View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by LJMU Research Online

Pattern Recognition Analysis of MR Spectra

Sandra Ortega-Martorell[1,2], Margarida Julià-Sapé[2,3], Paulo Lisboa[1] & Carles Arús[3,2]

[1] *Liverpool John Moores University, Liverpool, United Kingdom*
[2] *Centro de Investigación Biomédica en Red (CIBER), Cerdanyola del Vallès, Spain*
[3] *Universitat Autònoma de Barcelona (UAB), Cerdanyola del Vallès, Spain*

The need for multivariate analysis of magnetic resonance spectroscopy (MRS) data was recognized about 20 years ago, when it became evident that spectral patterns were characteristic of some diseases. Despite this, there is no generally accepted methodology for performing pattern recognition (PR) analysis of MRS data sets. Here, the data acquisition and processing requirements for performing successful PR as applied to human MRS studies are introduced, and the main techniques for feature selection, extraction, and classification are described. These include methods of dimensionality reduction such as principal component analysis (PCA), independent component analysis (ICA), non-negative matrix factorization (NMF), and feature selection. Supervised methods such as linear discriminant analysis (LDA), logistic regression (LogR), and nonlinear classification are discussed separately from unsupervised and semisupervised classification techniques, including *k*-means clustering. Methods for testing and metrics for gauging the performance of PR models (sensitivity and specificity, the 'Confusion Matrix', '*k*-fold cross-validation', 'Leave One Out', 'Bootstrapping', the 'Receiver Operating Characteristic curve', and balanced error and accuracy rates) are briefly described. This article ends with a summary of the main lessons learned from PR applied to MRS to date.

Keywords: feature selection, feature extraction, classification, supervised, unsupervised, independent test set, nosological image, visualization, performance evaluation

How to cite this article:
eMagRes, 2016, Vol 5: 945–958. DOI 10.1002/9780470034590.emrstm1484

Introduction

Magnetic resonance spectroscopy (MRS) is essentially a quantitative methodology. The areas and, with adequate constraints, the heights of individual resonances are directly proportional to the number or concentration of nuclei contributing to them in the sampled volume. Accordingly, a common way to obtain information of interest from a sample tissue is to quantify a number of substances that contribute to the spectral pattern. However, this quantification may not directly address the question posed by the clinician or the researcher. For example, even if there are statistically significant differences in the mean concentration measured between two experimental conditions of interest, their ranges can overlap so that no threshold can be used to reliably assign a new spectrum into one of the compared classes. In addition, in pathological conditions such as brain tumors, there are simultaneous changes in several resonances (Figure 1). This has led researchers to combine all of the changes using multivariate analysis methods in order to discriminate clinical conditions, based on either the whole spectral vector or selected regions from it after postprocessing or feature extraction.

One step further leads us into the pattern recognition (PR) domain. PR in MRS can be used for characterizing both in vivo spectra from animals or patients or ex vivo spectra from biopsies and cells and/or their extracts. We will restrict this article to the evaluation of in vivo MRS, although most of the PR principles and strategies described are also applicable to ex vivo data (e.g., to high-resolution magic angle spinning NMR analysis of biopsies).

This article will guide the reader through some of the strategies available to undertake a PR analysis of magnetic resonance (MR) spectra and will also discuss the requirements for obtaining a robust result.

Data

The use of PR methods with MRS data started in the early 1990s, when it was first shown that phosphorus (^{31}P) MR spectra from in vivo preclinical tumors could be successfully classified.[2] Subsequent studies on human brain tumors[3−5] also showed that in vivo ^{1}H MR spectra could be correctly assigned to different tumor types. Despite the early work on ^{31}P MRS, proton (^{1}H) MRS has historically been the most-used nucleus, with most studies concentrating in the categorization of tumor types and grades and their prognostic assessment, mostly for brain and prostate, but also in breast, lymphoma, and bone marrow. A few other studies have focused on the diagnosis or prognosis of various neurodegenerative diseases.

The reasons why these have been the most common questions addressed by PR are first, data sparsity, which is also a limiting factor for many other questions of potential interest; and second, data acquisition limitations, such as those related to magnetic field homogeneity or low signal-to-noise ratio (SNR). In addition, there is an inherent difficulty in the use of

© 2016 John Wiley & Sons, Ltd.

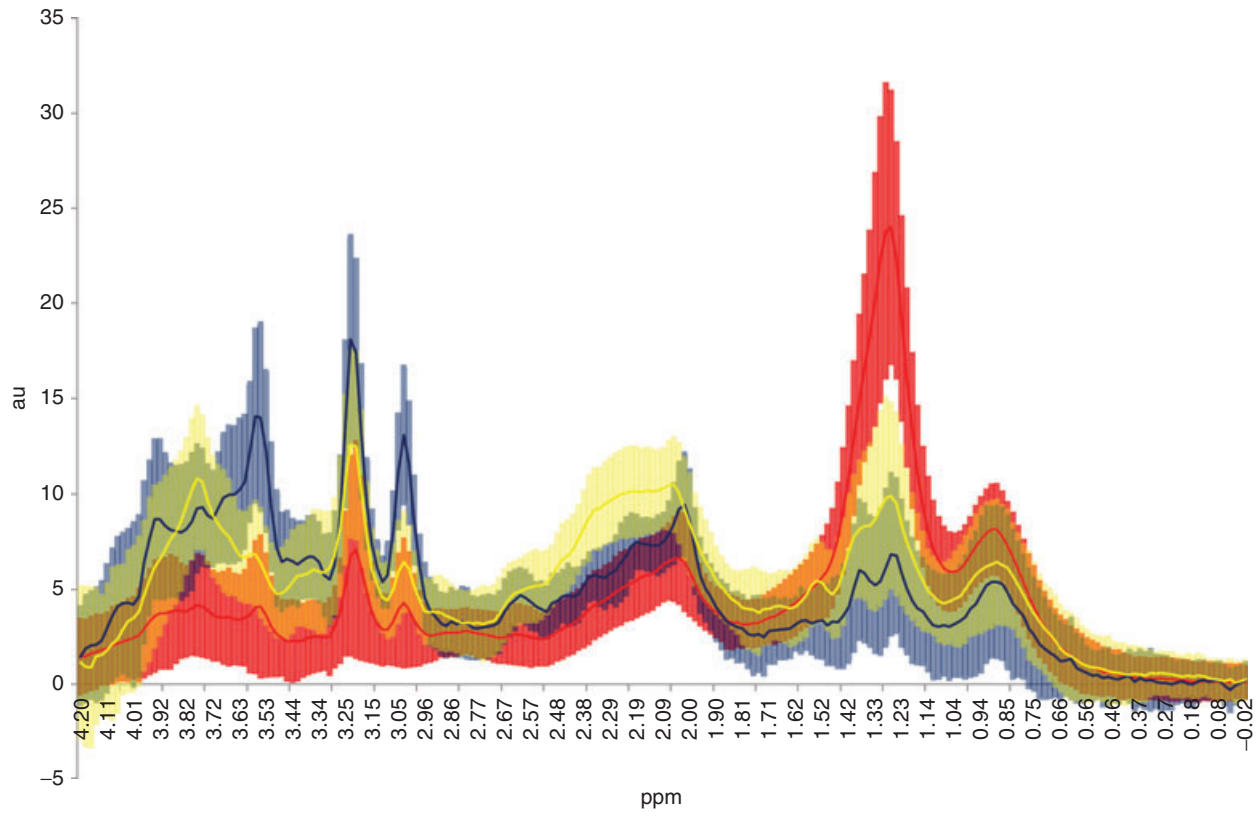

Figure 1. Mean (lines) and plus and minus one standard deviation (±1 SD; shaded areas) for the three most common types of brain tumor spectra included in the INTERPRET database,[1] at short TE (20–32 ms) at 1.5 T. Aggressive tumors (glioblastoma and metastasis), $n = 124$, are shown in red; low grade glial (astrocytoma, oligoastrocytoma and oligodendroglioma WHO grade II), $n = 35$ are in blue; and low grade meningioma, $n = 58$, are depicted in yellow. As can be seen, several peaks change, and their SDs also overlap, for example, the green shade results from overlap of the SDs of meningioma in yellow and low grade glioma in blue (see Ref. 1 for further details; au = arbitrary units for UL2-normalized mean spectra of different tumor types)

PR by clinicians and biological scientists, owing to problems ranging from its postprocessing requirements to the adequate visualization of the resulting classifications.

How Many Data?

Typically, MRS data sets are characterized by their small number of cases, especially when compared to the large number of spectral components, or 'features' that may be used for PR work. This makes computer-based automated classification particularly challenging. Most importantly, a large number of features can also preclude the straightforward interpretation of the results obtained, limiting their usability in a practical medical context in which interpretability is paramount and simplicity and robustness of the methods employed are essential.[6] For this reason, the features containing little discriminative power must be discarded, which means that we will need a large number of cases (about three to five times more cases than the number of features to extract) from which to extract those features.[7,8] However, in real life, the number of available cases will be limited by epidemiology and budget. This problem is called *the curse of dimensionality*.[7] It is common to try to compensate for this restriction with large multicenter studies[8] (see ***Clinical Trials of MRS Methods***), which brings an added problem: data compatibility in the face of slightly different acquisition conditions (field strength, localization pulse sequence, echo time (TE), and recycling time, among others). These factors will introduce variability or noise into the classifier training process.

Another source of variability originates from the labeling of the cases. There are many instances in which the 'gold standard' used for labeling can itself contain errors or ambiguities, which may create problems depending of the PR strategy being attempted. A well-known example of this is the labeling of astrocytic WHO grade III brain tumors (anaplastic astrocytomas) for which there is usually lack of agreement among pathologists,[9] compromising both classifier training and validation. Furthermore, many of the tissues investigated by single voxel (SV) MRS, for example, contain metabolic pattern heterogeneities at a scale smaller than the sampled voxel. This is particularly relevant for tumors.[10,11] Therefore, even if the gold standard label of two cases is the same, their metabolomes and spectral patterns may vary. The process of training robust classifiers will be hampered whenever intratissue (intraclass) heterogeneity is comparable to the variability in the population (interclass) heterogeneity.

One way out of this problem can be to use a smaller voxel size and perform a multivolume (MV) acquisition. This brings two possible bonuses, firstly, the heterogeneity of the case may be more apparent and can then be used for a combined assessment

 © 2016 John Wiley & Sons, Ltd.

of the functional question being analyzed and, secondly, a single case can produce hundreds of data vectors, highly correlated, but still different from each other, which can help to reduce the impact of the curse of dimensionality. It is common practice to obtain robust classifiers from only a few cases with the use of such highly correlated MV data matrices.[12]

Data Processing

Data processing has common steps for different nuclei. Fourier transformation of the time domain data is sooner or later required in the reconstruction pipeline, while there are specific steps required for certain nuclei, such as filtering of the partially suppressed water resonance in ^{1}H MRS. Some of these steps have a varying influence on the subsequent PR analysis. For example, some line broadening (signal smoothing) may help with visual appreciation of the instrumental quality of the spectral pattern by spectroscopy experts (via the assessment of SNR and the presence of artifacts), whereas this is not normally desirable before quantification of individual resonances (metabolites) in the frequency domain.[13] For proper PR studies, these individual metabolite quantifications should be transformed into data vectors, which can then be further analyzed (see section titled 'Analysis'). On the other hand, spectral alignment with respect to an internal reference is usually crucial for PR studies in which individual data points are used as the input for feature extraction.

Normalization is also an important issue. A common strategy is to normalize the spectrum by the unsuppressed water signal from the same volume of interest (VOI), be it for an SV spectrum or each of the individual voxels of an MV grid. This requires subsequent corrections, usually by assuming a percentage of water content in the tissue of interest or by the presence of various metabolites in the tissue when metabolite ratios are being quantified, and also the spin-lattice and spin–spin relaxation times (T_1 and T_2) of the water (if it is used as a reference) and the relevant functional groups of those metabolites, in the conditions being evaluated (e.g., control vs pathological). Any error in those assumptions will introduce noise into the PR system, although if it is coherent noise it may be tolerable. Another widely used option is normalization to one of the various unit length possibilities. An example is use of the 'UL2 norm' (see ***Clinical Trials of MRS Methods*** for a detailed explanation) because it maintains the sign of negative resonance intensities from lactate or alanine among other metabolites at 'long' TE values (135–144 ms). This type of normalization is not neutral and may introduce scaling effects of its own; the point here is whether those effects are beneficial for class/cluster discrimination or not.

Proper alignment must be checked and corrected if necessary, especially if we are using individual features or the whole spectral pattern instead of integrated peak areas. Internal references, such as total creatine, total choline, or mobile lipids in ^{1}H spectra, have been used for peak alignment because their chemical shift is pH independent in the biological range, and one of them will usually be visible in the tissue of interest. Some authors have also developed empirical alignment algorithms for cases in which the SNR for the potential reference resonances may vary strongly among samples.[14]

Analysis

Dimensionality Reduction. The reduction of the dimensionality of a data set can be seen as a process of selecting relevant features with the aim of removing redundant information from the data, while retaining the most informative features to improve the performance of predictive models.[15,16] Consequently, feature selection or feature extraction is often performed in MRS data sets before diagnostic classification. However, selecting the type of feature extraction or selection method is problem- and domain-dependent and thus requires some knowledge of the domain under consideration. Reducing the dimensionality of spectra can also be beneficial for reducing computational complexity and so allow algorithms to operate faster and more effectively, as well as reducing the risk of overfitting the data by having too many parameters in the model. In this section, however, we focus on how reduced feature sets can improve the accuracy of classification by briefly outlining some of the most commonly used methods for dimensionality reduction in MRS data analysis.

Historically, two general approaches to dimensionality reduction were adopted in early studies of MRS classification. One approach was to focus on spectral peaks with known metabolic significance, which could be normalized within each spectrum by taking the ratio of peak heights or integrated peak areas. This is still a useful and transparent method, which is developed later in this section with more data-driven methods to automatically search for discriminating parts of the spectrum. The alternative approach for dimensionality reduction for MRS was to exploit the correlation structure of the data with principal component analysis (PCA).

Principal Component Analysis (PCA). This is a natural methodology for entirely data-driven dimensionality reduction, as any data set with more covariates than the sample size is necessarily linearly separable, so that it is natural to apply linear methods first. The rationale for PCA is to capture as much of the variation in the data as possible using a small number of linear combinations of the predictive features, typically peak or individual data point heights. In this way, the correlation structure contained in the covariance matrix is optimally exploited to reduce the dimensionality to the selected number of principal component features: typically around a dozen suffices to describe about 90% of the variance in the data. This process is also known to result in some cleaning of noise from the observed variables.

From a statistical standpoint, linear principal components are obtained from a factorization of the sampled covariance matrix into a set of orthogonal directions called *eigenvectors*, which define the principal directions along which the data spreads.[16] PCA was successful for building early data-based models for differential diagnosis of tissue types or tumor grades. However, the interpretation of principal components is not trivial as they comprise both positive and negative linear combinations of the original variables. Nevertheless, they can be useful for exploring the variation in the data sets (Figure 2).

Independent Component Analysis (ICA). The purpose of this approach is to go beyond the correlation structure of the

© 2016 John Wiley & Sons, Ltd.

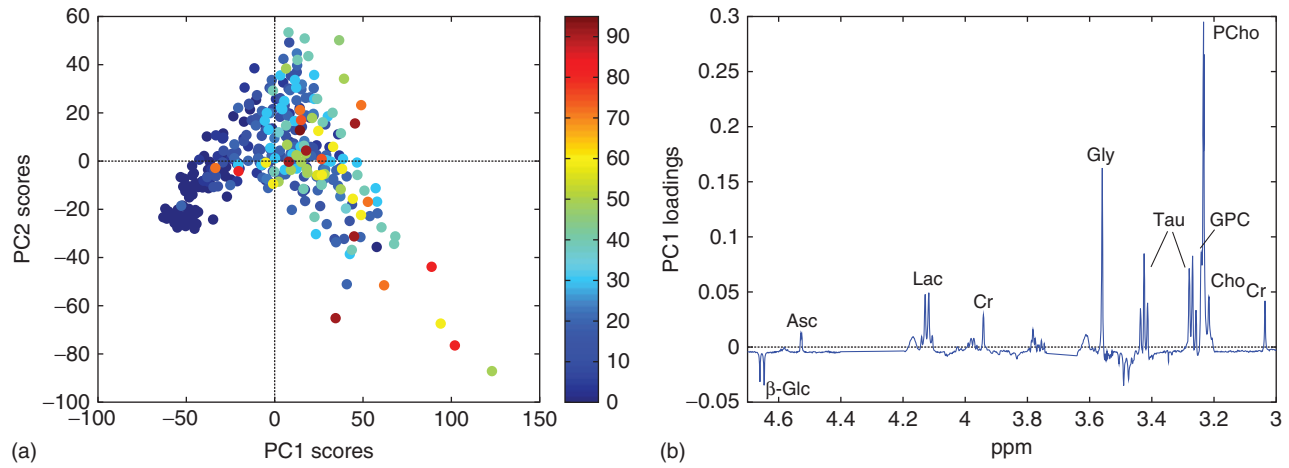

Figure 2. Variation in tumor cell content as described by PCA. (a) The score plot of the pre-processed spectra, colored based on tumor cell content (%, color scale at right) determined from corresponding biopsies. (b) The corresponding loading profile of PC1, explaining 40.1% of the total variation of the data. β-Glc, β-glucose; Asc, ascorbate; Lac, lactate; Cr, creatine; Gly, glycine; Tau, taurine; GPC, glycerophosphocholine; PCho, phosphocholine; Cho, free choline. (Reproduced from Ref. 17. © PLoS One, 2013)

spectra and focus instead on the latent tissue types that occur independently and together result in the mixtures measured in each voxel. In other words, to attempt to decompose the spectra in each voxel into the components (also named sources, latent variables, or independent components) of the constituent tissues present. ICA (independent component analysis) defines a generative model for the observed multivariate data, in which the data variables are assumed to be linear mixtures of some unknown latent variables such that the probability of each source being present is independent for the different sources. That is, the joint probability of all latent variables is assumed to factorize into a product of separate probabilities for each variable. The latent variables must be assumed to be non-Gaussian, as linearly mixing Gaussian distributions produce a single multivariate Gaussian distribution that cannot then be decomposed. The resulting latent variables are thus mutually independent and so are called the *independent components of the observed data*.[18,19]

Given an observed data matrix X (of dimensions $d \times n$, where d is the number of spectral features and n is the number of observations), factors S (of dimensions $d \times k$, namely the matrix of sources, latent variables, or independent components, where k is the number of components), and H (of dimensions $k \times n$, namely the mixing matrix) can be estimated using ICA, such that $X \approx SH$.

When applied to spectral peak heights, this method naturally separates normal from infiltrating tumoral tissue.[20] However, the independent components are not readily interpretable because they contain both positive and negative mixtures of the original variables, as was the case with PCA. It is then natural to seek positive-only mixtures and determine whether this permits a more detailed separation of component tissues.

Non-negative Matrix Factorization (NMF). In NMF (non-negative matrix factorization) methods,[21,22] the non-negative data matrix X (also of dimensions $d \times n$) is approximately factorized into two non-negative matrices: the matrix of sources or data basis S (of dimensions $d \times k$, where k is the number of sources, and $k < d$) and the mixing matrix H (of dimensions $k \times n$). The product of these two matrices provides a good approximation to the original data matrix in the form $X \approx SH$. There are different NMF variants, which mainly arise from using different cost functions for computing the divergence between X and SH.

While this methodology describes the observed data with positive-only mixtures of the latent variables or data sources, this does not apply to long echo times where spectral phase-related signal modulation frequently results in negative values in the lactate and alanine regions.[23] Convex non-negative matrix factorization (convex-NMF)[24] is a variant of NMF that imposes a restriction over the source matrix S to be a convex combination of the input data vectors. This restriction significantly improves the quality of data representation of S. Unlike standard NMF, convex-NMF applies to both non-negative and mixed-sign data matrices. What this means in practice is that (i) the data are described by positive-only mixtures of the sources; and (ii) the sources, or latent variables, are also positive-only mixtures of the data. In principle, this makes it easier to interpret both the mixing and unmixing processes, taking care to consider that the model coefficients represent the strength of linear mixtures but are not probabilities for the purpose of subsequent modeling.

Both NMF and convex-NMF have proved successful for classifying pathological tissue types into specific subgroups, e.g., by tumor type or grade[23,25] (Figure 3).

Feature Selection. Feature selection methods are also widely and successfully used in this area for reducing the dimensionality of a data set to a small number of relevant MRS features.[26,27] More details on this type of method are discussed in the following section.

Supervised Methods. Supervised learning, otherwise known as *regression* or *classification*, is the task of analyzing a set of variables or features to achieve known outcomes, which typically

© 2016 John Wiley & Sons, Ltd.

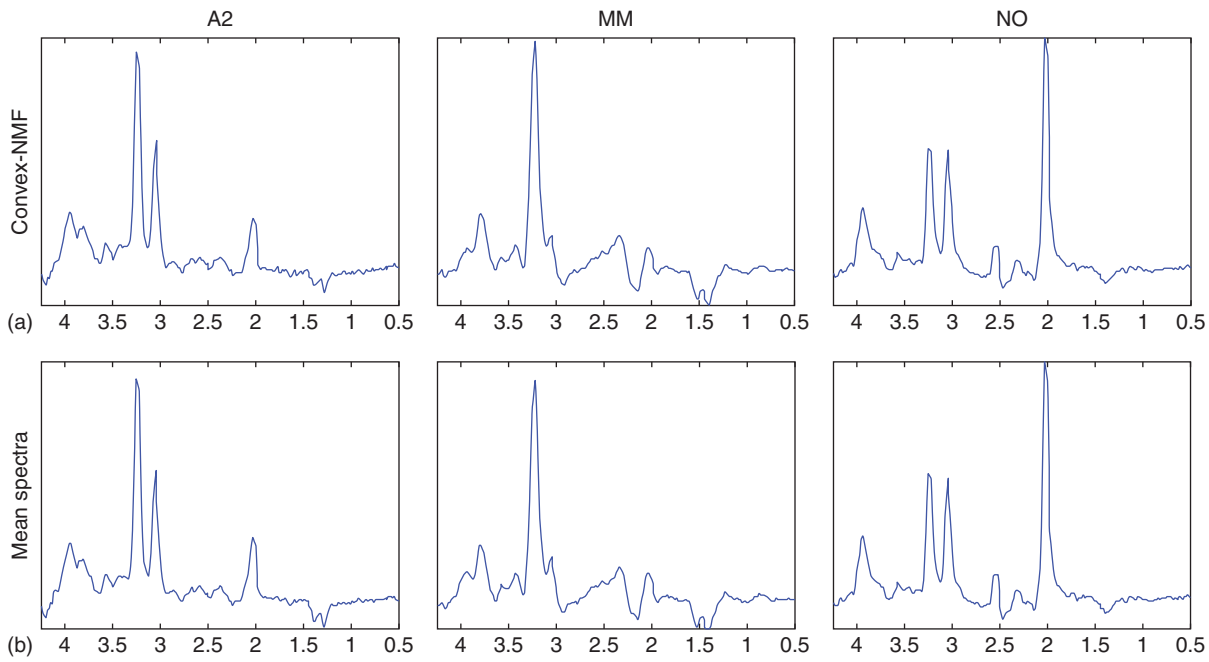

Figure 3. (a) Source signals obtained with convex-NMF in a data set containing SV, 1.5 T spectra, from 20 astrocytomas grade II (A2), 55 low-grade meningiomas (MM) and 15 normal brain parenchyma measurements from healthy controls (NO) at long TE (135–136 ms), extracted from the INTERPRET database.[1,23] (b) Mean spectra of A2, MM and NO, at long TE (135–136 ms). Note the resemblance between the extracted sources and the mean spectra

consist of continuous values or class labels. In MRS, the typical aim is to classify unseen test instances, that is, spectra other than the modeling data set. Although there is an extensive range of supervised learning algorithms available, there is no single learning algorithm that works best on all supervised learning problems. Broadly speaking, there are three groups of supervised models. First there are those which are linear in the modeling parameters, such as linear discriminant analysis (LDA) and logistic regression (LogR). Second there are semiparametric models that are generic nonlinear classifiers, such as neural networks; and finally, there are nonparametric approaches such as nearest neighbor classifiers. In this section, we briefly describe the most commonly used supervised learning algorithms in MRS data analysis.

Linear Discriminant Analysis (LDA). LDA is a classification method based on the assumption of a common variance of the classes. Fisher LDA is a reduced-rank version of LDA, which projects the variables into the lower dimensional subspace that maximizes the rate of the between-variance and the within-variance on the training set. As a simple illustration of this method, let us project a d-dimensional data set onto a line. It will usually produce a confused mixture of samples from all of the classes and thus poor recognition performance. However, by moving the line around, we might be able to find an orientation for which the projected samples are well separated.

Figure 4(a) illustrates this effect for a two-dimensional (2-D) example. The goal is to find the direction of this line. In brief, its formulation is as follows.[28] Suppose the observed data matrix X ($d \times n$, where d is the data set dimensionality and n the number of samples) has n_1 samples in the subset D_1, labeled ω_1, and n_2 samples in the subset D_2, labeled ω_2. A linear combination of the components of X is formed to obtain $Y = W^{\mathrm{T}}X$, and a corresponding set of n samples of Y divided into subsets, Υ_1 and Υ_2. The magnitude of W is of no real significance, as it merely scales Y. What is important is finding the best direction for W that will enable the most accurate classification of the two subsets.

The extent to which we create or learn a proper representation and how we quantify what is near and far apart will determine the success of a classifier. An additional desirable characteristic is[28] a small number of features, which might lead (i) to simpler decision regions and a classifier that is easier to train and (ii) to robust behavior, i.e., being relatively insensitive to noise or other errors.

Logistic Regression (LogR). The main downfall of LDA is that its predictions are binary class labels, without any indication of confidence in the predictions. To obtain this we need to turn to probabilistic classifiers. The one most commonly used is LogR. This model has a similar structure to the LDA, with an important difference, namely that now the linear combination of the components of X is linked directly to the odds ratio of class membership, i.e., $\log(\mathrm{OddsRatio}(\mathrm{class}|x)) = W^{\mathrm{T}}X$ where the $\mathrm{OddsRatio}(\mathrm{class}|\,x) = P(\mathrm{class}|\,X)/(1 - \mathrm{P}(\mathrm{class}|\,x))$. This results in models that are directly interpretable in terms of the components of X through the values of the linear coefficients and also very well calibrated to the actual probabilities of class membership.[29]

These models have a very natural Bayesian interpretation of the odds of a particular class being detected in terms of the individual contributions from each component

© 2016 John Wiley & Sons, Ltd.

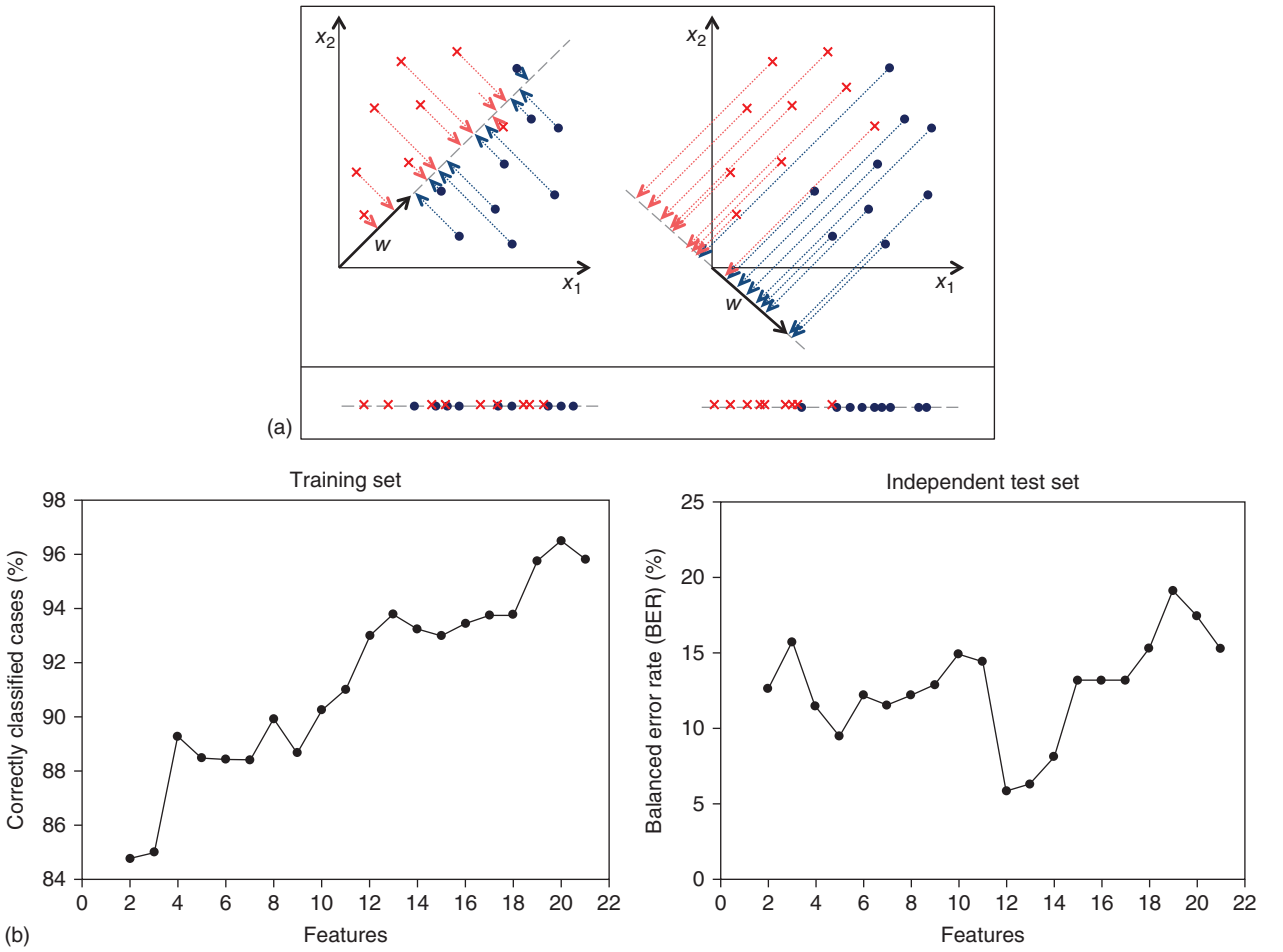

Figure 4. (a) Effect of choosing two different values for *w* for a 2-D example. Top: projection of samples onto two different lines. Bottom: representation of the projected samples. The example on the right shows greater separation between the red and blue projected points. (b) The ideal number of features for a classifier. (Reproduced with permission from Ref. 27. © John Wiley & Sons, Ltd., 2014.) An improvement in the percentage of correctly classified cases for response to therapy in preclinical ^{1}H MRS MV data with increasing number of features is observed in the training set. In the independent test set, overtraining occurs above 12 features after which the BER increases with increasing number of features

of $X = (X_1, X_2, \ldots, X_d)$ and the corresponding coefficients $(W_1, W_2, \ldots, W_d, W_0)$; the latter is related to the prior value, or prevalence of the class, as follows:

$$\text{OddsRatio}(\text{class}|x) = e^{W_1 X_1} e^{W_2 X_2} \ldots e^{W_d X_d} e^{W_0} \qquad (1)$$

Moreover, the probabilistic nature of the model makes it possible to guide a statistical feature selection process, to which we return to later in this section.

Nonlinear Classifiers. The linear structure of the previous models can be seen as a limitation if there are significant interactions between covariates. To test this hypothesis, two radically different methodologies can be applied. One is to exploit the interactions directly via the covariance matrix or by seeking independent latent variables, which were outlined in the previous section. The other approach is to model nonlinear decision boundaries directly. We now explore the latter method.

Perhaps the most straightforward approach to nonlinear modeling is to allocate spectra to the class membership of near-neighbors. However, for high-dimensional data such as MRS, the metric structure is nonintuitive because the pairwise distances between any one spectrum and all of the others is typically very similar, in marked contrast to intuition for 2-D or three-dimensional (3-D) spaces.

It is better to apply semiparametric models that may be purely discriminative with binary predictions, such as support vector machines (SVMs).[30] This algorithm is known as a *maximum margin classifier* because it seeks the largest gaps between classes using the spectra nearest to the boundary as reference points, known as *support vectors*. However, this approach is not calibrated to likelihood of class membership because it is not probabilistic. To achieve this, a probabilistic cost function based on least squares optimization may be used, hence least-squared support vector machine (LS-SVM).[31] Alternatively, a direct nonlinear model of the posterior class probability can be used, of which the most frequent is the multilayer perceptron, sometimes referred to as a *feedforward neural network*.

Different linear and nonlinear methods have been frequently benchmarked with similar classification performance. Occam's razor then suggests taking the simpler models: typically linear

© 2016 John Wiley & Sons, Ltd.

and preferably probabilistic. However, more value for clinical applications will typically be derived from interpretations of models that support clinical reasoning with data-based evidence. This is the focus of the latent variable models discussed earlier. Alternatively, if the aim is to study the predictive value of individual spectral values or features, it is then necessary to select the most informative ones.

Sequential Feature Selection. All classifiers may be applied to a predetermined set of spectral components or to features of known clinical relevance. An alternative approach for finding the most informative predictors is to carry out a systematic search using statistical feature selection methods. Sequential feature selection methods proceed either by successfully building a model one variable at a time, known as *forward selection*, or by starting with all available variables and pruning them one at a time, in backward selection.

For linear models such as LogR, there are well-understood models of classification error from which significance values can be generated for each model parameter. This automatically ends the process of selecting variables at a particular significance threshold set by a traditional p-value. Moreover, forward selection tests at each step for variables that can be dropped from, as well as added to the model.

In nonlinear models such as neural networks and SVM, no such model of the error exists; hence, the procedures for selecting variables are more pragmatic and sequentially increase or reduce the mode based on heuristic measures of the change in classification accuracy. Other related approaches use correlation-based heuristics to evaluate the worth or merit of features. These are called *correlation-based feature subset* (*CFS*) *evaluators*[32]; they evaluate and hence rank feature subsets rather than individual features. More complex approaches to take account of multiple testing using bootstrap resampling of linear models have also been published for spectral classification.[33]

Unsupervised Methods. Unsupervised learning studies the problem of learning from unlabeled data, based on the similarity of patterns, and creates a model that reflects the statistical structure of the overall collection of such data, assigning each pattern to a previously unknown class. Clustering or cluster analysis can be regarded as one of the most important unsupervised learning problems. It refers to the task of finding a structure in a collection of unlabeled samples or observations, and grouping them such that samples/observations in the same group (or cluster) are more similar (according to a particular criterion) to each other, than to those in other groups.

The clustering task can be accomplished by several algorithms that differ significantly in their definition of clusters and how to efficiently find them. The appropriate clustering algorithm would depend on the individual data set and the problem intended to be solved. A good clustering method would be one that creates clusters with high similarity intraclass or intracluster and low interclass similarity. From the wide range of clustering algorithms available, here, we briefly describe k-means clustering, one of the most commonly used methods in the analysis of MRS data.

k-Means Clustering. Arguably, the most widely used clustering algorithm, k-means is an iterative method to find k prototypes that best represent the variance of the data. This method optimizes a quadratic function, which is the sum of within-cluster distances. However, it requires that the number of prototypes is set in advance and can be prone to finding local minima of the cost function wherein different prototypes result from different random initializations. There are no generally accepted initialization methods.[34] In addition, the results of k-means clustering are strongly dependent on the choice of features selected to represent the spectra, as these determine the distances between data points.

Semisupervised Methods. A more advanced approach to unsupervised methods is to define a metric in the space of spectral features using a probabilistic classifier. This will ensure that the application of clustering methods with the estimated metric will naturally group together spectra with similar classification. The methodology proposed in Ref. 35 takes advantage of a useful feature of probabilistic models, namely that the change in information with respect to perturbations of the data defines a *bona fide* metric whose local properties are measured by the Fisher information matrix. This enables more advanced latent variable models such as NMF and convex-NMF, discussed earlier (see section titled 'Dimensionality Reduction'), to be applied when the sources sought are difficult to separate, for instance to separate the spectra of metastatic brain masses from those of glioblastomas.

Performance Evaluation. The evaluation of models of MRS has two distinct components. One is the interpretation of the model with respect to clinical expertise. This pertains to the structure of the latent variables identified in dimensionality reduction studies, the features selected by data-driven classifiers and the cluster structure in unsupervised studies. The second component is the predictive value of the models. For classifiers, the clinically accepted standard for performance evaluation is the receiver operating characteristic (ROC) framework. It is always necessary to estimate classification accuracy using data separate from the modeling data set, to mitigate the risk of overfitting. This also applies to linear models but especially to flexible or nonlinear classifiers. When the available sample size is low, hold-out samples can be used in the process of cross-validation. A more robust performance estimations strategy is the bootstrap, which requires repeated resampling of the data with replacement, if computational resources permit this.

In brief, the ROC permits several distinct views of the predictive performance to be characterized. First, there is the overall area under the receiver operating characteristic (AUROC) curve. This quantifies the discriminating ability of different classifiers without setting a threshold for classification, thus separating the tasks of discrimination and the selection of an operating point. From an operational point of view, the most relevant measures are the proportion of true cases detected (sensitivity) and the proportion of true negative cases excluded (specificity), the latter relating to the false positive detection rate. Neither of these measures depends on the prevalence of the data in each class, so they are robust evaluators for unbalanced

© 2016 John Wiley & Sons, Ltd.

data. They are sometime combined in a single equivalent accuracy weighted for balanced data, called the *balanced error rate* (*BER*). However, this can be misleading for practical purposes, as we see in the following.

In practice, arguably the most important single measure is a combination of the earlier indices and the prevalence to estimate the positive predictive value (PPV), which is the proportion of positive predictions that are correct. However, it is readily shown that the PPV will necessarily be limited by low prevalence. In order to derive clinical value from cases where the prevalence of the class to be detected is lower than about 10%, care must be taken to specify the classification model in a form that relates to an operational service model.

Evaluation methods allow the estimation of the predictive ability of the model, that is, the accuracy of the model when out-of-sample (i.e., new or previously unseen) but similar data are used. This validation is an essential part of the life cycle of the development of a classifier. Some of the most commonly used methods for evaluating MRS-based classifiers are listed and briefly described as follows.

- *Confusion matrix, sensitivity, and specificity*: The confusion matrix, also known as a *contingency table*, represents the count of a classifier's class predictions with respect to the actual outcome. Each row of the matrix represents the members in a predicted class, whereas each column represents the actual value of members in the original class. It leads naturally to the concepts of sensitivity and specificity. Sensitivity (true positive rate) measures the proportion of actual positives that are correctly identified as such, and specificity (false positive rate) measures the proportion of negatives that are correctly identified as such. A good predictor would be one with a high sensitivity and specificity.
- *k-fold cross-validation*: one round of cross-validation involves partitioning a data set into complementary subsets, performing the training on one subset, and validating the model on the other. In *k*-fold cross-validation, the original data set is partitioned into *k* subsamples. Of the *k* subsamples, a single subsample is retained for testing the model, and the remaining $k-1$ subsamples are used as training data. The cross-validation process is then repeated *k* times (as in *k*-fold), with each of the *k* subsamples used exactly once as test data. The *k* results from the *k*-fold cross-validations can then be averaged to produce a single estimation of prediction accuracy.[28]
- *Leave-one-out (LOO)*: this is a special and extreme case of a *k*-fold cross-validation. It uses a single case from the original data set for testing, and the remaining cases are used for training the model. This is repeated so that each case in the data set is used once as test data. This is the same as a *k*-fold cross-validation with *k* being equal to the number of cases in the original data set.
- *Bootstrapping*: it is implemented by constructing a number *n* of bootstrap cases of the observed data set (and of equal size to the training data set), each of which is obtained by random sampling with replacement from the original data set (there is almost always duplication of individual cases in a bootstrap data set). The *n* results from the bootstrap samples can then be averaged to produce a single estimation.[28] Bootstrapping may be better at estimating error rates in a linear discriminant problem compared to simple cross-validation.[36]
- *ROC curve:* it is a graphical plot of the sensitivity vs 1-specificity for a binary classifier system as its discrimination threshold is varied.[37] It is complemented numerically with the area under the curve (AUC), whose estimation can be interpreted as the probability that the classifier will assign a higher score to a randomly chosen positive sample than to a randomly chosen negative sample.
- *BER*: it is the average of the error rate of the classes, i.e., the average of the proportion of wrong classifications in each class. It is a useful measure for problems in which classes are imbalanced (not an unusual characteristic of MRS data sets) and the classifier could therefore be biased toward the most frequent class.
- *Balanced accuracy rate:* it is defined as the arithmetic mean of the sensitivity and specificity or the average accuracy obtained for the classes. As in the BER, it is useful because it avoids overstated performance estimates on imbalanced data sets. If the classifier performs equally well on all of the classes, this metric reduces to the conventional accuracy (that is, the number of correct predictions divided by the total number of predictions).[38]
- *Performance with out-of-sample test sets*: there is consensus in the literature that the use of a fully independent test set to validate the 'correctness' of a model is one of the most robust evaluation strategies (provided new cases are available) and is, to some extent, complementary to the methods described earlier. It assesses whether a model derived from an analysis of the original data set is transportable to similar cases in another location, providing an insight into the generalization applicability and validity of the model.[39] In favorable cases, it may even help to choose the optimal classifier for further evaluation (Figure 4b).

Which Pattern Recognition Technique is the Best?. As has been shown, the same classification problem can be analyzed by many different techniques for feature selection or classification. However, when dealing with the practical application of PR techniques to clinical problems, there are two important questions with respect to the presentation of PR results: namely, does PR method 'A' (for example, a novel method) perform significantly differently than, (i) existing PR method 'B' (for example, LDA) and (ii) technique 'C' (for example, the conventional accepted MRS analysis method)?

When comparing the performance of two different methods ('A' and 'B') on a two-class classifier for the same sample of cases, the McNemar test for two proportions on related groups[40] is appropriate. If there are more than two classes, then Cochran's *Q* test[41] can be used. If we just want to measure agreement among different PR techniques, we can use Cohen's Kappa.[40] Sometimes, classifier results are given as ROC curves; in these cases, the Hanley McNeil test for comparing two AUCs is the test of choice.[42]

© 2016 John Wiley & Sons, Ltd.

Lessons Learned from Previous Studies

An approach used to overcome the curse of dimensionality has sometimes been to create superclasses (aggregates of classes) that have similar spectral patterns.[8] This requires either prior knowledge of the diseases of study and their spectral patterns or a pilot PR study to determine that these spectral patterns are in effect too similar to be easily separable.

One of the most important lessons learned from the brain tumor studies that have dealt with classification was that easy discriminations normally yield similar performances, no matter the choice of the feature selection or extraction, and

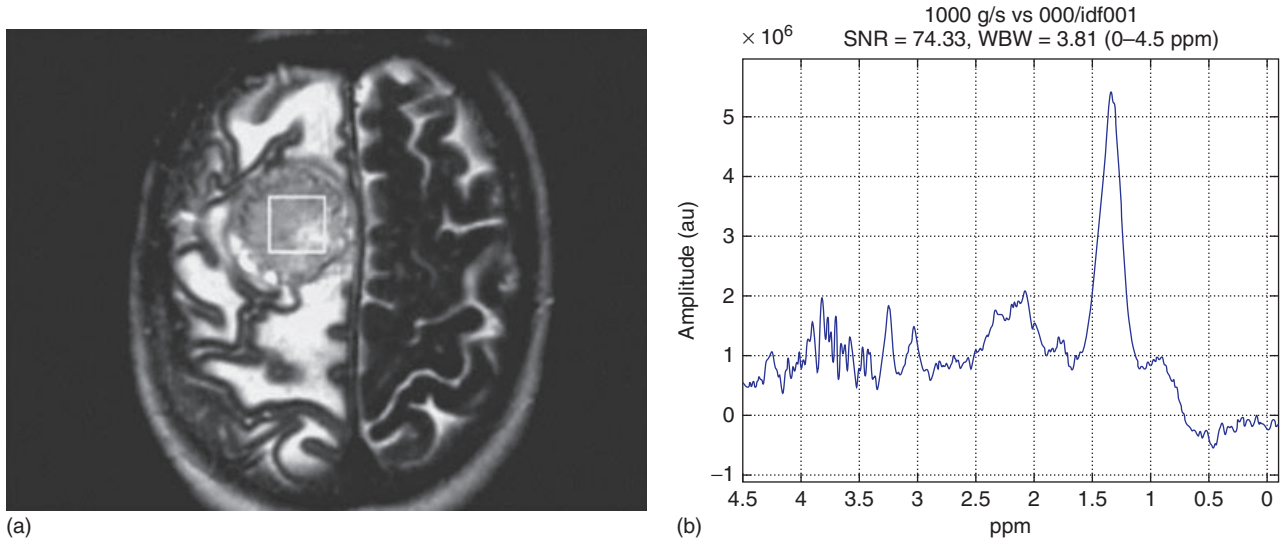

Figure 5. Case I0009 from the INTERPRET database.[1] The histopathology-validated diagnosis was low-grade meningioma. (a) T_2 weighted scan. (b) Short TE spectrum, in which lipids at 0.9 and 1.3 ppm are observed and an abnormal choline-to-creatine ratio for a low-grade meningioma (atypical for low-grade meningioma). The case was categorized as an outlier by two independent studies that used different PR techniques.[45,46] The explanation is that the superior limit of the voxel was close to the skull (see the brain circumvolutions), where contamination from subcutaneous fat was likely

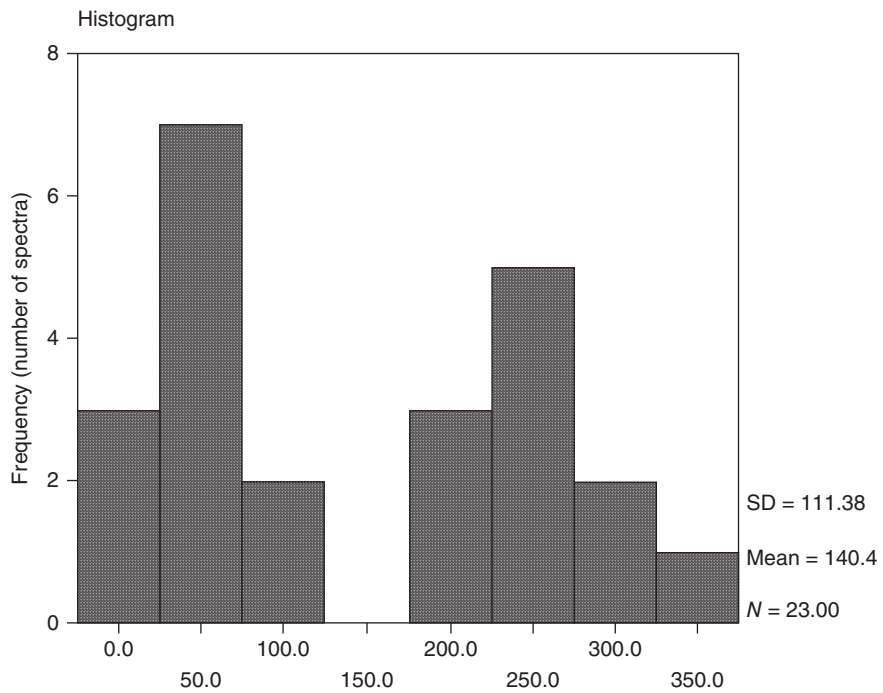

Figure 6. Histogram of the data point at 1.326 ppm for the glioblastoma group showing a bi-modal distribution of the intensity at this point. (Reproduced with permission from Ref. 48. © John Wiley & Sons, Ltd., 1998.)

© 2016 John Wiley & Sons, Ltd.

classification techniques.[43] Despite the fact that classes sometimes have overlapping spectra, linear and nonlinear classifiers usually perform similarly[44] for these types of discriminations. While this may seem disappointing for PR practitioners, it has an interesting corollary: spectra that are consistently misclassified by different PR techniques will be, most probably, outliers (Figure 5).

In contrast, if the problem is a difficult discrimination, such as distinguishing glioblastoma from metastasis with ^{1}H SV MRS, or among types of low-grade glial brain tumors, caution is required as good results obtained using a particular data set may not generalize to independent data sets,[8] for instance acquired from patients at a different hospital. Therefore, if we would like to apply a particular classifier to unseen patient MRS data or in any relevant application, we must ensure that it has undergone a previous external validation.[47] An important fact to consider as well is that not all classes are homogeneous, that is to say, they may not follow a normal distribution. One example is the spectral pattern of glioblastoma, as assessed by ^{1}H MRS (Figure 6).

There is an important difference between a classifier in the hands of a mathematician and the same classifier used as a routine tool by third parties (e.g., radiologists) on their laptop computers. While we can develop a classifier to distinguish disease 'A' from 'B', in a real-life situation MR spectra belonging to 'C', 'D', and 'E' will also be submitted to that classifier. It is crucial that the final user of the classifier is made aware of the classes and data types that it is able to deal with. Otherwise, the classifier in question would not only be unfairly tested but, more critically in clinical MRS, risk harm to patients if it leads to errors in patient management. For example, a classifier developed with 1.5 T ^{1}H MRS of adults should not be tested on 7 T spectra unless it has first been demonstrated that both data types would be compatible. Unfortunately, this may require a lengthy and expensive data collection process and validation, which may be less attractive unless the potential for improvement at 7 T were high and the performance at 1.5 T were sub-par, for example.

Finally, despite their attractiveness from the PR point-of-view, two-class classifiers will find a clinical application only in particular situations and clinicians will normally demand multiclass classifiers. This raises the question of what the main purpose of using advanced PR methods should be for clinical MRS. PR may be suitable as an initial screen among multiple classes, but for class labeling, it may be best suited to differential diagnosis between two classes where an ambiguity remains after information from other pertinent modalities has

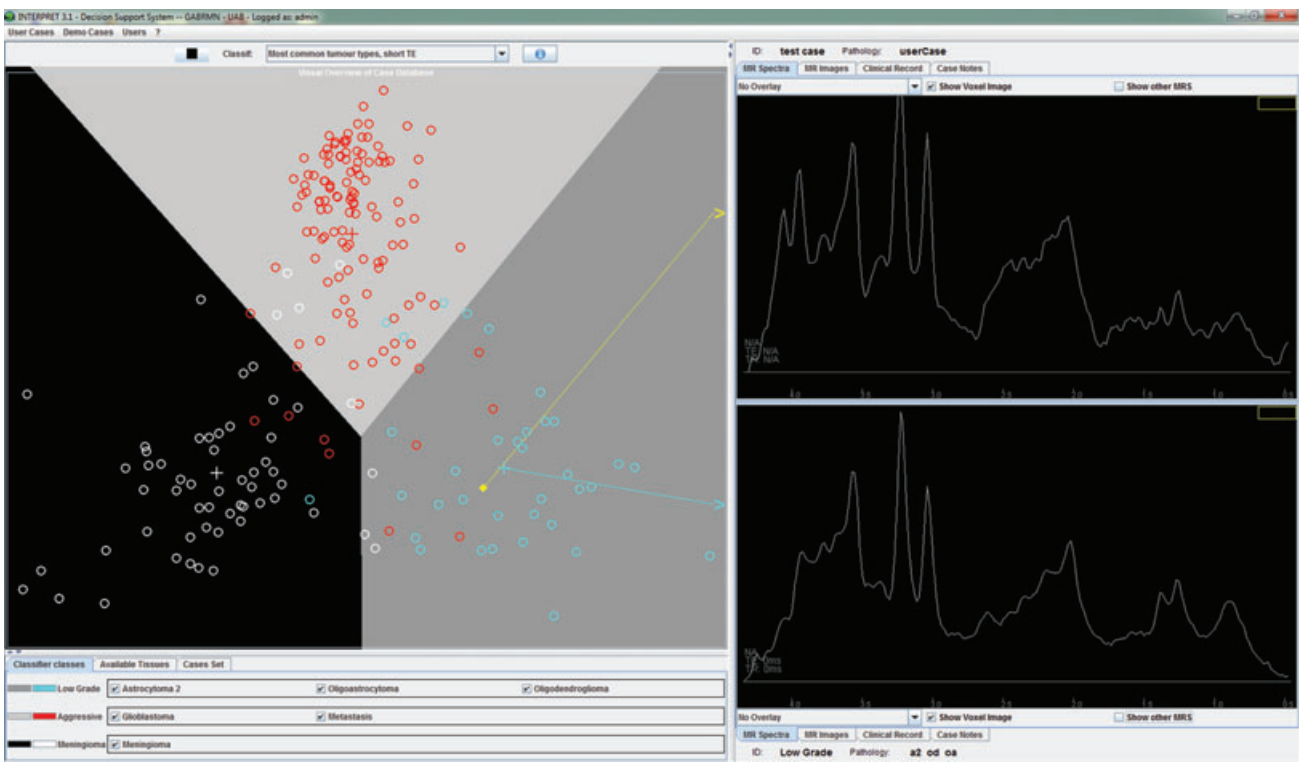

Figure 7. The INTERPRET DSS version 3.1 screen for the 'low grade glial vs aggressive tumors vs low grade meningioma' short-TE classifier. The screen is divided in two main parts, left and right. The overview space of cases in the database is displayed on the left side, where each case is a colored circle (see legend on the bottom left). The position of each case is determined by the result of the PR algorithm, in this case sequential forward feature selection followed by LDA, as described in Ref. 14. The right side has two panels (top and bottom) for comparison of the spectra from two different individual cases. The top right panel displays the short-TE spectrum of an astrocytoma of WHO grade II from the study,[8] which is displayed as a yellow symbol in the overview space, correctly classified as low-grade glioma. The bottom right panel displays the mean TE spectrum of the corresponding class from the INTERPRET validated database

© 2016 John Wiley & Sons, Ltd.

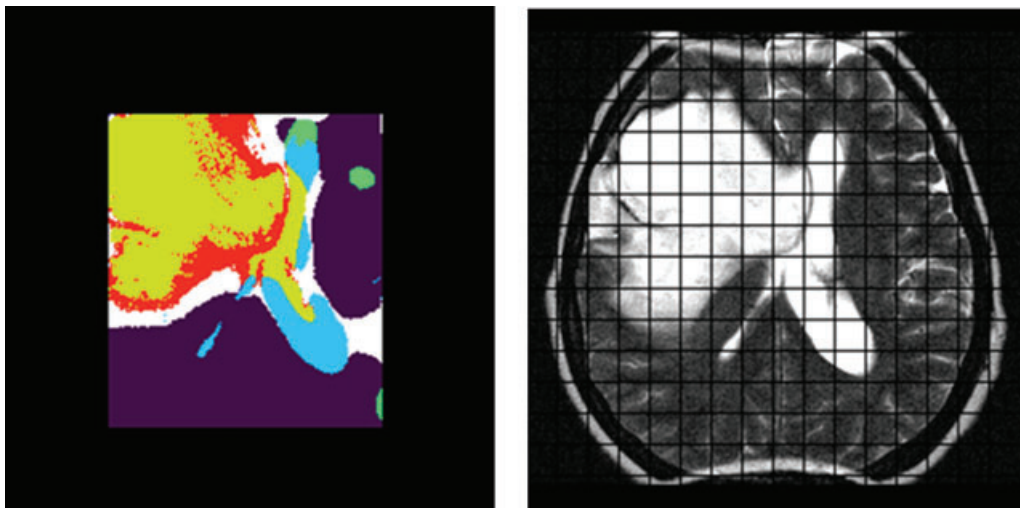

Figure 8. Nosological image. (Adapted from Ref. 5.) (a) Nosological image (yellow, necrosis; red, high-grade glioma; light blue, cerebrospinal fluid; purple, normal brain; light green, meningioma). The tumor was a high-grade glioma. (b) Morphological T_2-weighted image with the spectroscopic grid overlaid (black lines)

been exhausted. These concerns, as well as the need for user-friendly data input and evaluation of classifier output, have led to the development of decision-support systems.[8] It is in this step where the final clinical or biomedical user will experience classifiers. Having such systems will then require, for example, well-documented data sets, processing pipelines, and classifiers developed by PR practitioners. Unfortunately to date, most findings on PR with MRS have not yet reached that stage of maturity.

An important issue for the final users of a classifier is the visualization of their results. The possibility of seeing each spectrum as a symbol in a 2-D or 3-D space, where its position is determined by the PR algorithm, has been shown to be a successful approach, and it has even been implemented in a decision-support system (Figure 7).[8,14,49] On the other hand, for MV data, the most successful representation has been the 'nosologic image' concept, where each pixel or voxel is colored according to the predicted class (Figure 8)[5,50]. An interesting approach related to visualization of MV data was developed for classifying voxels in an MV study of brain tumor patients, as belonging to the investigated classes or to 'unknown' or 'undecided' (Figure 9), by applying a threshold based on Mahalanobis inter- and intraclass distances.[12]

With respect to the tools to perform PR analyses, the MRS specialist has the choice to either use mathematical platforms, such as R or Matlab, or more user-friendly programs that facilitate performing PR of their MRS data. A commercial alternative is SIMCA (http://www.umetrics.com/products/spectroscopyskin), which provides PCA and other techniques such as partial least squares (PLS) and is widely used by the '-omics' community. A free software suite that integrates PCA, sequential feature selection, and LDA is SpectraClassifier; it has been shown to perform robustly with ^{1}H MRS from brain tumors.[26,27]

Acknowledgments

Authors are funded by MOLIMAGLIO (SAF2014-52332-R) from MINECO (ES). M.J. and C.A. are also funded by CIBER-BBN [Centro de Investigación Biomédica en Red – Bioingeniería, Biomateriales y Nanomedicina (http://www.ciber-bbn.es/en)], an initiative of the Instituto de Salud Carlos III (Spain) co-funded by EU FEDER funds. SOM is funded by the European Union under the 7th FP, 'Marie Curie' FP7-PEOPLE-2012-IEF.

Biographical Sketches

Sandra Ortega-Martorell, b. 1981, BEng 2004, PhD 2012. She has published in the areas of blind source separation techniques, generative topographic models, and decision support systems for statistical analysis, all with application in biomedicine, particularly diagnosis of brain tumors. Research interest: using mathematics to solve real problems in the areas of biomedicine and bioinformatics.

Margarida Julià-Sapé, b. 1969, ''Licenciada'' in Biology 1994, PhD 2006. Joined the ''Universitat Autònoma de Barcelona'' in 2000 to be part of the team in a multicenter trial on PR of MRS data of brain tumors and, was later involved in other multi-center projects involving MRS with the data manager role. Research interests: applying pattern recognition techniques to MRS data of brain tumors to improve their diagnosis and follow-up. Other interests: MRS data curation, facilitating the uptake of MRS by clinicians through decision support.

Paulo J.G. Lisboa, b. 1958, BSc 1979, PhD 1983 (Liverpool, UK). First studied the discriminant structure of metabolites from brain MRS data in 1998. This led to the identification of underlying sources with independent components analysis in 2000, which were verified with a Bayesian model of signal separation in 2003. In 2011, new matrix factorization approaches enabled better delineation of functional tissues, changing the methodology from single-voxel studies of class membership to nosological analysis of tumor heterogeneity. Present interests in delineation of in vivo tissue composition with spectral methods and inferring data structure from complex probabilistic classifiers.

Carles Arús, b. 1954, BSc 1976, PhD 1981 (Barcelona, ES). First involved with protein NMR in 1979 in Naples (IT). Postdoctoral work

© 2016 John Wiley & Sons, Ltd.

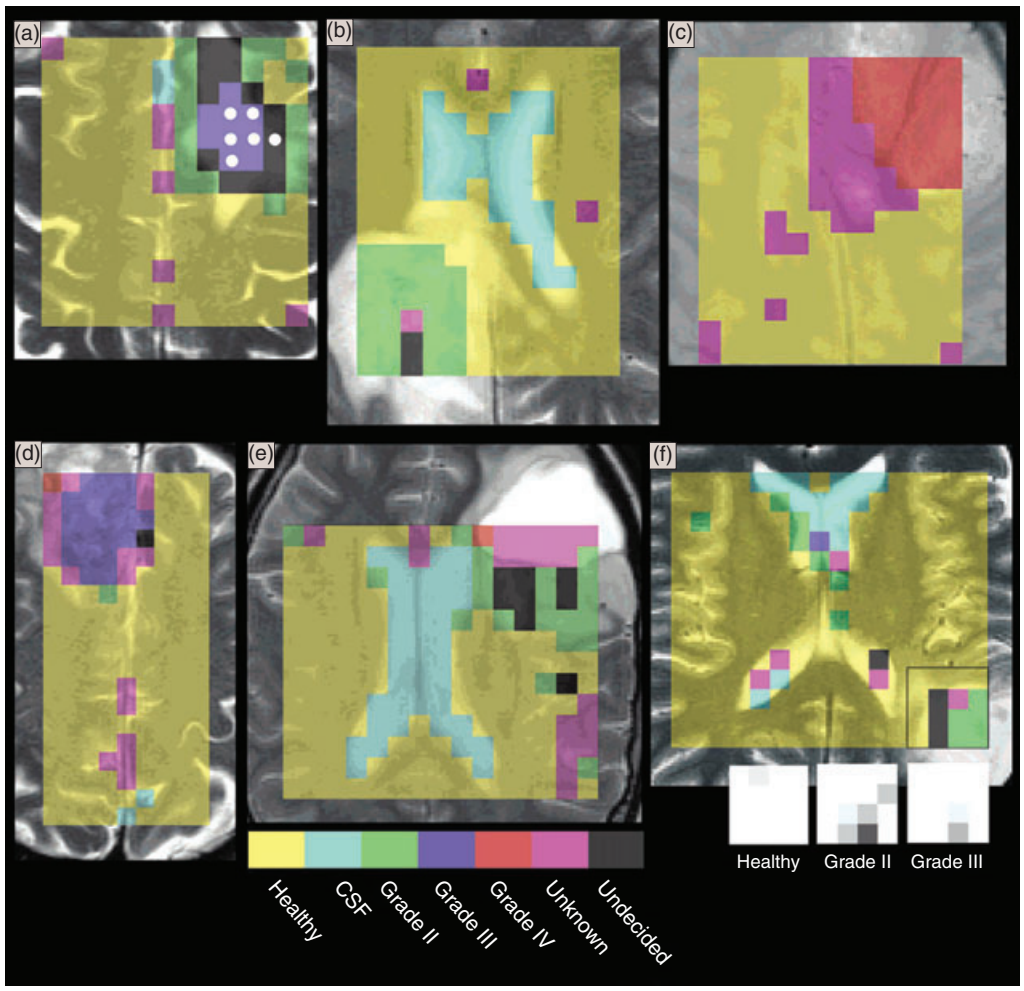

Figure 9. Classification result of six brain tumor patients from the INTERPRET project data set. Please access the INTERPRET database[1] and http://gabrmn.uab.es/interpretvalidateddb/ for additional patient information. The classification is projected over T_2-weighted (except part c, which is proton-density weighted) MRIs with a contrast that clearly shows the tumorous region: (a) Patient I1229, pathology: grade III; (b) Patient I1275, grade II; (c) Patient I1285, grade IV; (d) Patient I1212, grade III; and (e) Patient I1318 (42/195), grade II. Part (f) is from Patient I1281, with grade II–III pathology (no radiological consensus so voxels were not used for training classifiers). Note that a voxel classified as 'undecided' can be either grade II or a grade III tumor tissue. The inset shows a section of three probability maps for healthy, grade II, and grade III classifications (black, high probability; white, low probability). (Reprinted with permission from Simonetti AW, Melssen WJ, van der Graaf M, Postma GJ, Heerschap A, Buydens LM, A chemometric approach for brain tumor classification using magnetic resonance imaging and spectroscopy, Anal Chem., 2003; 75(20): 5352–61. © 2003 American Chemical Society.)

in muscle NMR with Michael Bárány and John L. Markley (1982–1985, USA). Since 1985 at the Department of Biochemistry and Molecular Biology of UAB (Barcelona, ES) and leading the GABRMN since then (http://gabrmn.uab.es/). Present interests in the field of MR-based molecular imaging of brain tumors, for diagnosis, prognosis, and therapy planning.

Related Article

Clinical Trials of MRS Methods

References

1. M. Julia-Sape, D. Acosta, M. Mier, C. Arus, and D. Watson, *Magn. Reson. Mater. Phy.*, 2006, **19**, 22.
2. S. L. Howells, R. J. Maxwell, F. A. Howe, A. C. Peet, M. Stubbs, L. M. Rodrigues, S. P. Robinson, S. Baluch, and J. R. Griffiths, *NMR Biomed.*, 1993, **6**, 237.
3. M. C. Preul, Z. Caramanos, D. L. Collins, J. G. Villemure, R. Leblanc, A. Olivier, R. Pokrupa, and D. L. Arnold, *Nat. Med.*, 1996, **2**, 323.
4. J. P. Usenius, S. Tuohimetsa, P. Vainio, M. Ala-Korpela, Y. Hiltunen, and R. A. Kauppinen, *Neuroreport*, 1996, **7**, 1597.
5. F. S. De Edelenyi, C. Rubin, F. Esteve, S. Grand, M. Decorps, V. Lefournier, J. F. Le Bas, and C. Rémy, *Nat. Med.*, 2000, **6**, 1287.
6. P. J. G Lisboa, E. Romero , A. Vellido, M. Julia-Sape, C. Arus, The Seventh International Conference on Machine Learning and Applications (ICMLA'08). 2008, 613.
7. R. L. Somorjai, B. Dolenko, and R. Baumgartner, *Bioinformatics*, 2003, **19**, 1484.

© 2016 John Wiley & Sons, Ltd.

8. A. R. Tate, J. Underwood, D. M. Acosta, M. Julia-Sape, C. Majos, A. Moreno-Torres, F. A. Howe, M. van der Graaf, V. Lefournier, M. M. Murphy, A. Loosemore, C. Ladroue, P. Wesseling, J. Luc Bosson, M. E. Cabañas, A. W. Simonetti, W. Gajewicz, J. Calvar, A. Capdevila, P. R. Wilkins, B. A. Bell, C. Rémy, A. Heerschap, D. Watson, J. R. Griffiths, and C. Arús, *NMR Biomed.*, 2006, **19**, 411.
9. S. W. Coons, P. C. Johnson, B. W. Scheithauer, A. J. Yates, and D. K. Pearl, *Cancer*, 1997, **79**, 1381.
10. A. Sottoriva, I. Spiteri, S. G. Piccirillo, A. Touloumis, V. P. Collins, J. C. Marioni, C. Curtis, C. Watts, and S. Tavaré, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 4009.
11. A. Sottoriva, H. Kang, Z. Ma, T. A. Graham, M. P. Salomon, J. Zhao, P. Marjoram, K. Siegmund, M. F. Press , D. Shibata, and C. Curtis, *Nat. Genet.*, 2015, **47**, 209.
12. A. W. Simonetti, W. J. Melssen, M. van der Graaf, G. J. Postma, A. Heerschap, and L. M. Buydens, *Anal. Chem.*, 2003, **75**, 5352.
13. S. W. Provencher, *Magn. Reson. Med.*, 1993, **30**, 672.
14. A. Perez-Ruiz, M. Julia-Sape, G. Mercadal, I. Olier, C. Majos, C. Arus, and The INTERPRET, *BMC Bioinformatics*, 2010, **11**, 581.
15. I. Guyon and A. Elisseeff, *J. Mach. Learn. Res.*, 2003, **3**, 1157.
16. T. Hastie, R. Tibshirani, and J. H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn, Springer: New York, 2009.
17. T. F. Bathen, B. Geurts, B. Sitter, H. E. Fjosne, S. Lundgren, L. M. Buydens, I. S. Gribbestad, G. Postma, and G. F. Giskeødegârd, *PLoS One*. 2013;**8**(4):e61578.
18. C. Jutten and J. Herault, *Signal Process.*, 1991, **24**, 1.
19. A. Hyvärinen and E. Oja, *Neural Netw.*, 2000, **13**, 411.
20. Y. Y. B. Lee, Y. Huang, W. El-Deredy, P. J. G. Lisboa, C. Arus, and P. Harris, *IEE Proc. Sci. Meas. Technol.*, 2000, **147**, 309.
21. D. D. Lee and H. S. Seung, *Nature*, 1999, **401**, 788.
22. P. Paatero and U. Tapper, *Environmetrics*, 1994, **5**, 111.
23. S. Ortega-Martorell, P. J. Lisboa, A. Vellido, M. Julia-Sape, and C. Arus, *BMC Bioinformatics*, 2012, **13**, 38.
24. C. Ding, T. Li, and M. I. Jordan, *IEEE Trans. Pattern Anal. Mach. Intell.*, 2010, **32**, 45.
25. S. Ortega-Martorell, P. J. G. Lisboa, A. Vellido, R. V. Simoes, M. Pumarola, M. Julià-Sapé, and C. Arús, *PLoS One*. 2012;**7**(10):e47824.
26. S. Ortega-Martorell, I. Olier, M. Julia-Sape, and C. Arus, *BMC Bioinformatics*, 2010, **11**, 1.
27. T. Delgado-Goni, M. Julia-Sape, A. P. Candiota, M. Pumarola, and C. Arus, *NMR Biomed.*, 2014, **27**, 1333.
28. R. O. Duda, P. E. Hart, and D. G. Stork, in Pattern Classification, 2nd edn, eds R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, Inc: New York; Chichester, 2001, 654pp.
29. P. Degraeuwe, G. Jaspers, N. Robertson, and A. Kessels, *Syst. Rev.*, 2013, **2**, 96.
30. C. Cortes and V. Vapnik, *Mach. Learn.*, 1995, **20**, 273.
31. J. A. K. Suykens and J. Vandewalle, *Neural Process. Lett.*, 1999, **9**, 293.
32. M. A. Hall, Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. Proceedings of the Seventeenth International Conference on Machine Learning. 657793: Morgan Kaufmann Publishers Inc.; 2000, pp. 359–366.
33. P. J. Lisboa, S. P. Kirby, A. Vellido, Y. Y. Lee, and W. El-Deredy, *NMR Biomed.*, 1998, **11**, 225.
34. S. J. Chambers, I. H. Jarman, T. A. Etchells, and P. J. G. Lisboa, *Int. J. Biomed. Eng. Technol.*, 2013, **13**, 323.
35. S. Ortega-Martorell, H. Ruiz, A. Vellido, I. Olier, E. Romero, M. Julià-Sapé, J. D. Martín, I. H. Jarman, C. Arús, and P.J.G. Lisboa, *PLoS One*. 2013;**8**(12):e83773.
36. B. Efron and R. J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall: New York, London, 1993.
37. T. Fawcett, *Pattern Recognit. Lett.*, 2006, **27**, 861.
38. K. H. Brodersen, O. Cheng Soon, K. E. Stephan, J. M. Buhmann. 20th International Conference on Pattern Recognition (ICPR). 2010, 3121.
39. D. G. Altman and P. Royston, *Stat. Med.*, 2000, **19**, 453.
40. A. Petrie and C. Sabin, Medical Statistics at a Glance, 3rd edn, Chichester: Wiley-Blackwell, 2009.
41. W. G. Cochran, *Biometrika*, 1950, **37**, 256.
42. J. A. Hanley and B. J. McNeil, *Radiology*, 1983, **148**, 839.
43. J. M. Garcia-Gomez, J. Luts, M. Julia-Sape, P. Krooshof, S. Tortajada, J. V. Robledo, W. Melssen, E. Fuster-García, I. Olier, G. Postma, D. Monleón, A. Moreno-Torres, J. Pujol, A. P. Candiota, M. C. Martínez-Bisbal, J. Suykens, L. Buydens, B. Celda, S. Van Huffel, C. Arús, and M. Robles, *Magn. Reson. Mater. Phy.*, 2009, **22**, 5.
44. C. L. C. Ladroue, Pattern Recognition Techniques for the Study of Magnetic Resonance Spectra of Brain Tumours, University of London: London, 2004.
45. J. M. Garcia-Gomez, S. Tortajada, C. Vidal, M. Julia-Sape, J. Luts, A. Moreno-Torres, S. Van Huffel, C. Arús, and M. Robles, *NMR Biomed.*, 2008, **21**, 1112.
46. A. Vellido, E. Romero, F. F. González-Navarro, L. A. Belanche-Muñoz, M. Julià-Sapé, and C. Arús, *Neurocomputing*, 2009, **72**, 3085.
47. D. G. Altman, Y. Vergouwe, P. Royston, and K. G. Moons, *Br. Med. J.*, 2009, **338**, b605.
48. A. R. Tate, J. R. Griffiths, I. Martinez-Perez, A. Moreno, I. Barba, M. E. Cabanas, D. Watson, J. Alonso, F. Bartumeus, F. Isamat, I. Ferrer, F. Vila, E. Ferrer, A. Capdevila, and C. Arús, *NMR Biomed.*, 1998, **11**, 177.
49. M. Julià-Sapé, C. Majós, A. Camins, A. Samitier, M. Baquero, M. Serrallonga, S. Doménech, E. Grivé, F. A. Howe, K. Opstad, J. Calvar, C. Aguilera, and C. Arús, *NMR Biomed.*, 2014, (**9**), 1009.
50. J. Luts, T. Laudadio, A. J. Idema, A. W. Simonetti, A. Heerschap, D. Vandermeulen, J. A. Suykens, and S. Van Huffel, *NMR Biomed.*, 2009, **22**, 374.

© 2016 John Wiley & Sons, Ltd.