Pattern Recognition and Classification in High-Resolution Magnetic Resonance Spectra

Markus T. Wenzel, Bernd Merkel, Matthias Althaus, Heinz-Otto Peitgen

MeVis Research GmbH, Bremen, Germany
Email: wenzel|merkel|althaus|peitgen@mevis.de

Abstract. We show the impacts of various signal preprocessing techniques – dimensionality reduction and transformations – for high-resolution NMR spectra on the classification accuracy of different breast cancer tissue. Our results show that some preprocessing algorithms that are widely used nowadays will not reduce the data dimensionality in an information-preserving way: the classification accuracy drops. Besides showing the most successful preprocessing steps, we can report excellent results on a challenging classification problem.

1 Introduction

Despite growing research efforts on the identification of good prognostic factors for breast cancer, only few of them are proving clinically useful for identifying patients at minimal risk of relapse, patients with a worse prognosis, or patients likely to benefit from specific treatments. Traditional prognostic factors as lymph node status and tumor size are insufficiently accurate. Better or supplementary predictors of high-risk and treatment response are needed. Today, a number of new experimental methods are being explored to improve diagnostic and prognostic information on the genetic, protein or metabolite level, such as gene expression arrays, protein arrays and magnetic resonance spectroscopy (MRS), respectively. The MRS method gives a comprehensive window into tissue biochemistry and interrogates cancer tissue for diagnostic and prognostic markers. MRS of tissue specimens is an ex vivo technique with very high spectral resolution and signal-to-noise ratio. To explore the complex nature of such spectra with high reproducibility, automated classification schemes have to be implemented. There are many ways of processing [1, 2] and classifying [3, 4] NMR spectroscopic datasets. Our effort is to aid physicians in the everyday clinical routine of cancer diagnosis by automated high-resolution MR spectra classification.

2 State of the art and new contribution

In previous work carried out by Derr [5] various simple approaches for dimensionality reduction on high-resolution MRS spectra are compared. It is suggested to firstly refine the alignment of the spectra iteratively before performing piecewise integration such that neighbouring integration intervals overlap with a ratio

that is selected to account for natural chemical shifts of metabolites. In a paper by Baumgartner and co-workers [6] a method reducing spectral data with approximately 1500 dimensions is presented by employing a genetic algorithm that is constrained to deliver a maximum of 30 regions. This promising approach is to be explored for its ability to scale well with the dimensions of the data, which is work-in-progress.

Although a pathologist's expertise can determine the tumor grading of tissue biopsates with high accuracy, there are several disadvantages which the MRS diagnosis of tissue can overcome: the need for rather voluminous biopsates, the time-consuming procedure as such and therefore the elongated period of unwanted remaining in uncertainty for the patient. Additionally, not only the state can be determined from the spectra, but also a quantitative assessment of cancer indicating metabolites is possible. Moreover, the results of the basic research conducted here have the potential to be transferred to examinations of tissue by in-vivo MRS-Imaging methods, thus limiting the need for biopsy to cases where automatic classification results are ambiguous.

From the computer assisted diagnosis perspective, the first step is to identify methods to deal with the high dimensionality of the data in question. Therefore the impacts of various preprocessing steps on the final classification results are shown and compared. We believe this to be a major contribution to the systematic analysis of processing methods for high dimensionality data in general and for the assessment of spectral information based on MRS in particular.

3 Methods

Breast tissue biopsates have been WHO graded by pathology and measured by MRS. The complete dataset consists of 91 high resolution NMR spectra acquired on a Bruker Avance 600 spectrometer with a spectral width of 9kHz (see [7] for details). Of this dataset, 41 were rated to be of grading 0, two of grading 1, 26 of grading 2, 21 of grading 3 and one of grading 4. All classification results reported within this paper improve considerably (about 3%) when omitting the 3 samples of grading 1 and 4. Within an experimental pipeline approach consisting of three steps, we systematically modify the processing.

3.1 Alignment of data

Upon acquisition of spectral data, reference points are manually set in the data. Therefore, instances are not correctly aligned with each other in general. This is a common problem in all research done where NMR spectral data is to be handled automatically. To overcome this unfavorable situation, we implemented an algorithm to estimate and correct this data misalignment. Regions present in all training spectra instances are identified manually. These regions are cut out and convolved with a a slightly broader region in all other examples. The peak of the convolution shows the respective "best fit" positions of the region in the test instance.

It is acknowledged, that all alignment performed by a rigid shift according to the displacement of only one section will only produce displacements in other regions. Still we achieved systematically increased classification rates after alignment, which may be due to fact that more important regions were now aligned with each other.

Advanced shifting algorithms – dynamic time warping (DTW) and correlation optimized warping (COW) – were also employed recently [8, 9]. Both were introduced for real-time processing of speech data, but were found to be useful in chemometrics as well. Our results with COW are more promising than those with DTW, but both require further research as to assess the side-effects, introduced with respect to the robustness of the overall system.

3.2 Transformations and data reduction

We performed data reduction with the following transformations on aligned and unaligned data, plus unchanged as control, ending up with four output datasets per input.

In subsampling every 5^{th} and also every 6^{th} data point with different starting points was kept.

For threshold-guided cutting we reduced the data to regions with signals above a threshold τ and longer than a threshold ρ. Only one configuration with $\tau = 800$ and $\rho = 0.034$ ppm was further considered, because the information loss seemed to worsen disproportionate to increasing τ and ρ, whereas the desired dimensionality reduction effect deteriorates disproportionate to decreasing thresholds. For the thresholds given above, the dimensionality approximately halves.

Additionally, we decomposed the data with predominantly biorthogonal wavelets, which were chosen based on prior experiences on mass spectrometric data. For further processing we used both detail and approximation coefficients from level 3 to 6, but also other wavelets like symlets and Daubechies were used.

Exhaustively searching all possible attribute combinations – although guaranteed to find the optimal solution – is clearly not feasible computationally on datasets of tens of thousands of attributes. A very common alternative is a feature selection, based on correlation, although it is doubted in [10]. The "Best First" forward selection (FS) method adds single best attributes iteratively unless some optimality criterion stops improving. FS is guaranteed to converge, but not necessarily to the optimal solution, because it will not combine individually inferior attributes, which may however perform better if combined. For our experiments, we used the WEKA implementation of this algorithm [11]. Also, we employed a Genetic Algorithm (GA) guided selection from the same toolbox, using its default parameters.

3.3 Evaluation by classification

To evaluate the performance of the data alignment, data transformation, and dimensionality reduction techniques on our data, we chose the classification accu-

Table 1. Best classification results with forward selection

Method	# Attributes / Accuracy			
	Unaligned		Aligned	
Subsampling (5^{th})	42	74.7%	30	80.2%
Threshold+Subsampling	29	70.3%	35	71.4%
`Bior3.7`-Approx.3	24	70.3%	33	74.7%
`Bior3.7`-Det.3	37	76.9%	34	73.6%
Only FS	72	79.1%	66	81.3%
Only FS w/o grad.I+IV	78	83.0%	68	84.1%

racy as the performance measure. The input of this pipeline step are the datasets produced from the above preprocessing.

We compare the results of all feature reduction algorithms by applying a Random Forest classification algorithm to the reduced data [12]. Random Forests are collections of Random Trees built from randomly selected subsets of all training subjects, where the split at each node is performed based on a random selection of attributes. The parametrization of the number of random attributes used for each split in the trees was chosen based on suggestions of Breiman [12]. The number of random trees to build was determined by our experiments and finally fixed at 100 trees.

4 Results

In our experiments we found that subsampling worsens the classification accuracy while not substantially reducing the feature number. If, however, FS is applied on the subsampled data, 80.2% on the aligned data resulted.

Threshold-guided cutting in combination with subsampling the spectra performed worse, although the classification accuracy increases with FS. Doing a classification only on the wavelet-transformed spectra was not successful, but here with FS better accuracies were achieved.

Table 1 summarizes our results from the main set of experiments where the pipeline steps were varied. We give only the best accuracies together with the according configuration. As the GA guided selection performed worse on almost every approach, no results are in here.

Reducing the problem to a two-class-problem (benign vs. malignant), we achieved 92.0% sensitivity and 95.1% specificity on the aligned data by applying only FS without any previous transformation or data reduction.

5 Discussion

Our comparison of feature selection algorithms showed the unexpected superiority of Forward Selection over all competing approaches. The results generally improved after a coarse alignment of the instance vectors motivating further

research in this area. A wavelet transformation did not improve classification results as expected. Nevertheless, since the accuracy did not drop significantly and since wavelet decomposition is a fast and widely used approach to dimensionality reduction, we will also explore these topics in the future.

Since we only examined the spectra from the lipid phase of the breast tissue we expect an improved classification result when the water soluble phase is also taken into account. In general, the classification on lipid-spectra is more challenging due to the high demand on the measurement accuracy [7], suggesting the possibility to generalize our approach on this data.

We wish to cross-check our results with other classifications schemes. In our ongoing research we implement projective classification schemes which promise to provide dimensionality reduction by projection and classification in a joint approach [9]. Besides, we are currently evaluating established methods of spectral analysis to be able to compare our findings better with widely acknowledged "ground truth" methods.

References

1. Vanhamme L, Sundin T, van Hecke P, van Huffel S. MR spectroscopic quantitation: A review of time-domain methods. NMR Biomed 2001;14:233–246.
2. Mierisová S, Ala-Korpela M. MR spectroscopy quantitation: a review of frequency domain methods. NMR Biomed 2001;14:247–259.
3. Hagberg G. From magnetic resonance spectroscopy to classification of tumors: A review of pattern recognition methods. NMR Biomed 1998;11:148–156.
4. Menze BH, Lichy MP, Bachert P, Kelm BM, Schlemmer HP, Hamprecht FA. Optimal classification of long echo time in vivo magnetic resonance spectra in the detection of recurrent brain tumors. NMR Biomed 2006;19:599–609.
5. Derr T. Medizinisch-diagnostische Anwendung neuronaler Netzwerke zur Analyse NMR-spektroskopischer Daten von Körperflüssigkeiten. Ph.D. thesis. Universität Bremen; 1997.
6. Baumgartner R, Ho TK, Somorjai R, Himmelreich U, Sorrell T. Complexity of magnetic resonance spectrum classification. Data Complexity in Pattern Recognition 2005.
7. Beckonert O, Monnerjahn J, Bonk U, Leibfritz D. Visualizing metabolic changes in breast-cancer tissue using 1H-NMR spectroscopy and self-organizing maps. NMR Biomed 2003;16:1–11.
8. Tomasi G, van den Berg F, Andersson C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. J Chemometrics 2004;18:231–241.
9. Wenzel MT, Merkel B, Althaus M, Peitgen HO. PCNSA for NMR spectroscopy breast tissue classification. In: ISMRM DP Spect WS; 2006.
10. Nikulin AE, Dolenko B, Bezabeh T, Somorjai RL. Near-optimal region selection for feature space reduction: Novel preprocessing methods for classifying MR spectra. NMR in Biomedicine 1998;11:209–216.
11. Witten IH, Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed. San Francisco: Morgan Kaufmann; 2005.
12. Breiman L. Random Forests. Machine Learning 2001;45:5–32.