Technical report from Multidimensional Image Processing, IWR, University of Heidelberg
http://klimt.iwr.uni-heidelberg.de/mip/

Automatische Lokalisation von Tumoren in ^{1}H-NMR-spektroskopischen *in vivo* Aufnahmen

M.Sc. Dipl.-Ing. Michael Kelm, M.Sc. Dipl.-Phys. Björn Menze, Prof. Dr. Fred Hamprecht, Universität Heidelberg

Kurzfassung:

Die *in vivo* NMR-spektroskopische Bildgebung (MRSI) ist ein mit der Magnetresonanz-Tomographie (MRT) verwandtes Aufnahmeverfahren welches für die Tumordiagnostik eingesetzt werden kann. Dennoch erfährt die MRSI in der klinischen Praxis bisher nur eine geringe Nutzung, da die manuelle Analyse der Daten aufgrund ihres Volumens und ihrer Dimension schwierig und aufwendig ist. Statt eines Bildvolumens mit skalaren Grauwerten wird bei der MRSI in jedem Voxel ein mindestens 512-dimensionales Spektrum aufgenommen. Das sehr geringe Signal-zu-Rausch Verhältnis (SNR) bei *in vivo* Aufnahmen erschwert die automatisierte Datenanalyse und hat zur Folge, dass traditionelle Verfahren wie die Resonanzlinienquantifizierung plötzlich und dramatisch scheitern. Wir untersuchen daher die Anwendung robuster Methoden der Mustererkennung für die vollautomatische Vorverarbeitung von *in vivo* MRSI Daten. Im Gegensatz zu bisher vorgeschlagenen Methoden bestimmen wir nicht nur lokale Tumorwahrscheinlichkeiten sondern bewerten außerdem die Zuverlässigkeit der gemachten Angaben. Erst damit wird eine effiziente Integration in die klinische Routine ermöglicht.

1. Einleitung

Die *in vivo* NMR-Spektroskopie (nuclear magnetic resonance) ermöglicht die nichtinvasive Aufnahme spektraler Bilder, welche Aufschluss über das lokale Konzentrationverhältnis bestimmter Metaboliten geben. Da dieses in Tumoren verändert ist, lassen sich NMR-spektroskopische Bilder prinzipiell für die Detektion und Lokalisierung von Tumoren nutzen, insbesondere in Fällen in denen die gewöhnliche Bildgebung eine pathologische Gewebs-veränderung nicht von einer Verletzung unterscheiden kann, wie z.B. nach erfolgter Operation.

Leider eignet sich die MRSI nicht für einen routinemäßigen Einsatz im klinischen Alltag, da der anfallende Datenstrom bisher nur mit sehr hohem Aufwand und Expertenwissen ausgewertet werden kann. Trotz der technologischen Verfügbarkeit der MRS in den gängigen MR Tomographen kann das erhebliche Potenzial der MRS für die routinemäßige Diagnostik somit nicht genutzt werden. Erst eine *vollautomatische* Vorverarbeitung der Daten ermöglicht die effiziente Auswertung der spektralen Bilder und macht die NMR-spektroskopische Bildgebung sowohl für die Diagnose als auch Therapie attraktiv. Für die klinische Praxis von besonderer Bedeutung ist dabei die Zuverlässigkeit der gewonnen Informationen bzw. eine Aussage über die Sicherheit, mit welcher die spektralen Daten ausgewertet wurden. Neben der

routinemäßigen Überwachung diagnostizierter bzw. operierter Tumore, kann die MRSI dann auch z.B. in der Bestrahlungsplanung eingesetzt werden.
Ziel unserer Arbeit ist es, die Detektion, Lokalisation und Gradierung von Hirntumoren und Prostatakarzinomen vollautomatisch vorzunehmen. Dazu setzen wir statistische Methoden aus den Bereichen Mustererkennung und Machine Learning ein, um lokale Tumorwahrscheinlichkeiten anzugeben und die Zuverlässigkeit der Wahrscheinlichkeitsaussagen zu bewerten.

2. Daten und experimenteller Aufbau

Von unserem klinischen Partner wurden uns ^{1}H-NMR-spektrale Aufnahmen der Prostata zur Verfügung gestellt, welche im Augenblick Gegenstand medizinischer Studien und damit von besonderem Interesse sind. Die spektralen Bilder wurden auf einem klinischen 1.5 Tesla Scanner (Siemens MAGNETOM Sonata) mit einer Standard CSI Sequenz aufgenommen [6]. Für jeden der insgesamt 36 Patienten wurden Spektren in einem Volumen von 16x16x16 Voxeln nebst den üblichen T2-gewichteten MR Bildern akquiriert. Bei 12 dieser Patienten waren die Aufnahmen aufgrund technischer Probleme so schlecht, dass diese vom Radiologen nicht diagnostisch ausgewertet werden konnten. Diese Patienten wurden in der vorliegenden Studie nicht berücksichtigt.
Von einigen Patienten lagen histopathologische Schnitte nach radikaler Prostatektomie vor. Diese konnten für qualitative Vergleiche als ‚Gold Standard' genutzt werden. Als quantitative Grundlage unserer Untersuchungen wurden die Ergebnisse einer standardmäßigen, semimanuellen Auswertung der spektralen Daten herangezogen [5, 7, 10]. Basis sind die relativen Konzentrationsverhältnisse von Cholin (Cho), Kreatin (Cr) und Citrat (Ci). Insgesamt wurden so 76 Schichten mit jeweils 256 Voxeln hinsichtlich der Klassenzugehörigkeit (gesund, unentschieden, krank) und der Datenqualität (nicht auswertbar, schlecht, gut) manuell klassifiziert, wobei für die Beurteilung der Datenqualität sowohl niedrige SNRs als auch Artefakte berücksichtigt wurden. Tabelle 1 gibt eine Übersicht des Datensatzes aus insgesamt 19456 Spektren. Für die Auswertung wurden in dieser Arbeit nur die Spektren mit guter und schlechter Qualität herangezogen.

Tabelle 1: Verteilung der Labels im Prostata-Datensatz (76 Schichten von 24 Patienten).

Qualität\Klasse	gesund	unentschieden	krank	insgesamt
nicht auswertbar	-	-	-	15268
schlecht	721	437	284	1442
gut	1665	629	452	2746
insgesamt	2386	1066	736	19456

2. Methoden

2.1 Resonanzlinienquantifizierung

Die übliche Vorgehensweise bei medizinischen Studien baut auf der Resonanzlinienquantifizierung auf [8]. Dazu wird das komplexwertige NMR Signal üblicherweise als Summe von K Modellkomponenten modelliert:

$$\hat{y}(t_n) = \sum_{k=1}^{K} a_k \exp[i\phi_k] \exp\left[-d_k t_n + i\omega_k t_n - g_k t_n^2\right] \quad (1)$$

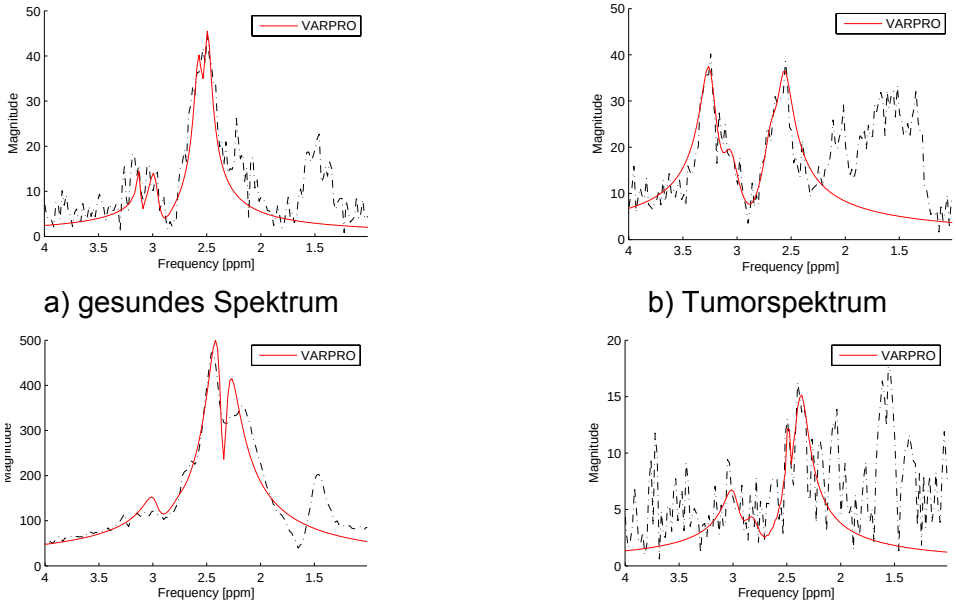
Jede einzelne Komponenten ist also eine Voigtkurve, die für $d_k = 0$ zur Gauss- bzw. für $g_k = 0$ zur Lorentzkurve wird. Die Parameter werden mittels nichtlinearer least squares Methoden im Zeitbereich geschätzt, d.h. es wird das quadratische Fehlerfunktional minimiert.

Im Falle von Lorentzmodellen können effiziente Algorithmen wie HSVD (Hankel Singular Value Decomposition) oder LPSVD (Linear Prediction SVD) verwendet werden um die optimalen Parameter zu bestimmen. Allerdings lässt sich damit Vorwissen über die zu erwartenden Resonanzsignale nur schwer berücksichtigen. Iterative Methoden wie z.B. VARPRO (Variable Projection) und AMARES bieten hinsichtlich dessen weit mehr Möglichkeiten und erlauben auch beliebige Kombinationen verschiedener Modellkomponenten. Damit eignen sich HSVD/LPSVD insbesondere für Fälle in denen wenig über die zu beschreibenden Signalkomponenten bekannt ist und VARPRO/AMARES für eine exakte und robuste Quantifizierung unter Einbeziehung von viel Vorwissen. Auf so quantifizierten Signalen werden in medizinischen Studien dann empirische Regeln aufgestellt, welche karzinogenes von normalem Gewebe unterscheiden sollen. Üblich ist dabei insbesondere die Orientierung an Konzentrationsverhältnissen wie Cho+Cr zu Ci, aber auch andere Ratios finden hier Verwendung [5].

Nachteil dieser Vorgehensweise ist, dass die Qualität der vorangegangenen Quantifizierung in die diagnostische Entscheidung in keinster Weise einfließt. Die Quantifizierung versagt aber zwangsläufig bei Spektren, welche aufgrund zu starken Rauschens oder aufgrund von Artefakten unbrauchbar sind. Dass solche Spektren bei der MRSI sehr häufig vorkommen ist aus Tabelle 1 ersichtlicht. In Bild 1 c) und d) sind zwei Beispiele dargestellt in denen VARPRO versagt. In c) wird ein Übersteuerungsartefakt gefunden, welches an den überhöhten Magnituden und in dem hier nicht abgebildeten zugehörigen MR T2-Bild zu identifizieren ist, wohingegen in d) ein Signal in reinem Rauschen gefunden wird.

Sicherlich könnte nun versucht werden, alle Arten von auftretenden Artefakten zu kategorisieren und deren Natur zu ergründen, um danach Regeln zu erstellen, wann und in welchem Maße einer darauf basierenden diagnostischen Entscheidung Glauben geschenkt werden kann. Zum einen wird dies vermutlich zu einem sehr komplexen Entscheidungsbaum mit vielen zu optimierenden Parametern führen und zum anderen wird man keinerlei Garantie für die optimale Wahl der elementaren Entscheidungsregeln haben. Einfacher ist es, das Diagnoseproblem, nämlich Merkmalsextraktion (denn nichts anderes stellt die Resonanzlinien-

quantifizierung dar) und Klassifikation, als Gesamtproblem zu betrachten und direkt von den spektralen Daten mittels statistischer Verfahren auf die Klassenzugehörigkeit und deren Vertrauenswürdigkeit schließen. Dabei legen wir besonderen Wert auf die Robustheit und Transparenz der verwendeten Verfahren, da diese ja Grundlage für therapeutische Entscheidungen sein sollen.



a) gesundes Spektrum b) Tumorspektrum

c) Artefakt d) Rauschen

Bild 1: Vier Beispiele für Magnitudenspektren nebst robuster VARPRO Quantifizierung aus einer einzigen MRSI Schicht; a) gesundes Spektrum, b) krank, c) mit Artefakt, d) verrauscht.

2.2 Generalized Linear Models und Partial Least Squares

In vorangegangenen Studien [3] konnte für Single Voxel Spektren aus dem Gehirn gezeigt werden, dass für die Klassifikation qualitativ guter Spektren nichtlineare Klassifikatoren wie z.B. neuronale Netze, Support Vector Machines oder Random Forests keinen Verbesserung gegenüber regularisierten linearen Methoden bringen. Vielmehr verliert man beim Einsatz nichtlinearer Verfahren in der Regel die Möglichkeit einer durchsichtigen Interpretation. So lassen sich Plausibilitätsprüfungen, die ja gerade für die medizinische Anwendung sehr wichtig sind, nur schwer durchführen. Daher beschränken wir uns im Folgenden auf die Betrachtung linearer Methoden, wobei insbesondere der Partial Least Squares (PLS) [9] Ansatz betrachtet werden soll. Diese ursprünglich für gewöhnliche lineare Regression entworfene Methode wurde in [2] für Generalized Linear Models (GLM) [4] verallgemeinert. GLMs bieten den Vorteil dass damit u.a. auch die logistische Regression durchgeführt werden kann. Im Allgemeinen werden GLMs über zwei Zusammenhänge definiert. Zum einen besteht zwischen erklärter Variable Y und erklärenden Variablen X ein deterministischer Zusammenhang mittels der Linkfunktion G(.) und zum anderen ist Y|X eine Zufallsvariable, welche einer Verteilung aus der exponentiellen Familie gehorcht:

$$\eta = G(E[Y \mid X]) = X\beta$$
$$f(Y \mid \eta, \psi) = \exp\left[\frac{y\,d(\eta) - b(\eta)}{a(\psi)} + c(y, \Psi)\right] \quad (2)$$

In diesen Modellen können die Parameter β aus einem Datensatz mit n Beobachtungen von Paaren (Y, X) mittels des Iteratively Reweighted Least Squares (IRLS) Algorithmus effizient geschätzt werden. Darüber hinaus kann für GLMs die Verteilung des Maximum Likelihood Schätzers unter zuhilfnahme des zentralen Grenzwertsatzes bestimmt werden, womit auch Vertrauensintervalle für die geschätzten Parameter β berechnet werden können.

Die Partial Least Squares Methode versucht nun latente Variablen $Z=C^TX$ als Linearkombination von erklärenden Variablen X so zu konstruieren, dass möglichst wenige dieser Variablen ausreichen um Y zu erklären. Die Methode ist mit der bekannteren Principal Component Analysís (PCA) eng verwandt und wird zur Regularisierung eingesetzt. Des Weiteren liefert die PLS aber auch die Richtungen im Merkmalsraum X ("loadings", PCA Hauptkomponenten), welche die wichtigsten diskriminierenden Muster darstellen.

Ergebnisse und Diskussion

In allen Spektren wurden zunächst die ungewollten Signale wie Reste der Wasser- und Lipid-/ Laktatresonanzen entfernt. Für die Klassifikation mittels Magnitudenspektren wurde der interessierende Bereich zwischen 2,3ppm und 3,4ppm ausgeschnitten wohingegen für die Klassifikation mittels Quantifizierung Cho und Cr jeweils mit einer Lorentzkurve und Citrat mit zwei Lorentzkurven modelliert wurden. Die Modellparameter wurden mittels VARPRO geschätzt wobei die relativen Phasen aller Resonanzen gleich gesetzt und auch die Dämpfung und die maximale Frequenzverschiebung eingeschränkt wurden um maximale Robustheit zu erreichen.

Alle numerischen Ergebnisse wurden 8-fach kreuzvalidiert [1], wobei die 8 Gruppen so aufgeteilt wurden, dass alle Spektren eines Patienten in derselben Gruppe landen. Damit wird verhindert, dass wegen der erheblichen Korrelation von Spektren desselben Patienten die geschätzte Fehlerstatistik aufgrund von "overfitting" überoptimistisch wird.

In der ersten Reihe von Bild 2 sind die Hauptmuster gezeigt, wie sie von der PLS Methode identifiziert wurden. Die erste und wichtigste Komponente bildet eine Differenz zwischen den Cho+Cr Resonanzen und der Ci Resonanz. Spektrale Muster, welche mit dieser Komponente stark korrelieren weisen nur noch Cho+Cr auf (zweite Reihe, links) und die Muster, welche am stärksten antikorreliert sind weisen eine ausgeprägte Citratresonanz auf (dritte Reihe). Die erste Komponente ist semantisch daher mit dem quantifizierten Cho+Cr/Ci Verhältnis der üblichen Vorgehensweise zu vergleichen. Der zweiten Komponente wird weniger aber immer noch beachtliches Gewicht für die Klassifizierung zugerechnet. Sie lässt sich am ehesten mit Baseline-Effekten in verbindung bringen. Die dritte Komponente ist eindeutig als Komponente zu identifizieren, die eine Frequenzverschiebung der Daten beschreibt. Eine solche Verschiebung ist zumeist auf Inhomogenitäten des Magnetfelds bei der Aufnahme

zurückzuführen und macht daher auch keine Aussagen über die Klassenzugehörigkeit. Die PLS misst dieser sowie der letzten Komponente so gut wie kein diskriminierendes Gewicht bei. Es handelt sich also lediglich um Variationen im X Raum, wie sie auch mittels PCA identifiziert worden wären.

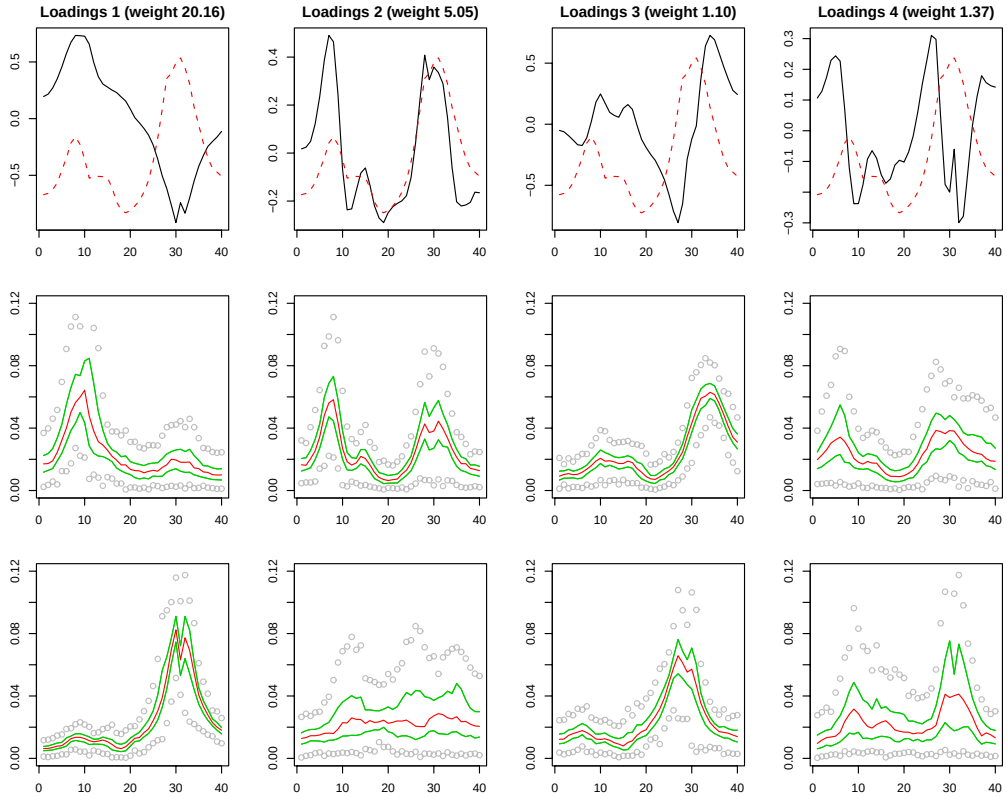


Bild 2: Die für die Diskriminierung zwischen gesunden und Tumorspektren wichtigsten Komponenten, wie sie mit der PLS Methode gefunden wurden. Die erste Zeile zeigt die „loadings“, die zweite die oberen 5% und die dritte Zeile die unteren 5% der Spektren entsprechend ihrer „score”.

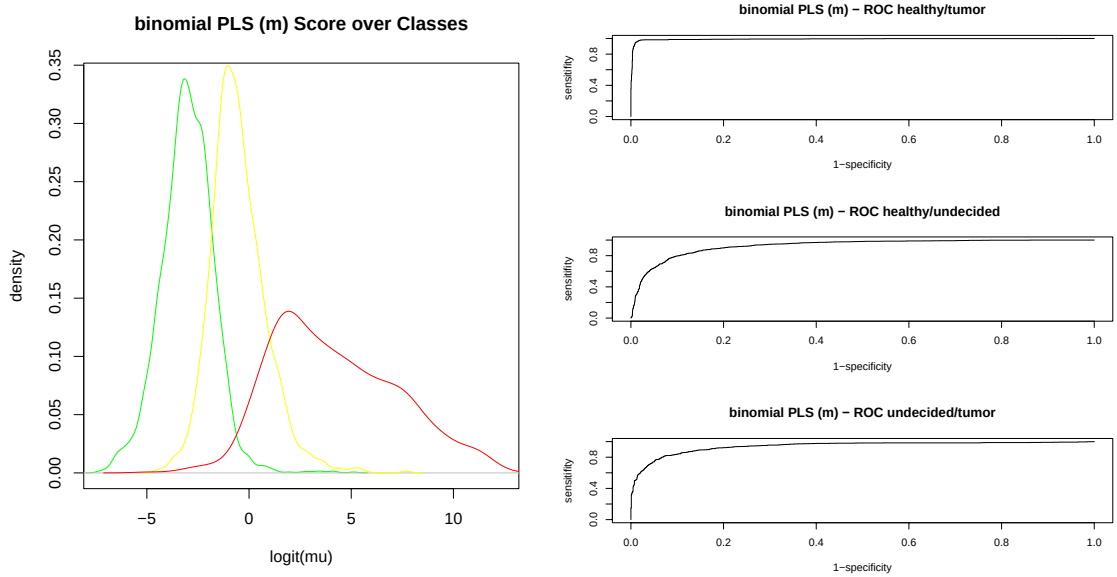


Bild 3: Links - Verteilung der „scores“ für die drei Klassen gesund (links), unentschieden (Mitte) und krank (rechts). Rechts – Die zugehörigen ROC Kurven.

Ein übliches Maß für den Vergleich verschiedener Klassifikatoren ist die Fläche unter der ROC (Receiver Operator Characteristic), welche die Trennbarkeit zweier Klassen bezüglich einer bestimmten Statistik quantifiziert [1]. In Bild 3 ist links die Verteilung der geschätzten Tumorwahrscheinlichkeiten (bzw. deren logit) nach tatsächlicher Klassenzugehörigkeit getrennt aufgetragen. Zu jedem Entscheidungsschwellwert dieser Statistik kann die Rate der Richtignegativen (Spezifität) und der Richtigpositiven (Sensitivität) bestimmt werden. Diese werden in der ROC gegeneinander aufgetragen, weshalb die Fläche unter der ROC, die AUC (Area under Curve), ein Maß für den Überlapp der beiden verglichenen Klassen angibt und sich somit ausgezeichnet für den Vergleich von Klassifikatoren eignet.

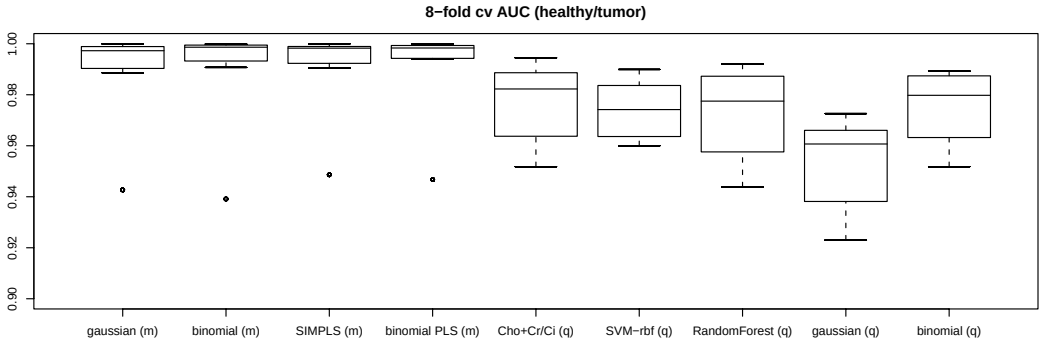


Bild 3: 8-fach kreuzvalidierte AUC für die getesteten Methoden. Die vier Klassifikatoren links benutzen das Magnitudenspektrum direkt, die rechten fünf bauen auf quantifizierten Resonanzlinien auf.

In Bild 3 ist ein Box-And-Whisker Diagramm der 8-fach kreuzvalidierten AUCs für verschiedene lineare Klassifikatoren auf den Magnitudenspektren und der Vergleich zu Klassifikationsergebnissen auf den quantifizierten Resonanzlinien dargestellt. Es ist zu erkennen, dass sowohl der Median als auch die Quartile der Ergebnisse unter Verwendung von Magnitudenspektren besser sind. Insbesondere die geringere Ausdehnung der Quartilen (ähnlich der Varianz) ist ein Hinweis darauf, dass der Verzicht auf die Quantifizierung zu erhöhter Robustheit führt, d.h. es gibt weniger Ausreißer. Des Weiteren erbringt die Verwendung der logistischen Regression (binomial) gegenüber der linearen (gaussian) auf den quantifizierten Spektren einen deutlicheren Vorteil als auf den Magnitudenspektren. Selbst die nichtlinearen Klassifikatoren RandomForest und SVM-rbf, die nach aktuellem Stand der Forschung zu den besten nichtlinearen Verfahren zählen, konnten in dieser Studie keine dem Cho+Cr/Ci Verhältnis bzw. der logistischen Regression überlegene Ergebnisse liefern. Sogar die völlig unregularisierten linearen Methoden auf den Magnitudenspektren (ganz links, gaussian/binomial) erbringen eine bessere Leistung.

Zusammenfassung und Ausblick

Es konnte gezeigt werden, dass der Verzicht auf die Resonanzlinienquantifizierung bei der diagnostischen Analyse von ^{1}H-NMR-spektroskopischen Aufnahmen der Prostata Vorteile sowohl bezüglich des Klassifikationsfehlers als auch bezüglich der Robustheit bringt. Anhand

der PLS wurde demonstriert, dass die Verwendung von Magnitudenspektren zusammen mit einer linearen Methode leicht interpretierbar ist und in keinster Weise eine „Blackbox“ darstellt. Beide Eigenschaften sind wichtig für die medizinische Anwendung, da die ausgewerteten NMR-Signale Grundlage therapeutischer Entscheidungen sein sollen. Die vorgestellten Methoden erlauben im Rahmen der GLMs die Angabe von Tumorwahrscheinlichkeiten (logistische Regression) und von Vertrauensintervallen bei neu klassifizierten Beispielen.

Die Ergebnisse bestätigen die Resultate in [3], wo aufgrund des sehr viel kleineren Datensatzes mit Aufnahmen des Gehirns gezeigt werden konnte, dass die Quantifizierung zumindest keine Vorteile bringt.

Die hier vorgestellten Verfahren wurden nur auf Spektren hoher Signalqualität durchgeführt. Zukünftige Methoden sollen über die Auswertung vorausgewählter Spektren hinausgehen und die automatische Erkennung und Abtrennung nicht auswertbarer Spektren vornehmen. Erste Einsätze nichtlinearer Verfahren zeigten hierbei bereits gute Ergebnisse.

Literatur

[1] T. Hastie et al. The elements of statistical learning. Springer, New York, 2001.

[2] B. D. Marx. Iteratively reweighted partial least squares estimation for generalized linear regression, Technometrics, 38(4): 374-381, 1996.

[3] B. H. Menze et al. Classification of in vivo magnetic resonance spectra. In Proceedings of the GfKl 2004. Springer, 2005.

[4] M. Müller. Generalized Linear Models. In: Handbook of Computational Statistics (Volume I). Concepts and Fundamentals, Springer-Verlag, Heidelberg, 2004.

[5] S. M. Noworolski et al. Dynamic contrast-enhanced MRI in Normal and Abnormal Prostate Tissues as Defined by Biopsy, MRI, and 3D MRSI. Magnetic Resonance in Medicine, 53(2):249-55, Feb 2005.

[6] T. W. Scheenen et al. Fast acquisition-weighted three-dimensional proton MR spectroscopic imaging of the human prostate. Magnetic Resonance in Medicine, 52(1):80-88, July 2004.

[7] P. Swindel et al. Pathologic characterization of human prostate tissue with proton MR spectroscopy, Radiology, 228(1):144-51, Jul 2003.

[8] L. Vanhamme. Advanced time-domain methods for nuclear magnetic resonance spectroscopy data analysis. PhD thesis, Nov 1999.

[9] H. Wold. Partial Least Squares. in Samuel Kotz and Norman L. Johnson, eds., Encyclopedia of Statistical Sciences, 6:581-91, New York: Wiley, 1985.

[10] K. L. Zakian et al. Transition zone prostate cancer: metabolic characteristics at ^{1}H MR spectroscopic imaging – initial results. Radiology, 229(1):241-7, Oct 2003.