Classification of *in vivo* magnetic resonance spectra

Björn H. Menze[1], Michael Wormit[2], Peter Bachert[2], Matthias Lichy[2,3], Heinz-Peter Schlemmer[2,3], and Fred A. Hamprecht[1]

[1] Multidimensionale Bildverarbeitung,
Interdisziplinäres Zentrum für Wissenschaftliches Rechnen (IWR),
Universität Heidelberg, 69120 Heidelberg, Germany
[2] Deutsches Krebsforschungszentrum (dkfz), 69120 Heidelberg, Germany
[3] Radiologische Diagnostik, Universitätsklinik, 72076 Tübingen, Germany

Abstract. We present the results of a systematic and quantitative comparison of methods from pattern recognition for the analysis of clinical magnetic resonance spectra. The medical question being addressed is the detection of brain tumor. In this application we find regularized linear methods to be superior to more flexible methods such as support vector machines, neural networks or random forests. The best preprocessing method for our spectral data is a smoothing and subsampling approach.

1 Introduction

The use of magnetic resonance (MR) is a well established and widespread standard in medical imaging. Less known is the use of magnetic resonance spectroscopy (MRS) for the *in vivo* analysis of the cell metabolism.
Metabolites evoke a specific spectral pattern, which is characteristic for a number of tissue types. Changes in this spectral signature allow for a diagnosis of certain pathophysiologies.

For the extraction of diagnostic information a further processing of the data is indispensable. Due to their high potential of automation, we focus on methods of *pattern recognition* and *machine learning*.

Our aim is the detection of recurrent tumors after radiotherapy by means of MRS. In this application standard imaging methods usually fail. Remaining brain lesions cannot be diagnosed reliably based on the intensity images provided by ordinary imaging methods. Intracranial biopsies, being considered as *gold standard*, do not guarantee a fully reliable result either and are associated with a considerable lethal risk of up to one percent. As a consequence, biopsies are not applicable in routine follow-up examinations and any other complementary information such as the one inherent to MRS signals is desirable (Howe and Opstad (2003)). While nearly all clinical MR scanners are able to acquire MR spectra, the know-how for interpretating these data is still rare among radiologists. A reliable automated method has the potential of making MRS accessible to a wider group of clinical practitioners.

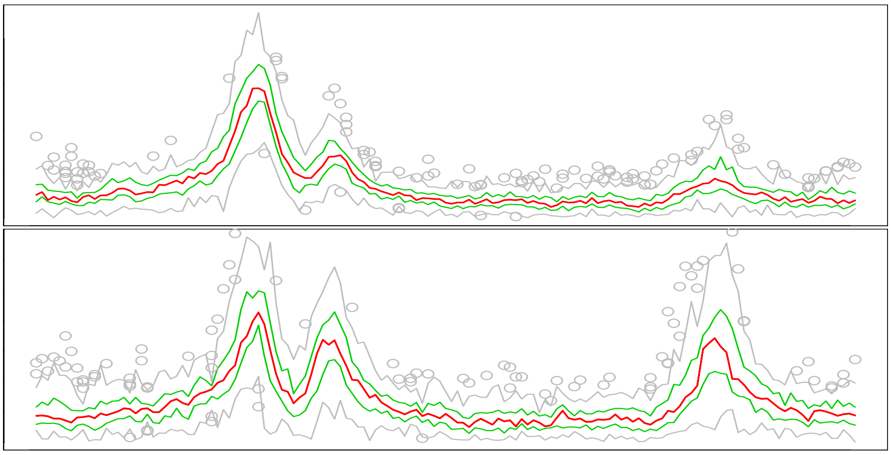


Fig. 1. Spectral pattern. Top: tumor group, Bottom: non-tumor group; red: median, green: quartiles, grey: outlier

2 Data

2.1 General features

In vivo magnetic resonance spectra share a number of features with other spectral data. A high correlation of the P spectral channels is typical for this kind of data. As a consequence, the intrinsic dimensionality is low. In our case, only three to five resonance lines are observable. As the acquisition time is limited in clinical practice, the signal-to-noise-ratio is poor. Some spectra are additionally corrupted by technical artifacts or uneven baselines. Generally, as in most medical studies of this kind, the number of observations N is much smaller than the number of explanatory variables P. In our data set, we have $N = 58$ and $P \leq 350$, depending on the preprocessing.

2.2 Details

The data set used in our survey comes from a retrospective study on the use of MRS in the evaluation of suspicious brain lesions after stereotactic radiotherapy (Schlemmer et al. (2001)). The spectra were acquired on a 1.5 Tesla MR Scanner at the German cancer research center (dkfz), Heidelberg, with long echo time (TE=135ms) by single-voxel-MRS sequences.

The study comprises a total of 58 spectra, recorded from 56 patients in a time span of several years. (Two patients participated twice in the study, at different time points.) The spectra fall into two classes: 30 of them stem from recurrent tumors, the remaining 28 from non-tumorous brain lesions. The final assignment of a spectrum to either of these classes was confirmed by clinical follow-up examinations.

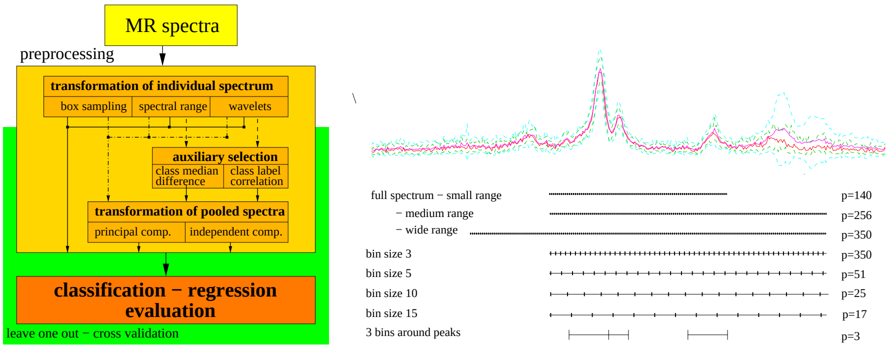


Fig. 2. Left: Flowchart of the different algorithms in the preprocessing step. All transformations involving the whole data set (such as the selection of the best ranked parameters from the auxiliary selection step) were part of the cross-validation process. Right: Plot of the tumor group, with an indication of the spectral regions chosen for the corresponding preprocessings, number of parameters P. Wavelet transformations were performed on the medium spectral range data set. Colors: magenta/blue – mean/variance, red/green – median/quartiles

Apart from filtering out the water resonance line and normalizing to the cumulative sum from the spectral region of the three most prominent peaks (choline, creatine and NAA resonances), no further preprocessing was performed on the absolute MR spectra.

3 Methods

We divide the algorithms applied to the data in two groups: preprocessing algorithms and classification algorithms. The algorithms in the first group disregard any class label information, while the latter use the group membership as an integral part.

Conceptually, the process of pattern recognition is often described as a sequence of feature extraction, feature selection and subsequent classification. For the purpose of feature extraction, the data are often transformed to a new space, the basis of which can be chosen independently of the data (as in wavelets) or dependent on the pooled observations (such as independent component analysis). Feature selection can either be explicit, as in a univariate preselection step, or implicit in the final classification. In feature extraction, a number of optional preprocessings were evaluated in a combinatorial way (cf. Fig. 2) In the end, we had about 50 differently preprocessed representations of the initial data set to evaluate the classifiers and regression methods on.

The question how to properly evaluate a benchmark of different classifiers on a small data set is yet unanswered, and hence the assessment of our results is anything but straightforward. Even for our limited problem it is not clear, how to deduce general results from a wide range of combinations of preprocessings and classifiers without getting trapped in overfitting or overmodelling.

3.1 Evaluated algorithms

For the preprocessing, we applied a number of transformations individually to each spectrum. Some of these preprocessing algorithms were as simple as discarding certain spectral ranges. The resulting data sets varied only in the number of spectral channels and the number of peaks visible in the graph of the spectral pattern. We also performed a smoothing and subsampling operation, called binning: it entails an accumulation of all the values from a certain number of adjacent spectral channels within a bin of predefined width, e.g. 3, 5, 10 or 15 channels. Another standard procedure in spectral preprocessing is to accumulate the spectral parameters into a single value within certain predefined spectral regions, e.g. around single resonance lines. For this we defined three bins around the three most prominent peaks. In addition, spectra were expressed in a dyadic wavelet basis (notably Daubechies-4 wavelets). Finally, continuous wavelets and wavelet packages were evaluated as possible preprocessing steps (for an overview see Fig. 2).

Also, there were transformations as adapted from the full data set. Principal component analysis (PCA) was used in conjunction with a follow-up regression step (principal component regression), while independent component analysis (ICA) was optionally performed on all data sets obtained from other preprocessing steps.

If necessary, an auxiliary selection was applied beforehand, in order to reduce the number of variables P approximately to the number of samples N. A ranking was performed according to the class label difference or the class label correlation of the single variables. In particular, it was optionally applied to the wavelet transformed data and the medium and wide range spectral vector data.

In the classification step, we evaluated fourteen different classification or regression methods. For the latter, a threshold was adjusted to obtain a binary result from the predicted values. Standard classifiers under study were *linear discriminant analysis* (Rao's LDA) and *k-nearest-neighbours* (knn). The first was also applied as stepwise LDA, with a f-value criterium. Regression methods using pooled data information were *principal component regression* (PCR) and *partial least squares* (PLS). Besides ordinary *multivariate linear regression* and *logistic regression* we used regularized multivariate linear regression methods, namely: *ridge regression* (here: being equivalent to *penalized discriminant analysis*), the *lasso*, *least angle regression* (LARS) and *forward selection.* From the classifiers, we evaluated *support vector machines* (using radial basis functions), feed-forward *neural networks* (nnet) and *random forests.* The hyperparameter of most of these methods was varied from

$\lambda = 1..12$ (see Table 1). Support vector machines, neural networks and random forests were evaluated with parameters varied in a grid search according to (Meyer et al. (2003)).

All computations were performed using the R computing language. For the algorithms mentioned above, we used the implementations available from the CRAN R repository (cran.r-project.org).

3.2 Benchmark settings

The optimization of real-world problems is rarely amenable to a one-stage-solution. A good representation of the problem in a low-level description usually has to be found in a first step, in order to obtain the desired high-level information in the following.

Proper benchmarking of different classifiers, even on one single data set of adequate size, is still an open question and subject to debate (Hothorn et al. (2003)). It is even more difficult to obtain results from an evaluation of processes that are composed of two essential parts.

Considering the small size of the data set, a naive selection of the best classifiers will easily result in overfitting or -modelling in spite of the cross-validation. So, having the size of the data set in mind, quantitative values should not be taken too literally and only allow for conclusions of a qualitative nature.

Within the given parameter range of each method, we assessed the classification error using the leave-one-out cross-validation. For the regression methods, we also evaluated the area under the receiver operator characteristic (ROC AuC), and the area under the precision-recall curve (PR AuC) as a performance measure.

The performance of each algorithm was measured using the best value obtained within the parameter range under study. This is probably overly optimistic, but can be understood in the light of the intrinsically low dimensional binary classification problem (see also Fig. 3). If possible (e.g. for the regression methods) top performing classifiers were checked for dimensionality and spectral interpretability.

To get a rough comparison of different preprocessing paradigms, we have pooled results as follows: at first we determined a subset of well performing classifiers (PCR, PLS and ridge regression showed to have the best overall performance with respect to our three measures). Then, we determined their optimal hyperparameters for each preprocessing scheme (binning with different sizes, wavelets, etc., cf. Fig.2). The number of correct and incorrect predictions in the leave-one-out cross-validation were evaluated for each method and interpreted as realizations from identical Bernoulli distributions. For each regression method, a Binomial distribution was fitted to these outcomes. Samples were drawn from the three distributions and concatenated into one list. Finally, the distribution of values in this list was visualized by the box-and-whisker plots in Fig. 3.

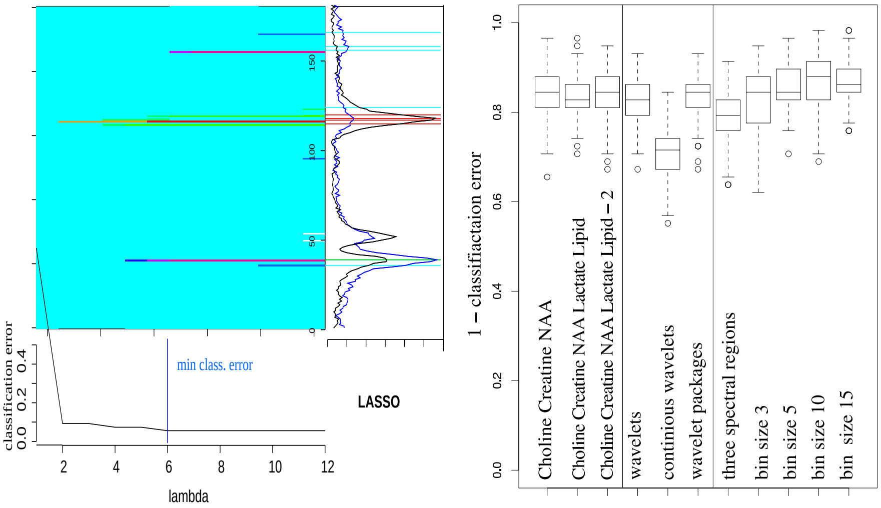
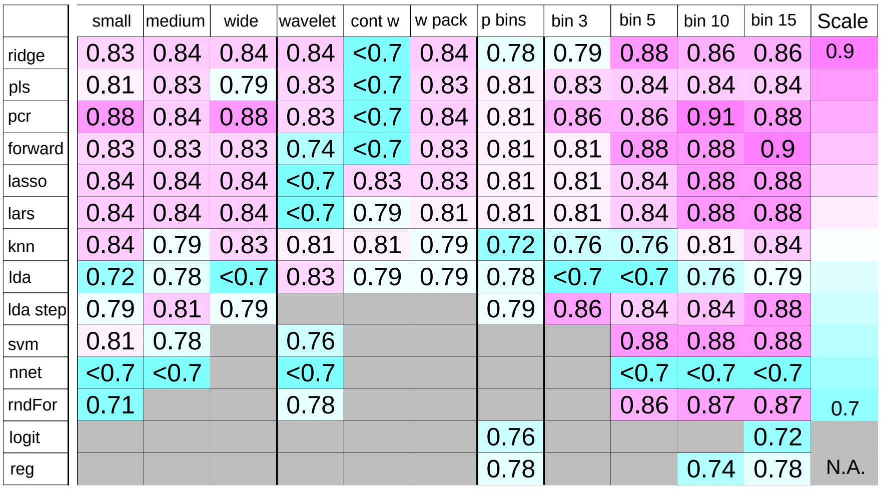


Fig. 3. Left: Spectral channels as chosen by lasso regression with corresponding classification error along hyperparameter lambda within tuning range. - Varying lambda values as determined by a - hypothetic - inner cross-validation loop would hardly affect the overall classification result, since the classification performance is nearly constant over lambda. Right: Classification performance pooled over the three top performing classifier for the given preprocessings. The only preprocessing to outperform plain spectral input (as in 'small' spectral range) is binning with an optimized bin width.

4 Results

Standard classifiers such as linear discriminant analysis or k-nearest- neighbours show rather moderate results and are outperformed by most of the other algorithms. Unconstrained regression methods only work well on the relatively low dimensional data sets. All kinds of shrinkage/regularized regressions perform considerably better, but a differentiation is difficult. On our data set principal component regression seems to perform best. Regardless of our wide grid search in the optimization of neural networks, we were seemingly not able to initialize this method correctly. Performance was bad throughout all data sets. Random forests and RBF-kernel support vector machines performed reasonably, especially after a prior dimensionality reduction. To summarize standard linear methods like PCA, PLS and ridge regression perform notably well throughout all three measures (classification error (cf. Fig.4), ROC AuC, PR Auc). As measured by the ROC and PR, these three performed best on nearly all preprocessings under study.

All three spectral ranges yield a similar classification accuracy: Neither the use of lipid/lactate (as included in the medium range), nor the extension of the spectral region to the water peak (as in the wide range data set) changed the overall classification result.



	small	medium	wide	wavelet	cont w	w pack	p bins	bin 3	bin 5	bin 10	bin 15	Scale
ridge	0.83	0.84	0.84	0.84	<0.7	0.84	0.78	0.79	0.88	0.86	0.86	0.9
pls	0.81	0.83	0.79	0.83	<0.7	0.83	0.81	0.83	0.84	0.84	0.84	
pcr	0.88	0.84	0.88	0.83	<0.7	0.84	0.81	0.86	0.86	0.91	0.88	
forward	0.83	0.83	0.83	0.74	<0.7	0.83	0.81	0.81	0.88	0.88	0.9	
lasso	0.84	0.84	0.84	<0.7	0.83	0.83	0.81	0.81	0.84	0.88	0.88	
lars	0.84	0.84	0.84	<0.7	0.79	0.81	0.81	0.81	0.84	0.88	0.88	
knn	0.84	0.79	0.83	0.81	0.81	0.79	0.72	0.76	0.76	0.81	0.84	
lda	0.72	0.78	<0.7	0.83	0.79	0.79	0.78	<0.7	<0.7	0.76	0.79	
lda step	0.79	0.81	0.79				0.79	0.86	0.84	0.84	0.88	
svm	0.81	0.78		0.76					0.88	0.88	0.88	
nnet	<0.7	<0.7		<0.7					<0.7	<0.7	<0.7	
rndFor	0.71			0.78					0.86	0.87	0.87	0.7
logit							0.76				0.72	
reg							0.78			0.74	0.78	N.A.

Fig. 4. Partial overview of the results. Scale: 1 - classification error. Classifier & regression methods: see text; preprocessings: small/medium/wide spectral ranges; wavelets, continuous wavelets, wavelet packages; bins around peaks, binning width 3/5/10/15. N.A.: no results available.

For an optimal bin width, smoothing and subsampling as performed by the binning approach, proved to be the best preprocessing.

The manual selection of bins around the visible peaks is somewhat worse than using the full spectral vector and seems to be – in our case – an inappropriate way of including *a priori* knowledge into the preprocessing.

The application of a wavelet transformation does not alter the classification performance compared to the raw spectrum. Without a preselection, the continuous wavelet transformation shows poor results. However, the use of wavelets that are smoother than the Daubechies 4 type we used, might result in better performance.

Generally, a preselection (from the auxiliary selection step) either on the wavelet transformed data or on the wide spectral ranges does not impair the classification performance. Nevertheless, it does not increase it over the performance of the respective smaller data sets (small spectral range, normal wavelet transform) either. A difference between the two univariate preselectors cannot be found.

The application of ICA does not improve the results, regardless of the number of mixing sources. Neither do the loadings of the independent components found by the algorithm help to interprete the data better than PCA, nor does the use of the ICA scores improve the following classification step compared to the performance obtained by PCA.

classifier/	parameter	preprocessing											function	from R
regression	range	S	M	W	dw	cw	wp	bp	b3	b5	b10	b15	name	package
ridge	$\lambda = 2^{-12..12}$	-4	-2	0	-3	-8	-3	-10	-1	-2	-1	2	gen.ridge	fda
PLS	$n = 1..12$	1	2	2	1	5	1	2	2	2	2	1	pls	pls.pcr
PCR	$n = 1..12$	9	3	8	6	8	6	2	3	5	7	2	pcr	pls.pcr
lasso	$n = 1..12$	6	6	6	8	5	8	3	7	10	5	5	lars	lars
lars	$n = 1..12$	6	6	6	7	7	7	3	7	9	5	5	lars	lars
forward	$n = 1..12$	6	6	6	8	8	8	3	5	9	6	6	lars	lars
knn	$k = 1..12$	6	8	8	6	3	6	3	2	1	10	7	knn	class
svm	$c = 2^{-5..12}$	0	0	-	2	-	-	-	-	2	4	4	svm	e1071
	$\gamma = 2^{-10..12}$	-8	-8	-	-10	-	-	-	-	-5	-10	-5		
nnet	$s = 1..5$	5	5	-	5	-	-	-	-	5	5	5	nnet	nnet
	$d = 0.1..1$	0.2	0.2	-	0.2	-	-	-	-	0.2	0.2	0.2		
randForest	$t = 25..200$	25	-	-	50	-	-	-	-	75	25	75	random- Forest	random- Forest
	$m = 1..7$	5	-	-	6	-	-	-	-	6	5	6		
	$ns = 1..12$	10	-	-	4	-	-	-	-	6	2	10		
stepw.LDA	$n = 1..8$	2	5	2	-	-	-	2	4	3	3	2	(lda)	(MASS)
LDA	-												lda	MASS
regression	-												lm	base
logit	-												glm	glm

Table 1. Parameter range under study and optimal parameters on the differently preprocessed data (compare Fig. 4).

5 Conclusions

In preprocessing, the application of binning, a smoothing along the spectral vector in conjunction with a dimensionality reduction by subsampling, improves the overall result. Regularized regression methods perform well on this binary and balanced problem. We cannot find a need to use nonlinear, 'black-box' type models. This is of some importance, as in medical applications an interpretability of the diagnostic helper is of high value.

References

SCHLEMMER, H.-P., BACHERT, P., HERFATH, K. K., ZUNA, I., DEBUS, J., and VAN KAICK, G. (2001): Proton MR spectroscopic evaluation of suspicious brain lesions after stereotactic radiotherapy. *American Journal of Neuroradiology*, 22:1316–1324.

HOWE, A. F. and OPSTAD, K. (2003): ^{1}H MR spectroscopy of brain tumour and masses. *NMR in Biomedicine*, pages 123–131.

MEYER, D., LEISCH, F., and HORNIK, K. (2003): The support vector machine under test. *Neurocomputing*, 55:169–186.

HOTHORN, T., LEISCH, F., ZEILEIS, A., and HORNIK, K. (2003): The design and analysis of benchmark experiments. Technical report, SFB Adaptive Informations Systems and Management in Economics and Management Science, TU Vienna.