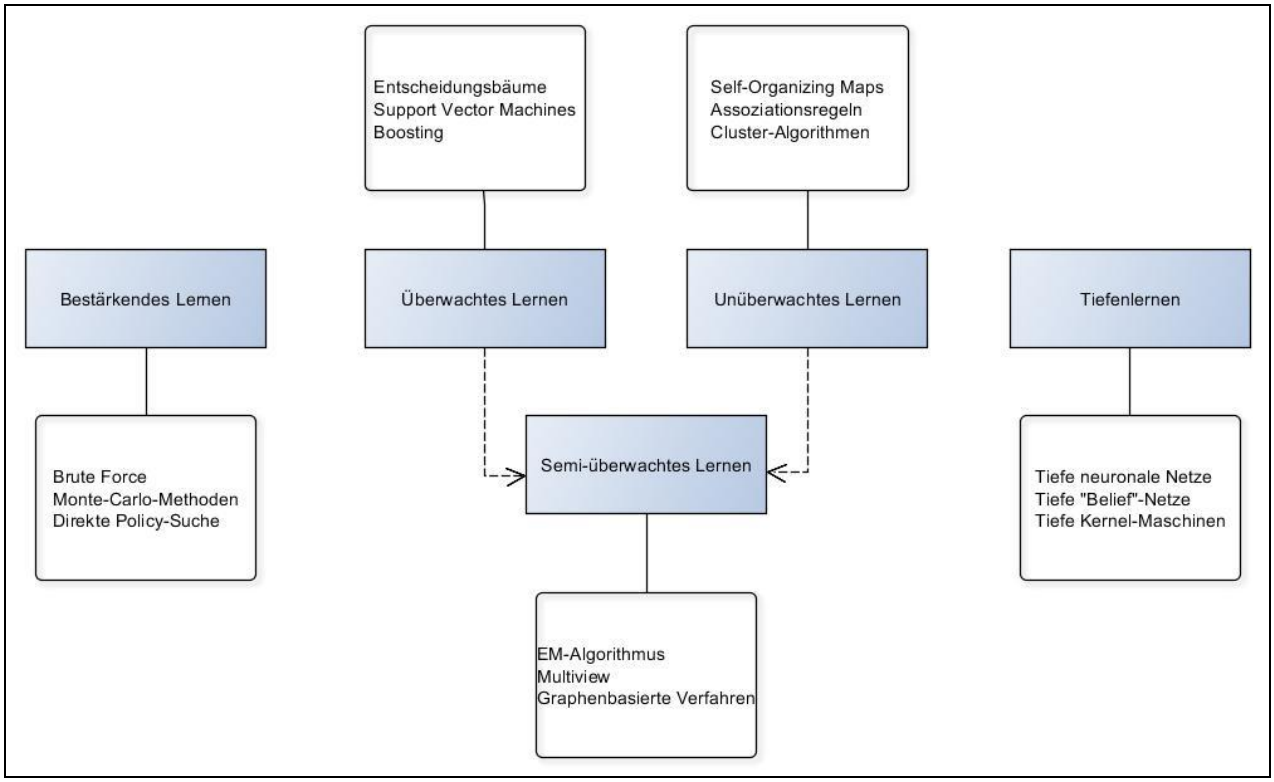

Abbildung 1: Übersicht über maschinelle Lernverfahren mit jeweils 3 beispielhaften Vertretern

1.3. Zusammenführen und Deduplizieren von biomedizinischen Daten

Das Problem der Identifizierung von Entitäten bei Zusammenführung und Deduplizierung von Daten besitzt in allen der Empirie zugewandten oder hinzuarbeitenden Disziplinen eine gewisse Relevanz (eine ähnliche Aussage findet sich schon in der Dissertation des Verfassers dieser Habilitationsschrift, 2009). Das Problem entsteht dadurch, dass einerseits ein und dasselbe Objekt in verschiedenen Datenquellen durch verschiedene Merkmalsdaten repräsentiert sein kann (Synonymfehler) und andererseits unterschiedliche Objekte, welche identische beziehungsweise sehr ähnliche Merkmalsausprägungen aufweisen, als gleich betrachtet werden (Homonymfehler).

Fehler in den Daten können sowohl bei der Registrierung als auch beim Verarbeiten der identifizierenden Informationen dadurch entstehen, dass

- Initial falsche oder unvollständige Angaben gemacht werden oder
- Fehler bei der Erfassung und Speicherung der Daten geschehen.

Zusätzlich zu diesen Fehlern können Entitäten (beziehungsweise Personen) aufgrund tatsächlich stattgefundener Änderungen durch unterschiedliche Daten repräsentiert sein. Alle diese potentiellen Fehlerquellen verdeutlichen, dass das Problem auch dann besteht, wenn künstliche und eindeutige Bezeichner für die Entitäten, z.B. global gültige und eindeutige Bezeichner, existieren. Die schwierigere und praktisch häufigere Situation ist jedoch diejenige, in welcher keine global-eindeutigen Bezeichner vorhanden sind.

Das Problem der Feststellung, ob Merkmalsdaten in unterschiedlichen Quellen auf dasselbe Objekt hinweisen, wird in den unterschiedlichen Fachgebieten der Epidemiologie, Statistik, Datenbank-Forschung und des maschinellen Lernens mit unterschiedlichen Namen belegt und unter verschiedene Aufgaben subsumiert. Zu den wichtigen Bezeichnungen gehören unter anderem: Record Linkage (Newcombe, Fair und Lalonde 1992), Data Cleaning (Rahm und Do 2000; Liu et al. 2007), Fuzzy Matching (Ananthakrishna, Chaudhuri, und Ganti 2002), Entity Resolution (Benjelloun et al. 2009) und Instance Identification Problem (Wang und Madnick 1989). Zumeist wird die Bezeichnung „Record Linkage“ (RL) genutzt, wenn es sich um die Bereinigung im Rahmen der Zusammenführung von Daten aus unterschiedlichen Quellen handelt, und „Deduplizierung“, wenn es um die Bereinigung in einer Datenquelle geht. Diese Unterscheidung ist soweit nicht strikt, als man das Record Linkage auch in Form der Deduplizierung durchführen kann, wenn man die Datensätze zentral zusammenführt.

Neben den klassischen stochastischen Ansätzen für RL, die zumeist auf dem Modell in (Fellegi und Sunter 1969) beruhen und sich in vielen Anwendungsfeldern der Epidemiologie als Standardverfahren etabliert haben, gibt es eine Reihe von maschinellen Lernverfahren für dieses Problem. Fasst man RL als Klassifikations-problem auf, kommen prinzipiell alle Klassifikationsverfahren zur Lösung des RL-Problems in Betracht. Ein wichtiger Vertreter dieser Methodenklasse ist das

Entscheidungsbaumverfahren, das von Breiman et al. (Breiman et al. 1984) in Form des CART-Verfahrens eingeführt wurde (für Alternativen vgl. (Salzberg 1994)). Angewendet werden (diese auf überwachtem Lernen basierende) Entscheidungsbäume für die Zusammenführung und das Deduplizieren von Daten beispielsweise in (Tejada, Knoblock und Minton 2001). Ein weiterer Klassifikations-algorithmus ist die Nearest-Neighbour-Methode, in der Daten zu Klassen zusammen-gefasst werden, wenn die gemäß einem Distanzmaß gemessenen Abstände von Merkmalsausprägungen im Vergleich zu den anderen Daten am geringsten sind (Christen 2008). In diesem Zusammenhang sind auch Support Vector Machines (SVM) als ein Klassifikationsverfahren zu nennen. Die Hauptüberlegung bei diesem (lineare Trennhyperebenen berechnenden Verfahren) besteht darin, dass in niedrigen Dimensionen des ursprünglichen Merkmalsraumes oft nur durch sehr komplexe Funktionen eine Trennung der Daten in nicht-überlappende Gruppen gelingt. Dagegen wird in höheren Dimensionen die Trennung von in diese Dimensionen projizierten Daten durch einfache Hyperebenen möglich (Burges 1998). Für die Objektidentifizierung kommt diese Idee eher selten zum Tragen (siehe auch Christen 2008).

1.4. Analyse von hochdimensionalen biomedizinischen Daten

Vor allem durch den Einsatz von hochparallelen Sequenziertechniken gibt es mittlerweile eine Vielzahl von Datenbeständen, wie SNP-, Gen-, oder Protein-Datenbanken, die zur Vorhersage, Erklärung und Beschreibung von patientenassoziierten Sachverhalten genutzt werden können (Amaratunga 2014; Bühlmann und Geer 2011). Hochdimensional bedeutet dabei im Allgemeinen, dass viel mehr Attribute existieren als Beobachtungen (also zum Beispiel über 10.000 gemessene SNP-Werte bei einigen hunderten Patienten). Herkömmliche stochastische Modelle, wie verallgemeinerte lineare Modelle, und damit verbundene Schätzalgorithmen sind nicht in der Lage, in solchen Situationen zu stabilen Parameterschätzungen zu gelangen. Eine weitverbreitete Umschiffung dieses Problems besteht auch heute noch häufig darin (siehe zum Beispiel (Roessler et al. 2014)), univariate Tests statt multiple Regressionsmodelle zu nutzen. Das impliziert jedoch einen gewissen Informationsverlust, da das sich gegenseitige Beeinflussen der Variablen im Hinblick auf den Zielwert nicht berücksichtigt und potentielle

verstärkende oder abschwächende Effekte des Inkludierens von Variablen nicht in Rechnung gestellt werden kann. Zudem verliert man an Effizienz, da eine Adjustierung für das multiple Testen notwendig wird, wenn man zu Signifikanzaussagen gelangen möchte.

Auch wenn sich für Klassifikationsaufgaben auf hochdimensionalen Daten (z.B. Determinierung eines Patienten als suszeptibel für eine gewisse Krankheit anhand gewisser genetischer Marker) die Anwendung einer Vielzahl von maschinellen Lernverfahren anbietet, haben sich im klinischen Kontext solche Verfahren durchgesetzt, die vorhandene stochastische Modelle mit maschinellen Lernverfahren so erweitern, dass eine Modellbildung im hochdimensionalen Raum möglich wird. Dies hat seine Gründe unter anderem darin, dass: (a) man in der klinischen Forschung häufig zu Resultaten kommen will, die mit einer statistischen Signifikanz belegt werden können. Dies erreichen klassische maschinelle Lernverfahren meistens nur indirekt über das Simulieren einer empirischen Nullverteilung ohne die klassischen Annahmen und Werkzeuge in statistischen Tests; (b) die entstehenden Modelle genauso gut interpretierbar sein sollen wie die bekannten Regressions- und Klassifikationsmodelle; und (c) es sehr viel mehr klassisch ausgebildete Statistiker gibt als solche mit einem profunden Hintergrund im maschinellen Lernen.

Wichtige Vertreter in dieser vornehmlich generalisierte lineare Modelle adressierende und auf überwachtem Lernen basierende Verfahrensklasse sind: LASSO (Tibshirani 1994; Tibshirani 2011), SCAD (Xie und Huang 2009; Fan und Li 2001), Adaptive LASSO (Zou 2006), Gradient Boosting (Friedman 2000), componentwise likelihood-based boosting (Tutz und Binder 2006), Dantzig selector (Candes und Tao 2007) und Elastic Net (Zou und Hastie 2005). Zur Verdeutlichung der Hauptidee sei beispielhaft das für alle anderen Verfahren paradigmatische LASSO-Verfahren betrachtet: Die Ziele bestehen darin, alle Variablen in der Schätzprozedur zu berücksichtigen und für alle irrelevanten Variablen einen Koeffizientenschätzer von Null zu erzwingen (was zu einer Variablenselektion führt). Das wird dadurch erreicht, dass eine obere Schranke für die Summe der absoluten Werte der Koeffizientenschätzer als Nebenbedingung in die Schätzung einfließt, was auch als L1-Penalisierung bezeichnet wird. Das sich ergebende Optimierungsproblem erfordert eine iterative Schätzung und die Lösung ist im Allgemeinen nicht explizit (als Formel) darstellbar.

Die Determinierung der Schranke ist dabei ein kritisches Problem und erfolgt zumeist über Kreuzvalidierung (Shao 1993).

Im Gegensatz zum klinischen Kontext, wo häufig konkrete Ziele und Hypothesen vorhanden sind, herrscht in der biomedizinischen Grundlagenforschung häufig ein exploratives Herangehen an hochdimensionale Daten vor, um Hypothesen zu generieren. Hier sind daher in erster Linie Data-Mining-Techniken gefragt. Dazu gehören unter anderem folgende Verfahren: Cluster-Analyse (Eisen et al. 1998; Arabie, Hubert und Soete 1996), Korrelationsanalyse, Bayessche Netzwerke und Dimensionsreduktion (Györfi et al. 2002). Eine wichtige Anwendung ist die Vorhersage von Gen -oder Proteinfunktionen auf Basis von Sequenzdaten und assoziierten Umweltfaktoren.

1.5. Zielsetzungen

Die Weiterentwicklung, die Anwendung und der Vergleich von ausgewählten maschinellen Lernverfahren für die Zusammenführung/Deduplizierung und die Analyse von biomedizinischen Daten stellen die Hauptziele dieser Habilitationsschrift dar. Diese Ziele wurden mit den in diese Schrift eingeflossenen Arbeiten in folgende Teilziele heruntergebrochen:

- Untersuchung einer verfeinerten Active-Learning-Strategie bei Anwendung von ausgewählten maschinellen Lernverfahren für die Zusammenführung und Deduplizierung von Daten.

- Entwicklung einer frei verfügbaren Software für das Zusammenführen und Deduplizieren von Daten auf Basis einer Vielzahl von Algorithmen, die vor allem dem Kontext des maschinellen Lernens entstammen.

- Entwicklung eines maschinellen Lernverfahrens, das sowohl generalisierte lineare Modelle auf den hochdimensionalen Fall erweitert als auch im Hinblick auf die Modellsparsamkeit flexibel ist.

- Konstruktion einer Kombination von maschinellen Lernverfahren zum Auffinden von relevanten Interaktionen (die mit geringen/keinen marginalen Effekten assoziiert sind) in hochdimensionalen Daten.

- Verdeutlichung des Unterschieds zwischen dem Einsatz von maschinellen Lernverfahren für das niederdimensionale RL-Problem und für die Analyse von hochdimensionalen molekulargenetischen Daten.

2. Eigene Originalarbeiten

In der vorliegenden kumulativen Habilitationsschrift werden zunächst maschinelle Lernverfahren für das niederdimensionale Record-Linkage-Problem, also für das Zusammenführen und Deduplizieren von Daten betrachtet. Als Erstes wird gezeigt, wie eine eigens zu diesem Zweck in R entwickelte Software in der Praxis verwendet werden kann (Sariyar und Borg 2010). In dieser Software sind unter anderem eine Reihe von Entscheidungsbaumverfahren, Support Vector Machines, so wie Parameterschätzungen auf Basis eines verfeinerten Expectation-Maximization-Algorithmus und der Extremwertstatistik implementiert. Es ist dabei wichtig, zwischen der Forschungsleistung bezüglich der in der Software implementierten Methoden und derjenigen bezüglich der Softwareentwicklung an sich zu unterscheiden. Anschließend wird eine Active-Learning-Strategie vorgestellt, die auf Basis von Entscheidungsbäumen und einer erweiterten Entropiefunktion zur Bestimmung einer „optimalen“ Trainingsmenge genutzt wird (Sariyar, Borg und Pommerening 2012b). Final werden dann für das Record-Linkage neuartige und vor allem aus dem Repertoire an Entscheidungsbaumalgorithmen stammende maschinelle Lernverfahren eingesetzt und untersucht, ob eine einfache Active-Learning-Strategie für diese Verfahren zu besseren Resultaten führt als eine zufällige Auswahl von größeren Trainingsmengen (Sariyar und Borg 2012).

Anschließend wird dargestellt und diskutiert, wie man unter Berücksichtigung von hochdimensionalen Daten aus der molekularen Biologie Modelle auf Basis von maschinellen Lernverfahren entwickeln kann, die sowohl interpretierbar als auch in solchen Situationen zu Schätzungen von Parametern fähig sind, in denen herkömmliche stochastische Verfahren häufig scheitern. Zu diesem Zweck wurden stochastische Überlebenszeitmodelle auf den hochdimensionalen Fall angepasst und Fähigkeiten von bestimmten maschinellen Lernverfahren genutzt, um zu zusätzlichen Einsichten in den Daten zu gelangen. Hierzu wurde auf dem aus dem maschinellen Lernen stammende Prinzip des *componentwise likelihood-based boosting* aufgebaut (Tutz und Binder 2007; Binder und Schumacher 2009), welches die *stagewise-forward-regression* (Efron et al. 2004) auf generalisierte lineare Modelle und das Cox proportional hazards-Modell erweitert. Diese Verfahrensklasse erlaubt die Synchronisation von Variablenselektion und Parameterschätzung.

In dem ersten Anwendungsfall werden zum Zwecke der Vorhersage der Überlebenswahrscheinlichkeit von 283 Nierenkrebspatienten (Hakimi et al. 2013; Sandoval et al. 2011) klinische Phänotypen um über 400.000 CpG-Methylierungsmesswerte erweitert. Um den Zusatznutzen dieses Methylierungsdatensatzes auf Basenniveau sowie aggregiertem Gen- und Chromosomenniveau einzuschätzen, wurde das *componentwise likelihood-based boosting* dahingehend erweitert, dass eine Anpassung der Sparsamkeit des Models an die vorliegenden Daten vorgenommen wird. Eine Simulationsstudie wurde vor Anwendung auf die realen Daten durchgeführt, um wichtige Eigenschaften des neuen Verfahrens zu eruieren (Sariyar, Schumacher und Binder 2014).

Im zweiten hochdimensionalen Szenario wurden Microarraydaten von 1) 240 Patienten mit diffusem großzelligem B-Zell-Lymphom (Rosenwald et al. 2002) und 2) 279 Neuroblastom-Patienten betrachtet und untersucht, ob zusätzlich zu den Haupteffekten die Berücksichtigung von allen potentiellen 2er Interaktionen zu einer Verbesserung der Effektdetektion und der Vorhersage der Überlebenswahrscheinlichkeit führen. Dabei wurde, neben dem *componentwise likelihood-based boosting* für die Haupteffekts- und Endmodellbestimmung, Resampling, Random Forests und das Prinzip der Orthogonalisierung von Daten für die Interaktionsdetektion genutzt. Eine umfassende Simulationsstudie wurde auch hier vorgeschaltet, um das Potential dieses neuen Verfahrens einschätzen zu können (Sariyar, Hoffmann und Binder 2014).

2.1. Das R-Softwarepaket RecordLinkage

Murat Sariyar, Andreas Borg

***The R Journal*. 2010; 2(2):61-67**

Um etablierte und vom Verfasser dieser Schrift entwickelte Verfahren zum Record-Linkage der Forschergemeinschaft frei zur Verfügung stellen zu können und die eigene Arbeit in diesem Bereich zu erleichtern, wurde das R-Softwarepaket RecordLinkage entwickelt. Das Paket bietet die Möglichkeit, Verfahren anzuwenden, sie weiterzuentwickeln und miteinander zu vergleichen. Es sind sowohl stochastische als auch maschinelle Lernverfahren implementiert. Die Innovationen bei ersteren bestehen in einem erweiterten und auf dem EM-Algorithmus basierenden Schätzverfahren und in der Anwendung der Extremwertstatistik zur Bestimmung von Schrankenwerten (zur Trennung der Matches von Non-Matches). Als maschinelle Lernverfahren wurden Entscheidungsbäume, Bagging, Boosting, Neuronale Netze und Support Vector Machines berücksichtigt. Das Paket bietet darüber hinaus die Option, verschiedene Ähnlichkeitsmetriken und Blocking anzuwenden. Das im nächsten Abschnitt behandelten Active-Learning ist nicht implementiert. Neben der Erweiterung des Pakets um solche Methoden soll in nächsten Versionen die Datenbereinigung vor dem eigentlichen Record Linkage intensiver unterstützt werden.

Verweis auf Originalarbeit (Seiten 14-20 der Habilitationsschrift):

The RecordLinkage Package: Detecting Errors in Data

Murat Sariyar, Andreas Borg

***The R Journal*. 2010; 2(2):61-67**

http://journal.r-project.org/archive/2010-2/RJournal_2010-2_Sariyar+Borg.pdf

2.2. Active-Learning-Strategien für die Deduplizierung von personenbezogenen Daten

Murat Sariyar, Andreas Borg, Klaus Pommerening

***Journal of Biomedical Informatics*. 2012; 45(5):893-900**

Überwachte maschinelle Lernverfahren benötigen gelabelte Trainingsdaten. Die Labels sind häufig nicht unmittelbar gegeben und müssen daher zumeist durch eine manuelle Sichtung der Daten gewonnen werden. Active-Learning-Strategien haben das Ziel, die Anzahl der zu sichtenden Datensätze so gering wie möglich zu halten. Dies wird erreicht, indem der Informationsgehalt von Datensätzen und die potentielle Relevanz für die anzuwendenden Methoden durch Informationsmaße wie die Entropie ermittelt werden. In dieser Arbeit wurde ein auf Entscheidungsbäumen basierendes Active-Learning-Verfahren weiterentwickelt und untersucht, wie sich dieses Verfahren auf zu deduplizierende Daten bei binären Vergleichen als auch bei Verwendung von stetigen Ähnlichkeitsmaßen verhält. Zum empirischen Vergleich der unterschiedlichen Strategien wurden fehlerbehaftete reale Datensätze aus einem medizinischen Register mit einer unterschiedlichen Anzahl an Attributen betrachtet. Die Ergebnisse zeigten, dass Active-Learning mit 9 Attributen und binären Vergleichen zu den besten Deduplizierungs-Resultaten führten. Bei Reduzierung der Anzahl der Attribute auf unter 6 führte die Nutzung von stetigen Ähnlichkeitswerten zu den besten Resultaten. In jedem der Fälle hat man nicht mehr als 200 Trainingsdaten labeln müssen, was angesichts von über 5 Mio. zu klassifizierenden Datensatzpaaren als gering erachtet werden kann. Wie sich eine einfache Active-Learning-Strategie auf andere maschinelle Lernverfahren verhält und was man berücksichtigen muss, um optimale Ergebnisse bei der Deduplizierung von Daten zu erhalten, wurde in der nachfolgenden Arbeit untersucht.

Verweis auf Originalarbeit (Seiten 22-29 der Habilitationsschrift):

Active learning strategies for the deduplication of electronic patient data using classification trees

Murat Sariyar, Andreas Borg, Klaus Pommerening

***Journal of Biomedical Informatics*. 2012; 45(5):893-900**

http://dx.doi.org/10.1016/j.jbi.2012.02.002

2.3. Bagging, Bumping, Multiview, und Active-Learning für das Record Linkage/Deduplizieren von personenbezogenen Daten

Murat Sariyar, Andreas Borg

***Computer Methods and Programs in Biomedicine*. 2012; 108(3):1160-1169**

In dieser Arbeit wurden zwei maschinelle Lernverfahren für das Record Linkage beziehungsweise Deduplizieren von Daten eingeführt: Bumping und eine spezielle Form des Multiview. Bumping steht für „bootstrap umbrella of model parameters“ und ist ein Entscheidungsbaumverfahren, das die Vorteile von Bagging hinsichtlich der Stabilisierung von Ergebnissen mit einer besseren Interpretierbarkeit verknüpft. Das hier vorgestellte Multiview-Verfahren kombiniert stochastisches Record Linkage, Support Vector Machines und Bagging, um eine Deduplizierung der Daten zu erreichen. Zusammen mit der getrennt angewendeten Bagging-Methode wurden diese beiden Verfahren in einer empirischen Evaluation auf drei unterschiedlichen Trainingsdaten miteinander verglichen. Einer der Trainingsmengen wurde dabei durch eine einfache Active-Learning-Strategie bestimmt. Die Resultate zeigen, dass die mit Active-Learning erreichte kleinste Trainingsmenge zu den besten Deduplizierungsergebnissen führte. Man erzielte mit Multiview nur auf einer zufällig gezogenen Trainingsmenge bessere Ergebnisse als die anderen Verfahren. Ansonsten führte die Verwendung von Entscheidungsbaumverfahren zu den besten Resultaten. Dabei gab es keine wesentlichen Unterschiede zwischen Bumping und Bagging. Daraus lässt sich vorsichtig ableiten, dass die Niederdimensionalität des Problems eine Verbesserung von Ergebnissen durch „ausgefeilte“ maschinelle Lernverfahren eher unwahrscheinlich macht. Die nächsten beiden Abschnitte zeigen, wie wichtig maschinelle Verfahren sind, wenn hochdimensionale und komplexere Probleme gelöst werden sollen.

Verweis auf Originalarbeit (Seiten 31-40 der Habilitationsschrift):

Bagging, bumping, multiview, and active learning for record linkage with empirical results on patient identity data

Murat Sariyar, Andreas Borg

Computer Methods and Programs in Biomedicine. 2012; 108(3):1160-1169

http://dx.doi.org/10.1016/j.cmpb.2012.08.003

2.4. Ein Boosting-Ansatz zur Anpassung der Sparsamkeit von Modellen bezüglich unterschiedlich aggregierter molekularer Daten

Murat Sariyar, Martin Schumacher, Harald Binder

***Statistical Applications in Genetics and Molecular Biology*. 2014;13(3):343-57**

Da moderne molekularbiologische Technologien immer kostengünstiger potentiell relevante Daten produzieren, ist es wichtig, Modelle zur Risikovorhersage um hochdimensionale molekulare Attribute anzureichern, beispielsweise um SNP- oder Methylierungsmessungen. Zur Gewährleistung der Interpretierbarkeit sollten die erzeugten Modelle dabei möglichst sparsam sein, also auf wenige zentrale molekulare und klinische Faktoren deuten. Die Vielzahl an molekularen Größen und die vorhandenen Aggregationsmöglichkeiten können es erforderlich machen, dass der Grad an Sparsamkeit des Modells an die jeweiligen Daten angepasst werden muss. Zudem besitzen viele Lösungen und Verfahren für hochdimensionale Daten das Problem, die wahren Effekte zu unterschätzen, gerade wenn immer die gleiche Annahme hinsichtlich der notwendigen Sparsamkeit zugrunde liegt. Aus diesem Grund wurde in dieser Arbeit ein Boostingverfahren vorgestellt, das die Sparsamkeit von Risikovorhersagemodellen in Abhängigkeit von den vorliegenden Daten automatisch anpasst. Damit kann man beide Probleme (Grad Sparsamkeit und Unterschätzung von Effekten) in einem integralen Ansatz angehen. Eine umfassende Simulationsstudie zeigte, dass dieser Ansatz die Unterschätzung der wahren Effekte reduziert, wenn hohe Sparsamkeit erforderlich ist. Außerdem führte in Szenarien, die weniger Sparsamkeit erfordern, dieser neue Ansatz zu einer adäquateren Modellgröße als das korrespondierende nicht adaptierende Boostingverfahren. Die Anwendung des Verfahrens auf ein reales Szenario mit DNA-Methylierungsdaten von Nierenkrebspatienten bestätigte die Simulationsstudie. Unterschiedliche Aggregierungsebenen führten zu einer Anpassung der Sparsamkeit des Modells, die gegenüber dem korrespondierenden nicht adaptierenden Boostingverfahren die Vorhersageleistung verbesserte und die Variablenselektion stabilisierte. Die folgende und letzte Arbeit dieser Habilitationsschrift zeigt, dass es Szenarien gibt (bspw. wenn mehr als eine Mio. potentielle Attribute zur Verfügung stehen), in denen ein Verfahren oft nicht ausreicht und daher eine Verfahrenskombination sinnvoll ist.

Verweis auf Originalarbeit (Seiten 42-56 der Habilitationsschrift):

A boosting approach for adapting the sparsity of risk prediction signatures based on different molecular levels

Murat Sariyar, Martin Schumacher, Harald Binder

***Statistical Applications in Genetics and Molecular Biology*. 2014;13(3):343-57**

http://dx.doi.org/10.1515/sagmb-2013-0050

2.5. Kombination von maschinellen Lernverfahren zur Selektion und Bewertung von Interaktionstermen in hochdimensionalen Überlebenszeitdaten

Murat Sariyar, Isabell Hoffmann, Harald Binder

***BMC Bioinformatics*. 2014; 15:58—16**

Für multivariate Risikovorhersagemodelle gibt es eine Vielzahl von Verfahren, die es erlauben, zu Schätzungen auf hochdimensionalen Daten zu gelangen. Dabei ist eines der Schlüsselprinzipien die Synchronisation von Parameterschätzung und Variablenselektion. Da Interaktionen zwischen molekularen Größen zu erwarten sind, ist es für eine realitätsnahe Modellierung wichtig, solche Interaktionen zu berücksichtigen. Eine Berücksichtigung von Interaktionen findet jedoch häufig deshalb nicht statt, weil die Anzahl an Attributen im hochdimensionalen Setting so stark ansteigt, dass eine Schätzung von Parametern immer schwieriger wird. In dieser Arbeit wird gezeigt, wie man Modellbausteine so kombinieren kann, dass eine Berücksichtigung von Interaktionen in multivariaten Risikovorhersagemodellen möglich wird, ohne alle möglichen Kombinationen von Attributen betrachten zu müssen. Es werden folgende Bausteine genutzt: (1) Resampling, (2) Random Forests und (3) Orthogonalisierung der Daten. Die Kombination dieser Modell-bausteine dient der Prä-Selektion von Interaktionen, bevor ein Boostingverfahren für die Modellschätzung zur Anwendung kommt. Zunächst wurde eine Simulationsstudie durchgeführt und gezeigt, dass alle Modellbausteine wichtig sind, um die wahren Haupt- und Interaktionseffekte zu finden. Die Resultate auf zwei realen Microarray-Überlebenszeitdaten von Krebspatienten zeigten, dass die Interaktionen gegenüber den Haupteffekten oft mit geringen Effekten assoziiert sind, obwohl sie potentiell biologisch relevant sind. Insgesamt sind die Ergebnisse vielversprechend und zeigen, dass es neben der Neuentwicklung von Verfahren auch wichtig ist, vorhandene Verfahren geeignet miteinander zu kombinieren, um neue und verbesserte Lösungen zu erhalten.

Verweis auf Originalarbeit (Seiten 58-73 der Habilitationsschrift):

Combining techniques for screening and evaluating interaction terms on high-dimensional time-to-event data

Murat Sariyar, Isabell Hoffmann, Harald Binder

***BMC Bioinformatics*. 2014; 15:58—16**

http://dx.doi.org/10.1186/1471-2105-15-58

3. Diskussion

3.1. Maschinelle Lernverfahren für das Zusammenführen und Deduplizieren von biomedizinischen Daten

Es wurde zunächst gezeigt, wie man eine Vielzahl von Algorithmen für das Record-Linkage in einem R-Paket verfügbar machen kann. Das *RecordLinkage*-Paket war dabei in erster Linie zur vergleichenden Evaluation von verschiedenen Verfahren vorgesehen, wird aber immer mehr auch für das Zusammenführen von größeren Datenmengen genutzt (Sadinle 2014; Sariyar und Borg 2010). Das Paket erleichtert den Einsatz von maschinellen Lernverfahren in der Praxis, sollte jedoch für einen das Record-Linkage umfassend begleitenden Einsatz um weitere zusätzliche Techniken und Maßnahmen ergänzt werden, beispielsweise:

- Schnittstellen zu datenhaltenden Systemen.
- Post-Linkage-Algorithmen, die für eine Gruppe von einer Entität zugeordneten Datensätzen einen deduplizierten Datensatz erzeugen. Hierzu eignen sich Algorithmen aus der linearen Programmierung (Jaro 1989).
- Prä-Linkage-Standardisierungsalgorithmen, um die zu vergleichenden Daten in einen einheitlichen Standard zu bringen. Dazu gehören auch Verfahren aus dem Bereich Ontologie-Matching, gerade wenn semantische Ambiguitäten existieren (Scharffe und Jérôme Euzenat 2011).

Des Weiteren ist es häufig wichtig, einen realistischen Goldstandard zu besitzen, um die anzuwendenden Modelle zu evaluieren. Dazu wurde eine aus dem Krebsregisterkontext stammenden Datensatz unter https://archive.ics.uci.edu/ml/datasets/Record+Linkage+Comparison+Patterns zur Verfügung gestellt.

Anschließend wurde vorgestellt, wie man Active-Learning nutzen kann, um mit einer geringen Anzahl von adäquat ausgesuchten Trainingsdaten zu guten Ergebnissen bei der fehlerminierenden Zusammenführung und dem Deduplizieren von Daten zu kommen (Sariyar, Borg und Pommerening 2012b). Hat man ausreichend viele Attribute (≥ 5) mit einem größeren Wertbereich, reichen dabei binäre Vergleiche zwischen Datensatzpaaren und damit binäre Trainingsdaten. Bei weniger Attributen zeigten die Resultate, dass die Verwendung von String-Metriken zum

Ähnlichkeitsvergleich wichtig für das Active-Learning und die angewendeten maschinellen Lernverfahren ist. Generell muss beachtet werden, dass es nicht immer möglich ist, alle Daten in einem manuellen Review zu labeln. In diesem Fall kann man beispielweise versuchen, Daten mit einem semi-überwachten Verfahren (Chapelle, Scholkopf und Zien 2010; Zhu und Goldberg 2009) zu labeln. Abschließend lässt sich bezüglich der Resultate in (Sariyar, Borg und Pommerening 2012b) sagen, dass die erarbeiteten Vorschläge zu Active-Learning-Strategien die notwendige Anzahl von zu labelnden Trainingsdaten erheblich reduzieren und auf den betrachteten Daten zu sehr guten Ergebnissen führen.

Hat man ausreichend viele und gut befüllte Attribute (das Problem von fehlenden Werten wird zum Beispiel in (Sariyar, Borg und Pommerening 2012a) behandelt), so reicht für die Trainingsdatenbestimmung bei binären Vergleichsmustern, sich Vertreter der zu den unterschiedlichen Vergleichsmustern korrespondierenden Datensatzpaare anzuschauen. Der Vergleich von verschiedenen maschinellen Lernverfahren, die auf per Active-Learning ausgewählten Trainingsdaten trainiert wurden, zeigte jedoch, dass man Vergleichsmuster mit einem ausgewogenen Mix an Gleichheits- und Ungleichheitswerten (schwer zu klassifizierende Fälle) überproportional häufig auswählen sollte. Häufig werden nur so erzeugte Modelle den schwierigen Fällen des Anwendungsszenarios gerecht (Sariyar und Borg 2012). Die Niederdimensionalität des Record-Linkage-Problems führt oft dazu, dass Ensemble-Verfahren wie Bagging, Bumping oder Multiview zu keinen besseren Resultaten führen als einfache Entscheidungsbäume. Dies führte zu dem Schluss, dass auf Basis von Active-Learning-Strategien trainierte Entscheidungsbäume zu den geeignetsten Record-Linkage-Verfahren gehören, die es derzeit gibt (vgl. auch (Sariyar, Borg und Pommerening 2009)). Eine kurz vor der Veröffentlichung stehende Arbeit macht jedoch deutlich, was passiert, wenn zu viele schwierig zu klassifizierende Fälle in die Trainingsmenge eingehen. In solchen Szenarien fehlen Entscheidungsbäumen oft die Pfade, welche zu offensichtlichen Matches oder Non-Matches führen. Eine Lösung dieses Problems besteht darin, deterministisches Linkage mit einfachen Entscheidungsregeln vorzuschalten.

3.2. Maschinelle Lernverfahren für die Analyse von hochdimensionalen biomedizinischen Daten

Um multivariable Regressionsmodelle auf hochdimensionale Daten anzupassen, ist es notwendig eine so genannte Regularisierung vorzunehmen. Dieses Konzept findet Anwendung, wenn man schlecht konditionierte Probleme vor sich hat, die durch Hinzunahme von Parametern beziehungsweise von externen Hilfsmitteln zu einem gut lösbaren Problem transformiert werden können. In der auf hochdimensionale Daten zurückgreifenden klinischen Forschung sind solche Regularisierungstechniken zentral und ermöglichen Parameterschätzungen in einem Regressionsmodell. Für das LASSO-Verfahren oder die *componentwise likelihood-based-boosting*-Methode bedeutet dies zum Beispiel eine relativ geringe Anzahl von Parameterschätzwerten ungleich Null. Die so erzeugten sparsamen Modelle deuten auf potentiell zentrale Marker (SNPs, Gene oder auch Microarray-Features), die zumeist biologisch validiert werden müssen. Ein Problem vieler maschineller Lernverfahren, auch im Hinblick auf die originäre Version des hier verwendeten *componentwise likelihood-based-boosting*-Verfahrens, besteht häufig darin, zu Unterschätzung der Effekte zu kommen. Dies tritt insbesondere dann auf, wenn die wahren Effekte als groß anzunehmen sind. Zudem will man alle bedeutsamen Einfluss-faktoren und ihre Korrelationen erfassen und nicht nur jeweils einen Repräsentanten aus einer Gruppe von korrelierten Variablen. Beide Probleme (Unterschätzung von Effekten und fehlende Inkludierung von korrelierten Variablen) werden durch unterschiedliche Ansätze und zumeist getrennt voneinander angegangen (siehe zum Beispiel (Fan, Feng und Wu 2009) oder (Yuan und Lin 2006)).

Mit AdaptiveBoost wurde ein auf *componentwise likelihood-based-boosting* basierender integrativer Ansatz vorgestellt, der beide Probleme zugleich angeht (Sariyar, Schumacher und Binder 2014). In dem schrittweise vorgehenden Ursprungsverfahren werden Koeffizientenschätzer über eine gewisse Anzahl von Boostingschritten aufgebaut, wobei in jedem Boostingschritt für jeweils eine Kovariate ein Update des Koeffizientenschätzers vollzogen wird. Wie groß jeweils ein Update sein soll, wird durch einen so genannten Step-size-Faktor bestimmt. Dieser Faktor wird in AdaptiveBoost nach jedem Update mit einem Step-size-Modifikationsfaktor so modifiziert, dass eine Anpassung an die vorliegenden Daten

hinsichtlich des angemessenen Niveaus an Sparsamkeit vorgenommen werden kann. In anderen Worten: Abhängig von der Richtung und dem Ausmaß der Modifikation, kann entweder die Unterschätzung von Koeffizientenschätzern reduziert oder ein Modell mit einer reduzierten Sparsamkeit anvisiert werden. Eine Schätzung dieses Sparsamkeitsparameters erfolgt im Rahmen einer Kreuzvalidierung zusammen mit der Bestimmung der Schrittzahl. Die Simulationsstudie im Überlebenszeit-Setting zeigte, wie gut die Sparsamkeitsanpassung an die Realität erfolgt, wenn eine variierende Anzahl von Kovariaten eine Relevanz für die Überlebenszeitvorhersage besitzt. Ist das wahre Modell selbst sparsam, so gewinnt man gegenüber dem ursprünglichen Boosting-Ansatz eine bessere Identifizierung der relevanten Variablen und eine Reduzierung der Unterschätzung des Effektenschätzers. Besteht das wahre Modell aus einer relativ hohen Anzahl an Variablen, so reduziert sich die Variabilität der Schätzer der vielen Variablen, die zu Recht ins Modell aufgenommen werden, aber die Anzahl an falsch-positiven Selektionen erhöht sich. In beiden Fällen (sparsame und nicht sparsame wahre Modelle) steigert sich die Vorhersageleistung.

Das reale Datenbeispiel mit CpG-Methylierungsdaten veranschaulicht das Verhalten von AdaptiveBoost auf unterschiedlichen Aggregationsebenen. Der automatisiert angepasste Sparsamkeitsparameter zeigte folgendes an: je höher der Aggregationsgrad, desto höher das selektierte Niveau an Sparsamkeit. Auf der granularsten molekularen Ebene (Basenniveau) wurden mehr Boostingschritte und mehr Variablen ausgewählt als im korrespondierenden ursprünglichen *componentwise likelihood-based-boosting*-Algorithmus. Auf Gen- und Chromosomenebene wurden weniger Schritte ausgesucht. Auf allen Ebenen erhöhte sich jedoch die Inklusionshäufigkeit von Variablen, wodurch sich Informationsextraktion aus den Daten stabilisierte. Eine systematische Verbesserung der Vorhersageleistung konnte nicht gezeigt werden, was sich zum Teil durch die Neuartigkeit der CpG-Methylierungsmessung (Ziller et al. 2013), die Berücksichtigung von schon viel erklärenden klinischen Kovariaten, als auch durch die Erhöhung der Anzahl an zu optimierenden Parametern, ähnlich wie beim Elastic Net oder SCAD (Benner et al. 2010), erklären lässt.

Die Berücksichtigung von Interaktionen in generalisierten linearen Modellen ist ein weiteres wichtiges Forschungsfeld im Bereich der hochdimensionalen Daten (Sariyar, Hoffmann und Binder 2014). Bei über tausenden Variablen, kann man in klassischen Modellen mit einigen hunderten Beobachtungen die Haupteffekte nicht simultan schätzen – und damit auch nicht die Interaktionen. Der hier vorgestellte integrative Ansatz für dieses Problem nutzt die gute Haupteffektselektion vom *componentwise likelihood-based-boosting* zusammen mit der Eigenschaft von Random Forests, nichtlineare Effekte und Interaktionen zu berücksichtigen (Leo Breiman 2001). Die Simulationsstudie zeigt, dass man mit Random Forests in der Tat relevante Interaktion finden kann, es aber für Interaktionen mit nur geringen marginalen Effekten einen Zwischenschritt geben muss, in welchem alle Variablen um die bis dato gefundenen Haupteffekte bereinigt werden, was einer Orthogonalisierung gleich kommt. Ohne diesen Zwischenschritt maskieren die Haupteffekte relevante Interaktionen. Zudem zeigte die Simulationsstudie, dass auch moderate Variablen-Inklusionshäufigkeiten von 10-30% noch auf relevante Interaktionen deuten können. Ob die Berücksichtigung von Interaktionen wirklich zu einer verbesserten Effektschätzung führt oder nicht, lässt sich in der Praxis häufig schwer beantworten. Dies haben auch die betrachteten realen Beispiele mit Microarraydaten gezeigt. Auch wenn auf potentiell relevante Interaktionen gedeutet wurde, ist eine biologische Validierung ein wichtiger weiterer Schritt, um richtige Fälle von falsch-positiven zu unterscheiden.

Zusammenfassend im Hinblick auf (Sariyar, Hoffmann und Binder 2014) heißt das: es wurde gezeigt, wie man 2er-Interaktionen in einem integrativen Ansatz in ein multivariates Risikovorhersagemodell integrieren kann. Dazu hat sich das Screening mittels Random Forests angeboten, welches für die Zukunft auch erlaubt, andersartige Interaktionen zu berücksichtigen. Insoweit kann (Sariyar, Hoffmann und Binder 2014) als Machbarkeitsstudie betrachtet werden. Alle drei Komponenten in dem Verfahrensvorschlag waren zentral: Subsampling, Random Forests und Orthogonalisierung. Statt Random Forests kann man auch Verfahren wie die logic regression (Kooperberg et al. 2007) nutzen; erste interne Ergebnisse dazu erscheinen vielversprechend. Das generelle Fazit lautet, dass es zusätzlich zur Etablierung von neuen Verfahren wichtig ist, die Potentiale von vorhandenen Verfahren durch geeignete Prozessierungsschritte auszureizen.

4. Zusammenfassung und Ausblick

Maschinelle Lernverfahren gewinnen in vielen Bereichen an Bedeutung, aber jeweils aus unterschiedlichen Gründen. Während für niederdimensionale und damit für relativ einfache Probleme in erster Linie verbesserte Ausnutzung der Strukturen für Effizienzgewinne beispielsweise in der Trainingsauswahl zu konstatieren ist, erlauben maschinelle Lernverfahren im hochdimensionalen Kontext überhaupt erst, für viele anstehende Probleme eine erste Lösung zu finden. Dies spiegelt sich auch in dieser Habilitationsschrift wider. Da für das Zusammenführen und Deduplizieren von Daten schon sehr früh Arbeiten darauf hindeuteten (Sariyar, Borg und Pommerening 2009), dass maschinelle Lernverfahren zu keinen wesentlich besseren Klassifikationsergebnissen führen als herkömmliche stochastische Verfahren, hat sich der Fokus auf Effizienzgewinne in der Prozessierung des Record-Linkage konzentriert. Das zeigt sich mehr oder weniger auch in den Arbeiten, die zu diesem Thema in diese Habilitationsschrift aufgenommen wurden. Auch zukünftig werden vor allem maschinelle Lernverfahren zur verbesserten Datenbereinigung vor der eigentlichen Durchführung des Record-Linkage zentrale Anwendungsbereiche in diesem Kontext sein.

Im Gegensatz dazu gibt es im Bereich der hochdimensionalen und insbesondere der molekulargenetischen Daten viele Probleme, die erst mit maschinellen Lernverfahren adressiert werden können. Diese Habilitationsschrift lag der Fokus dabei auf Problemen, die im Rahmen von generalisierten linearen Modellen formuliert werden konnten. Die Anpassung der Modellsparsamkeit und das Detektieren von Interaktionen sind wichtige Ziele der Nutzung von hochdimensionalen Daten. Gerade die Anwendung von maschinellen Lernverfahren (z.B. Random Forests für das Ziel der Interaktionsdetektion) haben sich hierbei als vielversprechend erwiesen. Nichtdestrotz gab es schon sehr früh Stimmen, die darauf hindeuteten, dass Resultate häufig nicht reproduzierbar sind (Ransohoff 2004). Dies erfordert Anstrengungen im Hinblick auf die Erhöhung der Validität und Nachvollziehbarkeit von Ergebnissen. Während es für klassische statistische Verfahren ausreichen mag, Koeffizientenschätzer und Konfidenzintervalle anzugeben, müssen gerade bei Verwendung von maschinellen Lernverfahren alle wichtigen Entscheidungen hinsichtlich der festzulegenden Parameter, der angewandten Software und der Prä-

wie auch Post-Prozessierung gut dokumentiert sein. Andernfalls hat man eine sehr schnelllebige Forschungslandschaft, in der interessante Resultate erzeugt werden, ohne zu einer Konsolidierung der Resultate und damit einer Translation in die klinische Praxis zu kommen.

Auf die wissenschaftstheoretischen Überlegungen zu Anfang der Arbeit zurückkehrend, kann man im Rückblick feststellen, dass die Methodenvielfalt im Bereich des maschinellen Lernens groß genug ist, um eine Vielzahl von Problemen zu adressieren. Sowohl eher einfache niederdimensionale als auch neuartige komplexe Probleme lassen sich gut mit den vorhandenen Verfahren angehen und lösen. Ein großer Unterschied zwischen diesen Problemfeldern besteht darin, dass für niederdimensionale Probleme oft auch Lösungen existieren, die nicht aus dem Kontext der maschinellen Lernverfahren stammen, wohingegen für viele hochdimensionale Probleme erst die Anwendung von solchen Verfahren zu ersten Lösungsmöglichkeiten führt.

Zusammenfassend lässt sich sagen: eine methodenbasierte Herangehensweise impliziert zwar eine Beschränkung auf bestimmte Probleme; ist die verwendete Verfahrensklasse jedoch vielschichtig genug, so hat solch eine Beschränkung keine relevanten Auswirkungen auf die Innovationsfähigkeit und die Qualität der Ergebnisse. Alternativen sollten jedoch nie aus den Augen verloren werden, damit die initiale Beschränkung auf eine bestimmte Methodenklasse in keine Sackgasse mündet.

Literatur

Amaratunga, Dhammika. 2014. Exploration and Analysis of DNA Microarray and Other High-Dimensional Data. Auflage: 2. Wiley.

Ananthakrishna, Rohit, Surajit Chaudhuri, and Venkatesh Ganti. 2002. “Eliminating Fuzzy Duplicates in Data Warehouses.” In Proceedings of the 28th International Conference on Very Large Data Bases, 586–97.

Arabie, Phipps, Lawrence J. Hubert, and Geert de Soete. 1996. Clustering and Classification. World Scientific.

Bell, Jason. 2014. Machine Learning: Hands-On for Developers and Technical Professionals. Auflage: 1. Auflage. Indianapolis, Ind.: John Wiley & Sons.

Benjelloun, O., H. Garcia-Molina, D. Menestrina, Q. Su, S.E. Whang, and J. Widom. 2009. “Swoosh: A Generic Approach to Entity Resolution.” VLDB Journal 18 (1): 255–76.

Benner, Axel, Manuela Zucknick, Thomas Hielscher, Carina Ittrich, and Ulrich Mansmann. 2010. “High-Dimensional Cox Models: The Choice of Penalty as Part of the Model Building Process.” Biometrical Journal. Biometrische Zeitschrift 52 (1): 50–69.

Binder, H. and M. Schumacher. 2009. “Incorporating Pathway Information into Boosting Estimation of High-Dimensional Risk Prediction Models.” BMC Bioinformatics 10.

Bishop, Christopher M. 2007. Pattern Recognition and Machine Learning. New York: Springer.

Boulesteix, Anne-Laure, and Matthias Schmid. 2014. “Machine Learning versus Statistical Modeling.” Biometrical Journal 56 (4): 588–93.

Breiman, Leo. 2001. “Random Forests.” Machine Learning 45 (1): 5–32.

Breiman, L, J Friedman, R Olshen, and C. Stone. 1984. Classification and Regression Trees. Belmont, CA: Wadsworth.

Bühlmann, Peter and Sara van de Geer. 2011. Statistics for High-Dimensional Data: Methods, Theory and Applications. Heidelberg; New York: Springer.

Burges, C.J.C. 1998. “A Tutorial on Support Vector Machines for Pattern Recognition.” Data Mining and Knowledge Discovery 2: 121–67.

Candes, Emmanuel and Terence Tao. 2007. “The Dantzig Selector: Statistical Estimation When P Is Much Larger than N.” The Annals of Statistics 35 (6): 2313–51.

Chapelle, Olivier, Bernhard Scholkopf, and Alexander Zien. 2010. Semi-Supervised Learning. Auflage: 1. Cambridge, Mass.; London: MIT Press.

Christen, Peter. 2008. “Automatic Record Linkage Using Seeded Nearest Neighbour and Support Vector Machine Classification.” In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 151–59.

Cleophas, Ton J., and Aeilko H. Zwinderman. 2013. Machine Learning in Medicine. New York: Springer.

Deng, Li and Dong Yu. 2014. Deep Learning: Methods and Applications. Now Publishers Inc.

“Dimension Reduction Techniques.” 2002. In: A Distribution-Free Theory of Nonparametric Regression, 448–58. Springer Series in Statistics. New York: Springer.

Dua, Sumeet, U. Rajendra Acharya, and Prerna Dua. 2013. Machine Learning in Healthcare Informatics: 56. Berlin; Heidelberg: Springer.

Efron, Bradley, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. 2004. “Least Angle Regression.” The Annals of Statistics 32 (2): 407–99.

Eisen, Michael B., Paul T. Spellman, Patrick O. Brown, and David Botstein. 1998. “Cluster Analysis and Display of Genome-Wide Expression Patterns.” Proceedings of the National Academy of Sciences 95 (25): 14863–68.

Fan, Jianqing, Yang Feng, and Yichao Wu. 2009. “Network Exploration via the Adaptive LASSO and SCAD Penalties.” The Annals of Applied Statistics 3 (2): 521–41.

Fan, Jianqing, and Runze Li. 2001. “Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties.” Journal of the American Statistical Association 96 (456): 1348–60.

Fellegi, I.P., and A.B. Sunter. 1969. “A Theory for Record Linkage.” Journal of the American Statistical Association 64 (328): 1183–1210.

Friedman, Jerome H. 2000. “Greedy Function Approximation: A Gradient Boosting Machine.” Annals of Statistics 29: 1189–1232.

Györfi L, Kohler M, Krzyzak A, Walk H. 2002. A Distribution-Free Theory of Nonparametric Regression. Auflage: 2. New York: Springer.

Hakimi, A. Ari, Irina Ostrovnaya, Boris Reva, Nikolaus Schultz, Ying-Bei Chen, Mithat Gonen, Han Liu, et al. 2013. “Adverse Outcomes in Clear Cell Renal Cell Carcinoma with Mutations of 3p21 Epigenetic Regulators BAP1 and SETD2: A Report by MSKCC and the KIRC TCGA Research Network.” Clinical Cancer Research: An Official Journal of the American Association for Cancer Research 19 (12): 3259–67.

Heisenberg, Werner. 1971. Physics and Beyond. Translated by A. J. Pomerans. London: HarperCollins Publishers Ltd.

Jaro, M.A. 1989. “Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa/Florida.” Journal of the American Statistical Association 89: 414–20.

Kooperberg, C., J.C. Bis, K.D. Marciante, S.R. Heckbert, T. Lumley, and B.M. Psaty. 2007. “Logic Regression for Analysis of the Association between Genetic Variation in the Renin-Angiotensin System and Myocardial Infarction or Stroke.” American Journal of Epidemiology 165 (3): 334–43.

Kuhn, T.S. 1996. The Structure of Scientific Revolutions. Auflage: 3. Chicago: University Of Chicago Press.

Liu, Y.Q., M. Zhang, R.W. Cen, L.Y. Ru, and S.P. Ma. 2007. “Data Cleansing for Web Information Retrieval Using Query Independent Features.” Journal of the American Society for Information Science and Technology 58 (12): 1884–98.

Luhmann, N. 1992. Die Wissenschaft Der Gesellschaft. Auflage: 6. Frankfurt am Main: Suhrkamp Verlag.

Luhmann, N. 1996. Social Systems. Stanford: Stanford University Press.

Mitchell, Tom. 1997. Machine Learning. New York: McGraw-Hill.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. 2012. Foundations of Machine Learning. Cambridge, MA: MIT Press.

Muslea, I., S. Minton, and C.A. Knoblock. 2006. “Active Learning with Multiple Views.” Journal of Artificial Intelligence Research 27: 203–33.

Newcombe, H.B., M.E. Fair, and P. Lalonde. 1992. “The Use of Names for Linking Personal Records.” Journal of the American Statistical Association 87 (420): 1193–1204.

Rahm, Erhard and Hong H. Do. 2000. “Data Cleaning: Problems and Current Approaches.” IEEE Data Eng. Bull. 23 (4): 3–13.

Ransohoff, David F. 2004. “Rules of Evidence for Cancer Molecular-Marker Discovery and Validation.” Nature Reviews Cancer 4 (4): 309–14.

Roessler, Jessica, Ole Ammerpohl, Jana Gutwein, Doris Steinemann, Brigitte Schlegelberger, Veronika Weyer, Murat Sariyar, et al. 2014. “The CpG Island Methylator Phenotype in Breast Cancer Is Associated with the Lobular Subtype.” Epigenomics, October, 1–13.

Rosenwald, Andreas, George Wright, Wing C. Chan, Joseph M. Connors, Elias Campo, Richard I. Fisher, Randy D. Gascoyne, et al. 2002. “The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma.” The New England Journal of Medicine 346 (25): 1937–47.

Sadinle, Mauricio. 2014. “Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach.” The Annals of Applied Statistics 8 (4): 2404–34.

Salzberg, Steven L. 1994. “C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993.” Machine Learning 16 (3): 235–40.

Sandoval, Juan, Holger Heyn, Sebastian Moran, Jordi Serra-Musach, Miguel A. Pujana, Marina Bibikova, and Manel Esteller. 2011. “Validation of a DNA Methylation Microarray for 450,000 CpG Sites in the Human Genome.” Epigenetics: Official Journal of the DNA Methylation Society 6 (6): 692–702.

Sariyar, M. and A. Borg. 2010. “The RecordLinkage Package: Detecting Errors in Data.” The R Journal 2 (2): 61–67.

Sariyar, M. and A. Borg. 2012. “Bagging, Bumping, Multiview, and Active Learning for Record Linkage with Empirical Results on Patient Identity Data.” Computer Methods and Programs in Biomedicine 108 (3): 1160–69.

Sariyar, M., A. Borg, and K. Pommerening. 2009. “Evaluation of Record Linkage Methods for Iterative Insertions.” Methods of Information in Medicine 48 (5): 429–37.

Sariyar, M., A. Borg, and K. Pommerening. 2012a. “Missing Values in Deduplication of Electronic Patient Data.” Journal of the American Medical Informatics Association: JAMIA 19 (e1): e76–82.

Sariyar, M., A. Borg, and K. Pommerening. 2012b. “Active Learning Strategies for the Deduplication of Electronic Patient Data Using Classification Trees.” Journal of Biomedical Informatics, Text Mining and Natural Language Processing in Pharmacogenomics, 45 (5): 893–900.

Sariyar, Murat, Isabell Hoffmann, and Harald Binder. 2014. “Combining Techniques for Screening and Evaluating Interaction Terms on High-Dimensional Time-to-Event Data.” BMC Bioinformatics 15 (1): 58. doi:10.1186/1471-2105-15-58.

Sariyar, Murat, Martin Schumacher, and Harald Binder. 2014. “A Boosting Approach for Adapting the Sparsity of Risk Prediction Signatures Based on Different Molecular Levels.” Statistical Applications in Genetics and Molecular Biology 13 (3): 343–57.

Scharffe, François, and Jérôme Euzenat. 2011. “Linked Data Meets Ontology Matching: Enhancing Data Linking through Ontology Alignments.” KEOD 2011: 279–84.

Schrom, John. 2014. Machine Learning for Healthcare. Auflage: 1. O’Reilly.

Settles, Burr. 2009. Active Learning Literature Survey. 1648. http://www.cs.cmu.edu/~bsettles/pub/settles.activelearning.pdf.

Shao, Jun. 1993. “Linear Model Selection by Cross-Validation.” Journal of the American Statistical Association 88 (422): 486–94.

Shmueli, Galit. 2010. “To Explain or to Predict?” Statistical Science 25 (3): 289–310.

Tejada, S., C.A. Knoblock, and S. Minton. 2001. “Learning Object Identification Rules for Information Integration.” Information Systems 26 (8): 607–33.

Tibshirani, Robert. 1994. “Regression Shrinkage and Selection Via the Lasso.” Journal of the Royal Statistical Society, Series B 58: 267–88.

Tibshirani, Robert. 2011. “Regression Shrinkage and Selection via the Lasso: A Retrospective.” Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73 (3): 273–82.

Tuia, D., M. Volpi, L. Copa, M. Kanevski, and J. Munoz-Mari. 2011. “A Survey of Active Learning Algorithms for Supervised Remote Sensing Image Classification.” IEEE Journal of Selected Topics in Signal Processing 5 (3): 606–17.

Tutz, G. and H. Binder. 2007. “Boosting Ridge Regression.” Computational Statistics & Data Analysis 51 (12): 6044–59.

Tutz, Gerhard and Harald Binder. 2006. “Generalized Additive Modeling with Implicit Variable Selection by Likelihood-Based Boosting.” Biometrics 62 (4): 961–71.

Wang, Y. Richard and Stuart E. Madnick. 1989. “The Inter-Database Instance Identification Problem in Integrating Autonomous Systems.” In Proceedings of the Fifth International Conference on Data Engineering, 46–55.

Wiering, Marco, Martijn Otterlo, and Martijn van Otterlo. 2012. Reinforcement Learning. Berlin; Heidelberg: Springer.

Xie, Huiliang and Jian Huang. 2009. “SCAD-Penalized Regression in High-Dimensional Partially Linear Models.” The Annals of Statistics 37 (2): 673–96.

Yuan, Ming, and Yi Lin. 2006. “Model Selection and Estimation in Regression with Grouped Variables.” Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68 (1): 49–67.

Zhu, X. and A. Goldberg. 2009. Introduction to Semi-Supervised Learning. Morgan & Claypool Publishers.

Ziller, Michael J., Hongcang Gu, Fabian Müller, Julie Donaghey, Linus T.-Y. Tsai, Oliver Kohlbacher, Philip L. De Jager, et al. 2013. “Charting a Dynamic DNA Methylation Landscape of the Human Genome.” Nature 500 (7463): 477–81.

Zou, Hui. 2006. “The Adaptive Lasso and Its Oracle Properties.” Journal of the American Statistical Association 101 (476): 1418–29. doi:10.1198/016214506000000735.

Zou, Hui, and Trevor Hastie. 2005. “Regularization and Variable Selection via the Elastic Net.” Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67 (2): 301–20.

Danksagung

Mein besonderer Dank gilt Herrn Prof. Tolxdorff, der mein Habilitationsvorhaben unterstützt und mir viele Hilfestellungen gegeben hat. Seine Unterstützung war zentral für das Angehen der Habilitation.

Bei Herrn Prof. Pommerening und bei Prof. Binder möchte ich mich herzlich dafür bedanken, meine wissenschaftlichen Schwerpunkte geprägt zu haben. Ich habe sehr viel von beiden gelernt und bin für das mir gebotene produktive und angenehme Klima sehr dankbar.

Weiterhin möchte ich mich bei Andreas Borg, Isabell Hoffmann und Prof. Schumacher für die reibungslose, lehrreiche und teilweise sehr intensive Zusammenarbeit bedanken. Ich habe von allen viel gelernt und bin sehr froh, mit Menschen zusammengearbeitet zu haben, die nicht nur fachlich, sondern auch persönlich prägend waren.

Mein Dank gilt auch allen Mitarbeitern am IMBEI in Mainz und vor allem der Direktorin, der guten Seele dieser Einrichtung. Es war eine wunderbare Zeit in Mainz, die an Vielseitigkeit und Erfahrungsintensität kaum übertroffen werden kann.

Dr. Krister Helbing, Dr. Elke Witt und Isabell Hoffmann möchte ich für die sehr hilfreichen Kommentare bezüglich der Ausformulierungen in dieser Habilitationsschrift danken.

Erklärung

§ 4 Abs. 3 (l) der HabOMed der Charité

Hiermit erkläre ich, dass

- weder früher noch gleichzeitig ein Habilitationsverfahren durchgeführt oder angemeldet wird bzw. wurde,

- die vorgelegte Habilitationsschrift ohne fremde Hilfe verfaßt, die beschriebenen Ergebnisse selbst gewonnen sowie die verwendeten Hilfsmittel, die Zusammenarbeit mit anderen Wissenschaftlern/Wissenschaftlerinnen und mit technischen Hilfskräften sowie die verwendete Literatur vollständig in der Habilitationsschrift angegeben wurden,

- mir die geltende Habilitationsordnung bekannt ist,

- mir die Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis bekannt ist und ich mich zur Einhaltung dieser Satzung verpflichte.

....

Datum Unterschrift