Aus dem Institut für Medizinische Biometrie und Statistik

der Universität zu Lübeck

Direktor: Univ.-Prof. Dr. rer. nat. Andreas Ziegler

Wahrscheinlichkeitsmaschinen: Konsistente maschinelle Lernverfahren zum Schätzen bedingter Wahrscheinlichkeiten

Inauguraldissertation

zur

Erlangung der Doktorwürde

der Universität zu Lübeck

- Aus der Sektion Medizin -

vorgelegt von

Jochen Kruppa

aus Uelzen

Lübeck 2013

Prüfer

1. Berichterstatter :	Univ.-Prof. Dr. rer. nat. Andreas Ziegler
2. Berichterstatter:	Priv.-Doz. Dr. phil. Hans-Jürgen Rumpf
Tag der mündlichen Prüfung:	06. August 2013

Zum Druck genehmigt: Lübeck, den 06. August 2013

- Promotionskommission der Sektion Medizin-

Inhaltsverzeichnis

Wahrscheinlichkeitsmaschinen: Konsistente maschinelle Lernverfahren zum Schätzen bedingter Wahrscheinlichkeiten

Einleitung

In der vorliegenden Niederschrift werden maschinelle Verfahren für das Schätzen bedingter Wahrscheinlichkeiten beschrieben: Wahrscheinlichkeitsmaschinen (eng. „Probability machines“). Als parametrisches Standardverfahren wird häufig die logistische Regression mit recht strengen und oft ignorierten Annahmen verwendet. Eine Alternative hierzu bieten maschinelle Verfahren, die in großem Maße im Bereich der Klassifikation und nicht für die Schätzung von Wahrscheinlichkeiten genutzt werden. Der zentrale Gedanke dieser Arbeit beruht darauf, die Wahrscheinlichkeitsschätzung als ein nichtparametrisches Schätzproblem anzusehen.

Zu Anfang der Niederschrift soll auf die Unterschiede zwischen der Klassifikation und der Wahrscheinlichkeitsschätzung eingegangen werden. Tabelle 1 zeigt die Unterschiede der Klassifikation zur Schätzung von Wahrscheinlichkeiten in den verschiedenen Anwendungsgebieten. Im Bereich der Diagnose verschiebt sich die Fragstellung von der reinen Bestimmung des Krankheitsstatus zu der Eintrittswahrscheinlichkeit für eine Person krank zu sein. Im Gegensatz zu der Klassifikation stellt sich hier die Frage, mit welcher Wahrscheinlichkeit ein Patient von einer Krankheit betroffen ist. Daher stellt sich nicht die Frage nach dem reinen Krankheitsstatus, krank oder gesund, sondern nach der Wahrscheinlichkeit krank zu sein.

Warum ist die Wahrscheinlichkeitsschätzung vorzuziehen? Zum einen liegt der Klassifizierung häufig eine Wahrscheinlichkeitsschätzung inne, daher werden Wahrscheinlichkeiten berechnet, die dann über eine Entscheidungsregel in eine Klassifizierung des Patienten nach zum Beispiel gesund oder krank umgewandelt werden. Dadurch kommt es zu einem Verlust an Information, die genutzt werden können. Es tritt das Problem der Dichotomizierung auf. Patienten, die Nahe an der Entscheidungsgrenze liegen, können durch leichte Änderungen der gemessen Variablen in die eine oder andre Klasse, gesund oder krank, fallen. In diesem Fall ist die Wahrscheinlichkeit mit mehr Informationen versehen.

Tabelle 1 Unterschiede zwischen der Klassifikation und der Wahrscheinlichkeitsschätzung in unterschiedlichen Anwendungsgebieten.

Thema	**Klassifikation**	**Wahrscheinlichkeitsschätzung**
Diagnose	Ist die Person erkrankt?	Wie hoch ist die Wahrscheinlichkeit, dass die Person erkrankt ist?
Prognose	Wird die Person in einem Jahr erkrankt sein?	Wie hoch ist die Wahrscheinlichkeit, dass die Person in einem Jahr erkrankt ist?
Therapie	Wird die Behandlung beim Patienten anschlagen?	Wie hoch ist die Wahrscheinlichkeit, dass die Behandlung beim Patienten anschlägt?
Wettervorhersage	Wird es morgen regnen?	Wie hoch ist die Regenwahrscheinlichkeit für morgen?
Kreditbewertung	Wird der Kunde den Kredit zurückzahlen?	Mit welcher Wahrscheinlichkeit wird der Kunde den Kredit zurückzahlen?

Das Standardverfahren: logistische Regression

Als ein Standardverfahren für die Wahrscheinlichkeitsschätzung wird häufig die logistische Regression und verwandte Methoden als parametrisches Verfahren verwendet. Vereinfacht wird dabei die kategoriale Zielvariable über eine $logit$-Funktion in eine stetige Zielvariable [0,1] umgewandelt. Ein Nachteil der logistischen Regression sind die teilweise strengen Annahmen. Zum einen müssen alle Variablen im Model korrekt spezifiziert, alle Interaktionen zwischen den Variablen berücksichtigt und schlussendlich das gesamte Model korrekt eingeben werden. Dies ist bei einem kleinen Datensatz mit wenigen Variablen noch machbar, ist jedoch bei genomweiten Assoziationsstudien begrenzt bis unmöglich.

Die Alternative: nichtparametrische Verfahren

Die vorrausgehend beschriebenen Nachteile der logistischen Regression und verwandten Methoden können zu verzerrten Schätzern oder einer schlechten bis falschen Vorhersage der Wahrscheinlichkeiten führen. Im ungünstigsten Fall konvergiert das logistische Model nicht und eine Wahrscheinlichkeitsschätzung ist nicht möglich. Eine Lösung stellt das maschinelle Lernen als

ein nichtparametrisches Verfahren dar. Nichtparametrische Verfahren oder „verteilungsfreie Verfahren“ unterschieden sich von parametrischen Verfahren, wie die logistische Regression, dadurch, dass die Modelstruktur oder Parameter nicht *a priori* festgelegt sind sondern aus den Daten ermittelt werden.

Im Folgenden wird die Wahrscheinlichkeitsschätzung unter Zuhilfenahme maschineller Lernverfahren beschrieben. Dabei bieten maschinelle Verfahren alle Vorteile eines nichtparametrischen Modells. Es werden keine Verteilungsannahmen an die Variablen gestellt und die Anzahl der potenziellen Variablen ist ebenfalls unbegrenzt. Schlussendlich muss kein Startmodell definiert werden um mit der Analyse zu beginnen. In allen in dieser Niederschrift vorgestellten Simulationsstudien und Anwendungen wird immer zwischen einem Trainingsdatensatz und einem Testdatensatz unterschieden. Dabei wird auf dem Trainingsdatensatz das Model aufgebaut und an dem Testdatensatz die Güte der Schätzung bewertet. Sollten keine externe oder temporale Validierungsdatensätze vorliegen, wird meist der vorliegende Datensatz in einem Verhältnis 2:1 in Training- und Testdaten aufgeteilt.

Auf dem Gebiet des maschinellen Lernens gibt es eine Breite an entwickelten Maschinen. In dieser Niederschrift stehen Zufallswälder im Fokus der Analysen. Bevor auf die Zufallswälder näher eingegangen wird, soll die Konsistenz von maschinellen Lehrverfahren im Allgemeinen und im speziellen für Zufallswälder betrachtet werden.

Konsistenz maschineller Lernverfahren

Die Konsistenz stellt die minimale Forderung an Schätzfunktionen dar. Mit steigender Fallzahl soll der Schätzer des Parameters sich dem wahren Wert annähern. Im Idealfall ist bei unendlich großer Fallzahl der Schätzer gleich dem wahren Parameter. Man unterscheidet zwischen schwacher Konsistenz wie beschrieben, der starken Konsistenz und quadratischer Konsistenz. Im Folgenden wird die Konvergenz für die Zufallswälder dargestellt auf denen der Fokus der nachfolgenden Publikationen liegt.

Im Folgenden soll eine Stichprobe von *n* Individuen mit einer dichotomen abhängigen Variablen $y_i = 1$, wenn die Person erkrankt ist und $y_i = 0$ wenn die Person gesund ist, betrachtet werden. Die Kovariablen der Person *i* werden mit $\mathbf{x}_i$ bezeichnet. Weiter soll die Wahrscheinlichkeit $P(y_i = 1|\mathbf{x}_i)$ geben der Variablen $\mathbf{x}$ geschätzt werden. Dabei gilt für $P(y_i)$ das

$$P(y_i) = \begin{cases} 1 & \text{wenn Person erkrankt ist} \\ 0 & \text{anderseits} \end{cases} \quad (1)$$

Da gilt das $p(\mathbf{x}_i) = P(y_i = 1|\mathbf{x}_i) = E(y_i = 1|\mathbf{x}_i)$ ist, ist das Problem der Wahrscheinlichkeitsschätzung identisch mit dem Schätzproblem der nichtparametrischen Regression $f(\mathbf{x}) = E(y|\mathbf{x})$. Daher wird jede Maschine, die gute Ergebnisse beim Lösen des nichtparametrischen Schätzproblems $f(\mathbf{x})$ liefert auch gute Ergebnisse bei der Schätzung der bedingten Wahrscheinlichkeit $p(\mathbf{x}_i)$ liefern. Allgemeiner gesprochen ist die Schätzung einer nichtparametrischen Regression $f(\mathbf{x})$ konsistent, wenn der mittlere quadratische Fehler zu null konvergiert, daher wenn $\lim_{n\to\infty} E\left(\hat{f}(\mathbf{x}_i) - f(\mathbf{x}_i)\right)^2$ gilt.

Im Folgenden wird die Konsistenz von Zufallswäldern näher betrachtet. Insbesondere bei großen Datensätzen, wo die Anzahl der Variablen $\mathbf{x}$ sehr viel größer ist als *n*, wie bei genomweiten Assoziationsstudien, ist es wichtig ein sparsames Model zu schätzen. Das heißt ein Model zu schätzen, welches nur die bedeutenden Variablen $\mathcal{S}$ – daher die informativen und nicht null enthaltenden – einschließt. Die Konvergenzgeschwindigkeit in Zufallswäldern hängt nur von der Anzahl der bedeutenden Variablen $\mathcal{S}$ ab und nicht von der Dimension von *x*. Unter der Annahme die informativen Variablen schnell zu erkennen, liegt die Konvergenzgeschwindigkeit bei $n^{\frac{-0.75}{\mathcal{S}\log 2+0.75}}$. Daraus ergibt sich, dass die Konvergenzrate strikt schneller ist als die gebräuchliche Minimaxrate. Dies gilt sobald $\mathcal{S} \leq \lfloor 0.54x \rfloor$ ist, wobei $\lfloor \cdot \rfloor$ den Integer Teil der Funktion beschreibt mit $x = \frac{1}{S}$. Zufallswälder lassen sich daher sparsam und konsistent zum Schätzen von bedingten Wahrscheinlichkeiten einsetzen.

Zufallswälder

Zufallswälder (eng. „Random Forests“) wurden Anfang des Jahrhunderts aus der CART Methode (eng. ***CL**assification and **R**egresssion **T**rees*) entwickelt. Es zeigte sich, dass ein Ensemble von einzelnen Bäumen („Wald“) einen großen Zugewinn an Genauigkeit in der Vorhersage brachte. Ein weiterer Wichtiger Aspekt ist der Zufall bei der Generierung der einzelnen Bäume, hierbei spielt zweifach der Zufall eine Rolle. Zum einen werden auf der Ebene der Individuen zufällig Personen ausgewählt mit denen die Bäume generiert werden. Dieses Durchmischen der Personen und zufälligem Ziehen mit Zurücklegen, wird als Bagging bezeichnet, eine implizite Ensemble-Methode. Das Wort Bagging ist ein Akronym für ***B**ootstrap **Agg**regat**ing.*** Zum ande-

ren werden auf der Ebene der Variablen zufällig einige Variablen ausgewählt um einen Baum aufzubauen. Daher setzt sich der Name Zufallswald aus dem Ensemble von einzelnen Bäumen („Wald“) und dem zufälligen Ziehen („Zufall“) von Personen und Variablen für die Genierung einzelner Bäume zusammen.

Im Weiteren werden Zufallswälder noch in zwei Kategorien anhand ihrer Regeln zum Verzweigen der einzelnen Bäume in einem Wald unterschieden. Das Ziel einer Verzweigung ist immer möglichste reine Tochterknoten zu erhalten. In einem optimalen Fall befinden sich in den Tochterknoten nach einer Verzweigung nur Kranke oder Gesunde. Dabei wird je Verzweigung ein Gütemaß berechnet, welches die Reinheit der Folgeknoten misst und die Bedeutsamkeit der für die Verzweigung herangezogenen Variable misst. Hierbei unterschiedet man zwischen Klassifikation Zufallswälder, die den Gini Index als Maßzahl nutzen und die Regression Zufallswälder, die die Verringerung des MSE (eng. „**m**ean **s**quare **e**rror“) für die Verzweigung berücksichtigen.

Wie funktioniert das Schätzen von Wahrscheinlichkeiten in Zufallswäldern? Im Zweiklassenfall, Patient erkrankt [1] oder Patient gesund [0], wird das Verhältnis der Einsen im Terminalknoten bestimmt und über alle Bäume des Ensembles gemittelt, als Wahrscheinlichkeit erkrankt zu sein, wiedergeben. Dabei ist von Bedeutung wie viele Personen schlussendlich in einen Terminalknoten fallen. Daher sollte die Größe der Terminalknoten an den Trainingsdaten optimiert werden. In dieser Niederschrift wird sich auf den Zweiklassenfall konzentriert.

Random Jungle: Eine Implementierung von Zufallswäldern

Im Rahmen der vorliegenden Arbeit wurde die Implementierung der Zufallswälder, Random Jungle um eine bedeutende Option erweitert. Die Implementierung Random Jungle zeichnet sich zum einen durch den sparsamen Speicherverbrauch und der schnellen Rechenleistung durch Parallelisierung der Rechenprozesse aus. Überdies wurde ein generelles Rahmenwerk für die weitere Integrierung von Varianten von Zufallwäldern geschaffen. Als eine zusätzliche Option wurde die in folgenden Publikationen diskutierte bedingte Wahrscheinlichkeitsschätzung in Random Jungle intrigiert. Dadurch ist es möglich die Wahrscheinlichkeiten aus Zufallswäldern auch in sehr großen Datensätzen zu schätzen und das volle Rahmenwerk von Random Jungle zu nutzen.

Probability machines: consistent probability estimation using nonparametric learning machines

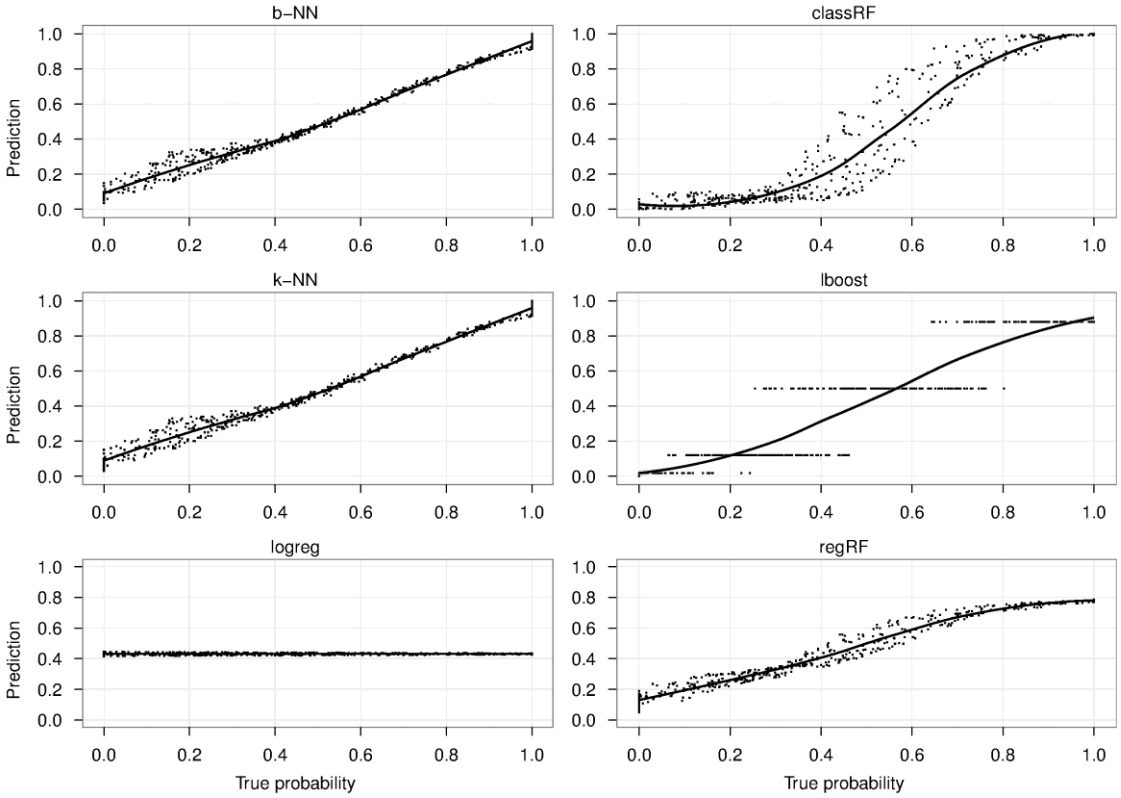
Malley, J. D., **Kruppa, J.**, Dasgupta, A., Malley, K. G. und Ziegler, A. (2012) Probability machines: consistent probability estimation using nonparametric learning machines. *Methods Inf Med* **51**, 74-81.

Methoden

Im Rahmen zweier Simulationsstudien mit Fall/Kontrollstatus wurden in zum einen verschiedene maschinelle Verfahren untereinander und gegen die logistische Regression, als Standardverfahren, verglichen. Zum einen wurden Zufallswälder (Klassifikation und Regression), k nächste Nachbarn, bagged nächste Nachbarn, der LogitBoost und die logistische Regression auf zwei simulierte Datensätze angewandt: dem Mease-Model, einem zweidimensionalen Kreismodel und dem Sonar-Model. Abschließend wurden alle Methoden auf zwei klinische Datensätze angewandt: die Diagnose von Appendicitis und den Diabetes Status von Pima Indianern. Die Vorhersagegüte wurde anhand ROC-Kurven (eng. „**r**eceiver **o**perating **c**haracteristic“), der AUC (eng. „**a**rea **u**nder the **c**urve“), dem MSE (eng. „**m**ean **s**quare **e**rror“), dem Brier Score und Hosmer-Lemeshow Abbildungen bewertet. Abschließend wurde die Konsistenz von maschinellen Lernverfahren beim Schätzen von bedingten Wahrscheinlichkeiten beschrieben.

Ergebnisse

In beiden Simulationsstudien zeigten die nächste Nachbarn Klassifikation, sowie die Zufallswälder sehr gute Ergebnisse bei Schätzen der bedingten Wahrscheinlichkeiten. Abbildung 1 zeigt die Effizienz der Methoden beim Schätzen der Wahrscheinlichkeiten für das Mease-Model. Die logistische Regression und der LogitBoost konnten keine guten Ergebnisse bei der Wahrscheinlichkeitsschätzung liefern. Ebenso waren die Brier Scores und die AUC Werte für die Zufallswälder weit besser als für die logistische Regression und dem LogitBoost. Es zeigte sich, dass die Regression Zufallswälder eine geringere Streuung zeigten als die Klassifikation Zufallswälder. Angewandt auf die klinischen Daten zeigte sich der Zufallswald als beste Methode zum Schätzen der Wahrscheinlichkeit zu erkranken, dargestellt in Tabelle 2.



Abblidung 1 Vorhergesagte (eng. „Prediction“) gegen wahre Wahrscheinlichkeit (eng. „True probability“) in dem Mease Model für alle sechs angewandten Methoden unter Verwendung der Testdaten. b-NN: bagged nächste Nachbarn, classRF: Klassifikation Zufallswald, k-NN: k nächste Nachbarn, lboost: LogitBoost, logreg: Logistische Regression, regRF: Regression Zufallswald.

Dabei schnitt der Zufallswald in beiden Diagnosedaten ähnlich gut ab. Die Regression Zufallswälder liefern zusammen mit den nächste Nachbarn Methoden die besten Brier Scores und die besten AUC Werte. Im Weiteren zeigten die Hosmer-Lemeshow Abbildungen eine sehr gute Streuung der Wahrscheinlichkeitsschätzung für große und kleine Wahrscheinlichkeiten insbesondere bei den Zufallswäldern. Dabei zeigte sich das Klassifikation Zufallswälder leicht schlechtere Wahrscheinlichkeitsschätzungen im Vergleich zu den Regression Zufallswäldern liefern. Als ein weiteres Ergebnis stellte sich heraus, dass die Methoden teilweise sehr lange für das Schätzen der bedingten Wahrscheinlichkeiten benötigten. Hieraus ergab sich die Frage, ob es eine schnellere und bessere Implementierung der Maschinen, als in der angewandten Umgebung, gibt. Insbesondere im Bezug auf die Zufallswälder, die einen großen Speicherbedarf aufgezeigt hatten.

Table 2 Fläche unter der ROC Kurve (AUC), Brier Score, und nicht parametrisches 95% bootstrap Konfidenzinterval (in Klammern) für den Appendicitis und den Pima Indian diabetes Datensatz.

	Appendicitis Daten		**Pima Indian diabetes Daten**	
	AUC	**Brier score**	**AUC**	**Brier score**
b-NN	0.847 (0.672 – 1.000)	0.102 (0.066 – 0.145)	0.819 (0.779 – 0.858)	0.180 (0.167 – 0.197)
classRF	0.931 (0.846 – 0.900)	0.075 (0.038 – 0.121)	0.952 (0.853 – 0.913)	0.163 (0.147 – 0.184)
lboost	0.976 (0.928 – 0.900)	0.043 (0.023 – 0.073)	0.863 (0.825 – 0.897)	0.173 (0.155 – 0.198)
logreg	0.853 (0.672 – 0.900)	0.088 (0.050 – 0.136)	0.839 (0.802 – 0.875)	0.160 (0.145 – 0.181)
k-NN	0.844 (0.694 – 0.969)	0.106 (0.066 – 0.149)	0.843 (0.777 – 0.855)	0.182 (0.168 – 0.199)
regRF	0.976 (0.934 – 0.982)	0.061 (0.037 – 0.088)	0.971 (0.862 – 0.919)	0.163 (0.151 – 0.179)

Schlussfolgerung

In der Publikation wird in zwei Simulationsstudien mit Fall/Kontrollstatus und zwei kleinen Beispieldaten für klinische Diagnosestudien die konsistente Schätzung von Wahrscheinlichkeiten gezeigt. Der zentrale Gedanke der Niederschrift, dass das Schätzen der Wahrscheinlichkeiten durch maschinelle Lernverfahren als ein Problem des Schätzens von Wahrscheinlichkeiten durch nichtparametrische Verfahren angesehen werden kann, wird erläutert. In den Simulationsstudien und der Auswertung der klinischen Datensätzen stellen sich Zufallswälder als eine konsistente und effiziente Möglichkeit heraus bedingte Wahrscheinlichkeiten zu schätzen. Im Besonderen zeigte der Regression Zufallswald eine bessere Wahrscheinlichkeitsschätzung als der Klassifikation Zufallswald. In der folgenden Publikation werden daher verschiedene Implementierungen von Zufallswäldern betrachtet, mit dem Ziel die Ergebnisse dieser Veröffentlichung auch auf größere Daten, wie genomweiten Assoziationsstudien, effizient und speicherschonend anzuwenden.

Overview of Random Forest methodology and practical guidance with emphasis on computational biology and bioinformatics

Boulesteix, A.-L., Janitza, S., **Kruppa, J.** und König I. R. (2012) Overview of Random Forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *WIREs Data Mining Knowl Discov* **2**, 493-507.

Methoden

In der vorherigen Publikation wurden Zufallswälder als eine effiziente Methode zum Schätzen bedingter Wahrscheinlichkeiten beschrieben. In der vorliegenden Übersichtsarbeit werden verschiedene Implementierungen von Zufallswäldern beschrieben und die unterschiedlichen Parameter und Limitierungen von Zufallswäldern diskutiert und näher erläutert. Im Besonderen wird der Algorithmus des Zufallswaldes beschrieben und erläutert. Überdies wird eine praktische Anleitung gegeben für die Wahl der Parameter in einem Zufallswald. Eine ausführliche Besprechung der Implementierungen zeigt die Vorteile und Nachteile einzelner Programme auf und gibt abschließend Empfehlungen für zwei Implementierungen. Zwei Beispiele für die Anwendung mit Beispielcode von Zufallswäldern runden diesen Überblick ab. In einer Literaturübersicht wird auf die bisherigen Anwendungen von Zufallswäldern eingegangen und das breite Anwendungsgebiet beschrieben.

Ergebnisse

Zufallswälder werden in einem breiten Spektrum der Wissenschaft angewandt: von der Bioinformatik über biomedizinischen Anwendungen zu genomweiten Studien und Expressionsstudien bis hin zu ökologischen Fragestellungen. Dennoch werden bis heute Zufallswälder zu großen Teilen fast ausschließlich auf dem Gebiet der Klassifikation genutzt. Insbesondere die Erweiterung der Zufallswälder auf dem Gebiet der Wahrscheinlichkeitsschätzung war theoretisch möglich, zeigte aber Limitierungen und verzerrte Schätzer. Probleme des Algorithmus, wie fehlende Werte, lassen sich durch importieren von Werten lösen. Stark korrelierte Einflussvariablen lassen sich durch bedingte Bedeutsamkeitsmaße gewichten und so deren verzerrende Einfluss auf Schätzer verringern.

Zufallswälder werden durch verschiedene Parameter gesteuert. Zum einen die Anzahl an zu wachsenden Bäumen, diese sollte je größer sein, desto mehr Variablen in den Daten vorhanden sind. Die Größe der Terminalknoten ist im Besonderen bei der Wahrscheinlichkeitsschätzung von Bedeutung und sollte an den Trainingsdaten optimiert werden. Ebenso die Anzahl an ausgewählten Variablen je Verzweigung sollte optimiert sein. Abschließend können stark nicht balancierte Daten zu verzerrten Schätzern führen, dies kann über Upsampling und/oder Downsampling Methoden behoben werden. In Tabelle 2 sind aktuelle Implementierungen von Zufallswäldern dargestellt.

Tabelle 3 Überblick über Implementierungen von Zufallswäldern. Nur RF: drückt aus ob das Programm nur für Zufallswälder geschrieben wurde (ja) oder Teil eines größeren Softwarepaketes ist (nein). MT: Paralleles Rechnen möglich.

Name	Nur RF	MT	System	Code
ALGLIB	Nein	Nein	Win/Unix	C++
cforest function in R	Nein	Ja	All	C++/S
FastRandomForest	Ja	Ja	All[a]	Java
Orange	Nein	Nein	Win/Unix/Mac	C++/Python
PARF - Parallel RF Algo-	Ja	Ja	All[a]	F90
Random forest	Ja	Nein	All[a]	F77
Randomforest-matlab	Ja	Nein	All[a]	C/C++
randomForest-R package	Nein	Ja	All	C++/S
Random Jungle	Ja	Ja[a]	Win/Unix	C++
RT-Rank	Ja	Ja	Unix[a]	C++/Python
Waffles	Nein	Nein	Win/Unix/Mac	C++
Weka	Nein	Nein	All[a]	Java

[a]Ein Kompiler wird vorrausgesetzt

[b]Nur für UNIX Rechner vorhanden

Es zeigte sich, dass viele Implementierungen nur klassifizieren konnten und somit die erfolgreiche Regressionsimplementierung von Zufallswäldern nicht anwendbar war. Im Weiteren waren auch viele Implementierungen in anderen Softwarepaketen integriert und hatten nicht den vollen

Parameterumfang zu Verfügung. Einige Implementierungen wurden von den Entwicklern modifiziert ohne die genauen Modifikationen zu benennen.

Bei der abschließenden Empfehlung stellte sich das Programm Random Jungle als eine leistungsstarke Implementierung der Zufallswälder heraus, welche ausreichend dokumentiert ist. Insbesondere die Möglichkeit speicherschonend und parallel Daten zu verarbeiten, erlaubt die Anwendung von Random Jungle auf große Datensätze wie genomweite Assoziationsstudien. Die Möglichkeit bedingte Wahrscheinlichkeiten zu schätzen war in der dargestellten Version noch nicht möglich, wurde aber im Rahmen der Niederschrift in das Programm implementiert.

Schlussfolgerung

Die Publikation zeigt einen Überblick über die Implementierungen von Zufallswäldern. Verschiedenste Implementierungen von Zufallswäldern wurden aufgezeigt und diskutiert. Abschließend wurde eine Empfehlung wird für die R Implementierung `randomForest` und die eigenständige Implementierung Random Jungle gegeben. Beispielcode für beide Verfahren wurde vorgestellt und das breite Anwendungsgebiet von Zufallswäldern diskutiert. Es zeigte sich, dass Zufallswälder im Bereich der Klassifikation und weniger zu Wahrscheinlichkeitsschätzung genutzt werden. Auch wurden die Limitierungen, wie zum Beispiel korrelierte Daten und fehlende Werte aufgezeigt und Lösungen, wie die bedingten Bedeutsamkeitsmaße und das Importieren von fehlenden Werten beschrieben. Es zeigte sich, dass große Datensätze sehr effizient und mit geringer Rechenzeit von Random Jungle analysiert werden können. Insbesondere liefert Random Jungle für genomweite Assoziationsstudien eine einzigartige Möglichkeit des speicherschonenden Wachstums von Zufallswäldern. Es lässt sich eine eindeutige Empfehlung für Random Jungle bei der Analyse von genomweite Assoziationsstudien aussprechen.

Risk estimation and risk prediction using machine-learning methods

Kruppa, J., Ziegler, A. und König, I. R. (2012) Risk estimation and risk prediction using machine-learning methods. *Hum Genet* **131**, 1639-1654.

Methoden

Diese abschließende Publikation enthält einen systematischen Literaturüberblick über maschinelle Lernverfahren im Kontext von genomweiten Assoziationsstudien. Die Erkenntnisse aus den vorherigen Publikationen werden nun auf größere Datensätze angewandt. Ebenso wird der Aufbau einer Regel für die Klassifikation oder das Schätzen einer Wahrscheinlichkeit, deren Evaluierung und Validierung vorgestellt. Im Weiteren werden Regression Zufallswälder auf einen Beispieldatensatz angewandt zum Thema Rheumatische Arthritis. Ein Ziel beim Konstruieren einer Vorhersageregel ist es ein möglichst sparsames Modell zu erhalten. Ein statistisches Modell ist sparsam, wenn es nach Möglichkeit nur die informativen Variablen aus den Daten erhält, die für eine Vorhersage der Daten notwendig sind. Ein sparsames Zufallswaldmodel kann durch eine Rückwärtsselektierung, bei der immer nur ein bestimmter Prozentsatz von bedeutenden Variablen behalten wird, ermöglicht werden.

Überdies zeigte der Datensatz viele fehlende Werte, was bei einer genomweiten Assoziationsstudie nicht ungewöhnlich ist. Die fehlenden Werte wurden über einen gängigen Importierungsalogrithmus unter Berücksichtigung des Kopplungsungleichgewichts ergänzt. Ebenso zeigten sich stark korrelierte Marker, die durch das bedingte Bedeutsamkeitsmaß beim Wachstum der Bäume gewichtet wurden. Im Weiteren wurden die Daten auf Populationsstratifikation untersucht und abweichende Individuen gefiltert. Auf der Ebene der Marker wurden monomorphe und Marker mit einer Abweichung vom Hardy-Weinberg-Gleichgewicht ausgeschlossen.

Verschiedene Methoden wurden angewandt um die Wahrscheinlichkeit zu Schätzen an rheumatischer Arthritis erkrankt zu sein. Dabei wurden gängige Methoden im Besonderen mit den Regression Zufallswäldern verglichen. Zum einen wurden die Risikoallele über alle Marker bei jeder Person gezählt (Allele count) und zum anderen die gewichteten Marker aus einer Einzelmarkeranalyse (logOR) bestimmt. Als eine weitere Methode wurde Lasso (eng. „**l**east **a**bsolute **s**hrinkage and **s**election **o**perator“), die logistische Regression (LogReg) und die Re-

gression Zufallswälder in der Random Jungle Implementierung (RJ-Reg) betrachtet. Dabei stellte sich heraus, dass nur Lasso und Zufallswälder in der Lage waren bei großer Markeranzahl eine Wahrscheinlichkeitsschätzung zu liefern. Die gängigen Methoden konnten bei großer Markeranzahl keine Wahrscheinlichkeiten schätzen.

Ergebnisse

Tabelle 1 zeigt die Ergebnisse der systematischen Literatursuche. Dabei wurden 509 Artikel identifiziert und nach dem Bewerten der Zusammenfassungen und der Titel schließlich 115 relevante Artikel evaluiert. Es zeigte sich das der Median der Marker per Studie bei 39 Marker lag. Es wurden also bisher sehr kleine Datensätze untersucht. In den folgenden Analysen stellten große Markersets auch Probleme an die gängigen Methoden und waren teilweise nicht mehr im angemessenen Zeitrahmen zu berechnen. Nur zwei Artikel aus der Literatursuche beschrieben die genomweite Analyse mit maschinellen Lernverfahren und konzentrierten sich auf die Klassifikation. Zusammenfassend lässt sich schließen, dass viele Artikel nicht klar zwischen Klassifikation und Wahrscheinlichkeitsschätzung formal unterscheiden.

Tabelle 1 Ergebnisse von der PubMed Suche unter ncbi.nlm.nih.gov/sites/entrez?db=PubMed am 1. September 2011.

Suchterm	**# Treffer**
genome wide association machine learning	41
genome wide association random forest	15
genome wide association support vector	55
genome wide association boost*	24
genome wide association neural network	10
genome wide association logic regression	2
genome wide association MDR	15
snps machine learning	120
snps random forest	35
snps support vector	246
snps boost*	37
snps neural network	51
snps logic regression	21

In Abbildung 4 finden sich die ROC Kurven von der Wahrscheinlichkeitsschätzung in vier ausgewählten Markersets. Zum einen konnten alle beschriebenen Methoden nur auf die kleinste Subgruppe angewendet werden, was nur ~0.01% der gesamten Marker entspricht. Insbesondere die logistische Regression als das Standardverfahren konnte nur in der kleinsten Subgruppe eine Wahrscheinlichkeitsschätzung liefern. Mit steigender Markeranzahl ist nur die Random Jungle Implementierung in der Lage gute Ergebnisse zu liefern. In allen dargestellten Marker Sets sind Zufallswälder den anderen Methoden zu mindestens numerisch überlegen. Die Ergebnisse der ROC Kurven decken sich ebenfalls für die Brier Scores (Abbildung 5). Auch hier zeigt sich, dass Random Jungle der Lasso Methode überlegen war.

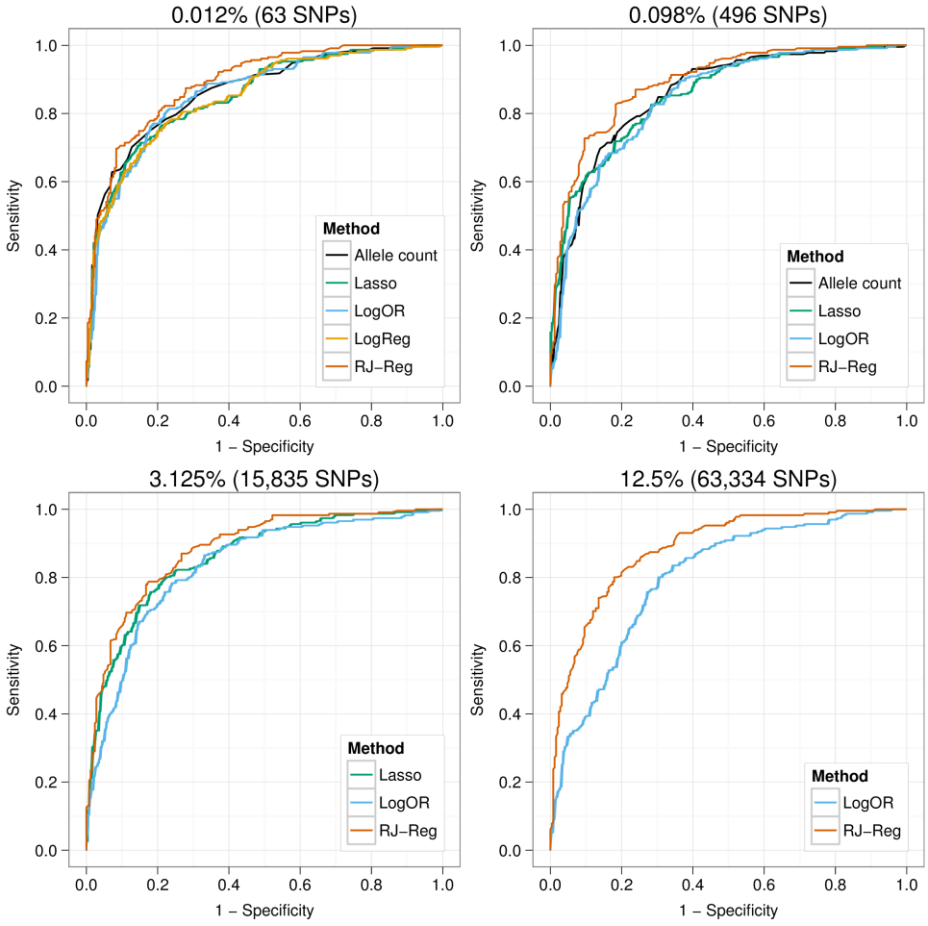


Abbildung 4 ROC Kurven für alle Methoden in ausgewählten Markersets. Allele Count: Risikoallele über alle Marker bei jeder Person, Lasso: least absolute shrinkage and selection operator, LogOR: gewichtete Marker aus einer Einzelmarkeranalyse, LogReg: Logistische Regression, RJ-Reg: Random Jungle Regression Zufallswälder.

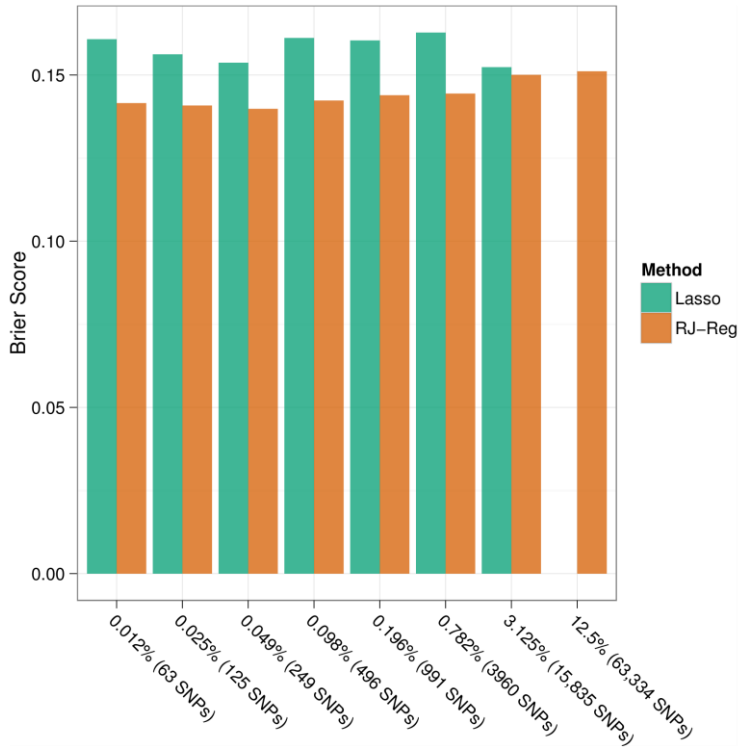


Abbildung 5 Brier Scores für Lasso und Random Jungle Regression Zufallswälder für die Test Daten.

Schlussfolgerung

In der abschließenden Publikation wu merkrden Zufallswälder unter zur Hilfe der Implementierung Random Jungle verwendet um genomweite Assoziationsstudien auszuwerten. Um die großen Datenmengen zu verarbeiten und ein sparsames Model zu erreichen werden die Datenmengen rekursive in kleinere Datensätze, gewichtet nach den bedingten Bedeutungsmaßen, aufgeteilt. Es konnte gezeigt werden, dass Zufallswälder den anderen vorgestellten Verfahren in allen SNP Sets zu mindestens numerisch überlegen waren. Teilweise waren Zufallswälder auf große SNP Sets das einzige, in vertretbarer Rechenzeit, anwendbares Verfahren. Ebenso wurde eine umfangreiche Literaturübersicht über maschinelle Lernverfahren im Kontext von genomweiten Assoziationsstudien dargestellt. Es stellte sich heraus, dass maschinelle Lernverfahren nur für Klassifikation und nicht für die Wahrscheinlichkeitsschätzung genutzt werden. Abschließend werden Regeln für das Aufstellen für die Klassifikation und Wahrscheinlichkeitsschätzung erläutert und diskutiert.

Zusammenfassung

In der einzureichenden Niederschrift werden verschiedene maschinelle Lernverfahren für die bedingte Wahrscheinlichkeitsschätzung vorgestellt. Insbesondere werden Regression Zufallswälder betrachtet.

In einer Simulationsstudie mit Fall/Kontrollstatus und zwei einfachen Beispieldaten die konsistente Schätzung von Wahrscheinlichkeiten gezeigt. Der zentrale Gedanke der Niederschrift, dass das Schätzen der Wahrscheinlichkeiten durch maschinelle Lernverfahren als ein Problem des Schätzens von Wahrscheinlichkeiten durch nichtparametrische Verfahren angesehen werden kann, wird erläutert. Dabei stellten sich Zufallswälder als eine konsistente und effiziente Methode zum Schätzen von bedingten Wahrscheinlichkeiten heraus. Im Folgenden wurden verschiedene Implementierungen von Zufallswäldern betrachtet, um die Ergebnisse der Arbeit auch auf größere Daten anzuwenden.

Verschiedene Implementierungen von Zufallswäldern wurden diskutiert und eine Empfehlung die eigenständige Implementierung Random Jungle gegeben. Es zeigte sich, dass Zufallswälder mehr zur Klassifikation und weniger zu Wahrscheinlichkeitsschätzung genutzt wurden. Auch wurden die Limitierungen, wie zum Beispiel korrelierte Daten und fehlende Werte aufgezeigt und Lösungsansätze wie die bedingten Bedeutsamkeitsmaße und das Importieren von fehlenden Werten beschrieben. Insbesondere die Möglichkeit, Random Jungle speicherschonend auch für sehr große Datensätze anzuwenden, gab dieser Implementierung für die weitere Arbeit den Vorrang. Random Jungle wurde im Rahmen der Niederschrift um die Möglichkeit der bedingten Wahrscheinlichkeitsschätzung erweitert.

Abschließend wurden Zufallswälder unter zur Hilfe der erweiterten Implementierung Random Jungle verwendet, um genomweite Assoziationsstudien (GWA). Um die großen Datenmengen von GWA-Studien zu verarbeiten und ein sparsames Model zu erreichen, wurden die Datenmengen rekursiv in kleinere Datensätze, gewichtet nach der bedingten Bedeutungsmaßen, aufgeteilt. Es konnte gezeigt werden, dass Regression Zufallswälder den anderen vorgestellten Verfahren in allen Variablen-Sets zumindest numerisch überlegen waren. Ebenso wird eine umfangreiche Literaturübersicht über maschinelle Lernverfahren im Kontext von GWA-Studien dargestellt. Abschließend wurden Regeln für das Aufstellen für die Klassifikation und Wahrscheinlichkeitsschätzung erläutert und diskutiert.

Anhang

Danksagung

Mein Dank gilt zu allererst Herrn Univ.-Prof. Dr. rer. nat. A. Ziegler als Doktorvater der vorliegenden Arbeit für die Bereitstellung des spannenden und herausfordernden Themas, umfassender wissenschaftlicher Betreuung, hilfreicher Diskussionen und konstruktiver Kritik. Weiterer Dank gehört Frau Univ.-Prof. Dr. rer. biol. hum. I. R. König für umfangreiche wissenschaftliche Unterstützung, fast uneingeschränkter Bereitschaft der offenen Tür und zur Diskussion mannigfaltiger Probleme.

Im Weiteren möchte ich allen Co-Autoren, der in dieser Niederschrift vorgestellten Artikel für ihre ausgezeichnete und interessante wissenschaftliche Zusammenarbeit danken. Insbesondere möchte ich J. D. Malley, Ph.D. für ausgiebige und sehr lehrreiche Diskussion über die Zufallswälder danken.

Weiter danke ich Herrn Dr. rer. hum. biol. A. Schillert für die technische und wissenschaftliche Unterstützung beim Programmieren und für die formelle und technische Umsetzung dieser Niederschrift.

Und was wäre diese Arbeit ohne die kurzweiligen Diskussionen und Spaziergänge? Daher mein herzlicher Dank an Dipl. Math. C. Loley und Dipl. Math. T. Holste.

Abschließend möchte ich meiner Ehefrau Olga danken, ohne deren Unterstützung und Geduld diese Arbeit in dieser Art niemals zustande gekommen wäre.

Lebenslauf

Zur Person

Jochen Kruppa

Dornbreite 7

23556 Lübeck

geb. am 23.05.1983 in Uelzen

Ausbildung

10/2009 – 03/2013	Promotionsvorhaben zum Doktor der Humanbiologie, Universität zu Lübeck
09/2010	Master of Sience Pflanzenbiotechnologie (Abschluss mit Auszeichnung)
10/2007 – 09/2009	Studium der Pflanzentechnologie im Master mit dem Schwerpunkt Biostatistik
10/2004 – 10/2007	Studium der Pflanzenbiotechnologie im Bachelor mit dem Schwerpunkt Quantitative Genetik
06/1997 – 06/2003	Schulausbildung zur allgemeinen Hochschulreife, Herzog Ernst Gymnasium Uelzen

Zivildienst

07/2003 – 04/2004	Zivildienst, Upmeier Altenpflegeheim, Wrestedt

Berufliche Tätigkeit

seit 10/2009	Wissenschaftlicher Mitarbeiter, Institut für Medizinische Biometrie und Statistik, Universität zu Lübeck mit dem Schwerpunkten genomweiten Assoziationsstudien, Expressionsstudien und maschinellem Lernen.

Publikationsverzeichnis

Zeitschriftenartikel

1. Boulesteix, A.-L., Janitza, S., **Kruppa, J.** und König I. R. (2012) Overview of Random Forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *WIREs Data Mining Knowl Discov* **2**, 493-507.

2. **Kruppa, J.**, Ziegler, A. und König, I. R. (2012) Risk estimation and risk prediction using machine-learning methods. *Hum Genet* **131**, 1639-1654.

3. Malley, J. D., **Kruppa, J.**, Dasgupta, A., Malley, K. G. und Ziegler, A. (2012) Probability machines: consistent probability estimation using nonparametric learning machines. *Methods Inf Med* **51**, 74-81.

4. Schäfer, A., Emmert, S., **Kruppa, J.**, Schubert, S., Tzvetkov, M., Mössner, R., Reich, K., Berking, C., Volkenandt, M., Pföhler, C., Schön, M. P., Vogt, T., König, I. R. und Reichrath, J. (2012) No association of vitamin D metabolism-related polymorphisms and melanoma risk as well as melanoma prognosis: a case-control study. *Arch Dermatol Res* **304**, 353-361.

5. Schurmann, C.*, Heim, K.*, Schillert, A.*, Blankenberg, S., Carstensen, M., Dörr, M., Endlich, K., Felix, S. B., Gieger, C., Grallert, H., Herder, C., Hoffmann, W., Homuth, G., Illig, T. **Kruppa, J.**, Meitinger, T., Müller, C., Nauck, M., Peters, A., Rettig, R., Roden, R, Strauch, K., Völker, U., Völzke, H., Wahl, S,. Wallaschofski, H., Wild, P.S., Zeller, T., Teumer, A.*, Prokisch, H.* und Ziegler, A.* (2012) Analyzing Illumina gene expression microarray data from different tissues: Methodological aspects of data analysis in the MetaXpress consortium. *PLOS ONE* **7**, e50938.

Kongressbeiträge

Publizierte Kurzfassungen

1. **Kruppa, J.**, König, I. R. und Ziegler, A. (2012) Using Random Forests for consistent probability estimation in whole genome association studies. *European Mathematical Genetics Meeting*, Göttingen. *Ann Hum Genet* **76,** 410 – 433.

2. **Kruppa, J.**, König, I.R und Ziegler, A. (2012) Using Random Forests for consistent probability estimation in whole genome association studies. *Abstracts from the annual meeting of the international genetic epidemiology society*, Portland, USA. *Genet Epidemiol* **36**, 1098-2272.

Vorträge

3. **Kruppa, J.** und Ziegler, A. (2011) Using R and Random Jungle for probability estimation. *CEN*, Zürich, Schweiz.

4. **Kruppa, J.** und König, I. R. (2012) Statistik in der Lehre – Mediziner für Biometrie begeistern? *IBS AG Workshop „Neue Konzepte und Ideen zur Biometrie in der Lehre“*, Heidelberg.

5. **Kruppa, J.**, König, I. R. und Ziegler, A. (2012) Using Random Forests for probabilty estimation. *58. Biometrisches Kolloquium*, Berlin.

6. **Kruppa, J.** und Ziegler, A. (2012) Probability estimation for personalized medicine using machine learning methods. *57. GMDS / Informatik*, Braunschweig.

7. Ziegler, A., **Kruppa, J.** und König, I. R. (2012) Probability machines: estimating individual probabilities using machine learning methods. *36. GfKl*, Hildesheim.

8. Ziegler, A., **Kruppa, J.** und König, I. R. (2012) Probability machines: estimating individual probabilities using machine learning methods. Seoul, Südkorea.

9. Müller, C., Schurmann, C., Heim, K., Schillert, A., Herder, C., **Kruppa, J.**, Homuth, G., Meitinger, T., Carstensen, M., Peters, A., Illig, T., Wild, P. S., Blankenberg, S., Roden, M., Teumer, A., Prokisch, H., Ziegler, A., Felix, S., Völker, U. und Zeller, T. (2012) MetaXpress: Populations-basierte Transkriptomanalysen von HypertonieParametern. *5. Deutscher Atherosklerosekongress*, München.

Poster

10. Müller, C., Schurmann, C., Heim, K., Schillert, A., Herder, C., **Kruppa, J.**, Homuth, G., Meitinger, T., Carstensen, M., Peters, A., Illig, T., Wild, P. S., Blankenberg, S., Roden, M., Teumer, A., Prokisch, H., Ziegler, A., Felix, S., Völker, U. und Zeller, T. (2012) MetaXpress: Populations-basierte Transkriptomanalysen von HypertonieParametern. *5th NGFN-Plus*, Heidelberg.

11. Schurmann, C., Heim, K., Schillert, A., Blankenberg, S., Carstensen, M., Dörr, M., Endlich, K., Felix, S. B., Gieger, C., Grallert, H., Herder, C., Hoffmann, W., Homuth, G., Illig, T. **Kruppa, J.**, Meitinger, T., Müller, C., Nauck, M., Peters, A., Rettig, R., Roden, R, Strauch, K.,,Völker, U., Völzke, H., Wahl, S,. Wallaschofski, H., Wild, P.S., Zeller, T., Teumer, A., Prokisch, H. und Ziegler, A. (2012) Analyzing Illumina gene expression microarray data from different tissues: Methodological aspects of data analysis in the MetaXpress consortium. *5th NGFN-Plus*, Heidelberg.