Neurocomputing 73 (2010) 622–632

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Feature and model selection with discriminatory visualization for diagnostic classification of brain tumors

Félix F. González-Navarro [a], Lluís A. Belanche-Muñoz [a,*], Enrique Romero [a], Alfredo Vellido [a], Margarida Julià-Sapé [b,c], Carles Arús [c,b]

[a] Dept. de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
[b] Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), Cerdanyola del Vallès, Spain
[c] Grup d'Aplicacions Biomèdiques de la RMN (GABRMN), Dept. de Bioquímica i Biología Molecular (BBM), Unitat de Biociències, Universitat Autònoma de Barcelona (UAB), Cerdanyola del Vallès, Spain

ARTICLE INFO

Available online 3 December 2009

Keywords:
Proton magnetic resonance spectroscopy
Brain tumors
Feature selection
Data visualization
Medical decision support systems

ABSTRACT

Machine Learning (ML) and related methods have of late made significant contributions to solving multidisciplinary problems in the field of oncology diagnosis. Human brain tumor diagnosis, in particular, often relies on the use of non-invasive techniques such as *Magnetic Resonance Imaging* (MRI) and *Spectroscopy* (MRS). In this paper, MRS data of human brain tumors are analyzed in detail.

The high dimensionality of the MR spectra makes difficult both their classification and the interpretation of the obtained results, thus limiting their usability in practical medical settings. The use of *dimensionality reduction* techniques is therefore advisable. In this work, we apply feature selection methods and several off-the-shelf classifiers on various ^{1}H-MRS modalities: long and short echo times and an *ad hoc* combination of both. The introduction of bootstrap resampling techniques permits the obtention of mean performance estimates and their variability. Our experimental findings indicate that the feature selection process enhances the classification performance compared to using the full set of features. We also show that the use of *combined* information from the different echo times is a better strategy for small numbers of spectral frequencies; however, the use of ever greater numbers of short echo time frequencies permits the obtention of many models with similar performance. The final induced models offer very attractive solutions both in terms of prediction accuracy and number of involved spectral frequencies, which are also amenable to metabolic *interpretation*. A linear dimensionality-reduction technique that preserves class discrimination capabilities is used for *visualizing* the data corresponding to the selected frequencies.

1. Introduction

Over the last decade, ML has made significant inroads in the fields of bioinformatics and biomedicine. One particular application area that has attracted the attention of both medical practitioners and data analysts is that of human oncology [1]. In this work we are specifically interested in quantitative information in the form of patients' biological signals. We analyze data corresponding to different types of human brain tumors, obtained by single-voxel proton magnetic resonance spectroscopy (^{1}H-MRS), with the purpose of developing reliable tools for the support of medical expert diagnostic decision making. Decisions in this area are extremely sensitive and are usually based on information obtained by non-invasive measurement techniques.

The analyzed data belong to a multi-center international database that contains cases of a number of brain tumor pathologies [2]. MRS provides a detailed metabolic fingerprint of the tumor-affected tissue that varies according to the *echo time* of the acquisition and can be used to characterize these pathologies. The echo time is a relevant parameter of ^{1}H-MRS measurement, given that, at short-echo times (SET), some of the metabolites are better resolved—although numerous overlapping resonances exist, making the spectra difficult to interpret [3]. The use of a long echo time (LET) yields less clearly resolved metabolites but also less baseline distortion, resulting in a more readable spectrum.

The available data have been acquired both at SET and LET and bundled into three groups or super-classes as described below; they are scarce and of high dimensionality, making their discrimination a non-trivial undertaking. Therefore, the need arises for the use of dimensionality reduction methods (feature selection and/or extraction) in order to reduce the overall complexity of the problem. We use an *entropic filtering* algorithm for feature selection as a fast method to generate relevant subsets of spectral frequencies. An in-depth feature selection study is performed, not only in LET and SET data, but in a *combination* of

* Corresponding author.
E-mail address: belanche@lsi.upc.edu (L.A. Belanche-Muñoz).

0925-2312/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2009.07.018

both echo times. Bootstrap resampling techniques are used to yield mean performance estimates and their variability, and thus a more reliable measure of predictive ability. The combination of feature selection and classification aims at obtaining simple models (in terms of low numbers of features) capable of good generalization.

We report experimental results that support the practical advantage of combining robust feature selection and classification in this application, as accurate classification is obtained with parsimonious and interpretable subsets of spectral frequencies. We also aim to progress in the comparison of performances for MRS data acquired at different echo times, as well as in the comparison of these with data that combine both echo times.

Of special importance in a practical medical setting is the *interpretability* of the solutions in terms of these spectral frequencies, something that limits the applicability of methods such as PCA or ICA (whose solutions involve weighted combinations of frequencies, instead of individual frequencies). Moreover, even if interpretable by mere inspection of the involved features, the final selection of spectral frequencies may still provide few clues about the structure of the classes (tumor types). In this medical context, data visualization in a *low-dimensional* representation space may become extremely important, as it would help radiologists to gain insights into this complex and highly sensitive domain. A linear dimensionality reduction technique that provides a data projection—while preserving the class discrimination achieved by a classifier—is also used in our study. The goal of combining feature selection and visualization is to increase the intuitive interpretability of the classifier results.

2. Literature review

Early attempts to study ^{1}H-MRS data in assessing human brain tumors *in vivo* can be traced back two decades [4]. This pioneering research showed that spectra corresponding to normal brain spectra and tumors differ significantly in terms of the presence/absence of different metabolites. Even though no ML analysis of spectra was done in establishing these differences, it was concluded that ^{1}H-MRS may help to differentiate tumors for diagnostic and therapeutic purposes, limiting the need for invasive and risky diagnostic procedures such as biopsies. This same line of research was followed by several other studies, e.g. [5–8]. Further work has shown that it is possible to employ ML techniques successfully for the diagnosis and grading of adult brain tumors [9], and for distinguishing between different brain tumor pathologies [10,11]. As explained in the introduction, ^{1}H-MRS data can be acquired at different echo times. There are few studies comparing the classification potential of different data acquisition modalities, and the existing ones give, overall, a slight advantage to using SET information (see e.g. [3,12]). Only recent works have attempted to use the information contained in both echo times simultaneously [13,14].

Previous work analyzing the same LET ^{1}H-MRS data used in this study resorted to PCA followed by a linear discriminant (LDC) to distinguish only between high-grade malignant tumors and meningiomas, obtaining a mean AUC (area under the ROC curve) of 0.94, using the first 6 principal components (PC) and a (2/3, 1/3) train/test partition repeated 200 times [11]. The same method was used to distinguish between high-grade malignant tumors and astrocytomas Grade II (part of the low-grade gliomas superclass described in Section 3.1), obtaining a mean AUC of 0.92, also using the first 6 PCs. Two drawbacks of PCA in this setting are that all the spectra may participate in the PCs, and the fact that the linear combination may mix both positive and negative weights, which might partly cancel each other. A further disadvantage is the lack of physical meaning of the extracted components. In [9], LDC with 6 spectral frequencies (3.72, 3.04, 2.31, 2.14, 1.51 and 1.20 ppm) achieved a 83% of correct classification on one independent test set, this time using exactly the same three super-classes that we have analyzed in this study.

For SET data alone, previous existing work analyzing the same ^{1}H-MRS data using 10 PCs obtained at most 85% of correct classification [10]. In the task of separating high-grade malignant tumors from meningiomas, a mean AUC of 0.95, (slightly better than that for LET) with 4 PCs was achieved, and a mean AUC of 0.97 with 3 PCs for separating high-grade malignant tumors from astrocytomas Grade II [11]. In this study it is was also reported that kernel-based methods present good results even without any feature reduction, but in general there are no significative differences among the classification techniques. In [9], LDC with 5 spectral frequencies (3.76, 3.57, 3.02, 2.35, 1.28 ppm) yielded 89% accuracy in an independent test set.

In one of the first studies to explicitly *compare* LET and SET information [3], it was found that SET yielded better results (81% of accuracy with LDC) than LET (78%), in agreement with [11]. From this setting, a natural step forward is the *combination* of echo times, as a way to boost classification results. A recent such investigation, using PCA and Relief [15] to reduce dimensionality, reported experimental results (using LDC) achieving 88.7% of accuracy, using between 12 and 22 spectral frequencies; with LET only, 82.50% using between 7 and 14 frequencies; and with SET only, 88.82% using between 5 and 11 frequencies [12].

In previous work by the authors, the ^{1}H-MRS LET data were analyzed with the purpose of obtaining classification models showing good generalization ability after a strong dimensionality reduction process [16]. In the present study we are interested in performing a more in-depth feature selection study in both LET and SET types of data by the introduction of bootstrap resampling techniques as well as confidence intervals for generalization error. It is important to point out that LET and SET spectral points are considered as two separate sets of features. Specifically, we are interested in assessing which spectral representation (LET, SET or their combination) is the most adequate for prediction purposes. As stated in the introduction, the final goal is the obtention of a reduced set of spectral frequencies that can be amenable to visualization and interpretation for radiologists or oncologists.

3. Materials and methods

3.1. The ^{1}H-MRS data

^{1}H-MRS is by no means a novel technique for the exploration of the brain, but its use for the routine diagnostic examination of brain abnormal tissue is far for standard in clinical practice. Among the reasons to explain this situation is that a simple visual interpretation of ^{1}H spectra does not easily lead to a clear diagnosis. Moreover, few radiologists (to whom the diagnostic decision pertains) are trained to use and make sense of this technique [17]. Instead, they often resort to magnetic resonance imaging (MRI) for diagnostic characterization. MRI has excellent spatial resolution but bad frequency resolution. On the contrary, single-voxel ^{1}H-MRS has poor spatial resolution but excellent frequency resolution, making it a rich source of metabolic information.

The echo time is an influential parameter in ^{1}H-MRS spectra acquisition. In SET spectra (20–40 ms) some metabolites are better resolved (e.g. lipids, myo-inositol, glutamine and glutamate). However, there may be numerous overlapping frequencies (e.g. glutamate/glutamine at 2.2 ppm and NAA at 2.01 ppm), making metabolic interpretation difficult [3]. The use of LET

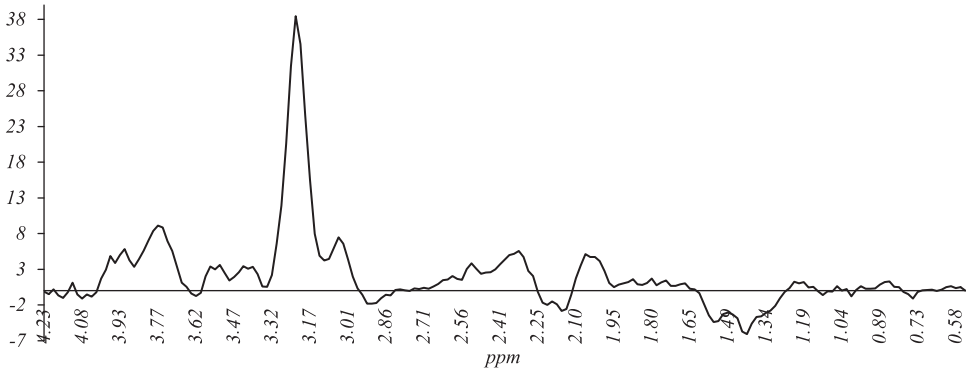

Fig. 1. Example mean spectra for a long echo time ^{1}H-MRS data set.

(270–288 ms) yields less clearly resolved metabolites but also less baseline distortion, resulting in a more readable spectrum. An example of a long echo time ^{1}H-MRS data set is depicted in Fig. 1.

The analyzed ^{1}H-MRS data were extracted from a database [2] resulting from the *International Network for Pattern Recognition of tumors Using Magnetic Resonance* (INTERPRET) European research project. For details on data acquisition and processing, and on database characteristics, please refer to [18] and [2]. Pathology (class) labeling was performed according to the World Health Organization (WHO) system for diagnosing brain tumors by histopathological analysis of a biopsy sample. The three data sets are detailed as follows:

- 217 SET (PRESS 30–32 ms) single voxel ^{1}H-MR spectra acquired in vivo from brain tumor patients. They include 58 meningiomas (*mm*), 86 glioblastomas (*gl*), 38 metastases (*me*), 22 astrocytomas grade II (*a2*), 6 oligoastrocytomas grade II (*oa*), and 7 (SET) and 5 (LET) oligodendrogliomas grade II (*od*).
- 195 LET (PRESS 135–144 ms) single voxel ^{1}H-MR spectra acquired in vivo from brain tumor patients. They include 55 meningiomas (*mm*), 78 glioblastomas (*gl*), 31 metastases (*me*), 20 astrocytomas grade II (*a2*), 6 oligoastrocytomas grade II (*oa*), and 5 oligodendrogliomas grade II (*od*).
- 195 items built by combination (through concatenation) of single voxel ^{1}H-MR spectra measured at two echo times (SET and LET). We call LSET the combined LET plus SET ^{1}H-MRS data set.

In the experiments included in this study, spectra were bundled into three groups or super-classes, namely: G1: *low grade gliomas* (*a2*, *oa* and *od*); G2: *high grade malignant tumors* (*me* and *gl*); and G3: *meningiomas* (*mm*). Only the clinically relevant regions of the spectra were analyzed. They consist of frequency intensity values measured in parts per million (ppm), an adimensional unit of relative frequency position in the data vector, starting at 4.25 ppm. These frequencies become the data features in all cases.

3.2. Entropic filtering

Information-theoretic measures have been used with success as a criterion for feature selection in ML tasks (see e.g. [19] for a recent compilation). Specifically, *mutual information* measures the mutual dependence of two random variables. In this work we use this concept embedded in a fast algorithm that computes the mutual information between a subset of variables and the class variable by generating first a "super-feature", obtained considering the concatenation of each combination of possible values of its forming features. In symbols, let $X=\{X_1,\ldots,X_n\}$ be the original feature set and consider a subset $\tau=\{\tau_1,\ldots,\tau_k\}$. A single feature $\mathcal{V}_\tau$ can be obtained uniquely, whose possible values are the concatenations of all possible values of the features in τ (for completeness, define $\mathcal{V}_\emptyset=\emptyset$). The conditional entropy between $\mathcal{V}\tau$ and the class feature Y is then:

$$H(Y|\tau_1,\ldots,\tau_k)=H(Y|\mathcal{V}_\tau)=-\sum_{v\in\mathcal{V}_\tau}\sum_{y\in Y}p(v,y)\log\frac{p(v,y)}{p(y)}. \quad (1)$$

Proceeding in this way, mutual information can be determined as a simple bivariate case: $I(\mathcal{V}_\tau;Y)=H(Y)-H(Y|\mathcal{V}_\tau)$. An *index of relevance* of the feature $X_i\in X$ to a class Y with respect to a subset $\tau\subset X$ is given by

$$R(X_i;Y|\tau)=\frac{I(X_i;Y|\mathcal{V}_\tau)}{H(Y|\mathcal{V}_\tau)}=\frac{H(Y|\mathcal{V}_\tau)-H(Y|X_i;\mathcal{V}_\tau)}{H(Y|\mathcal{V}_\tau)}. \quad (2)$$

This measure $R(X_i;Y|\tau)$ can be regarded as a conditioned *coefficient of constraint* [20,21]. It takes values between zero (no relevance) and one (maximum relevance). This way of calculating feature subset relevance is used to evaluate subsets of spectra, embedded into a filter *forward-search* strategy, conforming the *Entropic Filtering Algorithm* or EFA (Algorithm 1). The use of the super-feature allows faster computations, that are not essential for its understanding. A detailed implementation can be found in [22].

Algorithm 1. Entropic Filtering

$\Phi\leftarrow\emptyset$
repeat
  $Z\leftarrow\arg\max_{X_i\in X\setminus\Phi}\{R(X_i;Y|\Phi)\}$
  $\Phi\leftarrow\Phi\cup\{Z\}$
until $R(Z;Y|\Phi)=1$ **or** $\Phi=X$;

The algorithm outputs a collection of solutions Φ_i of the same size. This is because, in the last step of the loop, there may be more than one possibility of reaching maximum relevance. In this study it was decided to generate them all and choose that solution Σ with minimum redundancy, by defining the function:

$$\mathcal{I}(\Phi)=\sum_{a\neq b\in\Phi}I(a;b) \quad (3)$$

and setting $\Sigma=\text{argmin}_{\Phi_i}\mathcal{I}(\Phi_i)$. Note that no normalization is necessary, given that all the summations in (3) have the same number of terms.

In order to apply Algorithm 1, a discretization process is needed. Many dimensionality reduction studies use discretization schemes as a way to favor classification tasks (such as [23,24]). This change of representation does not often result in a significant loss of accuracy (sometimes significantly improves it); it also offers large reductions in learning time. The CAIM algorithm [25] is a discretization method that analyzes possible cut-points by

computing a metric that measures the interdependence between the target class and the discretized feature. It is herein selected because it is able to work with supervised data and does not require the user to specify the number of intervals for each feature.

3.3. The bootstrap

Bootstrap methods can be used to select the best model according to a certain prediction criterion [26]. Usually, model selection is associated with parameter estimation and the bootstrap samples can be used for both model selection and inference applied to the selected model [27]. Bootstrap methods are also well-suited for the construction of standard error estimates and confidence intervals (CI) when sample size is small, as in our case, or the distribution of the statistic is unknown. In particular, the *percentile method* uses the entire bootstrap distribution, allows for asymmetric distributions of the statistic and is invariant to transformations [28]. Assuming we have a data set $S=\{(x_i,y_i)\}_{i=1-p}$, we draw *bootstrap samples* $S_1^*,\ldots,S_B^*$ of size p by sampling S with replacement and refit a model to each of the $S_b^*, b=1-B$. A statistic of interest $\hat{\theta}=\theta(S)$ can be estimated in the usual way, e.g.:

$$\overline{\theta}^*=\frac{1}{B}\sum_{b=1}^{B}\theta_b^*;\quad Var(\theta^*)=\frac{1}{B-1}\sum_{b=1}^{B}(\theta_b^*-\overline{\theta}^*)^2 \tag{4}$$

are the mean and variance of the bootstrap distribution of $\hat{\theta}$, and $\theta_b^*=\theta(S_b^*)$. An estimation for the bias of $\hat{\theta}$ can be obtained as $\overline{\theta}^*-\hat{\theta}$. The leave-one-out (LOO) estimate of prediction error is, for the classification case:

$$Err^*=\frac{1}{p}\sum_{i=1}^{p}\frac{1}{|S^{-i}|}\sum_{b\in S^{-i}}\mathbb{I}[y_i\neq f_b^*(x_i)], \tag{5}$$

where $\mathbb{I}(z)$ is 1 when z is true and 0 otherwise, S^{-i} is the set of indexes of the bootstrap samples that do not contain observation x_i and f_b^* is the model developed in S_b^*. Similarly, the resubstitution error is estimated as

$$\overline{e}^*=\frac{1}{p}\sum_{i=1}^{p}\frac{1}{|S^{i}|}\sum_{b\in S^{i}}\mathbb{I}[y_i\neq f_b^*(x_i)], \tag{6}$$

where $S^i=\{1-B\}\backslash S^{-i}$. In both cases, if an index set is empty, the term is skipped and the formula is renormalized accordingly. The 0.632-bootstrap estimate is defined by

$$Err^{(0.632)}=0.368\overline{e}^*+0.632Err^*. \tag{7}$$

Intuitively, this formula pulls the LOO bootstrap estimate down toward the training error, thereby reducing its likely upward bias [27]. CIs can be obtained by the percentile method, as follows: let $\{\theta_b^*\}_{1-B}$ denote again the bootstrap distribution on the samples S_b^*. The $\alpha-$level CI is constructed by ordering these values in ascending order and choosing critical value observations $\theta_{(L)}^*,\theta_{(U)}^*$ as the endpoints of the CI, such that $Pr\{\theta_{(L)}^*\leq\theta\leq\theta_{(U)}^*\}=1-\alpha$. For instance, for $B=1000$, observations 26 and 975 are the endpoints of the 95% CI.

3.4. The classifiers

Several well-known classifiers were chosen: the *Naïve Bayes classifier* (NB), a *Linear Discriminant classifier* (LDC), a *Quadratic Discriminant classifier* (QDC), *Logistic Regression* (LR), and the *Support Vector Machine* (SVM) with *linear* kernel (SVM-L) and *quadratic* kernel (SVM-2), both with parameter C (the regularization constant) as well as with a *radial* kernel (SVM-R), that also needs the setting of the width $\gamma=\sigma^{-2}$. A description of all these techniques can be found in [29].

3.5. Low-dimensional data visualization with scatter matrices

As mentioned in the introduction, the high dimensionality of the analyzed spectra makes the interpretation of the results a non-trivial undertaking, which potentially limits their usability in a practical decision-making context such as brain tumor diagnosis. This may still be the case even after a feature selection process such as the one described in Section 3.2. In this medical context, data visualization in a low-dimensional representation space may become extremely important, helping radiologists to gain insights into what undoubtedly is a complex domain.

Low-dimensional visualization methods generally fall into three categories. Purely linear methods frequently utilize singular values spanning the largest variance in the data, for instance the widely used PCA-based bi-plots [30]. This approach is useful to visually verify known correlations between attributes, but it is generally the case that the first two or three PCs explain a relatively small proportion of the variance in the data, with the consequence that true compact groups of data (be them clusters or, if labels are available, classes) are severely mixed due to the loss of information incurred by the projection.

A second approach is to relax the linearity assumption and to define a non-linear projection to optimize the difference between distances in the original input space and the corresponding distances in the two-dimensional projections of the individual data points, such as in Multi-Dimensional Scaling (MDS) [31] or Sammon's mapping [32]. These maps can be too sensitive to noise in the data, radically altering the data projections even when only a small number of points vary or are added or removed from the data set.

A third approach generates topographic maps by projecting data onto a curved surface weaving through the data and cutting through noise, such as in Self-Organizing Maps (SOM) [33] or in Generative Topographic Mapping (GTM) [34]. Although these are powerful methods for the simultaneous clustering and visualization of intrinsically non-linear data (and, therefore, able to produce more faithful representations), their non-linearity can make the interpretation of the obtained results difficult.

The method proposed for the task at hand is linear in nature, making it easier to use in a real decision-making process that requires an intuitive representation of results. It is based on the decomposition of the scatter matrix, a method that has for long been known to be a valuable alternative for the visualization of data groups [35], with the remarkable property of maximizing the separation between the projections of compact groups of data (tumor classes, in this work). It has been recently improved by a process that involves the sphering of the data followed by a projection onto the space defined by the class means. It leads to the definition of low-dimensional projective spaces that preserve the discrimination between classes obtained by the classifiers, even when the data covariance matrix is singular [14]. A brief explanation follows.

It is well-known that the overall variance of the data, A_T, can be decomposed into the sum of two terms, known as the scatter matrices, which calculate the variance referred to the mean of each data group and between the group mean vectors [36,37] generating a within-cluster matrix, A_W, and a between-cluster matrix, A_B, namely: $A_T=A_W+A_B$. For a $d\times N$ data matrix $D=\{\mathbf{x}_i\}_{i=1-N}$ comprising d-dimensional data points of overall mean $\mathbf{m}$,

$$A_T=\sum_{i=1}^{N}\{(\mathbf{x}_i-\mathbf{m})(\mathbf{x}_i-\mathbf{m})^T\}, \tag{8}$$

$$A_W=\sum_{j=1}^{N_c}\sum_{i=1}^{N_j}\{(x_i^j-\mathbf{m}_j)(x_i^j-\mathbf{m}_j)^T\}, \tag{9}$$

$$A_B = \sum_{j=1}^{N_c} N_j\{(\mathbf{m}_j - \mathbf{m})(\mathbf{m}_j - \mathbf{m})^T\}, \quad (10)$$

where the data are partitioned into N_c groups (tumor classes in this study), each with N_j points and mean $\mathbf{m}_j$. Scalar merit figures for the separation between classes are readily obtained by taking the traces of the scatter matrices, defining sum-of-squares within- and between-classes. These partial sums are sensitive to linear transformations of the data (for instance, relative scaling of the axes), introducing an element of arbitrariness that is not necessary. This leads to the definition of an invariant scatter matrix $M = A_W^{-1}A_B$ and an invariant class separation index $J = tr(M)$. This merit figure suggests that the eigenvalues of the scatter matrix contain useful information about the structure of the data once partitioned into classes.

Given the importance of the class means as representatives for the classes themselves, it is natural to project the data onto the sub-space spanned by these means. This is readily achieved by defining an orthonormal set of basis vectors $\{\mathbf{c}_i\}_{i=1-N_c}$ (for instance, by Gram-Schmidt orthogonalization), generating the first compact projective representation in this method, $D^c = C^T D$, where $C = [\mathbf{c}_1, \ldots, \mathbf{c}_{N_c}]$. Reducing the dimensionality of the data to be visualized from its original value to N_c now requires a drop in rank by just one unity for the scatter matrix calculated in the space of class means, namely:

$$A_W^c = \sum_{j=1}^{N_c}\sum_{i=1}^{N_j}\{(\mathbf{x}_i^c - \mathbf{m}_j^c)(\mathbf{x}_i^c - \mathbf{m}_j^c)^T\}, \quad (11)$$

$$A_B^c = \sum_{j=1}^{N_c} N_j\{(\mathbf{m}_j^c - \mathbf{m}^c)(\mathbf{m}_j^c - \mathbf{m}^c)^T\} \quad (12)$$

and $M^c = (A_W^c)^{-1}A_B^c$. Diagonalization of the new scatter matrix M^c shows, typically, that the trace of the matrix is contained in the largest few eigenvalues, whose eigenvectors form the basis for a 2-D or 3-D visualization of the data. In addition, whitening (sphering) of the data prior to the projection onto the space spanned by the class means exactly preserves the separation index J.

4. Methodological setup

Feature selection can often be considered part of model selection and becomes an important step, specially when the number of observations roughly matches the number of features. Performing model selection in the joint space of features and parameters in this situation is at best a delicate task that entails a very high risk of overfitting. In this work feature selection and classifier selection are carried out in an interleaved way. First, *feature* selection is done in a classifier-independent way in the bootstrap samples; then a set of classifiers is developed on the bootstrap samples using the previously selected sets of features. The outcome is the selection of a specific *classifier* (and its parameters, if any) for each data type. A final *model* (the model parameters and the final feature subset) is obtained using again the bootstrap samples using an iterative procedure. The remainder of this section describes these steps in detail.

4.1. Selection of a set of frequencies

The three distinct ^{1}H-MRS data sets S are used separately to build $B = 1000$ bootstrap samples $S_1^*, \ldots, S_B^*$ for each data type, that will play the role of *training sets*. This procedure is done separately and independently for SET, LET and LSET data. We denote $T_b^* = S \backslash S_b^*$ the corresponding *test sets*. The EFA filter method (Section 3.2) is then applied to every bootstrap sample S_b^*, yielding a collection of solutions $\Sigma_1^*, \ldots, \Sigma_B^*$ for each data type. Recall that this algorithm is applied to the discretized ^{1}H-MRS data.

4.2. Selection of a classifier and its parameters

The tested classifiers include: NB, LDC, QDC, LR, SVM-L, SVM-2 (with $C = 10^k$, k running from -2 to 2 in steps of 0.25) and SVM-R (also with $\gamma = 2^k$, k running from -15 to 15).[1] These classifiers are built in the (bootstrap) training sets using the *original* continuous frequencies, in such a way that a model developed in S_b^* uses *only* the features in Σ_b^*. A single classifier (and its parameters, if any) per data type is selected: that having the smallest bootstrap estimate of prediction error as given by formula (7). We denote such selections $\mathcal{C}_{SET}^*$, $\mathcal{C}_{LET}^*$ and $\mathcal{C}_{LSET}^*$, respectively. For comparative purposes, the classifiers are also built using the full sets of frequencies.

4.3. Selection of specific models for the data types

The final step is the development of a specific *model* for every data type. Once the bootstrap feature sets $\Sigma_1^*, \ldots, \Sigma_B^*$ are obtained, there is an inherent difficulty: the selection process yields a different (though in many cases quite similar) solution for every sample. Stability analysis for the outcome of feature selection is an incipient field nowadays, and still there is no consensus on how to derive a single solution [39]. We develop in this work a specific strategy to obtain what we call a *bootstrap feature sequence*, as described next.

Let $\Im$ denote the indicator function, i.e., $\Im_A(z) = \mathbb{I}(z \in A)$. First create the set $\mathcal{S}^*$ as the union of all the Σ_b^* and define the *frequency* of a feature f as $\varphi(f) = B^{-1}\sum_{1 \le b \le B}\Im_{\Sigma_b^*}(f)$. Assuming that the feature selection algorithm has captured most of the relevant features (with the inevitable variability due to random subsampling), the main difficulty is the *redundancy* across the sets Σ_b^*. This is why a simple frequency-based greedy selection is likely to be suboptimal because it will capture too much redundancy, negatively affecting performance. Rather, the most dominant peaks in the frequency distribution of selected spectral points can be visually selected (Figs. 2–4). The idea is to generate a sequence $\{s_i\}$ of nested feature subsets, as $s_0 = \emptyset$ and $s_{i+1} = s_i \cup \{f_{i+1}\}$. For every element of the sequence, $|s_{i+1}| = |s_i| + 1$ and thus $|s_i| = i$. Associated to it, a second sequence $\{\varepsilon_i^*\}$ is formed, where ε_i^* is the bootstrap estimate of prediction error given by formula (7), evaluated using the features in s_i (this time *the same features* across all the bootstrap samples S_b^*).

The strategy devised to form the $\{s_i\}$ sequence is described next. Under the hypothesis that similar frequencies carry similar information, groups are formed around the peaks, configuring a partition $G_1, \ldots, G_K$ of $\mathcal{S}^*$. This is done differently and independently for the three types of data. Let f_1 be the feature showing the highest frequency across all the groups. Let $f_1 \in G_{i_1}$. Then choose f_2 as the feature showing the highest frequency across all groups except G_{i_1} ($f_2 \in G_{i_2}$); then choose f_3 as the feature showing the highest frequency across all groups except G_{i_1}, G_{i_2}, etc. This procedure is repeated until all groups have been visited. Then f_{K+1} can be chosen again freely among the groups. This selection is repeated until all the frequencies have been picked. We call this simple procedure *Iterative Frequency Selection* (IFS), depicted in Algorithm 2.

[1] For the experiments, we use a MATLAB implementation; specifically, for the SVMs we use the MATLAB interface to LIBSVM [38].

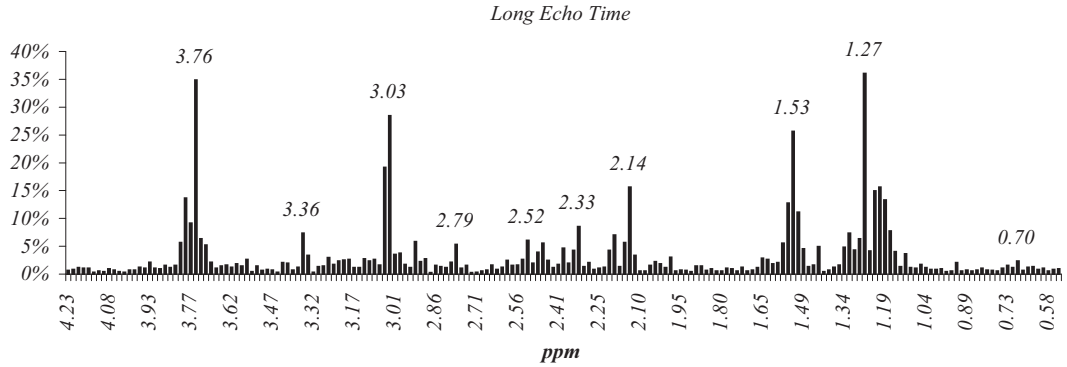

Fig. 2. Relative frequency distribution of spectral points selected in the bootstrap samples from ^{1}H-MRS LET data. Frequency values on top of the peaks are set only as a reference.

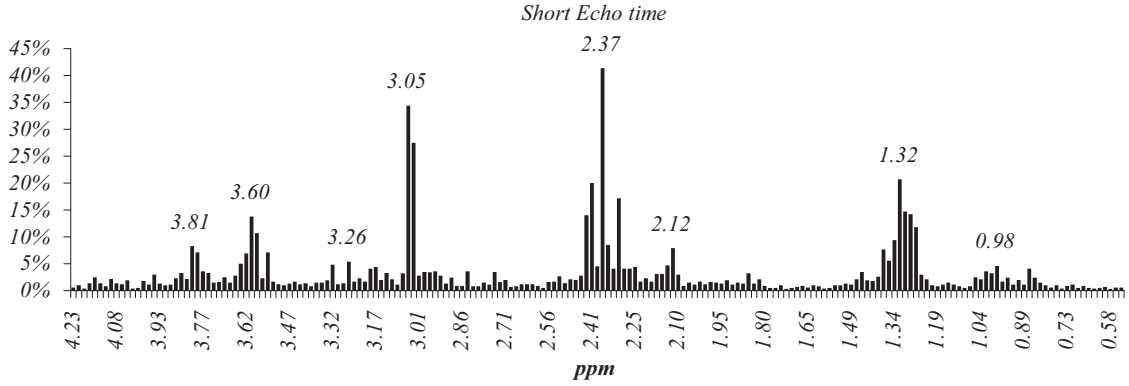

Fig. 3. Relative frequency distribution of spectral points selected in the bootstrap samples from ^{1}H-MRS SET data. Frequency values on top of the peaks are set only as a reference.

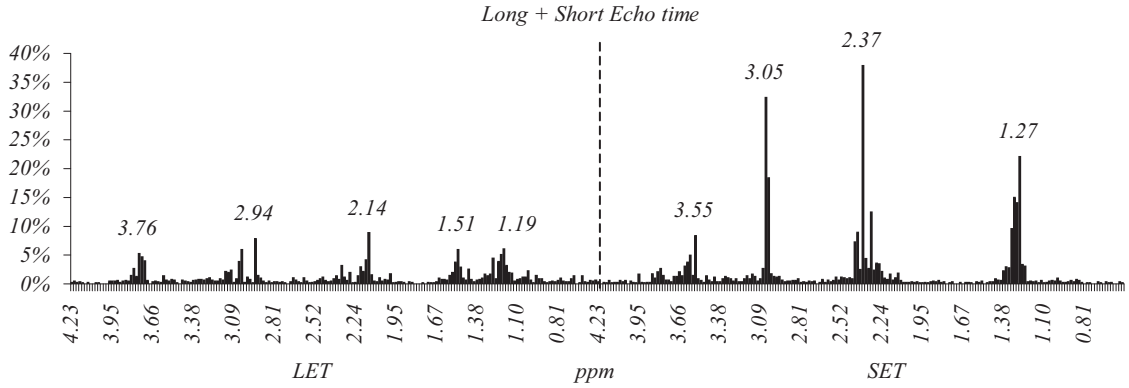

Fig. 4. Relative frequency distribution of spectral points selected in the bootstrap samples from ^{1}H-MRS LSET data. Frequency values on top of the peaks are set only as a reference. The vertical line separates LET from SET frequencies.

Algorithm 2. Iterative frequency selection

Procedure IFS $(\mathcal{S}^*, \{G_1, \ldots, G_K\})$
$s_0 \leftarrow \emptyset$; $i \leftarrow 0$
repeat
  $A \leftarrow \emptyset$
  for $g \leftarrow 1$ **to** $max\ (K, |\mathcal{S}^*| - |s_i|)$ **do**
    $f_{i+1} \leftarrow \arg \max_{f \in \mathcal{S}^* \setminus (\cup_{a \in A} a)} \varphi(f)$
    $A \leftarrow A \cup \left\{ \arg \max_{1 \le k \le K} \Im_{G_k}(f_{i+1}) \right\}$
    $s_{i+1} \leftarrow s_i \cup \{f_{i+1}\}$
    $i \leftarrow i+1$
until $|s_i| = |\mathcal{S}^*|$;

The IFS method was used in conjunction with the previously selected $\mathcal{C}^*_{SET}$, $\mathcal{C}^*_{LET}$ and $\mathcal{C}^*_{LSET}$ classifiers. The final model for each data type is that having a lower ε^*_i along its bootstrap feature sequence $\{s_i\}$.

5. Experimental results

The frequency distributions of spectral points selected by the EFA in the bootstrap samples (Section 4.1) are shown in Figs. 2–4.

The results on classifier construction using the previously selected sets of features (also developed in the bootstrap samples, see Section 4.2) are displayed in Table 1. These are the bootstrap

estimates of prediction error as given by formula (7), translated to accuracy percentages for ease of reading. Additionally, 95% CIs for the mean are reported in Table 2. As it turns out, the selected $\mathcal{C}^*_{SET}, \mathcal{C}^*_{LET}$ and $\mathcal{C}^*_{LSET}$ classifier is the SVM-R, for all the data types. The parameter values found by model selection are $C=4, \gamma=8$ (LET), $C=8, \gamma=8$ (SET) and $C=4, \gamma=8$ (LSET). The results of using these selected classifiers with the *full sets* of frequencies are also shown (column **NR** in Tables 1 and 2), for comparative purposes.

Boxplots of bootstrapped performance (Fig. 5) of the best classifier configuration (among those listed in Section 4.2) are reported, separately for the LET, SET and combined LSET ^{1}H-MRS data sets. Kolmogorov–Smirnov normality tests for the error distributions indicate that the hypothesis of normality cannot be sustained. Therefore, a non-parametric Wilcoxon signed-rank test is used for the (null) hypothesis that the median of the differences between the errors of these selected classifiers and another classifier's error is zero. This hypothesis has to be rejected at the 95% level when the SVM-R is compared against all other classifiers, remarkably for all three data types. Specifically, the greatest p-values found are 7.62e−91 for LET (against SVM-L), 1.45e−37 for SET (against SVM-2) and 2.09e−47 for LSET (against LDC). This ends step two of the process (Section 4.2). As explained in Section 4.3, the last step is the selection of specific *models* for all the data types. The results of running the IFS algorithm on the selected $\mathcal{C}^*_{SET}$ $\mathcal{C}^*_{LET}$ and $\mathcal{C}^*_{LSET}$ classifiers are shown in Fig. 7, with indication (numbers in brackets) of the size of the subset with best performance and its performance; 95% percentile CIs for these final accuracy results are 87.0–95.7 (LET), 87.6–95.4 (SET) and 87.0–95.1 (LSET). The obtained subsets of spectral frequencies are detailed in Table 3; these will be subject of visualization as well as metabolic interpretation in the following sections.

Table 1
0.632-Bootstrap classification performance (in percentage) for all the ^{1}H-MRS data sets (LSET refers to the combined LET plus SET data set); **NR** stands for no feature reduction using the SVM-R.

Data	NR	NB	LDC	QDC	LR	SVM-L	SVM-2	SVM-R
LET	72.1	80.9	81.7	82.0	82.8	83.0	82.8	85.5
SET	72.9	83.3	87.1	86.3	86.6	86.9	87.1	88.2
LSET	72.1	82.9	85.9	85.6	85.5	85.8	85.6	87.2

Table 2
95% CIs of classification performance for all three ^{1}H-MRS data sets (LSET refers to the combined LET plus SET data set); **NR** stands for no feature reduction using the SVM-R.

Data	NR	NB	LDC	QDC	LR	SVM-L	SVM-2	SVM-R
LET	66.0–77.4	74.1–87.0	74.8–87.8	75.3–88.1	75.4–89.0	74.1–89.5	75.8–89.4	78.4–91.3
SET	67.1–78.7	77.7–88.0	82.1–91.2	81.3–90.6	81.7–90.8	82.2–91.1	80.3–92.4	82.7–93.0
LSET	66.6–77.9	77.1–88.1	80.0–90.3	80.0–90.4	79.8–89.8	80.2–90.2	79.0–91.5	81.2–92.4

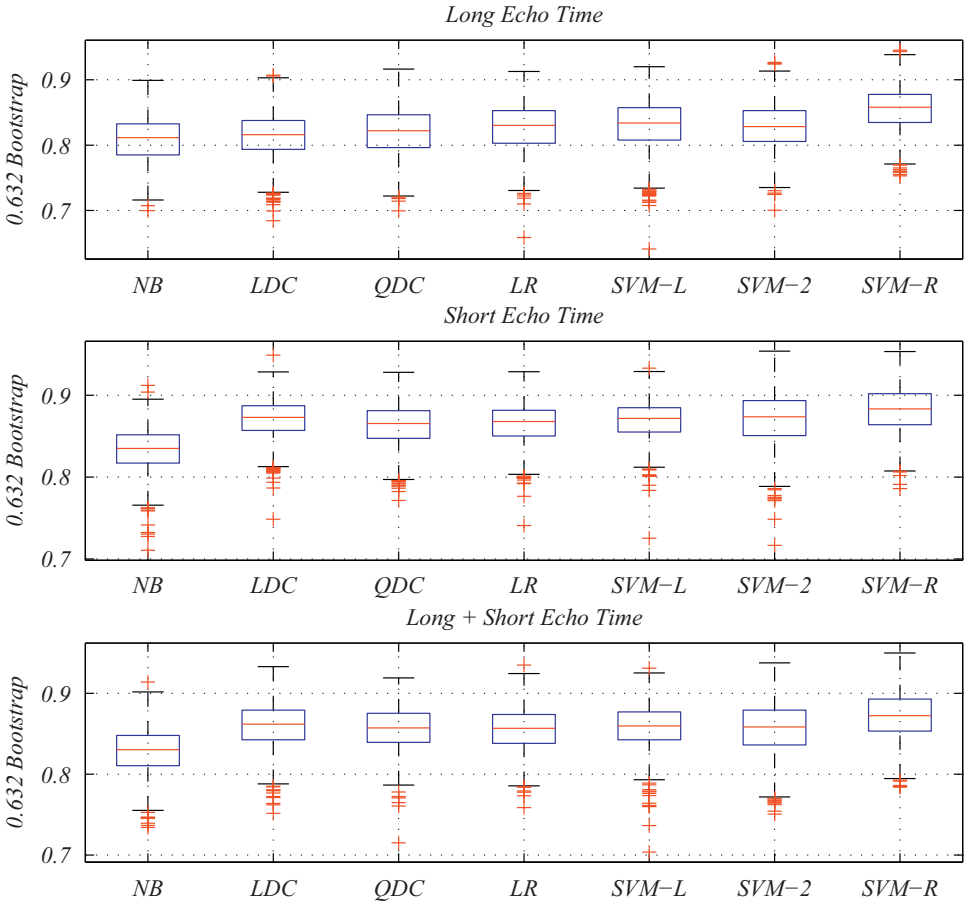

Fig. 5. Boxplots of 0.632-Bootstrap classification performance for each of the three ^{1}H-MRS data sets and all the classifiers.

Table 3
Final subsets of spectral points that yield the maximum 0.632-Bootstrap mean classification performance, separated for the three ^{1}H-MRS data sets.

Data	Subset size	Spectral frequencies (in ppm)
LET	10	1.27, 3.76, 3.03, 1.53, 2.14, 2.33, 3.36, 2.79, 0.70, 1.21
SET	18	2.37, 3.05, 1.32, 3.60, 3.81, 2.12, 3.26, 0.98, 2.84, 1.86 4.15, 2.41, 3.03,1.30, 3.58, 3.79, 2.14, 3.32
LSET	7	S2.37, S3.05, S1.27, L2.14, S3.55, L2.94, L1.19

Left-right and then top-bottom reading of the selected frequencies indicates their relevance ranking along their sequence s_i—see Fig. 7. For the LSET data set, the origin is indicated by a prefix (L- or S-).

5.1. Discriminatory visualization of selected features

At this point, data visualization is still a challenge for the obtained numbers of dimensions. The use of a dimensionality reduction strategy based on feature extraction prior to classification would bring about the inconvenience that each of the extracted features would still be a combination of the whole spectrum of frequencies. This would make its practical interpretation a very difficult undertaking, if possible at all. Moreover, if classification accuracy was the only relevant outcome of a brain tumor diagnosis—on the basis of the available data—then feature selection would become a redundant process. Indeed, the interpretability of the results is a compulsory requirement in this problem.

This possibility is made easier if the results can be explained in terms of a parsimonious subset of spectral frequencies, which is what feature selection achieves. Nonetheless, even if interpretable by mere inspection, the final selection of spectral frequencies may still provide few clues about the distribution of the classes (tumor types). It could be argued that visualization can be obtained regardless of whether feature selection is used or not; however, using feature selection as a starting point for the visualization method ensures that the latter will not be affected by the use of information that is not relevant in terms of classification.

The visualization of the data is herein achieved using the method described in Section 3.5 and illustrated by the plots in Fig. 6. These are scatter plots of 2-D projections of the three classes (using the first two eigenvectors of M^c).

5.2. Interpretation of the obtained metabolites

The selection of features resulting from the methodology described in Sections 3 and 4 is amenable to at least partial metabolic interpretation by an expert radiologist, making it useful in clinical diagnosis.

For the 10 selected LET frequencies listed in Table 3, we find, among others, five main regions of relevance: 3.76 corresponding to Glutamate/Glutamine-containing compounds and Alanine; 3.36 in the Taurine area; 3.03 corresponding to Creatine; 2.14 and 2.33 that belong to an area defining a different subtype of Glutamate/Glutamine compounds; 1.53 near the Alanine peak; and 1.21 and 1.27 corresponding to the presence of Lipids.

Out of the 18 selected SET features, we find Glutamate–Glutamine and Alanine at 3.79–3.81; myo-Inositol and Glycine at 3.58–3.60; Choline-containing compounds at 3.26; again Creatine at 3.03–3.05; Glutamate-Glutamine and *N*-acetylaspartate at 2.12, 2.14 and 2.41; Lactate and Mobile Lipids at 1.30 and 1.32; a narrow singlet at 2.37 ppm, most likely corresponding to Pyruvate with a possible Glutamate contribution; and a possible macromolecule at 2.84. The rest of selected frequencies do not have an immediate metabolic explanation.

An extraordinarily small subset of 7 spectral frequencies is selected for the combination of SET and LET data. This makes this subset especially suitable for its ease of interpretation in a real clinical context. It includes some of the most relevant features selected for each of the echo times separately (mainly for SET, which are also found to include the most relevant frequencies overall). At SET, Creatine, Glutamine, Lactate and Mobile Lipids are again present, while at LET we again find Glutamate/Glutamine compounds and Lipids, as well as a yet unidentified relevant frequency at 2.94 ppm.

5.3. Summary and discussion of findings

In view of all these experimental results, several findings are now summarily presented:

1. Feature selection appears to be a viable avenue for dimensionality reduction in this field: with less than a tenth of spectral frequencies, mean performance of the finally selected classifiers improves that achieved using the full frequency sets. This behavior is remarkable, both for computational and scientific reasons.
2. As stated in Section 2, most of the existing literature indicates an advantage in using SET information over LET [3,12]. The present work adds strong support for this finding, given that all the classifiers obtained markedly better results for SET data against LET—Tables 1 and 2 supply quite conclusive information in this respect. For SET data, Table 3 also indicates that some redundancy (presence of neighboring features) among the chosen frequencies is necessary to obtain an improvement on the performances reported in Table 1. This makes perfect sense from a radiologist point of view, given that contiguous spectral points in peak regions will usually be highly intercorrelated. Correspondingly, full frequency intervals of high relevance can be observed in Figs. 2–4.
3. It can be seen in Fig. 7 that, for all three data types, overall performance increases rather rapidly, stabilizes and then drops very gradually; in this sense, the IFS algorithm seems to work well, as it produces rather sensible results. We have found that, although the use of SET spectral frequencies is in general a better classification strategy, the *combined* use of small numbers of frequencies at both echo times permits the obtention of much simpler models with similar performance.
4. The resulting sets of selected spectral frequencies (shown in Table 3) have been subject of a medical interpretation in terms of known metabolites. Of special importance is the smoothness of relative frequency distribution of spectral points selected in the bootstrap samples, for all the data types (Figs. 2–4). This is consistent with the fact that neighboring frequencies may correspond to the same metabolite or group of metabolites.
5. The discrimination ability of these sets can also be subject of visual interpretation (which is paramount for the clinical use of these methods), according to the comparative plots in Figs. 6(a)–(f). Though discrimination is not perfect, in all three cases it is very similar to that obtained using the full sets of frequencies.

6. Conclusions and future work

MRS is yet to become a standard method for day-to-day clinical diagnosis of brain tumors. This is despite it being a non-invasive technique and one that provides rich information about the biochemistry of the tumor pathology. Instead, MRI is often the method of choice for diagnosis in practice, in spite of its limitations. To become mainstream, the diagnosis based on MRS

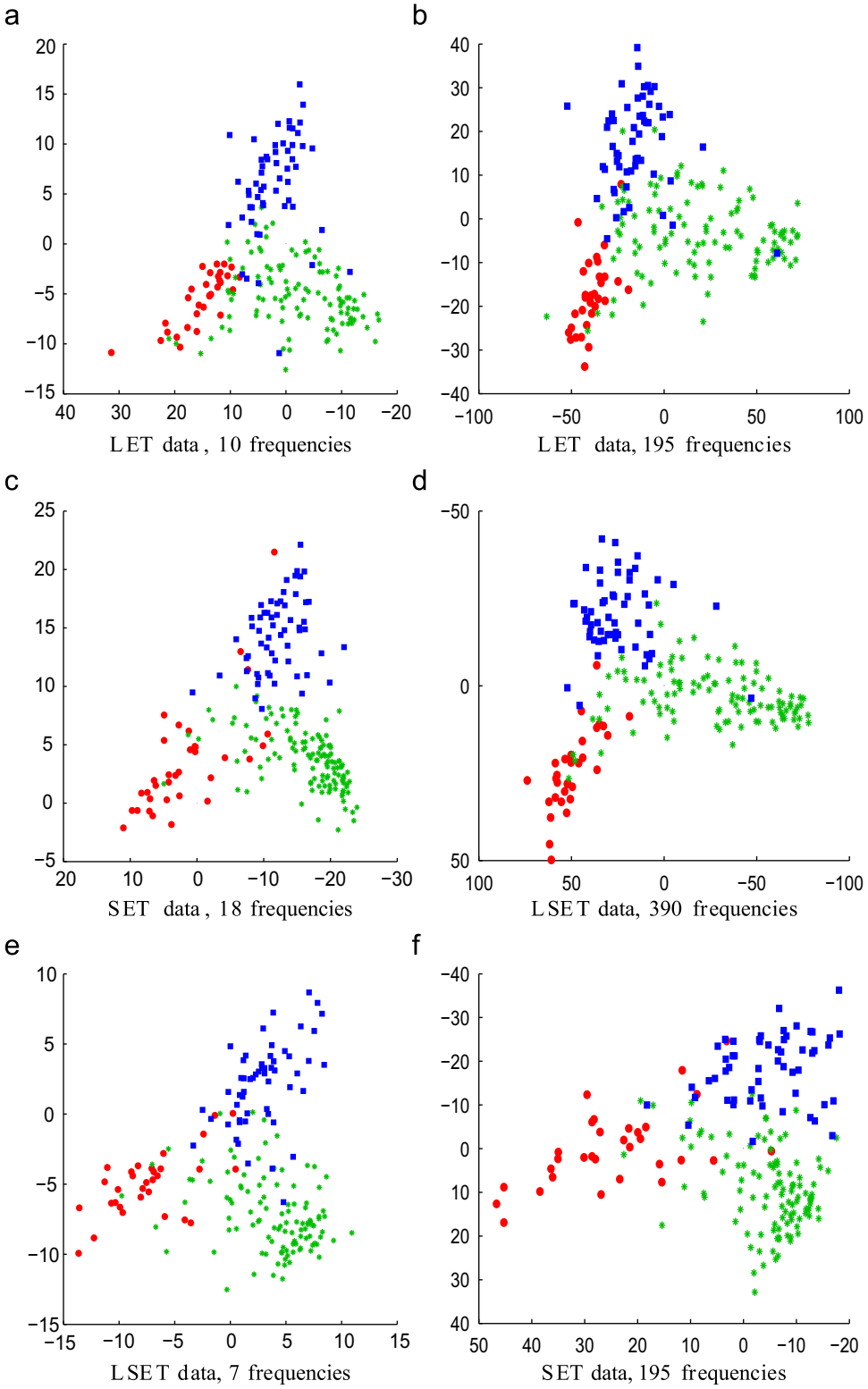

Fig. 6. Visualization of G1 (low grade gliomas, *circles*), G2 (high grade malignant tumors, *asterisks*) and G3 (meningiomas, *squares*). The two projection axes correspond to the first two eigenvectors of M^c. (a) LET data, 10 frequencies. (b) LET data, 195 frequencies. (c) SET data, 18 frequencies. (d) LSET data, 390 frequencies. (e) LSET data, 7 frequencies. (f) SET data, 195 frequencies.

must be sufficiently robust and, for that, reliable tools for spectral data analysis are required. In this study, algorithms for the selection of spectral frequencies have been applied to a set of bootstrap samples, in combination with several classifiers, for the final obtention of a reliable set of interpretable and accurate models of brain tumor diagnosis. The developed methodology has shown to be able to provide a drastic reduction in dimensionality while being competitive with or even improving on the performance obtained using the full set of spectral frequencies. This holds true for all the MRS acquisition modalities considered: short and long echo times, and the combination of both by concatenation of spectra. These results are extremely important

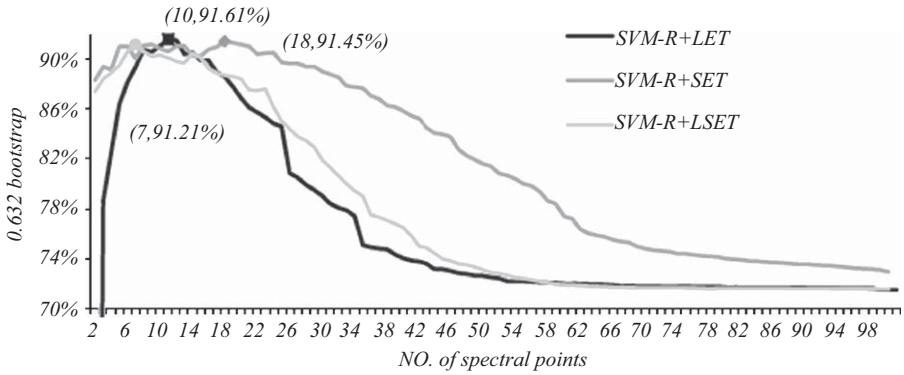

Fig. 7. Performance curves for the three ^{1}H-MRS data types. The horizontal axis is the size of the subset being evaluated, following the s_i sequence as described in the text. The vertical axis is the 0.632-Bootstrap mean classification performance of the three SVM-R models. The numbers in brackets are the size of the subset with best performance and its bootstrapped performance. Only the first 100 spectral features are shown.

as they make the diagnosis easily interpretable in terms of a handful of spectral frequencies, most of them associated to metabolites that are well-known in the biomedical field. We have reached a step beyond feature selection in improving the interpretability of the results by providing a visualization method that preserves the discrimination capability of the obtained classifier models. Our classification results also provide support to similar studies in the literature as they show the comparative advantage of using SET data or a combination of SET and LET data. Future research will extend the use of the proposed methodology to the analysis of other brain tumor classification problems involving different pathologies and pathology groupings.

Acknowledgments

Authors gratefully acknowledge the former INTERPRET partners (INTERPRET, EU-IST-1999-10310) and, from 1st January 2003, *Generalitat de Catalunya* (CIRIT SGR2001-194, XT 2002-48 and XT 2004-51 grants); data providers: Dr. C. Majós (IDI), Dr. A. Moreno-Torres (CDP), Dr. F.A. Howe and Prof. J. Griffiths (SGUL), Prof. A. Heerschap (RU), Dr. W. Gajewicz (MUL) and Dr. J. Calvar (FLENI); data curators: Dr. A.P. Candiota, Ms. T. Delgado, Ms. J. Martín, Mr. I. Olier, Mr. A. Pérez and Prof. Carles Arús (all from GABRMN-UAB). Authors also acknowledge funding for CICyT TIN2006-08114 and SAF2005-03650 projects, the Mexican CONACyT and Baja California University.

References

[1] A. Vellido, E. Biganzoli, P.J.G. Lisboa, Machine learning in cancer research: implications for personalised medicine, in: European Symposium on Artificial Neural Networks, d-Side pub., 2008, pp. 55–64.
[2] M. Julià-Sapé, D. Acosta, M. Mier, C. Arús, D. Watson, The interpret consortium: a multi-centre, web-accessible and quality control-checked database of in vivo MR spectra of brain tumour patients, Magn. Reson. Mater. Phys. 19 (2006) 22–33.
[3] C. Majós, M. Julià-Sapé, J. Alonso, M. Serrallonga, C. Aguilera, J. Acebes, J. Gili, C. Arús, Brain tumor classification by proton MR spectroscopy: comparison of diagnostic accuracy at short and long TE, Am. J. Neuroradiol. 25 (2004) 1696–1704.
[4] H. Bruhn, J. Frahm, M. Gyngell, K. Merboldt, W. Hanicke, R. Sauter, C. Hamburger, Noninvasive differentiation of tumors with use of localized ^{1}H-MR spectroscopy in vivo: initial experience in patients with cerebral tumors, Radiology 172 (1989) 541–548.
[5] H. Kugel, W. Heindel, R. Ernestus, J. Bunke, R. du Mesnil, G. Friedmann, Human brain tumors: spectral patterns detected with localized ^{1}H-MR spectroscopy, Radiology 183 (1992) 701–709.
[6] D. Ott, J. Hennig, T. Ernst, Human brain tumors: assessment with in vivo proton MR spectroscopy, Radiology 186 (1993) 745–752.
[7] Y. Kinoshita, H. Kajiwara, A. Yokota, Y. Koga, Proton magnetic resonance spectroscopy of brain tumors: an in vitro study, Neurosurgery 34 (4) (1994) 606–614.
[8] H. Shimizu, T. Kumabe, T. Tominaga, T. Kayama, K. Hara, Y. Ono, K. Sato, N. Arai, S. Fujiwara, T. Yoshimoto, Noninvasive evaluation of malignancy of brain tumors with proton MR spectroscopy, Am. J. Neuroradiol. 17 (4) (1996) 737–747.
[9] A. Tate, et al., Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra, NMR Biomed. 19 (2006) 411–434 (26 authors).
[10] C. Ladroue, Pattern recognition techniques for the study of magnetic resonance spectra of brain tumours, Ph.D. Thesis, St. George's Hospital Medical School, 2003.
[11] A. Devos, Quantification and classification of MRS data and applications to brain tumour recognition, Ph.D. Thesis, Katholieke University Leuven, Belgium, 2005.
[12] J. García, S. Tortajada, C. Vidal, M. Julià-Sapé, J. Luts, S. Van Huffel, C. Arús, M. Robles, On the use of long TE and short TE SV MR spectroscopy to improve the automatic brain tumor diagnosis, Technical Report 07-55, Katholieke University Leuven, Belgium, 2007.
[13] J.M. García-Gómez, S. Tortajada, C. Vidal, M. Julià-Sapé, J. Luts, À. Moreno-Torres, S. Van Huffel, C. Arús, M. Robles, The influence of combining two echo times in automatic brain tumor classification by Magnetic Resonance Spectroscopy, NMR Biomed. 21 (10) (2008) 1112–1125.
[14] P.J.G. Lisboa, I.O. Ellis, A.R. Green, F. Ambrogi, M.B. Dias, Cluster-based visualisation with scatter matrices, Pattern Recognition Lett. 29 (13) (2008) 1814–1823.
[15] K. Kira, L. Rendell, The feature selection problem: traditional methods and a new algorithm, in: Proceedings of the National Conference on Artificial Intelligence, 1992, pp. 129–134.
[16] F. González, Ll. Belanche, Feature Selection in *in vivo* ^{1}H-MRS single voxel spectra, in: Proceedings of the KES 2008 Conference, Lecture Notes in Computer Science, vol. 5178, Springer, Berlin, 2008, pp. 197–205.
[17] N. Sibtain, The clinical value of proton magnetic resonance spectroscopy in adult brain tumours, Clin. Radiol. 62 (2007) 109–119.
[18] INTERPRET: International Network for Pattern Recognition of Tumours Using Magnetic Resonance project 〈http://azizu.uab.es/INTERPRET〉.
[19] P.E. Meyer, C. Schretter, G. Bontempi, Information-theoretic feature selection in microarray data using variable complementarity, IEEE J. Selected Top. Signal Process. 2 (3) (2008).
[20] C.H. Coombs, R.M. Dawes, A. Tversky, Mathematical Psychology An Elementary Introduction, Prentice-Hall, Englewood Cliffs, NJ, 1970.
[21] H. Wang, Towards a unified framework of relevance, Ph.D. Thesis, University of Ulster, 1996.
[22] F. González, Ll. Belanche, Gene subset selection in microarray data using entropic filtering for cancer classification, Expert Systems 26 (1) (2009) 113–124.
[23] M. Ng, L. Chan, Informative gene discovery for cancer classification from microarray expression data, in: IEEE Workshop on Machine Learning for Signal Processing, IEEE, 2005, pp. 393–398.
[24] D. Le, S. Satoh, Robust object detection using fast feature selection from huge feature sets, in: 13th International Conference on Image Processing, IEEE, 2006, pp. 961–964.
[25] L. Kurgan, K. Cios, CAIM discretization algorithm, IEEE Trans. Knowl. Data Eng. 16 (2) (2004) 145–153.
[26] B. Efron, R. Tibshirani, Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, Stat. Sci. 1 (1986) 54–75.
[27] T. Hastie, R. Tibshirani, J.H. Friedman, The Elements of Statistical Learning, Springer, New York, 2001.
[28] C.Z. Mooney, R. Duval, Bootstrapping: a Nonparametric Approach to Statistical Inference, Sage Publications, 1993.

[29] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Wiley, New York, 2001.
[30] K.R. Gabriel, The biplot graphical display of matrices with applications to principal component analysis, Biometrika 58 (3) (1971) 453–467.
[31] T.F. Cox, M.A.A. Cox, Multidimensional Scaling, Chapman & Hall, UK, 2001.
[32] J.W. Sammon, A Non-linear mapping for data structure analysis, IEEE Trans. Comput. C-18 (1969) 401–408.
[33] T. Kohonen, Self-organized formation of topologically correct feature maps, Biol. Cybernet. 43 (1) (1982) 59–69.
[34] A. Vellido, P.J.G. Lisboa, Handling outliers in brain tumour MRS data analysis through robust topographic mapping, Comput. Biol. Med. 36 (10) (2006) 1049–1063.
[35] H.P. Friedman, J. Rubin, On some invariant criteria for grouping data, J. Am. Stat. Assoc. 62 (320) (1967) 1159–1178.
[36] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.
[37] H.P. Friedman, J. Rubin, On some invariant criteria for grouping data, J. Am. Stat. Assoc. 62 (320) (1967) 1159–1178.
[38] C.-C. Chang, C.-J. Lin, LIBSVM: a library for support vector machines, 2001, Software available at: 〈http://www.csie.ntu.edu.tw/%7ecjlin/libsvm〉.
[39] A. Kalousis, J. Prados, M. Hilario, Stability of feature selection algorithms: a study on high-dimensional spaces, Knowl. Inf. Syst. 12 (1) (2007) 95–116.

Félix F. González-Navarro is an Associate Professor at the Engineering Institute at Baja California State University, Méxicali, México. Currently, he is a Ph.D. student in the Departament de Llenguatges I Sistemes Informátics at the Universitat Politècnica de Catalunya (UPC), where he investigates in the areas of Pattern Recognition, Feature Selection Algorithms and Information Theory.

Lluís A. Belanche is an Associate Professor in the Software Department at the Universitat Politènica de Catalunya (UPC). He received his B.Sc. in Computer Science from the UPC in 1990 and an M.Sc. in Artificial Intelligence from the UPC in 1991. He joined the Computer Science Faculty shortly after, where he completed his doctoral dissertation in 2000. He has been doing research in neural networks and support vector machines for pattern recognition and function approximation, as well as in feature selection algorithms and their collective application to workable artificial learning systems.

Enrique Romero received his B.Sc. degree in Mathematics in 1989 from the Universitat Autònoma de Barcelona. In 1994, he received his B.Sc. in Computer Science from the Universitat Politècnica de Catalunya (UPC). In 1996, he joined the Software Department at the UPC, as an Assistant Professor. He received his M.Sc. degree in Artificial Intelligence and Ph.D. degree in Computer Science from the UPC in 2000 and 2004, respectively. His research interests include neural networks, support vector machines and feature selection.

Alfredo Vellido received his degree in Physics from the Department of Electronics and Automatic Control of the University of the Basque Country (Spain), in 1996. He completed his Ph.D. at Liverpool John Moores University (UK), in 2000. After a few years of experience in the private sector, he briefly joined Liverpool John Moores University again as research officer in a project in the field of computational neurosciences. He is now a Ramón y Cajal research fellow for the Technical University of Catalonia. Research interests include, but are not limited to, pattern recognition, machine learning and data mining, as well as their application in medicine, market analysis, ecology and e-learning, on which subjects he has published widely.

Margarida Julià-Sapé holds a B.Sc. Hon. in Biology from the Universitat de Barcelona (UB), Spain, 1994, as well as an M.Sc. in Biotechnology (1995) from the UB. She was awarded her Ph.D. in 2006 by the Universitat Autònoma de Barcelona (UAB), Spain. The author is currently a postdoctoral researcher with the Networking Research Center on Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), at UAB, Cerdanyola del Vallès, Spain.

Carles Arús was born in Barcelona (Spain), in 1954. B.Sc. in Biology from the Universitat Autònoma de Barcelona (UAB), Spain, in 1976. Ph.D. in Chemistry from UAB, in 1981 (Ph.D. advisor Prof. Claudi M. Cuchillo) on the subject of the sub-site structure of bovine pancreatic RNase A (enzyme kinetics, NMR spectroscopy). Best thesis award in the Faculty of Sciences of UAB in 1982. Postdoctoral work in the USA (1982-1985) on biomedical NMR with Prof. Michael Bárány (University of Illinois at Chicago, IL) and Prof. John L. Markley (Purdue University, IN). Since 1985, tenured assistant professor, and, since 2002, full Professor at the Department of Biochemistry and Molecular Biology of the UAB. His research group has carried out work on the application of NMR spectroscopy of tumours for diagnostic purposes and has also contributed to the investigation of human muscle bioenergetics by 31P MRS. His present interests in the field of tumour spectroscopy target the use of 1H MRS of human brain tumours, biopsies and cell models for diagnosis, prognosis and therapy planning. He has published 66 PubMed accessible articles since 1977.