Classification of Spectroscopic Images in the DIROlab Environment

F.O. Kaster[1,2], B.M. Kelm[3], C.M. Zechmann[1], M.A. Weber[4], F.A. Hamprecht[2], and O. Nix[1]

[1] Deutsches Krebsforschungszentrum, Department for Imaging and Radiooncology, Heidelberg, Germany
[2] Ruprecht-Karls-Universität Heidelberg, Heidelberg Collaboratory for Image Processing, Heidelberg, Germany
[3] Siemens AG, Corporate Technology, Erlangen, Germany
[4] Ruprecht-Karls-Universität Heidelberg, Radiological Clinic, Heidelberg, Germany

Abstract—**We present the magnetic resonance spectroscopy imaging (MRSI) analysis functionality of DIROlab, an integrated software platform for cancer diagnosis and therapy planning. Completely automated estimation of cancer probability from the spectral signature is achieved by state-of-the-art statistical classification techniques; furthermore an easy-to-use interface for spectrum labeling, classifier retraining and evaluation and the benchmarking and comparison of several alternative algorithms is currently under development. The effectiveness of this approach is exemplarily demonstrated by detecting adenocarcinoma in 1.5 Tesla MRSI measurements of the prostate.**

Keywords—**Magnetic resonance spectroscopy imaging, computer-assisted diagnostics, statistical classification, prostate adenocarcinoma, DIROlab.**

I. Introduction

Pattern recognition methods have proved their effectiveness for automated tumor detection based on Magnetic Resonance Spectroscopy Imaging (MRSI) data [1] [2] [3]: in MRSI, spatially resolved chemical shift spectra are measured, which indicate the concentrations of diagnostically relevant metabolites such as choline or citrate. Abnormal concentration ratios in a certain region are strong predictors for the presence of a tumor; especially for the detection of brain or prostate tumors MRSI can have a high diagnostic value [4].

In pattern recognition, training spectra are acquired for which it is known whether they correspond to healthy or malignant tissue (either from expert judgment or ideally from a histopathological gold standard). After some preprocessing, they are used for training a classifier, i.e. a statistical model which can then predict the tumor probability also for new test spectra, whose malignancy grade is unknown. Both for tumors of the brain [2] and of the prostate [3], extensive benchmarking studies of this approach were already conducted, which identified the best-performing combinations of a preprocessing filter and a classifier and showed that pattern recognition could achieve higher accuracy than the traditional approach of spectral quantification. It was also possible to predict the signal quality in high agreement with human experts, i.e. to predict whether a spectrum can be used for diagnostic purposes or whether it is degraded by measurement artifacts or excessive noise [5].

The classification methods with the highest accuracy were implemented in a clinical software prototype called CLARET [6], which can be used easily by non-experts in pattern recognition through an intuitive graphical user interface, which visualizes the classification results in form of color-coded probability maps and which allows the direct and flexible comparison with tumor scores derived from quantification. However, it is applicable only for MRSI data measured with a specific protocol (prostate measurements acquired with an endorectal coil at 1.5 Tesla with an echo time of 135 ms and a sampling interval of 0.8 ms). Retraining is only possible using both specialized tools and specialized knowledge about pattern recognition.

This functionality of automatic tumor probability estimation shall form an integral part of DIROlab, an integrated software platform for integrated cancer diagnosis and therapy planning based on multimodal image data (MRSI and MRI as well as diffusion- or perfusion-weighted imaging, CT and PET) which is currently under development. However, in this new environment an increased flexibility is required: since the platform shall serve as a general-purpose tool for the radiological assessment of cancer, it must be tunable to different organ systems or measurement settings also by non-experts. We report the necessary adaptations of the software and the validation results on 1.5 Tesla prostate adenocarcinoma data. Our final aim is to provide an intuitive user interface not only for classification of MRSI spectra (as already exists in CLARET or in the work of the INTERPRET collaboration [7]), but also for classifier retraining. Since the software is work in progress (especially the graphical user interface), we report the preliminary state.

II. Materials and methods

Our software is projected to enable the user to load MRI and MRSI DICOM images, to display the spectra of single voxels, assign labels both for voxel class (i.e. malignancy grade) and signal quality manually and save the labeling

O. Dössel and W.C. Schlegel (Eds.): WC 2009, IFMBE Proceedings 25/V, pp. 252–255, 2009.
www.springerlink.com

results together with the spectra to an HDF5 file. We use three-staged labels: “tumor” / “undecided” / “healthy tissue” or “good signal” / “poor signal” / “not evaluable signal”. For the voxel class training, the users may choose to use either only “good” signals or also “poor” signals. The data of several training data files may then be used for training one or several classifiers: those are then written to a classifier file, which can be loaded for classification of unlabeled test spectra. Besides classifying single unlabeled spectra (as in the day-to-day application of the software) it will also be possible to analyze a whole set of test spectra with known labels, to compare true and predicted labels and automatically compute accuracy measures, which is an important feature for quality control.

Before training or testing, the spectra can be preprocessed by water suppression with a Hankel singular value decomposition scheme [8], resampling with a B-spline interpolation scheme and discarding parts of the spectrum for dimensionality reduction. Classification is performed on mean-subtracted absolute magnitude spectra in the frequency domain which are normalized such that the sum of absolute values in a prescribed spectral interval equals one.

We provide the following five alternatives, which are all successful general-purpose classifiers and which in previous studies [2] [3] [5] also showed good performance for MRSI analysis: C-support vector machines (SVMs) with linear kernel [9], C-SVMs with radial basis function (RBF) kernel [9], random forests (RFs) [10], ridge regression (RR) [9] and principal components regression (PCR) [9]. Free hyperparameters (the slack penalty and the RBF kernel width for C-SVMs, the number of features per node for RFs, the L_2 norm penalty in RR and the number of principal components in PCR) can be chosen in advance by minimizing a characteristic score over a set of equispaced or log-equispaced proposal values (the generalization error as estimated by cross validation [9] for SVMs, the out-of-bag error [10] for RFs and the generalized cross-validation score [9] for RR and PCR). Since this is a time-expensive process, only a subset of randomly selected training spectra may be used. The first three algorithms can be run either in a native multiclass mode (in which the three label stages are used) or in a binary mode: in the latter case, intermediate labels can be assigned to either of the extreme classes, depending on whether a liberal or conservative classification is desired. RR and PCR are only available in the binary mode.

For the core algorithms of the classifiers, the popular libraries LIBSVM [11] (for SVMs) and VIGRA [12] (for RFs, RR and PCR) were incorporated. The FFTW library is employed for transferring the time-domain measurements in the frequency domain [13]. In order to meet the quality requirements for medical software, strong exception-safety was achieved by creating additional resource management classes following the Resource Acquisition Is Initialization (RAII) idiom [14]. We use the MeVisLab environment for visualization and user interaction.

Since our software shall serve as a platform for research studies, statistical functionality for benchmarking several algorithms is incorporated. The most informative performance measures of classification algorithms are the Receiver Operating Characteristics (ROC) and its Area Under Curve (AUC) value (i.e. its integral), since they are insensitive to the choice of the decision threshold. We use the algorithms described in [15] and estimate the standard deviation of the AUC value by the nonparametric bootstrap procedure recommended in [16]. Furthermore also common measures such as precision, recall, specificity, F-score or correct classification rate are provided, as well as functionality for assessing the significance of differences between algorithms in e.g. the correct classification rate: If the number of measurements is sufficiently large so that the data can be partitioned in a separate training and test set, we supply McNemar’s test for this purpose [17]. However, for small samples it is typically necessary to partition the data in several hold-out folds: then data from one fold are tested against a classifier trained on all the other folds and the relevant statistics are averaged over all hold-out steps. In this setting, the variances of the differences cannot be estimated without bias [18] and hence no principled significance test exists, so we implemented Grandvalet’s recommendation for a conservative *t*-test (which assumes that the maximal correlation between different hold-out folds is bounded from above). [19] Usually more than one single pair of classifiers is compared; we allow for that by adjusting the reported *p*-values either by Holm’s step-down method or by Hochberg’s step-up method [20].

We validate our implementation on two prostate data sets measured at 1.5 Tesla, which had already been studied previously and were hence ideal for comparison: one data set (DS1) of 36864 training and 45312 test spectra with signal quality labels available (with 101 magnitude channels as features, see [5] for details) and another data set (DS2) of 19456 spectra from 24 patients with both signal quality and voxel class labels available (with 41 magnitude channels as features, see [3] for details). No additional preprocessing steps were employed. In the latter case, for training the voxel class classifiers only the 2746 spectra with “good” signal quality were used. Furthermore we did not partition data set 2 in a separate test and training set, but used an eight-fold holdout scheme with each fold comprising the data of three patients. C-SVMs with RBF kernel have preliminarily been excluded from this analysis due to time constraints. The optimal free hyperparameters were selected from the proposal values in Table 1 (using ten-fold cross-validation for the C-SVMs with linear kernel).

Table 1 Classifier hyperparameter values compared in the experiments together with the finally selected values for signal quality (SQ) prediction based on data set 1 (DS1) and signal quality and voxel class (VC) prediction based on data set 2 (DS2)

Hyperparameter (Classifier)	Compared values	Finally selected values DS1(SQ)/DS2(SQ)/DS2(VC)
Slack penalty (C-SVM)	10^{-2},10^{-1},…,10^{3}	10^{1} / 10^{2} / 10^{2}
Number of features per node (RF)	4,6,…,16	16 / 14 / 16
L_2 norm penalty (RR)	10^{-3},10^{-2},…,10^{2}	10^{-1} / 10^{-1} / 10^{-2}
Number of principal components (PCR)	10,15,…,40	40 / 35 / 25

III. Results

Our focus lies in providing a software platform for the flexible training and evaluation of classification algorithms for MRSI analysis rather than in the comparison of algorithms themselves; hence the following results serve mainly as a quality assurance of our implementation: table 2 contains the results for data set 1 (signal quality prediction), while the results for data set 2 are given in tables 3 (signal quality) and 4 (voxel class). For data set 2 we quote the empirical standard deviation over the eight hold-out folds as a lower bound for the true standard deviation: since the hold-out folds overlap, this is an underestimation, which however cannot be avoided without further assumptions about the distribution of these statistics [18].

Table 2 Classification statistics for signal quality prediction with data set 1

	C-SVM	Random Forest	Ridge Regression	PCR
Precision	0.815	0.869	0.921	0.922
Recall=Sensitivity	0.913	0.913	0.797	0.802
Specificity	0.972	0.982	0.991	0.991
F-score	0.861	0.891	0.855	0.857
Correct classification rate	0.965	0.973	0.968	0.968

Table 3 Classification statistics for signal quality prediction with data set 2 (empirical standard deviations from eight hold-out folds)

	C-SVM	Random Forest	Ridge Regression	PCR
Precision	0.73(11)	0.832(57)	0.79(12)	0.79(12)
Recall=Sensitivity	0.57(18)	0.58(17)	0.42(17)	0.43(17)
Specificity	0.964(23)	0.9820(62)	0.980(18)	0.979(19)
F-score	0.621(14)	0.67(13)	0.53(15)	0.54(16)
Correct classification rate	0.905(37)	0.922(32)	0.899(37)	0.899(38)

Table 4 Classification statistics for voxel class prediction with data set 2 (empirical standard deviations from eight hold-out folds)

	C-SVM	Random Forest	Ridge Regression	PCR
Precision	0.908(76)	0.864(27)	0.966(39)	0.900(14)
Recall=Sensitivity	0.69(17)	0.753(16)	0.50(21)	0.50(21)
Specificity	0.983(23)	0.9771(87)	0.9966(39)	0.9928(78)
F-score	0.76(12)	0.79(11)	0.63(22)	0.63(21)
Correct classification rate	0.932(42)	0.937(42)	0.909(59)	0.909(62)

For the signal quality prediction with data set 1 (where we had sufficient data for an independent test set), we conducted McNemar's test in order to test for significant differences in the correct classification rate between the different algorithms. Both Holm's step-down and Hochberg's step-up methods were used for *p*-value adjustment, both of them yielded the same qualitative results: For signal quality prediction with data set 1, the correct classification rate of RF differed with high significance ($p < 10^{-6}$) from all other classifiers, C-SVM differed from PCR significantly ($p < 10^{-3}$), and the differences between RR and PCR and C-SVM and RR were barely significant ($p < 10^{-2}$). For both signal quality and voxel class prediction based on data set 2, no (even barely) significant differences could be detected by Grandvalet's conservative *t*-test with an assumed upper bound of 0.7 for the between-fold correlation (even without Holm's or Hochberg's correction.

Table 5 shows the AUC values with estimated standard deviations: no pair of classifiers differs significantly.

Table 5 Area under the ROC curve (with the standard deviation calculated as in [17]) for different tasks and classifiers

	Data set 1 (Signal quality)	Data set 2 (Signal quality)	Data set 2 (Voxel class)
C-SVM	0.989(14)	0.891(54)	0.97(15)
Random Forest	0.993(14)	0.946(57)	0.98(15)
Ridge Regression	0.990(14)	0.890(54)	0.96(15)
PCR	0.990(14)	0.890(54)	0.95(15)

IV. Discussion

Our experimental work does not aim to compare the merits of different classifiers for MRSI analysis (which has be done before), but to evaluate the correctness of the implementation and to compare it against the preceding scientific studies. Especially the AUC values achieved by the RF classifier for data set 1 are comparable to the findings of [5], while the AUC values found on data set 2 for the RF classifier are slightly worse than reported in [3]; however, the differences are not significant and in this study fewer observations were used (owing to stricter inclusion criteria).

Since the specificity is usually close to one, with the sensitivity being much smaller, the trained classifiers are conservative and yield very few false positives, but considerably more false negatives: this behavior is desirable for signal quality prediction (as no conclusions are drawn from borderline-quality spectra) but not for voxel class prediction (as small tumors might be missed). Since all these classifiers return continuous scores, it is trivially possible to impose any desired minimum sensitivity on the voxel classification task by lowering the decision threshold between the positive and the negative class until the required number of positive test observations is classified correctly (in this paper we dispensed therewith in order to keep the results of the different classification tasks comparable, but this will be accounted for in the clinical release version). It is also possible to assign different weights to the correct classification already in the training process, which might lead to better sensitivities for a given specifity and vice versa.

V. Conclusions

We make no claim of methodological novelty; all presented methods are well-established and their usefulness for detecting tumor signatures in MRSI data has already been confirmed. Our work has the main advantage that the same user-friendliness as for the classification of new test spectra with a fully trained classifier will be made available also for the training process. Previously, the adaptation to every new organ, magnetic field strength or MRSI sequence and the comparison and benchmarking of different classifiers for the new setting needed the joint manual intervention of an expert in pattern recognition (for the preprocessing and training) and an expert in radiology (for labeling the spectra). Now the whole process can be conducted by medical personnel only, so that the usefulness of the pattern recognition in many different settings can now be tested easily. While only the core parts have been finished yet, in the final version an intuitive GUI front-end will enable the users to also assign labels and set the classification parameters.

Acknowledgment

We thank Dr. Ullrich Köthe and Ralf Floca for helpful discussions concerning good software design practice.

References

1. Hagberg G (1998) From magnetic resonance spectroscopy to classification of tumors. A review of pattern recognition methods. NMR Biomed 11:148-156
2. Menze BH, Lichy MP, Bachert P et al (2006) Optimal classification of long echo time in vivo magnetic resonance spectra in the detection of recurrent brain tumors. NMR Biomed 19: 599-609
3. Kelm BM, Menze BH, Zechmann CM et al (2007) Automated Estimation of Tumor Probability in Prostate MRSI: Pattern Recognition vs. Quantification. Magn Res Med 57: 150-159
4. Gillies RJ, Morse DJ (2005) In Vivo Magnetic Resonance Spectroscopy in Cancer. Ann Rev Biomed Eng 7:287-326
5. Menze BH, Kelm BM, Weber MA et al (2008) Mimicking the human expert: Pattern recognition for an automated assessment of data quality in MRSI. Magn Res Med 59: 1457-1466
6. Kelm BM, Menze BH, Neff T et al (2006) CLARET: a tool for fully automated evaluation of MRSI with pattern recognition methods. In: Handels H, Ehrhardt J, Horsch A et al (eds.) Bildverarbeitung für die Medizin 2006. Informatik Aktuell. Springer, Berlin
7. Tate AR, Underwood J, Acosta DM et al (2006) Development of a decision support system for diagnosis and grading of brain tumours using *in vivo* magnetic resonance single voxel spectra. NMR Biomed 19(4): 411-434
8. Zhu G, Smith D, Hua Y (1997) Post-acquisition solvent suppression by singular-value decomposition. J Magn Res 124:286-289
9. Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning. Springer, New York
10. Breiman L (2001) Random Forests. Mach Learn 45(1): 5-32
11. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.tw/~cjlin/libsvm
12. Köthe U (2000) Generische Programmierung für die Bildverarbeitung. PhD thesis (Univ. Hamburg). Software available at http://kogs-www.informatik.uni-hamburg.de/~koethe/vigra
13. Frigo M, Johnson SG (2005) The Design and Implementation of FFTW3. Proc IEEE 93(2):216-231
14. Stroustrup B (2001) Exception Safety: Concepts and Techniques. In: Dony C, Knudsen JL, Romanovsky A et al. (eds.) Advances in Exception Handling Techniques. Springer, New York
15. Fawcett T (2006) An introduction to ROC analysis. Patt Recog Lett 27(8): 861-874
16. Bandos AI, Rockette HE, Gur D (2007) Exact Bootstrap Variances of the Area Under ROC Curve. Comm Stat Theor Meth 36:2443-2461
17. Dietterich TG (1998) Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. Neural Comput 10: 1895-1923
18. Bengio Y, Grandvalet Y (2004) No Unbiased Estimator of the Variance of K-Fold Cross-Validation. J Mach Learn Res 5:1089-1105
19. Grandvalet Y, Bengio Y (2006) Hypothesis Testing for Cross-Validation. Technical Report 1285, Département d'Informatique et Recherche Opérationelle, University of Montréal
20. Shaffer JP (1995) Multiple Hypothesis Testing. Ann Rev Psych 46:561-584

Corresponding author: Frederik Orlando Kaster
Institute: Deutsches Krebsforschungszentrum
Street: Im Neuenheimer Feld 280
City: Heidelberg
Country: Germany
Email: frederik.kaster@iwr.uni-heidelberg.de