Neurocomputing 72 (2009) 3085–3097

Contents lists available at ScienceDirect

Neurocomputing

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

Outlier exploration and diagnostic classification of a multi-centre ^{1}H-MRS brain tumour database

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

[a] Dept. de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya, C./ Jordi Girona, 1-3, 08034 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), Departament de Bioquímica i Biología Molecular (BBM), Unitat de Biociències Universitat Autònoma de Barcelona (UAB), Cerdanyola del Vallès, Spain

ARTICLE INFO

Article history:
Received 30 June 2008
Received in revised form
10 March 2009
Accepted 22 March 2009
Communicated by T. Heskes
Available online 9 April 2009

Keywords:
Proton magnetic resonance spectroscopy
Brain tumours
Outlier detection
Nonlinear dimensionality reduction
Feature selection
Medical decision support systems

ABSTRACT

Non-invasive techniques such as magnetic resonance spectroscopy (MRS) are often required for assisting the diagnosis of tumours. Radiologists are not always accustomed to make sense of the biochemical information provided by MRS and they may benefit from computer-based support in their decision making. The high dimensionality of the MR spectra obscures atypical aspects of the data that may jeopardize their classification. In this study, we describe a method to overcome this problem that combines nonlinear dimensionality reduction, outlier detection, and expert opinion. MR spectra subsequently undergo a feature selection process followed by classification. The impact of outlier removal on classification performance is assessed.

1. Introduction

Decision making in oncology is a sensitive matter, and even more so in the specific area of brain tumour oncologic diagnosis, for which the direct and indirect costs—both human and financial—of misdiagnosis are very high. In this area, in which most diagnostic techniques must be non-invasive, clinicians should benefit from the use of an at least partially automated computer-based medical decision support system (DSS).

AIDTumour (artificial intelligence decision tools for tumour diagnosis [1]) is a research project for the design and implementation of a medical DSS to assist experts in the diagnosis of human brain tumours on the basis of biological signal data obtained by magnetic resonance spectroscopy (MRS). This is a technique that can shed light on cases that remain ambiguous after clinical investigation. The MRS data used in AIDTumour and analysed in this paper belong to a complex multi-centre set containing cases of several brain tumour pathologies [16]. These data have undergone a rigorous pre-processing quality control that validates them from the viewpoint of the radiologists. Nevertheless, and for their use in an automated computer-based DSS, the various origins of these spectra and the complexity of their pre-processing make further data exploration advisable.

It might be problematic to include some of the spectra in an automated DSS without further ado for three different reasons. Firstly, some may contain measurement or acquisition artefacts that, even if not completely precluding diagnosis by visual inspection, might induce errors in computer-based diagnosis: these are what we call here *artefact-related outliers*. Secondly, atypical cases that do not contain artefacts but are nevertheless unrepresentative of the main distributions of the whole dataset: herein, these will be referred to as *distinct outliers* [33]. Thirdly, some cases with a clear biopsy-based diagnosis (tumour type attribution) may yield spectra that are quantitatively similar to those of other tumour types, misleading a computer-based classification system. Even if representative of the data as a whole, they are still unrepresentative of their own tumour type: these we will call *class outliers*. Note that these three outlier typologies are not always mutually exclusive.

Machine learning (ML) and related methods can play a useful role [35] in dealing with the uncertainty introduced by the presence of outliers in a diagnostic setting. Here, we show the effectiveness of a method to identify and characterize potentially conflicting MRS data that combines techniques of nonlinear

* Corresponding author.
E-mail address: avellido@lsi.upc.edu (A. Vellido).

0925-2312/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2009.03.010

dimensionality reduction (DR), exploratory visualization, and outlier detection, with expert knowledge. The introduction of the latter is paramount, as it will help to skim those cases truly conflictive out of those shortlisted by blind quantitative criteria. Dimensionality reduction is not trivial in this setting, as the available MRS data are scarce and high dimensional. Sammon's mapping [29] is used to this end. Generative topographic mapping (GTM [3]), a manifold learning model, is used to quantify the atypicality of spectra [33,34].

Overall, the aforementioned method is conceived as a preliminary step to data classification in the DSS, in which specific cases are tagged with information regarding their possible atypicality and its characteristics. The fact that the MRS data analysed in this study are scarce and of high dimensionality makes their computer-based automated classification a difficult undertaking. Most importantly, this high dimensionality also precludes the straightforward interpretation of the obtained results, limiting their usability in a practical medical setting. Consequently, dimensionality reduction, in the form of either feature selection or feature extraction, would help to reduce the complexity of the problem at hand. Feature extraction, though, may not comply with the interpretability requirement. The expert radiologists who are meant to be assisted by the medical DSS are not usually trained to make sense of new features extracted from the MRS frequencies. Instead, they often have knowledge of specific MRS frequencies related to metabolites of known significance for tumour type discrimination. Note also that one goal of exploratory studies of this kind is to understand where the variables selected by the model fit in relation to prior knowledge from the medical domain [23]. This may limit the practical applicability of methods such as PCA or ICA as used, for instance, in [19,13,24,37] for assisting brain tumour diagnosis. As an example, in analysing these type of data, ICA will often yield components that "would correspond with identifying the independent degrees of freedom in MRS, not with individual metabolites, but with characteristic tissue generators" [13].

An entropic filtering algorithm (EFA) is used in this study for feature selection as a fast method to generate a relevant subset of MR spectral frequencies. Bootstrap resampling techniques are used to obtain mean performance estimates and their variability. The main goal is obtaining simple models (in terms of low numbers of hopefully interpretable MR spectral frequencies) that generalize well. Outliers might still unduly bias the automated classification process in the DSS, even if for different reasons. We hypothesized that, by removing the cases labelled as outliers, classification accuracy would improve and feature selection would experiment significant variations. The experimental results reported in this paper provide partial support for the first hypothesis but not for the second.

The remaining of the paper is structured as follows. First, the ^{1}H-MRS dataset available for experimentation is briefly described. This is followed, in Section 3, by a description of the different analytical methods. Experimental results are presented in Section 4. The paper closes with a section summarizing our conclusions.

2. ^{1}H-MRS brain tumour data

The echo time is an influential parameter in ^{1}H-MRS data acquisition. In short-echo time (SET) spectra (typically acquired at 20–40 ms) some metabolites are better resolved (e.g. lipids, myo-inositol, glutamine and glutamate). However, there may be numerous overlapping resonances (e.g. glutamate/glutamine at 2.2 ppm and NAA at 2.01 ppm) which make the spectra difficult to interpret [26]. The use of a long-echo time (LET) yields less clearly resolved metabolites but also less baseline distortion, resulting in a more readable spectrum. There are a few studies comparing the classification potential of these two types of spectra (see, e.g. [26]). In this study, we focus on LET data.

The analysed data correspond to 195 LET single voxel ^{1}H-MR spectra acquired in vivo from brain tumour 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*). Following a common procedure [26,24], the clinically-relevant regions of the spectra were sampled to obtain 195 frequency intensity values (measured in parts per million (ppm), an adimensional unit of relative frequency position in the data vector), from 4.25 ppm down to 0.56 ppm. These frequencies become data attributes in the reported experiments and, as a result, the analysed data consist of 195 cases and 195 attributes.

These data are extracted from a database resulting from the *international network for pattern recognition of tumours using magnetic resonance* (INTERPRET) European research project [16]. The criteria for the selection of cases to be included in the original complete database (in which there are more tumour types than the ones analysed in this study as well as cases corresponding to normal tissue and abscesses) were: (a) that the case had a single voxel SET, 1.5 T spectrum acquired from a nodular region of the tumour; (b) that the voxel was located in the same region as where subsequent biopsy was obtained; (c) that the short-echo spectrum had not been discarded because of acquisition artefacts or other reasons and (d) that a histopathological diagnosis was agreed among a committee of neuropathologists. In those cases in which the spectra were obtained from normal volunteers without the pathology, or corresponded to abscesses or clinically proven metastases, biopsy was not required. For further details on data acquisition and processing, and on database characteristics, see, for instance, [15,16].

Class labelling was performed according to the World Health Organization (WHO) system for diagnosing brain tumours by histopathological analysis of a biopsy sample. For the analyses reported in this study, a subset of spectra from the database were bundled into three groups, namely: G1: *low-grade gliomas* (*a2*, *oa* and *od*); G2: *high-grade malignant tumours* (*me* and *gl*); and G3: *meningiomas* (*mm*). This type of grouping is justified [31] by the well-known difficulty in distinguishing between metastases and glioblastomas, due to their similar spectral pattern produced by the highly necrotic nature of these tumours.

3. Methods

3.1. Outlier characterization

3.1.1. MRS data dimensionality reduction and visualization through Sammon's mapping

There are several decisions involved in the choice of a dimensionality reduction method. To name just a few [22]: hard vs. soft DR; generative vs. non-generative methods; implicit vs. explicit mappings; or linear vs. nonlinear DR. For this study, a nonlinear DR method was preferred in principle (instead of a linear alternative such as PCA or classical Multi-Dimensional Scaling, for instance), as there existed no *a priori* reason to assume only linear dependencies. Given that DR in this study does not aim at providing generalization, an explicit mapping procedure was also preferred. A typical *desiderata* for the visual representation of data and knowledge can be formulated in terms of maximizing structure preservation and, therefore, a method with "in-built" preservation of inter-point distances was also preferred. The nonlinear Sammon's mapping method [29] fits all those

requirements and has been widely and successfully used in many application fields.

Sammon's mapping is constructed as to minimize the inter-point distortions it introduces, quantified by the error measure:

$$\frac{1}{\sum_{i<j}\delta_{ij}}\sum_{i<j}\frac{(\delta_{ij}-\xi_{ij})^2}{\delta_{ij}}, \tag{1}$$

where δ_{ij} is the Euclidean distance between spectra i and j in the original data space and ξ_{ij} is the Euclidean distance between the projections of these spectra in the 3-D space. A low Sammon error means that distances in the original space are preserved in the 3-D visualization space.

We must provide quantitative support to the preliminary choice of Sammon's mapping as a method for dimensionality reduction and visualization in a 3-D space. Its performance is therefore compared to that of an alternative linear method, PCA, in terms of the *Trustworthiness* and *Continuity* measures developed in [36]. Data neighbourhood relationships that are not preserved in the low-dimensional representation, hamper the continuity of the latter, while spurious neighbouring relationships in the low-dimensional representation that do not have a correspondence in the observed space limit its trustworthiness. PCA is chosen for comparison as it is commonly used for dimensionality reduction in oncology MRS studies [7,20].

Trustworthiness is formally defined as

$$T(K)=1-\frac{2}{NK(2N-3K-1)}\sum_{i=1}^{N}\sum_{x_j\in U_K(x_i)}(r(x_i,x_j)-K), \tag{2}$$

where $U_k(x_i)$ is the set of data points x_j for which $x_j\in\hat{C}_K(x_i)\wedge x_j\notin C_K(x_i)$ and $C_K(x_i)$ and $\hat{C}_K(x_i)$ are the sets of K data points that are closest to x_i in the observed data space and in the low-dimensional representation space, respectively. *Continuity* is in turn formally defined as

$$Cont(K)=1-\frac{2}{NK(2N-3K-1)}\sum_{i=1}^{N}\sum_{x_j\in V_K(x_i)}(\hat{r}(x_i,x_j)-K), \tag{3}$$

where $V_K(x_i)$ is the set of data points x_j for which $x_j\notin\hat{C}_K(x_i)\wedge x_j\in C_K(x_i)$. The terms $r(x_i,x_j)$ and $\hat{r}(x_i,x_j)$ are the ranks of x_j when data points are ordered according to their distance from the data vector x_i in the observed data space and in the low-dimensional representation space, respectively, for $i\neq j$.

The *Continuity* and *Trustworthiness* results for PCA and Sammon's mapping are reported in Table 1. They indicate the adequacy of the Sammon's mapping choice, specially for the preservation of the *Trustworthiness*.

3.1.2. Outlier detection using t-GTM

Generative topographic mapping [3] is a nonlinear latent variable model generating a mapping from K points in a low-dimensional latent space onto the multivariate data space. The mapping is carried through by a set of basis functions generating a constrained mixture density distribution. It is defined as a generalized linear regression model:

$$\mathbf{y}=\mathbf{\Phi W}, \tag{4}$$

where $\mathbf{\Phi}$ is a $K\times M$ matrix built with the images of M basis functions, which, in the original GTM formulation, for continuous data of dimension D, were chosen to be spherically symmetric Gaussians $\phi_m(\mathbf{u})=\exp\{-1/2\sigma^2\|\mathbf{u}-\mu_m\|^2\}$, with centres μ_m and common width σ; $\mathbf{W}$ is a matrix of adaptive weights w_{md} that defines the mapping and $\mathbf{u}$ is a point in latent space. To avoid computational intractability, a regular grid of K points $\mathbf{u}_k$ can be sampled from the latent space. Each of them is mapped, using (4), into a low-dimensional manifold nonlinearly embedded in the data space. Therefore, GTM can be considered as a manifold learning model. A probability distribution for the multivariate data $\mathbf{X}=\{\mathbf{x}_n\}_{n=1}^N$ can then be defined, leading to the following expression for a log-likelihood:

$$L=\sum_{n=1}^{N}\ln\left\{\frac{1}{K}\sum_{k=1}^{K}\left(\frac{\beta}{2\pi}\right)^{D/2}\exp\left\{\frac{-\beta\|\mathbf{y}_k-\mathbf{x}_n\|^2}{2}\right\}\right\}, \tag{5}$$

where a prototype $\mathbf{y}_k$ residing in the observed data space is obtained for each latent space point $\mathbf{u}_k$, using (4); and β is the inverse of the noise model variance. As for finite mixture models, of which GTM is a manifold-constrained instance, the expectation-maximization (EM) algorithm is a straightforward alternative to obtain the maximum likelihood estimates of the adaptive parameters of the model, namely $\mathbf{W}$ and β.

For the standard Gaussian GTM, the presence of outliers is likely to negatively bias the estimation of the adaptive parameters. In order to overcome this limitation, the GTM was recently redefined [34] as a constrained mixture of Student's t distributions: the t-GTM, aiming to increase the robustness of the model towards outliers. The mapping described by Eq. (4) remains, with the basis functions now being Student's t distributions and leading to the definition of the following mixture density:

$$p(\mathbf{x}|\mathbf{W},\beta,v_k)=\frac{1}{K}\sum_{k=1}^{K}\frac{\Gamma\left(\frac{v_k+D}{2}\right)\beta^{D/2}}{\Gamma\left(\frac{v_k}{2}\right)(v_k\pi)^{D/2}}\left(1+\frac{\beta}{v_k}\|\mathbf{y}_k-\mathbf{x}_n\|^2\right)^{v_k+D/2}, \tag{6}$$

where $\Gamma(\cdot)$ is the gamma function and the parameter $v=(v_1,\ldots,v_K)$ represents the degrees of freedom for each component k of the mixture, so that it can be viewed as a tuner that adapts the level of robustness (divergence from normality) for each component.

As a byproduct of this reformulation of GTM, a statistic quantifying to what extent t-GTM considers a data case $\mathbf{x}_n$ to be an outlier can be defined, following [28], as

$$O_n=\sum_k p(\mathbf{u}_k|\mathbf{x}_n)\beta\|\mathbf{y}_k-\mathbf{x}_n\|^2. \tag{7}$$

The larger the value of Eq. (7), the more likely the case is to be an outlier. Notice that $p(\mathbf{u}_k|\mathbf{x}_n)$ is the responsibility assumed by a latent point k for the data case n and, the same as for the standard GTM, it is obtained as part of the Maximum Likelihood estimation of the model's parameters, in the M-step of the EM algorithm.

3.1.3. Shortlisting outlier cases of interest

The process of shortlisting outlier cases of potential interest is structured in four stages:

- Sammon's mapping, as described in Section 2, is first used to produce a nonlinear dimensionality reduction of the data to three dimensions.

Table 1
Trustworthiness and *continuity* results for PCA and Sammon's mapping, for different neighbourhood sizes K.

	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=10$
Trustworthiness						
Sammon	0.910	0.908	0.911	0.913	0.916	0.922
PCA	0.887	0.894	0.893	0.897	0.899	0.900
Continuity						
Sammon	0.952	0.953	0.953	0.953	0.951	0.952
PCA	0.953	0.950	0.952	0.950	0.947	0.946

- The free software package KING [17] is then used to visualize in 3-D the Sammon's mapping of the spectra, enabling a preliminary data exploration by experts.
- The data projections obtained with Sammon's mapping are then modelled by *t*-GTM (using all the spectra for finding *artefact-related outliers* and *distinct outliers*, but only spectra belonging to, in turn, *G*1, *G*2 and *G*3 for finding *class outliers*), from which a value of O_n is obtained for each data case that quantifies the corresponding degree of atypicality. Histograms of O_n were generated to shortlist potentially conflictive cases of the three types described in the Introduction. Loose thresholds of the statistic were set for the selection of the lists of outlier candidates.
- Using all this information, two experts then singled out those spectra they considered to be truly atypical in any sense and compared them to the characteristic spectra corresponding to their tumour type. Cases were only accepted as outliers when both experts singled them out as truly atypical in any sense *and* any of them deemed that this would preclude the correct interpretation of the case by the diagnostic decision maker. When agreed to be *artefact-related outliers*, spectra were tagged in the database with information about the artefacts, and recommendations on the suitability of their use for classification were made. When agreed to be *distinct outliers*, they were tagged as such. When agreed to be *class outliers*, a warning was included in the tags so that it could be taken into account before attempting classification.

3.2. Feature selection and classification

3.2.1. Feature selection

Let $X = \{X_1, \ldots, X_D\}$ be the original feature set. Mutual information (MI) measures the mutual dependence of two random variables from the point of view of information theory. It has been used with success as a criterion for feature selection in machine learning tasks. In this study, we use this concept embedded in a fast algorithm that calculates the MI between the class variable and a set of variables $\tau = \{\tau_1, \ldots, \tau_k\}$ by computing the MI between the class variable and a "super-feature" $\mathscr{V}_\tau$, whose possible values are the concatenations of all possible values of the features in τ. The conditional entropy between τ and the class feature Y is then calculated as

$$H(Y|\tau_1, \ldots, \tau_k) = H(Y|\mathscr{V}_\tau) = -\sum_{v \in \mathscr{V}_\tau} \sum_{y \in Y} p(v,y) \log \frac{p(v,y)}{p(y)}. \quad (8)$$

In this way, the MI can be determined as a simple bivariate case: $I(\mathscr{V}_\tau; Y) = H(Y) - H(Y|\mathscr{V}_\tau)$. An *index of relevance* of the feature $X_i \in X$ to a class Y with respect to a subset $\tau \subset X$, inspired on [2], is given by

$$R(X_i; Y|\tau) = \frac{I(X_i; Y|\mathscr{V}_\tau)}{H(Y|\mathscr{V}_\tau)} = \frac{H(Y|\mathscr{V}_\tau) - H(Y|X_i; \mathscr{V}_\tau)}{H(Y|\mathscr{V}_\tau)}. \quad (9)$$

This measure $R(X_i; Y|\tau)$ can be regarded as a conditioned *coefficient of constraint* [4], taking values between zero (no relevance) and one (maximum relevance). A discretization process is required in order to compute the conditional entropies in Eq. (9). Many dimensionality reduction studies use discretization schemes as a way to favor classification tasks (such as [27,21]). This change of representation does not often result in a significant loss of accuracy (on the contrary, sometimes significantly improves it); it also offers large reductions in learning time. The CAIM algorithm [18] is here selected because it is able to work with supervised data and does not require the user to define a specific number of intervals for each feature.

Let $\Delta_{p\times(D+1)} = (d_{i,j})$ be a discrete data matrix described by D variables $X = \{X_1, \ldots, X_D\}$ (plus the class variable Y, in column $D+1$). The matrix Δ is first sorted using lexicographical order, which accelerates future computations (this is done only once). Then the conditional entropy of the class variable given a super-feature can be incrementally computed in only one pass over the observations. This way of calculating feature subset relevance is used to evaluate subsets of spectra, embedded into a fast filter forward-search strategy, conforming the *entropic filtering algorithm*. A detailed description of a fast implementation of this algorithm can be found in [11].

3.2.2. Classification

Cross-validation (CV) is often used for the estimation of prediction errors in classification, providing almost unbiased estimation. However, estimating misclassification error with small samples such as the one available for this study raises concerns over its performance, since CV presents large variability. One way around this potential problem involves combining the bootstrap with CV (by performing CV on each of the bootstrap samples), in a method called bootstrap CV or BCV [9]. Bootstrap methods are well-suited for the construction of standard error estimates and their confidence intervals (CIs) when sample size is small or the distribution of the statistic is unknown.

The ^{1}H-MRS dataset S was used to generate $B = 1000$ bootstrap samples $S_1, \ldots, S_B$ that played the role of *training sets*. Denote $T_i = S \backslash S_i$ the corresponding *test sets*. The training sets were used for the feature selection process itself (by the EFA), *a posteriori* classifier induction and model selection (by CV). The test sets were used to ascertain the generalization ability of the developed classifiers.

Seven different classifiers were first designed using the training sets by means of leave-one-out CV (LOOCV) and the full set of frequencies. They are: the nearest-neighbour technique with Euclidean metric (*kNN*) and parameter k (number of neighbours), the *Naïve Bayes classifier* (NB), a *linear discriminant classifier* (LDC), a *quadratic discriminant classifier* (QDC), *logistic regression* (LR), a *support vector machine with quadratic kernel* (SVM^2) and a *support vector machine with linear kernel* (SVM-L): both SVMs with a parameter C (regularization constant). EFA was applied to the discretized ^{1}H-MRS data (the training parts only) to obtain what we call Best Spectral Subsets (BSS). Note that the EFA, being a filter method, does not require an inducer. The classifiers are then built in the training sets using the original continuous frequencies (both in the full set and in the obtained BSSs) and evaluated in the corresponding test sets.

There is an unavoidable inconvenience resulting from the application of a feature selection algorithm on every boostrap sample: the process yields a different (though probably quite similar) solution for each and every sample. Bias-variance for feature selection is a promising research field, but there is no consensus yet on how to derive a single solution from multiple ones. In the present setting, this situation is aggravated by the fact that the EFA is capable of delivering *more than one solution*, if desired. This is so because there may be several possibilities of reaching maximum relevance by the addition of the last feature. Given the importance of a correct assessment of feature importance, it was decided in this study to track them all. Hence, the application of the EFA yields a collection of solutions $\Sigma_1, \ldots, \Sigma_B$, which are sets of feature subsets.

Every Σ_i was first collapsed into a single subset σ_i by computing the following function:

$$\mathscr{I}(\Sigma) = \sum_{a \neq b \in \Sigma} MI(a,b), \quad (10)$$

where *MI* is the mutual information between features a and b and setting $\sigma_i = \text{argmin}_{\Sigma \in \Sigma_i} \mathscr{I}(\Sigma)$. By construction of the EFA solutions,

all the Σ_i are of the same size, so normalization is not necessary, given that all the summations in (10) have the same number of terms. This way of proceeding ensures that the chosen feature subset has maximum relevance (because it was one of the sets delivered by the EFA) and hopefully minimum redundancy among its features—because it is the minimizer of (10).

Once the boostrap feature sets $\sigma_1, \ldots, \sigma_B$ are obtained, we explored in this work four different strategies to obtain a single bootstrap solution σ^*, as follows. First, we created the set $\mathscr{F}$ as the union of all the σ_i and defined the *frequency* of a feature as the number of times it belongs to any of the σ_i, divided by B.

$R = 1$: Elements in $\mathscr{F}$ are fed into a forward selection algorithm sorted by descending frequency, where the stopping condition is met when reaching the maximum relevance ($R = 1$).

20% *cum.*: For many events, 80% of the effects (viz. classification ability) come from 20% of the causes (viz. spectral frequencies). Following this Pareto principle, spectral points in $\mathscr{F}$ were included in the final subset until reaching approximately 20% of the normalized cumulative frequency.

20% *fea.*: Similarly to the previous strategy, the 20% of the most frequent spectral points in $\mathscr{F}$ are included.

Peaks: The elements in the sets $\sigma_1, \ldots, \sigma_B$ are considered as a distribution, which can be displayed. Some of the most dominant and interpretable peaks of the resulting histogram are then selected with the advice of the expert radiologist (see Figs. 6 and 8).

4. Experimental results and discussion

4.1. Outlier characterization

In this study, the minimization of the Sammon's error was performed by the Newton method. A collection of models was obtained by varying the initial points (100 different random values) and the step size (nine different values), for a total of 900 runs. The models with lowest Sammon's error were selected for further analysis. The visualization of the high-dimensional spectra through Sammon's nonlinear mapping is illustrated in Fig. 1 (left). *High-grade malignant* tumours are displayed in black, *low-grade gliomas* in white, and *meningiomas* in gray. Overall, these three groups look well-defined and show a reasonable degree of separation, but it is also clear that some cases do not conform to this behaviour and that some of the issues outlined in the Introduction should be considered. The level of distortion resulting from Sammon's mapping is illustrated in Fig. 1 (right). A perfect diagonal would correspond to a mapping with no distortion.

Several spectra of interest are also displayed in Fig. 2 for illustration, highlighted from different views of the 3-D data: two cases (I0105, a glioblastoma, and I1090, a meningioma) in Fig. 2 (top row) that are both *distinct* and *class outliers*. The meningioma I0009, in Fig. 2 (middle row), which is a *class outlier*, but not a *distinct outlier*. Finally, in Fig. 2 (bottom row), two *artefact-related outliers*, which the experts described as being contaminated by noise and affected by alignment (this without consensus between experts) and polispiculated artefacts (in the case of I0175, a glioblastoma), and affected only by a polispiculated artefact (I0420, a meningioma).

The histogram in Fig. 3 displays the distribution of the O_n measure in Eq. (7) for the complete MRS dataset. A threshold of $O_n = 15$ was set to shortlist outlier candidate spectra. This yielded 21 potential outliers that were inspected by the experts. The first expert decided that 18 of them qualified as such (3 *distinct outliers* and 15 *artefact-related outliers*. The three shortlisted potential outliers not considered as such by the first expert are represented and described in Fig. 4). The second expert also singled out 18 outliers (although not exactly the same: there was lack of coincidence in 2 cases, one proposed by each expert, which were therefore not included in the final list). Out of the 17 remaining candidates, both experts agreed on that 4 of them, even if atypical in some sense, could still be interpreted correctly by a trained decision maker. Consequently they were not included in the final list either.

The corresponding full outlier characterization is presented in Table 2. Interestingly, there is no *low grade glioma* amongst them on which the experts could agree upon and, instead, *high-grade malignant* outliers predominate (69% of all the agreed outliers, while only 56% of all data).

As mentioned in the Introduction, spectra can also be atypical specifically with respect to their group of tumours. These are what we call *class outliers* and we now turn our attention to them. Their corresponding histograms for statistic O_n are displayed in Fig. 5. Nine *low-grade gliomas*, seven *high-grade malignant tumours* and 10 *meningiomas* were shortlisted and inspected by the experts. Both experts considered that, out of these, none of the *low-grade gliomas* qualified as *class outliers*. There was more disagreement on *high-grade malignant tumours*: the first expert only accepted 2, whereas the second expert singled out these 2 plus another 4. Finally, the first expert singled out five *meningiomas*, all but one of

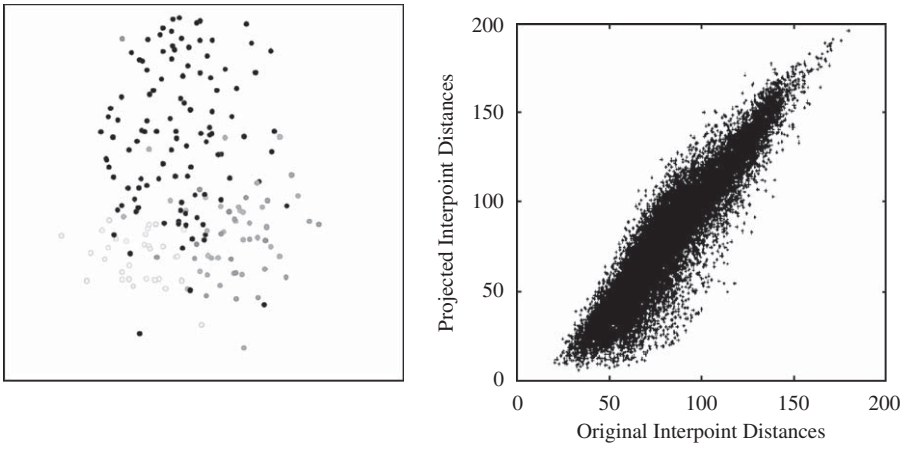

Fig. 1. A 3-D view of the projected MRS data obtained by Sammon's mapping (left: *High grade malignant* tumours are displayed in black, *low grade gliomas* in white, and *meningiomas* in gray) and the projected vs. original interpoint distances (right).

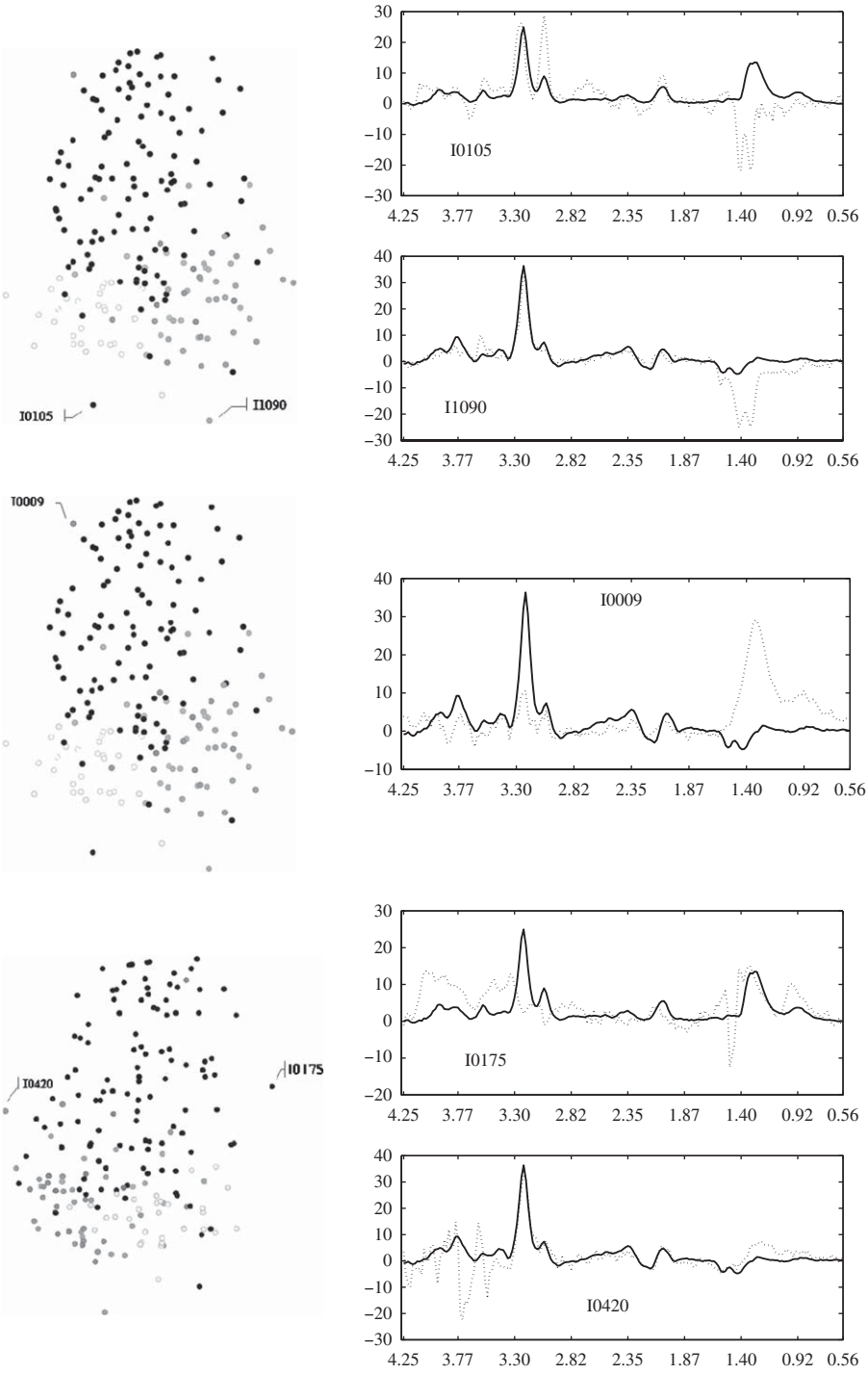

Fig. 2. 3-D Sammon's mapping view of several cases of interest (with groups of tumours displayed in black, gray and white, as in Fig. 1), on the left column, and their corresponding individual spectra (dotted lines) and mean spectra (solid lines) of the tumour groups they belong to, on the right column (case numbering as coded in the original INTERPRET database [14]). The abscissa axis displays frequency in ppm.

them agreed by the second expert. Some of these cases also contain artefacts, and they are characterized in full in Table 3. The decision to keep *class outliers* in the final list was, in this case, more conservative: they were not included if there was no agreement between the experts to single them out as truly atypical *or* if both of them deemed that atypicality would not preclude the correct interpretation of the case by the diagnostic decision maker. It is worth noting that the small number of *class outliers* identified in this dataset suggests the existence of quite compact and homogeneous tumour groups. This is most evident for *low-grade gliomas*, amongst which the experts did not agree upon any of the cases.

4.2. Feature selection and classification

For every feature selection experiment, the size of the corresponding BSSs, their test set performance, basic sample

statistics and bootstrap confidence intervals are reported in this section.

The average computation times for the different procedures (all carried out in a node of a computation cluster including 2 CPUs at 2.21 MHz) were as follows: for the application of EFA to one bootstrap sample: 65 s. For the obtention of one (final) subset of spectral points per bootstrap sample: 1.87 s for the slowest method. For the obtention of the final subset of spectral points (i.e. for the whole process): 4.04 h. The average times for the development of a classifier from one bootstrap sample (including model selection, when necessary) ranged between 0.3 and 4.2 s.

The spectral frequencies corresponding to the features in the final BSSs derived from the four strategies described in Section 3.2 (*R*1, 20% *cum.*, 20% *fea.* and *Peaks*) are reported in full in Table 4. Their relative frequency of selection is displayed in Fig. 6. They are also summarily depicted in Fig. 7, shown against the average spectra for all classes (tumour groups).

Five general regions of specially relevant spectral frequencies can be easily observed: the first, between 3.89 and 3.72 ppm, corresponds to the presence of glutamate/glutamine-containing compounds (^{2}CH-groups) and of alanine (^{2}CH-group), which mostly separates *meningiomas* from the other groups of tumours. The second has its centre in the creatine peak at 3.03 ppm; this metabolite plays a role in the maintenance of energy metabolism [7]. The third, from 2.52 to 2.18 ppm, corresponds again to glutamate and glutamine metabolites (this time $^4CH_2$-groups), with characteristically high values for *glioblastomas* and *meningiomas*. The fourth, from 1.67 to 1.32 ppm, is roughly located nearby the Alanine ($^1CH_3$-group) peak. Finally, the fifth region, from 1.38 to 1.15 ppm, covers the area where lactate and lipids are neatly identified; *high-grade malignant tumours* show very high values in this area, which indicate the existence of anaerobic metabolism resulting from lack of blood irrigation.

Notice the fact that the feature selection histogram in Fig. 6 is quite smooth, with several spectral frequencies out of those around the most frequently selected ones being also often picked up as relevant. Correspondingly, we also observe the existence of several frequency intervals of low relevance. All this is consistent with the fact that contiguous spectral points in a peak region will in most cases be highly inter-correlated [32].

Complete test-set performances are reported as follows. In Table 5, the five rows include information on the size of the used

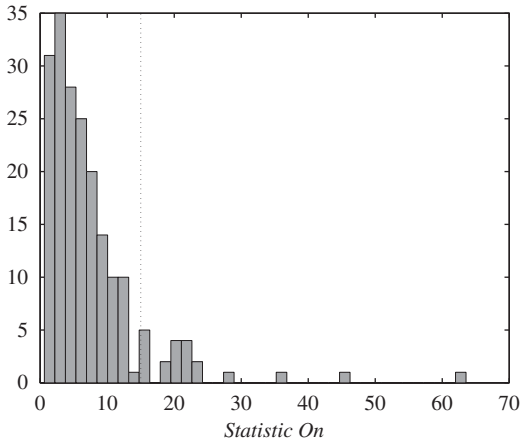

Fig. 3. Histogram of statistic O_n obtained using *t*-GTM for the whole dataset. The selected threshold at value 15 is represented as a vertical dotted line.

Table 2
Outlier characterization of the complete ^{1}H MRS LET dataset.

Id	*Tum*	*Dis*	Artefact-relat. outl.					
			noi	*wat*	*ali*	*bas*	*pol*	*edd*
I1061(R)	G1(*a2*)†				X^1			
I0062*	G2(*gl*)‡		X^1	X^2	X^1	X^1		
I0105*	G2(*gl*)‡	X^2						
I0172	G2(*gl*)‡		X^1	X^2		X^2		
I0175*	G2(*gl*)‡		X^2		X^1		X^2	
I0354*	G2(*gl*)‡			X^1			X^2	
I0428*	G2(*gl*)‡			X^1			X^2	
I1044**(R)	G2(*gl*)‡						X^2	
I1057**(R)	G2(*gl*)‡		X^2	X^2			X^2	
I1379*	G2(*gl*)‡		X^2					X^2
I0027	G2(*me*)‡		X^2		X^1			
I0368**(R)	G2(*me*)‡			X^1			X^2	
I1070	G2(*me*)‡	X^2						
I0055(R)	G2(*me*)†	X^1						
I0390**(R)	G3(*mm*)‡				X^1		X^2	
I0420*	G3(*mm*)‡						X^2	
I1074	G3(*mm*)‡		X^2					
I1090*	G3(*mm*)‡	X^2						
I1378	G3(*mm*)‡		X^2		X^1		X^2	

Columnwise, *Id* is an anonymized case identifier from the INTERPRET database [14]; star superscripts in this column indicate that there are artefacts that do not preclude the expert's correct interpretation of the case (one star if only one expert agrees with this; two stars if both experts agree); (R) indicates that the case is rejected from the final list. *Tum* refers to tumour type (see labels in Section 2); in this column, "†" indicates that only one expert selected this case, whereas "‡" indicates that both experts selected it. *Dis* refers to *distinct outliers*. Six types of artefacts were found: *noi* stands for noise; *wat* for bad water signal suppression; *ali* for alignment; *lin*, linebase; *pol* for the polispiculated artefact and *edd* for eddy currents. See main text for details. In the *Dis* column, and in all the artefact-related ones, the superscript figure indicates if only one expert or both of them identified the corresponding type of artefact.

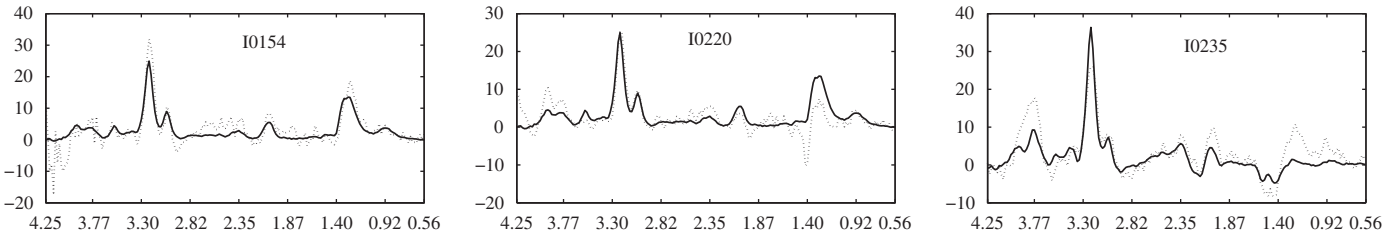

Fig. 4. Plots of the individual spectra (dotted lines) and mean spectra of the tumor groups they belong to (solid lines) for the three shortlisted potential outliers not considered as such by the first of the experts. The statistic O_n for all these three cases was very close to the selected threshold. (Left) A glioblastoma with an artefact in a very reduced range of frequencies over 4 ppm that do not correspond to metabolites of known relevance. (Centre) A clear glioblastoma with an unusual narrow inverted peak in the lactate area due to partial cancelation of the lactate/lipids doublet signals. (Right) An easily identifiable meningioma with unusually high-lipid/macromolecule levels.

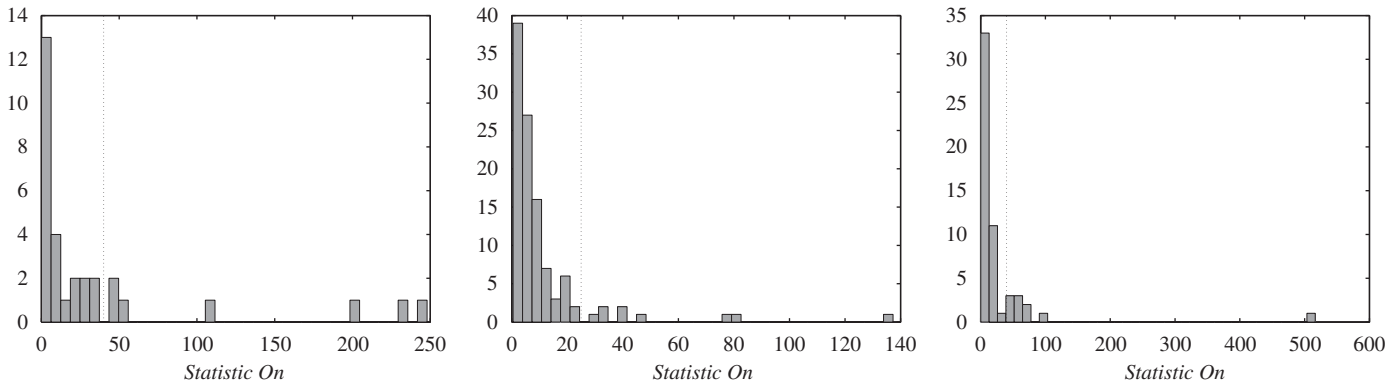

Fig. 5. Histogram of statistic O_n for each of the tumour groups in the dataset. (Left) *Low grade gliomas*; (centre) *High-grade malignant tumours*; (right) *Meningiomas*. The selected thresholds are represented as vertical dotted lines.

Table 3
Class outlier characterization of the ^{1}H MRS dataset, by groups of tumours.

Id	*Tum*	Artefacts					
		noi	*wat*	*ali*	*bas*	*pol*	*edd*
Low grade gliomas (G1)							
ø							
High-grade malignant (G2)							
I0175	*gl*†	X^1		X^1		X^1	
I0105*	*gl*‡						
I0172	*gl*†	X^1	X^1		X^1		
I0428	*gl*†		X^1			X^1	
I0055	*me*†						
I1070	*me*‡						
Meningiomas (G3)							
I0114*(R)	*mm*†						
I1090	*mm*‡						
I1378	*mm*‡	X^1				X^1	
I0002**(R)	*mm*‡						
I0009*	*mm*‡						

Label description as in previous table.

Table 4
Final subsets of spectral points (features) obtained by each strategy on the complete ^{1}H-MRS data set.

Reduction	$\|BSS\|$	ppm
$R=1$	24	3.81, 3.79, 3.77, 3.76, 3.74, 3.36, 3.05, 3.03, 2.94, 2.79, 2.52, 2.33 2.20, 2.14, 1.55, 1.53, 1.51, 1.32, 1.29, 1.27, 1.23, 1.21, 1.19, 1.17
20% *cum.*	4	3.76, 3.03, 1.53, 1.27
20% *fea.*	39	3.81, 3.79, 3.77, 3.76, 3.74, 3.72, 3.36, 3.05, 3.03, 3.00, 2.94, 2.79, 2.52, 2.48, 2.46, 2.39, 2.35, 2.33, 2.22, 2.20, 2.16, 2.14, 1.57, 1.55, 1.53, 1.51, 1.49, 1.44, 1.34, 1.32, 1.30, 1.29, 1.27, 1.25, 1.23, 1.21, 1.19, 1.17, 1.15
Peaks	8	3.76, 3.36, 3.03, 2.33, 2.14, 1.53, 1.32, 1.27

BSS and the mean classification performances for different classifiers in the form $\mu^* \pm \sigma^*/\sqrt{B}$, where μ^* and σ^* are the test set mean and standard deviation for accuracy in the bootstrap samples, respectively. These figures give a first impression of mean test-set performance and its stability. Each row corresponds

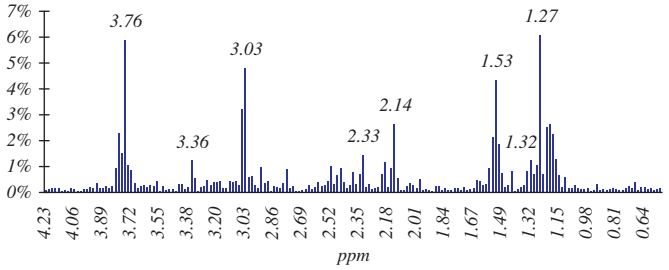

Fig. 6. Relative frequency distribution of spectral points selected by the EFA in the 1000 bootstrap samples from the complete ^{1}H-MRS data. Frequencies selected by the *Peaks* procedure are labelled for reference (3.76 ppm: glutamate/glutamine compounds; 3.36 ppm: unidentified; 3.03 ppm: creatine; 2.33 and 2.14 ppm: second type of glutamate/glutamine compounds; 1.53 ppm: nearby alanine; 1.32 ppm: lactate; 1.27: lipids).

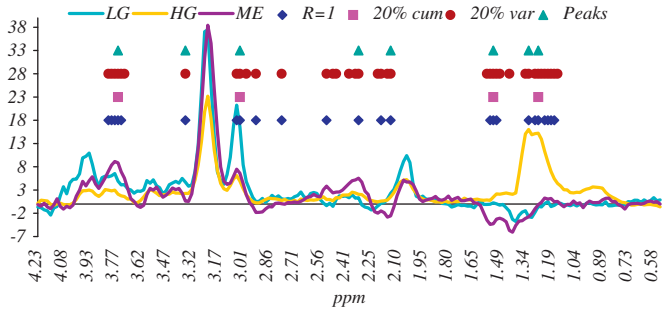

Fig. 7. Position of the final best spectral subsets per strategy in the complete ^{1}H-MRS long echo time dataset.

to a different method of choosing the BSS. More specifically, NR stands for "no reduction", and the other four are the strategies described in Section 3.2.

Test set confidence intervals, as shown in Table 6, can be obtained by the bootstrap percentile method, as follows: let $e^* = (e_1^*, \ldots, e_B^*)$ denote the error obtained by a given configuration (classifier plus reduction strategy) on the bootstrap samples $S_1, \ldots, S_B$. The CI is constructed by ordering e^* in ascending order and choosing critical value observations as the endpoints of the confidence intervals. For instance, for $B = 1000$, observations 26 and 975 are the endpoints of the 95% CI.

In addition, a statistical significance analysis of the observed differences between classification accuracies was performed using the Wilcoxon signed rank test (at the 95% level). We mainly aimed

Table 5
Bootstrap mean classification performance on the ^{1}H-MRS test sets.

Reduction	$\|BSS\|$	NB	kNN	LDC	QDC	LR	SVM^2	SVM-L
NR	195	83.2 ± 0.1	78.4 ± 0.1	*	*	72.0 ± 0.2	77.8 ± 0.1	84.7 ± 0.1
$R = 1$	24	85.0 ± 0.1	84.1 ± 0.1	82.9 ± 0.1	*	76.6 ± 0.2	79.4 ± 0.1	80.1 ± 0.1
20% *cum.*	4	82.4 ± 0.1	80.3 ± 0.1	82.4 ± 0.1	82.6 ± 0.1	84.4 ± 0.1	80.2 ± 0.1	84.2 ± 0.1
20% *fea.*	39	85.3 ± 0.1	83.6 ± 0.1	81.8 ± 0.1	*	76.7 ± 0.1	79.3 ± 0.1	82.1 ± 0.1
Peaks	8	84.1 ± 0.1	85.7 ± 0.1	85.6 ± 0.1	80.7 ± 0.1	86.1 ± 0.1	79.1 ± 0.1	84.5 ± 0.1

Results marked with (*) indicate numerical problems (number of observations equal or less than number of features). NR stands for "no reduction".

Table 6
Confidence intervals by percentile method of bootstrap mean classification performance on the ^{1}H-MRS test sets.

Reduction	NB	kNN	LDC	QDC	LR	SVM^2	SVM-L
NR	(74.7,90.3)	(68.9,86.7)	*	*	(61.2,81.1)	(68.9,86.3)	(76.5,91.7)
$R = 1$	(76.6,91.9)	(75.7,90.9)	(74.6,90.4)	*	(65.5,84.8)	(70.0,87.8)	(71.4,87.7)
20% *cum.*	(74.3,90.4)	(72.3,88.1)	(74.0,90.3)	(74.7,90.1)	(76.7,91.5)	(71.0,88.1)	(76.8,91.4)
20% *fea.*	(77.3,92.9)	(75.3,91.1)	(73.5,89.6)	*	(67.1,85.3)	(69.2,87.5)	(73.9,89.6)
Peaks	(76.9,91.6)	(76.0,90.8)	(76.7,91.3)	(71.2,88.6)	(76.1,91.5)	(71.0,87.9)	(75.7,91.7)

Results marked with (*) indicate numerical problems (number of observations equal or less than number of features). NR stands for "no reduction".

to ascertain whether the results obtained using the four reduction methods were significantly different to those of the NR (no reduction) method. All differences reported were found to be significant, with the single exception of the SVM-L classifier using the "Peaks" method.

Previous published work analysing similar ^{1}H-MRS data (but in a more simple setting that involves binary classification) used PCA followed by LDA to distinguish between *high-grade malignant tumours* and *meningiomas*, obtaining a mean AUC (area under the ROC curve) of 0.94, using six principal components [7]. The same method was used to distinguish between *high-grade malignant tumours* and astrocytomas grade II (part of the *low-grade gliomas* group), obtaining a mean AUC of 0.92, also using six principal components. (Note, though, that AUC results cannot be directly compared to classification accuracies such as the ones we report in our experiments.) In [32], a basic linear model (LDA) with six spectral frequencies (3.72, 3.04, 2.31, 2.14, 1.51 and 1.20 ppm) achieved a 83% of correct classification on an independent test set, this time using exactly the same multi-class setting with three groups of tumours that we have analysed in this study. Similar results were found in [24] for a combination of PCA and LDA and for different versions of support vector machines.

Our results, reported in Table 5, are very consistent with these, achieving a maximum 86% average correct classification with the *Peaks* procedure (eight mostly interpretable spectral frequencies, see Fig. 6) and a substantial 84% with just four interpretable frequencies, namely 3.76 (^{2}CH-groups of glutamate/glutamine-containing compounds), 3.03 (creatine), 1.53 (nearby the $^1CH_3$-group alanine peak) and 1.27 (lipids).

In view of the results reported in Figs. 6 and 7, and in Tables 4–6, several experimental findings can be summarized as follows:

1. Overall, all classifiers report stable test results, mostly in the region of 78–86% accuracy, which are consistent and slightly superior to others reported in the literature [32]. Naïve Bayes, LDC and SVM-L are, overall, the bests classifiers for these data, reaching up to 86% average accuracy (in the case of LDC) with only eight variables. The computational cost of SVM-L is higher, though, mainly due to the necessary model selection step performed for the latter. This gives an extra edge to simple models such as Naïve Bayes and LDC. It also suggests that linear models work better for these data.
2. The results justify the use of feature selection. All classifiers yield better results for at least some, if not all, the selection procedures than for the complete set of 195 frequencies (the *NR* procedure). All but one of these differences have been found to be statistically significant. Moreover, the selected features increase considerably the interpretability of the results. In the simplest scenario, Logistic Regression and SVM-L achieve 84% mean accuracy with only 4 frequencies that can be described in terms of the presence of metabolites in the tumour. The use of a selection procedure such as *Peaks*, partially guided by a human expert, finds justification in the maximum accuracy results it yields (86% for kNN, LDC and LR). This is specially important for the clinical implementation of these methods, in which the inclusion of expert prior knowledge is an almost compulsory requirement.

4.3. Assessing the effect of outliers on feature selection and classification

As stated in the Introduction, we hypothesize that the existence of outliers, and specially of *class outliers*, in the dataset will negatively affect the classification process, by forcing the classifiers to fit data which are atypical and, therefore, unrepresentative of either the overall population of brain tumours or the class they belong to.

In order to test this hypothesis, we repeat the experiments reported in the previous section for a subset of the complete dataset: one in which the 15 outliers (a 7.69% of cases: note that many *distinct* and *artefact-related outliers* are also *class outliers*) listed in Tables 2 and 3 are removed.

Again, for every feature selection experiment, the size of the corresponding BSSs, their test set performances, basic sample statistics and bootstrap confidence intervals are reported. The new spectral frequencies for the reduced dataset, corresponding to the features in the final BSSs derived from the *R*1, 20% *cum.*, 20% *fea.*, and *Peaks* strategies are reported in Table 7. Their relative frequency of selection is displayed in Fig. 8. They are also depicted in Fig. 9.

The same regions (with minor changes) of specially relevant spectral frequencies are obtained. Interestingly, the only region previously identified as relevant that loses relevance once the

Table 7
Final Best Spectral Subsets per strategy in ^{1}H-MRS long echo time dataset without outliers.

Reduction	$\|BSS\|$	ppm
$R = 1$	23	3.79, 3.76, 3.74, 3.05, 3.03, 2.94, 2.39, 2.35, 2.33, 2.31, 2.29, 2.20 2.16, 2.14, 1.57, 1.55, 1.53, 1.51, 1.32, 1.27, 1.23, 1.21, 1.19
20% *cum.*	5	3.76, 3.03, 2.14, 1.17, 1.23
20% *var.*	39	3.79, 3.77, 3.76, 3.74, 3.72, 3.05, 3.03, 2.94, 2.90, 2.52, 2.50, 2.48, 2.46, 2.39, 2.35, 2.33, 2.31, 2.29, 2.20, 2.16, 2.14, 1.57, 1.55, 1.53, 1.51, 1.49, 1.46, 1.44, 1.34, 1.32, 1.30, 1.29, 1.27, 1.25, 1.23, 1.21, 1.19, 1.17, 1.15
Peaks	6	3.76, 3.03, 2.33, 2.14, 1.51, 1.23

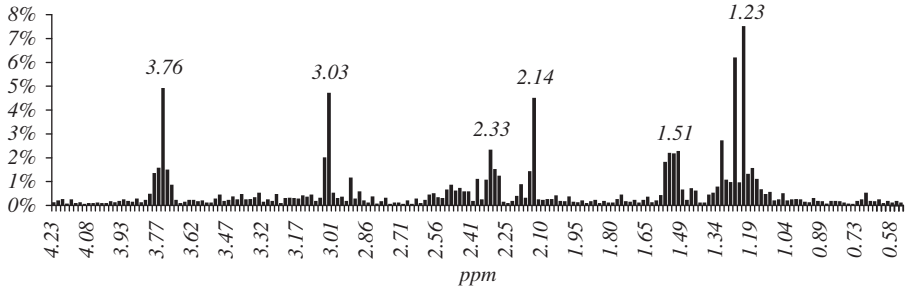

Fig. 8. Relative frequency distribution of spectral points selected by the EFA in the 1000 bootstrap samples from the ^{1}H-MRS data subset without outliers. Frequencies selected by the *Peaks* procedure are labelled for reference (The 3.76, 3.03, 2.14, 1.53 and 1.32 ppm frequencies remain, while the 2.33 ppm frequency in the second type of glutamate/glutamine compounds is now replaced by 2.39 ppm, and 1.27 ppm is replaced by 1.23 ppm in the lipids).

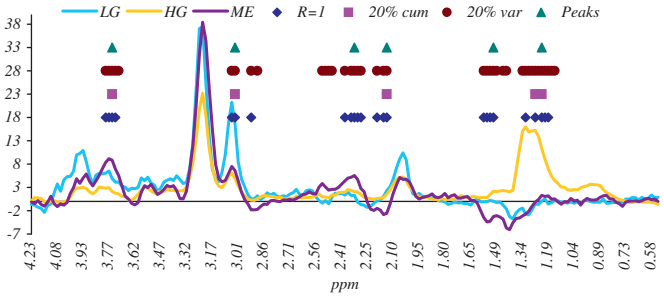

Fig. 9. Position of the final best spectral subsets per strategy in ^{1}H-MRS long echo time dataset without outliers in the whole spectrum.

outliers are removed is the one around 3.36 ppm, a frequency with unknown metabolical interpretation. This might indicate that the original relevance was unduly caused by the presence of outliers. In any case, these results refute one of the hypothesis stated in the introduction, according to which we expected the feature selection results to vary substantially in the absence of the identified outliers. The feature selection histogram in Fig. 8 is once again very smooth and makes the implementation of the *Peaks* procedure easy.

The classification results in Tables 8 and 9 are most interesting, as they reveal that the removal of the outlier cases improves the results for all classifiers overall and for most of the selection procedures. In summary:

1. Overall, all classifiers report very stable test results, mostly with over 80% accuracy. Again, Naïve Bayes and SVM-L yield, overall, some of the best results, this time accompanied by kNN (with LDC not far behind), reaching almost 88% average accuracy (in the case of LR and SVM-L) with only five variables and over 88% with six variables (with SVM-L). The accuracies improve in most of the settings on those obtained with the complete dataset, reported in Table 5, at least partially confirming the corresponding hypothesis formulated in the introduction section. This is reinforced by a statistical significance analysis (Wilcoxon signed rank test at the 95% level) that was carried out to qualify the differences of performance with and without outliers, for the four reduction methods plus the *no reduction* method. All differences were found to be significant, with the single exception of the SVM^2 classifier using the 20% *fea.* method. Such results at least partially justify the strategy of keeping the outliers out of the model.
2. The new results also confirm the usefulness of feature selection. All classifiers yield better results for at least some, if not all, the selection procedures. LR and SVM-L, as previously mentioned, achieve around 88% average accuracy with only five or six metabolically interpretable frequencies. A statistical significance analysis was again carried out to contrast the results obtained using the four reduction methods against the NR (no reduction) method. All differences were found to be significant, with the two exceptions of the SVM-L classifier using the *Peaks* method and the SVM^2 classifier using the $R = 1$ method.
3. The improvement in accuracy is somehow less noticeable for the most radical feature selection processes (20% *cum.* and *Peaks*). The possible reason for this is that by removing most of the features (leaving only 5 in 20% *cum.* and 6 in *Peaks*), we are likely to be removing most of the causes of outlierness for many spectra. Therefore, for extreme dimensionality reduction via feature selection, the strategy of keeping the outliers out of the model for classification becomes slightly less relevant.

Figs. 10 and 11 provide an illustrative general comparison of the classification results for the complete ^{1}H-MRS dataset and the subset without outliers using Naïve Bayes and SVM-L, two of the classifiers achieving best overall results. In this comparison,

Table 8
Bootstrap mean classification performance on the ^{1}H-MRS long echo time test sets without outliers.

Reduction	$\|BSS\|$	NB	kNN	LDC	QDC	LR	SVM^2	SVM-L
NR	195	85.7 ± 0.1	81.6 ± 0.1	*	*	74.6 ± 0.2	80.2 ± 0.1	88.2 ± 0.1
$R = 1$	23	88.2 ± 0.1	88.8 ± 0.1	86.6 ± 0.1	*	79.2 ± 0.2	80.1 ± 0.1	87.8 ± 0.1
20% *cum.*	5	84.8 ± 0.1	85.0 ± 0.1	85.4 ± 0.1	84.0 ± 0.1	87.6 ± 0.1	82.8 ± 0.1	87.8 ± 0.1
20% *var.*	39	88.3 ± 0.1	87.7 ± 0.1	84.9 ± 0.1	*	79.7 ± 0.1	79.6 ± 0.1	88.9 ± 0.1
Peaks	6	84.9 ± 0.1	86.5 ± 0.1	87.2 ± 0.1	85.8 ± 0.1	87.3 ± 0.1	82.6 ± 0.2	88.1 ± 0.1

NR stands for "no reduction".

Table 9
Confidence intervals by percentile method of bootstrap mean classification performance on the ^{1}H-MRS long echo time test sets without outliers.

Reduction	NB	kNN	LDC	QDC	LR	SVM^2	SVM-L
NR	(78.7,93.4)	(73.1,89.8)	*	*	(64.2,83.9)	(71.4,88.9)	(81.0,99.3)
$R = 1$	(80.6,95.3)	(81.7,95.4)	(79.5,92.9)	*	(62.5,89.1)	(70.8,88.7)	(79.1,95.2)
20% *cum.*	(77.9,92.4)	(77.6,91.9)	(78.1,92.4)	(75.7,91.8)	(79.7,94.3)	(72.3,91.0)	(80.0,94.9)
20% *fea.*	(80.9,95.3)	(80.6,93.9)	(76.1,92.5)	*	(70.3,88.2)	(70.0,87.9)	(81.3,95.5)
Peaks	(74.4,92.4)	(80.3,93.0)	(80.0,93.8)	(77.6,93.1)	(79.1,94.2)	(72.4,91.9)	(80.6,94.2)

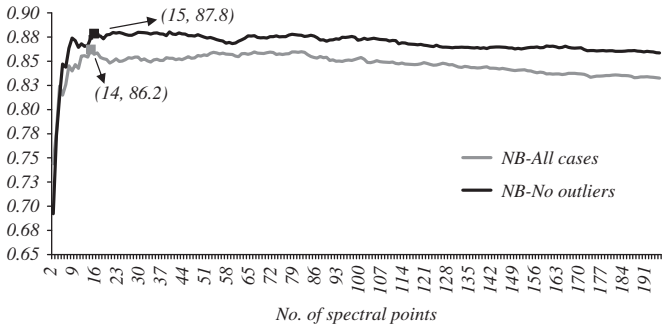

Fig. 10. Comparison of the bootstrap mean test classification performance for the ^{1}H-MRS long echo time test datasets using Naïve Bayes with and without outliers, using incremental subsets by adding one spectral point at a time, in order of relevance, and starting from the two most frequently selected by EFA. *X*-axis indicates the current size of the subset evaluated. An accuracy of almost 88% can be achieved with 15 frequencies.

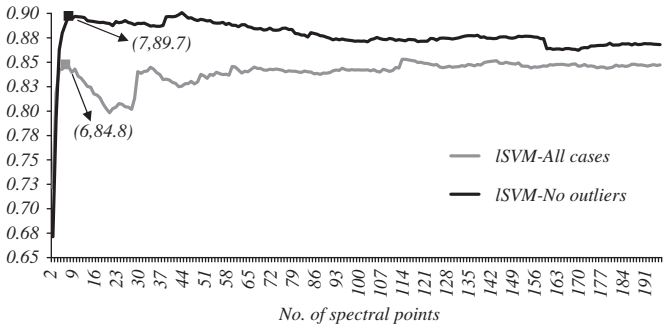

Fig. 11. Comparison of the bootstrap mean test classification performance for the ^{1}H-MRS long echo time test datasets using SVM-L with and without outliers, using incremental subsets by adding one spectral point at a time, in order of relevance, and starting from the two most frequently selected by EFA. *X*-axis indicates the current size of the subset evaluated. An accuracy of almost 90% can be achieved with as little as seven frequencies.

incremental frequency subsets are evaluated by adding one spectral point at a time (and it can, therefore, be considered as a generalization of the 20% *fea*. procedure). The subset without outliers outperforms the complete dataset throughout the whole range of subset sizes.

5. Conclusion

The diagnosis of brain tumours is a challenging medical area, where non-invasive techniques, such as those based on MR, play an important role. MR comes in two main flavors: imaging and spectroscopy. Most specialized doctors are accustomed to the imaging mode, but few are familiar with the nuances of spectroscopy, a technique that reveals the metabolic fingerprint of tumours. In this situation, many experts should benefit from the use of at least partially automated computer-based decision support.

In this study, we have defined a fairly general method for the semi-automated identification and characterization of atypical and potentially conflicting MR spectra corresponding to an international, multi-centre database of brain tumour pathologies. Atypical spectra can exist regardless of careful database quality control. This method combines nonlinear dimensionality reduction, exploratory visualization and automatic outlier detection techniques with expert knowledge.

A thorough feature selection process, based on an entropic filtering algorithm, followed by classification using a wide array of linear and nonlinear models, has revealed that a parsimonious and easily interpretable subset of spectral frequencies can provide significantly better accuracy than the complete high-dimensional spectrum. It has also been shown that the removal of the detected outliers from the analysed datasets can significantly improve classification, although this strategy becomes less relevant for the most extreme feature selection procedures. It is also interesting that, refuting our initial hypothesis, the removal of outliers from the dataset does not substantially modify the feature selection results. In conclusion, the reported results partially support the usefulness of the outlier identification and characterization procedure, as it is shown to yield models with similar or better generalization capabilities.

Future research should extent these experiments to short echo time ^{1}H-MRS data. These data provide more detailed metabolic information (due to a lesser attenuation of the signal) but at the expense of poorer peak resolution (i.e. more peak overlapping). Comparisons between the different echo times should be carried out. Ways of combining the information of both echo times should also be explored.

Acknowledgements

Authors gratefully acknowledge the former INTERPRET (EU-IST-1999-10310) European project partners. Data providers: Drs. C. Majós (IDI), À. Moreno-Torres (CDP), F.A. Howe and Prof. J. Griffiths (SGUL), Prof. A. Heerschap (RU), Drs. W. Gajewicz (MUL) and J. Calvar (FLENI); data curators: Dr. A.P. Candiota, Ms. T. Delgado, Ms. J. Martín, Mr. I. Olier and Mr. A. Pérez (all from GABRMN-UAB). C. Arús and M. Julià-Sapé are funded by the CIBER of Bioengineering, Biomaterials and Nanomedicine, an initiative of the *Instituto de Salud Carlos* III (ISCIII) of Spain. The authors acknowledge funding from M.E.C. projects TIN2006-08114 and SAF2005-03650, and, from 1 January 2003, Generalitat de Catalunya (Grant CIRIT SGR2005-00863). E. Romero acknowledges the use of the UPC Applied Math I department computing cluster system (URL: http://www.ma1.upc.edu/eixam/index.html). F.F. González-Navarro acknowledges funding from The Mexican National Council of Science and Technology CONACyT and Baja California University (UABC). A. Vellido is a Ramón y Cajal research fellow, funded by the Spanish Ministry of Science and Innovation.

References

[1] Artificial Intelligence Decision Tools for Tumour Diagnosis Research Project ⟨http://www.lsi.upc.edu/~websoco/AIDTumour⟩.
[2] D.A. Bell, H. Wang, A formalism for relevance and its application in feature subset selection, Mach. Learn. 41 (2) (2000) 175–195.
[3] C.M. Bishop, M. Svensén, C.K.I. Williams, The generative topographic mapping, Neural Comput. 10 (1) (1998) 215–234.
[4] C.H. Coombs, R.M. Dawes, A. Tversky, Mathematical Psychology: An Elementary Introduction, Prentice-Hall, Englewood Cliffs, NJ, 1970.
[7] A. Devos, Quantification and classification of MRS data and applications to brain tumour recognition, Ph.D. Thesis, Katholieke University, Leuven, Belgium, 2005.
[9] W. Fu, R. Carroll, S. Wang, Estimating misclassification error with small samples via bootstrap cross-validation, Bioinformatics 21 (9) (2005) 1979–1986.
[11] F. González, Ll. Belanche, Gene subset selection in microarray data using entropic filtering for cancer classification, in: Expert Systems, 2008, in press.
[13] Y. Huang, P.J.G. Lisboa, W. El-Deredy, Tumour grading from magnetic resonance spectroscopy: a comparison of feature extraction with variable selection, Stat. Med. 22 (2003) 147–164.
[14] INTERPRET project ⟨http://azizu.uab.es/INTERPRET⟩.
[15] INTERPRET project, Data Protocols ⟨http://azizu.uab.es/INTERPRET/cdap.html⟩.
[16] 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. MAGMA 19 (2006) 22–33.
[17] KING visualization software ⟨http://kinemage.biochem.duke.edu/software/king.php⟩.
[18] L. Kurgan, K. Cios, CAIM discretization algorithm, IEEE Trans. Knowledge Data Eng. 16 (2) (2004) 145–153.
[19] C. Ladroue, F.A. Howe, J.R. Griffiths, A.R. Tate, Independent component analysis for automated decomposition of in vivo magnetic resonance spectra, Magn. Reson. Med. 50 (2003) 697–703.
[20] C. Ladroue, Pattern recognition techniques for the study of magnetic resonance spectra of brain tumours, Ph.D. Thesis, St. George's Hospital Medical School, UK, 2004.
[21] D. Le, S. Satoh, Robust object detection using fast feature selection from hugh feature sets, in: 13th International Conference on Image Processing, IEEE, 2006, pp. 961–964.
[22] J. Lee, M. Verleysen, Nonlinear Dimensionality Reduction, Springer, Berlin.
[23] P.J.G. Lisboa, A review of evidence of health benefit from artificial neural networks in medical intervention, Neural Networks 15 (2002) 11–39.
[24] L. Lukas, A. Devos, J.A. Suykens, L. Vanhamme, F.A. Howe, C. Majós, A. Moreno-Torres, M. Van der Graaf, A.R. Tate, C. Arús, S. Van Huffel, Brain tumor classification based on long echo proton MRS signals, Artif. Intell. Med. 31 (2004) 73–89.
[26] C. Majós, M. Julià-Sapé, J. Alonso, M. Serrallonga, C. Aguilera, J.J. Acebes, C. Arús, J. Gili, Brain tumor classification by proton MR spectroscopy: comparison of diagnostic accuracy at short and long TE, Am. J. Neuroradiol. 25 (2004) 1696–1704.
[27] M. Ng, L. Chan, Informative gene discovery for cancer classification from microarray expression data, in: IEEE Workshop on Machine Learning for Signal Processing, 2005, pp. 393–398.
[28] D. Peel, G.J. McLachlan, Robust mixture modelling using the t distribution, Stat. Comput. 10 (2000) 339–348.
[29] J.W. Sammon Jr., A nonlinear mapping for data structure analysis, IEEE Trans. Comput. C-18 (1969) 401–409.
[31] A.R. Tate, C. Majós, A. Moreno, F.A. Howe, J.R. Griffiths, C. Arús, Automated classification of short echo time in vivo ^{1}H brain tumor spectra: a multicenter study, Magn. Reson. Med. 49 (2003) 29–36.
[32] A. Tate, J. Underwood, D.M. Acosta, M. Julià-Sapé, C. Majós, À. Moreno-Torres, F.A. Howe, M. van der Graaf, V. Lefournier, M.M. Murphy, A. Loosemore, C. Ladroue, P. Wesseling, J.L. 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, C. Arús, 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.
[33] A. Vellido, P.J.G. Lisboa, Handling outliers in brain tumour MRS data analysis through robust topographic mapping, Comput. Biol. Med. 36 (2006) 1049–1063.
[34] A. Vellido, P.J.G. Lisboa, D. Vicente, Robust analysis of MRS brain tumour data using *t*-GTM, Neurocomputing 69 (2006) 754–768.
[35] A. Vellido, E. Biganzoli, P.J.G. Lisboa, Machine Learning in cancer research: implications for personalised medicine, in: M. Verleysen (Ed.), Proceedings of the 16th European Symposium on Artificial Neural Networks (ESANN 2008), d-Side pub., Evere, Belgium, 2008, pp. 55–64.
[36] J. Venna, S. Kaski, Neighborhood preservation in nonlinear projection methods: an experimental study, in: G. Dorffner, H. Bischof, K. Hornik, (Eds.), Proceedings of the International Conference on Artificial Neural Networks (ICANN 2001), Springer, Berlin, 2001, pp. 485–491.
[37] A.J. Wright, C. Arús, J.P. Wijnen, A. Moreno-Torres, J.R. Griffiths, B. Celda, F.A. Howe, Automated quality control protocol for MR spectra of brain tumors, Magn. Reson. Med. 59 (2008) 1274–1281.

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.

Enrique Romero received a B.Sc. degree in Mathematics in 1989 from the *Universitat Autònoma de Barcelona* (UAB). In 1994, he received a B.Sc. degree in Computer Science from the *Universitat Politècnica de Catalunya* (UPC). In 1996, he joined the Department of *Llenguatges i Sistemes Informátics* at UPC, as an assistant professor. In 2004, he received the Ph.D. degree in Computer Science from the UPC. His research interests include Pattern Recognition, Neural Networks, Support Vector Machines and Feature Selection.

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-Muñoz is an associate professor in the *Departament de Llenguatges i Sistemes Informátics* at the *Universitat Politècnica de Catalunya* (UPC) in Barcelona, Spain. He received a B.Sc. in Computer Science from the UPC in 1990 and an M.Sc. in artificial intelligence in the UPC in 1991. He joined the Computer Science Faculty shortly after, where he completed his doctoral dissertation in 2000. His research involves neural networks and support vector machines for pattern recognition and function approximation, as well as feature selection algorithms, and their collective application to workable artificial learning systems.

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 (Univ. 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 ^{31}P MRS. His present interests in the field of tumour spectroscopy target the use of ^{1}H MRS of human brain tumours, biopsies and cell models for diagnosis, prognosis and therapy planning. He has published 66 PubMed accessible articles since 1977.