Classification, Dimensionality Reduction, and Maximally Discriminatory Visualization of a Multicentre ^{1}H-MRS Database of Brain Tumors

Paulo J.G. Lisboa
School of Computing and Mathematical Sciences
Liverpool John Moores University
Byrom St., L3 3AF Liverpool, United Kingdom
P.J.Lisboa@ljmu.ac.uk

Enrique Romero, Alfredo Vellido
Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
C. Jordi Girona, 1-3, 08034 Barcelona, Spain
eromero,avellido@lsi.upc.edu

Margarida Julià-Sapé[2,1], Carles Arús[1,2]
[1]Grup d'Aplicacions Biomèdiques de la RMN (GABRMN)
Universitat Autònoma de Barcelona (UAB), Cerdanyola del Vallès, Spain.
[2]CIBER-BBN, Cerdanyola del Vallès, Spain.
marga@carbon.uab.es,carles.arus@uab.es

Abstract

The combination of an Artificial Neural Network classifier, a feature selection process, and a novel linear dimensionality reduction technique that provides a data projection for visualization and which preserves completely the class discrimination achieved by the classifier, is applied in this study to the analysis of an international, multi-centre ^{1}H-MRS database of brain tumors. This combination yields results that are both intuitively interpretable and very accurate. The method as a whole remains simple enough as to allow its easy integration in existing medical decision support systems.

1. Introduction

Ideally, full patient bioprofiles should include disparate and heterogeneous information including, amongst other possible options, individual and family health history, genomic and proteomic data, and electronic medical records including biosignal and image recordings. The creation of standard databases of this type of bioprofiles is still far from becoming a reality even in advanced healthcare systems.

The creation and use of more restricted and goal-specific databases of bioprofiles, instead, is within closer reach. An example of this is the combination of Magnetic Resonance Spectroscopy (MRS) data with metabolomic and genomic profiles for brain tumor diagnostic assistance used in the e-Tumour European research project [1]. E-Tumour builds on the MRS decision support system (DSS) tool for tumour diagnosis previously delivered by the INTERPRET research project [2].

Decision making in oncology is a sensitive matter, and even more so in the specific area of brain tumour 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 should be non-invasive, clinicians might benefit from the use of an at least partially automated computer-based medical DSS. In this study, we analyze MRS data from the INTERPRET project, which are scarce and of very high dimensionality. This makes their computer-based automated classification a non-trivial undertaking. Most importantly, this high dimensionality also precludes the straightforward interpretation of the obtained results, limiting their usability in a practical medical context, in which interpretability is paramount and simplicity and robustness of the methods employed are compulsory requirements.

In this paper, we propose a combination of two methods: First, an Artificial Neural Network (ANN) classifier that, combined with a supervised feature selection (FS) procedure, yields very accurate results with a parsimonious subset of interpretable spectral MRS frequencies. Second, a novel linear dimensionality reduction technique that preserves in its integrity the class discrimination achieved by the classifier, while providing an intuitive visualization of the original dataset and making the results more interpretable. The experiments validate the usefulness of the proposed combination of methods, which could easily be integrated in a computer-based medical DSS.

978-0-7695-3495-4/08 $25.00 © 2008 IEEE
DOI 10.1109/ICMLA.2008.20

2. Data and methods

2.1. Brain tumour ^{1}H-MRS data

The data analyzed in this study were extracted from an international and multi-centre web-accessible database resulting from the International Network for Pattern Recognition of Tumours Using Magnetic Resonance (INTERPRET) European research project [2]. These data correspond to the combination (through concatenation) of single voxel ^{1}H-MR spectra measured at two echo times: a short-echo time (SET: PRESS 30-32 ms) and a long-echo time (LET: PRESS 135-144 ms). These spectra were acquired in vivo from 195 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). For further details on data acquisition and processing, and on database characteristics, see, for instance, [3, 10, 22].

Class labelling was performed according to the World Health Organization (WHO) system for diagnosing brain tumors by histopathological analysis of a biopsy sample. For the reported analysis, spectra were bundled into three groups, namely: *G1*: low grade gliomas (*a2*, *oa* and *od*); *G2*: high grade malignant tumors (*me* and *gl*); and *G3*: meningiomas (*mm*). This type of grouping is justified [21] by the difficulty in distinguishing between metastases and glioblastomas, which has its origin in the similar spectral pattern produced by the highly necrotic nature of both of these types of tumors. The clinically-relevant regions of both the SET and LET spectra were sampled to obtain 195 frequency intensity values from each (measured in parts per million (ppm), an adimensional unit of relative frequency position in the data vector), from 4.25 parts per million (ppm) down to 0.56 ppm. As a result, the data in the reported experiments consist of 195 cases and 390 features.

The approach of combining the SET and LET versions of these data through concatenation, in order to classify G1, G2 and G3 was also recently followed in [8].

2.2. Classification and feature selection using an artificial neural network

Models. Single-Layer Perceptron (SLP) ANNs with sigmoidal output units (one for each of the analyzed classes) were used both to obtain the test accuracy (within the learning process) and for the feature subsets evaluation criterion (within the FS process). The activation y_j of the output unit j for a d-dimensional input vector x is computed as

$$y_j = g\left(\sum_{i=1}^{d} x_i \cdot \omega_{ji} + b_j\right), \qquad (1)$$

where ω_{ji} is the weight that connects the input unit i with the output unit j, b_j is the bias of the output unit j, and $g(z)$ is the logistic function. The SLPs were trained in this study so as to minimize the sum-of-squares error.

There are several reasons for using SLPs instead of more complex ANN alternative models in this particular case. FS with Multi-Layer Perceptrons (MLP) would be computationally too expensive for the number of features of the analyzed ^{1}H-MRS brain tumour dataset (as described in Section 2.1) [16]. In addition, MLP parameters are more difficult to adjust. Alternatively, FS with linear Support Vector Machines (SVM) [9, 15] usually computes the saliency of the features as a function of the weights, as in our model (see below). However, the weights of a SLP are not necessarily a linear combination of the data, as for linear SVMs. Therefore, the saliency of every feature is likely to be more independent for SLPs than for linear SVMs.

In addition, linear models had shown quite good performance with these data in previous studies [22].

Feature subsets evaluation criterion. The relevance of a feature subset is computed as the sum of the individual saliencies of its features, where the saliency s_i of a feature i over O outputs is:

$$s_i = \sum_{j=1}^{O} |\hat{\omega}_{ji}|,$$

and where $\hat{\omega}_{ji}$ are the weights of the trained SLP.

This method is based on the hypothesis that irrelevant features produce smaller variations in the output values than relevant ones. Hence, a natural way to compare the relevance of two features is to compare the absolute values of the derivatives of the output function with respect to their respective input units in the trained model.

Formally, the derivative in the trained model of the output function y_j in (1) with respect to an input feature x_i is $\frac{\partial y_j}{\partial x_i} = g'\left(\sum_{i=1}^{d} x_i \cdot \hat{\omega}_{ji} + b_j\right) \cdot \hat{\omega}_{ji}$, and, for every j,

$$\frac{|\partial y_j / \partial x_{i_1}|}{|\partial y_j / \partial x_{i_2}|} = \frac{|\hat{\omega}_{ji_1}|}{|\hat{\omega}_{ji_2}|}.$$

Therefore, the variation (in absolute value) of the output function is smaller for input features with smaller weights (in absolute value), and they are the main candidates to be eliminated in a FS process. In summary, for linear discriminant functions such as SLPs, the magnitude of the weights corresponding to a feature is considered as an indicator of its importance.

Similar approaches have been applied elsewhere for other models. In [9] and [15], the squares of the weights in a trained linear SVM were used as a proxy of saliency. For hard-margin SVMs, this value is the variation in the cost function experimented when the feature is removed from the trained model.

For MLPs, more variations on this measure can be found in the literature. In [13] and [20], for example, the saliency of a feature is computed as the sum of the squares of the weights in the first hidden layer. In [12], [24] and [25], the absolute values of the weights in the first hidden layer are used to that end. The derivative of the output function with respect to the input features has also been widely used to compute their saliency, as in [14], [17], or [19].

FS search procedure. A backward selection procedure was used as an iterative selection process guided by the previously defined saliency measure. Starting from the complete set of available features, a subset of them was deleted at every step of the algorithm according to the evaluation criterion. Since the evaluation measure of a feature subset is computed as the sum of the saliencies of its features, the features to be deleted at every step are those with the smallest saliency. The number of features deleted at every step is a parameter of the system that controls the granularity of the selection and its computational cost.

The application of the standard backward selection procedure to FS (deleting one feature at every step) would involve the training of d networks. However, under the hypothesis that many of the features are not necessary to obtain a good classification performance (which is a reasonable hypothesis for the analyzed MRS dataset), a more aggressive strategy can be designed to save computational time: at every step, a fixed percentage of features is deleted. In order to control the granularity of the selection, the whole process can be repeated in different phases with different percentages at every phase (and different initial feature subsets). In the last phase, when only a few irrelevant features are supposed to remain in the dataset, the procedure can switch to deleting them one by one.

2.3. Low-dimensional data visualization with scatter matrices

As mentioned in the introduction, the high dimensionality of the analyzed spectra makes the interpretation of the obtained results a non-trivial undertaking, which potentially limits their usability in a practical medical decision making context such as brain tumor diagnosis. This may still be the case even after a FS process as the one described in the previous section. In these 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 Principal Component Analysis-based bi-plots [7]. While this approach is useful to visually verify known correlations between attributes, it is generally the case that the first two or three components 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 as in this study, 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 correspondence between nearest neighbour distances in the original input space and between the two-dimensional projections of the individual data points, such as in Multi-Dimensional Scaling (MDS) [4] or Sammon mapping [18]. However, 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. The non-linearity of the projections also makes them prone to misinterpretation.

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) [11] or in Generative Topographic Mapping (GTM) [23]. Although these are powerful methods for the simultaneous clustering and visualization of intrinsically non-linear data (and, therefore, able to produce more faithful representations of such data), their non-linearity can again make the interpretation of the obtained results difficult.

The method proposed here is linear in nature, making it easier to use in a real decision-making process that requires an intuitive representations of results. It is based on the decomposition of the scatter matrix, which is arguably a neglected method for dimensionality reduction with the remarkable property of maximizing the separation between the projections of compact groups of data (tumor classes, in this work). A new result is here described, which leads onto the definition of low-dimensional projective spaces with good separation between classes even when the covariance matrix of the data is singular.

It is well-known that the overall variance of the data, S_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 [5, 6] generating a within-cluster matrix, S_W, and a between-cluster matrix, S_B, namely: $S_T = S_W + S_B$. For a data matrix $\{x_i\}_{i=1}^{N}$comprising d-dimensional data points of overall mean m,

$$S_T = \sum_{i=1}^{N} \{(x_i - m)^T (x_i - m)\}$$

$$S_W = \sum_{j=1}^{N_c} \sum_{i=1}^{N_j} \{(x_i^j - m_j)^T (x_i^j - m_j)\}$$

$$S_B = \sum_{j=1}^{N_c} N_j \{(m_j - m)^T (m_j - m)\},$$

where the data are partitioned into N_c groups (tumor classes in this study), each with N_j points and mean 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 $M = S_W^{-1} S_B$ 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 $\{\hat{b}_i\}$, $i = 1 \ldots N_c$, for instance by Gram-Schmidt orthogonalization, generating the first compact projective representation in this method

$$x^c = \sum_{i=1}^{N_c} x \hat{b}_i^T .$$

However, the separating boundaries in the Voronoi decomposition created by the grouping will generally not be preserved by this projective map, causing mixing among classes. One way to preserve as much as possible of the class separation after compressing the data by projecting onto the space spanned by the means is to linearize the Mahalanobis metric in the original data space by whitening, or sphering, the data. This transformation is applied solely for the purpose of dimensionality reduction and visualisation.

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:

$$S_W^c = \sum_{j=1}^{N_c} \sum_{i=1}^{N_j} \{(x_i^c - m_j^c)^T (x_i^c - m_j^c)\}$$

$$S_B^c = \sum_{j=1}^{N_c} N_j \{(m_j^c - m^c)^T (m_j^c - m^c)\}$$

and $M^c = (S_W^c)^{-1} S_B^c$. Consequently, a diagonalization of the new scatter matrix M^c shows, typically, that the trace of the matrix is contained in the largest few eigenvalues. These eigenvectors form the basis for a two- or three-dimensional visualization of the data.

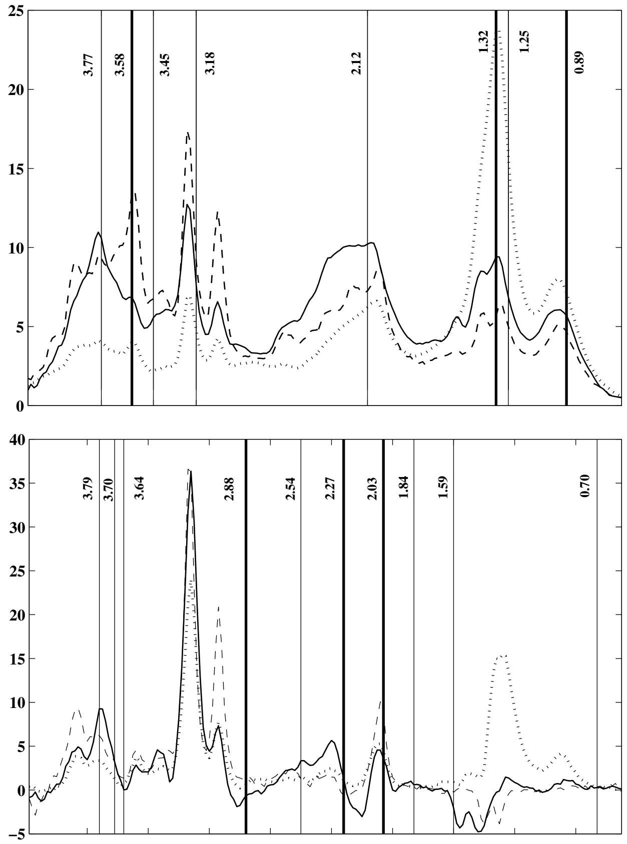

Figure 1. (Top): Representation of the 8 selected SET spectral frequencies (as vertical lines, with their value in ppm tagged by their side) out of the subset of 18 selected frequencies. The 3 most relevant out of the 8 are represented by thicker lines. Mean SET spectra of each class are represented as a dashed line (low grade gliomas), a dotted line (high grade malignant tumors), and a solid line (meningiomas). (Bottom): Similar representation of the 10 selected LET spectral frequencies.

3. Results and discussion

The SLP-based FS process and classification described in section 2.2 were carried out as follows. For the FS process, a first phase was performed where 50% of the features were deleted at every step of the backward selection procedure. In a second phase, a 20% of the remaining features were deleted. Finally, features were deleted one by one. This *ad hoc* procedure allowed keeping the computation time within reasonable bounds while threading more carefully through the sensitive final third stage of the feature selection process. In order to compute the saliencies of

Dataset	Test acc.	Nf	Subset of selected features
SET + LET	98.46%	18	L2.88 L2.27 S0.89 S3.58 L2.03 L2.54 L3.64 L1.59 S1.32 L3.79 S3.77 L1.84 S3.45 L3.70 S1.25 S2.12 L0.70 S3.18
SET + LET	91.08%	9	L2.27 L2.88 S2.12 L1.59 L2.54 L3.79 S3.58 S1.25 L2.03

Table 1. Classification results. First column: Dataset description. Second column: Average test accuracy of a 5-fold cross-validation procedure repeated 5 times. Third column: Number of features (N_f) selected. Fourth column: Frequency (in ppm) of the selected features, ranked according to their relative relevance; the frequency is preceded by a letter indicating the frequency ascription to either SET (S) or LET (L).

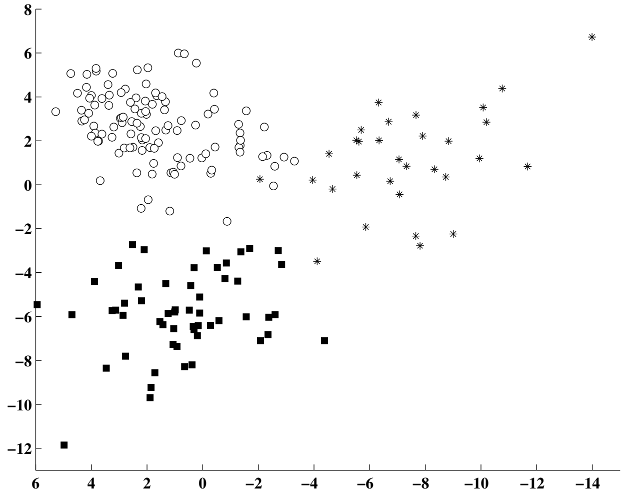

Figure 2. Visualization of low grade gliomas (asterisks), high grade malignant tumors (circles), and meningiomas (squares) using the first two eigenvectors of M^c, with the method described in the main text.

the features, the SLPs were trained with the whole data set. For the classification results, a 5-fold cross-validation procedure was performed 5 times for every feature subset obtained during the FS process, producing the results reported in Table 1.

Up to a 98.46% average test accuracy was achieved using a parsimonious subset of 18 data features, 8 of them belonging to SET and 10 to LET. The results reported in the second row of Table 1 illustrate how the test accuracy deteriorates as we detract features from the selected 18 subset: a parsimonious selection of 9 features is still able to retain over 91% of the average test classification in the multi-class classification problem involving *low grade gliomas*, *high grade malignant tumors*, and *meningiomas*. These results are a neat improvement on the average 88.71% accuracy obtained in [8] for the same problem, using stepwise FS with Linear Discriminant Analysis (LDA), and the just over 90% obtained using 8 principal components in PCA with LDA.

For illustration, the 18 selected spectral frequencies are displayed together with the mean spectra of the three classes, both for SET (Figure 1, top) and LET (Figure 1, bottom) datasets. Many of them have a clear interpretation in terms of metabolites and molecules often reported in the MRS literature as descriptors of brain tumor pathologies, such as frequencies associated to lipids, lactate, and myo-inositol in SET, N-acetyl-aspartate in LET, and Glutamine-Glutamate-GABA complex at both echo times.

A selection of 18 spectral frequencies, out of the 390 originally available, already simplifies the interpretation of results considerably. Even though, the intuitive visualization of the data is still out of bounds in a 18-dimensional data space. At this point, we would like to visualize the data without loosing any of the class discrimination obtained by the classifier. This is fully achieved using the method described in section 2.3, as illustrated by Figure 2. This is a 2-D view of a 3-D projection of the three classes (using only the first two eigenvectors of M^c). The discrimination is near-perfect by itself and fully preserves the classifier results, as $J = tr(M^c) = 1$. Bear in mind that this is a linear dimensionality reduction method and, therefore, the projection results could also be straightforwardly interpreted in terms of the initial subset of 18 selected features.

4. Conclusions

A very accurate classification of brain tumor typologies based on MRS information has been accomplished. However accurate, though, classification by itself does not suffice to ensure the clinical applicability of the results. Feature selection, as applied in this study, simplifies the results and makes them far more interpretable. Interpretation becomes more intuitive through data visualization, which can only be the result of further dimensionality reduction. Most dimensionality reduction techniques do not preserve class separation. Here, we have described one such technique that does preserve class separability in its integrity in the visualization of MRS brain tumor data.

Acknowledgements. Authors gratefully acknowledge the former INTERPRET (EU-IST-1999-10310) European project partners. Data providers: Dr. C. Majós (IDI), Dr. À.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 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.

References

[1] e-Tumour project. URL: http://www.etumour.net/.

[2] International Network for Pattern Recognition of Tumours Using Magnetic Resonance (INTERPRET) project. URL: http://azizu.uab.es/INTERPRET.

[3] International Network for Pattern Recognition of Tumours Using Magnetic Resonance (INTERPRET) project: Data protocols. URL: http://azizu.uab.es/INTERPRET/cdap.html.

[4] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman and Hall, UK, 2001.

[5] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley, NY, 1973.

[6] H. P. Friedman and J. Rubin. On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62(320):1159–1178, 1967.

[7] K. R. Gabriel. The biplot graphical display of matrices with applications to principal component analysis. *Biometrika*, 58(3):453–467, 1971.

[8] J. M. García-Gómez, S. Tortajada, C. Vidal, M. Julià-Sapé, J. Luts, A. Moreno-Torres, S. Van Huffel, C. Arús, and M. Robles. The influence of combining two echo times in automatic brain tumor classification by Magnetic Resonance Spectroscopy. *NMR in Biomedicine*, 2008. Accepted for publication.

[9] I. Guyon, J. Weston, S. Barnhill, and V. N. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422, 2002.

[10] M. Julià-Sapé, D. Acosta, M. Mier, C. Arús, D. Watson, and the INTERPRET Consortium. A multi-centre, web-accessible and quality control-checked database of in vivo MR spectra of brain tumour patients. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 19(1):22–23, 2006.

[11] T. Kohonen. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1):59–69, 1982.

[12] H. Lee, K. Mehrotra, C. Mohan, and S. Ranka. Selection procedures for redundant inputs in Neural Networks. In *INNS World Congress on Neural Networks*, volume 1, pages 300–303, 1993.

[13] K. Messer and J. Kittler. Choosing an optimal neural network size to aid a search through a large image database. In *British Machine Vision Conference*, volume 1, pages 235–244, 1998.

[14] K. L. Priddy, S. E. Rogers, D. W. Ruck, and G. L. Tarr. Bayesian Selection of Important Features for Feedforward Neural Networks. *Neurocomputing*, 5(2-3):91–103, 1993.

[15] A. Rakotomamonjy. Variable Selection using SVM-based Criteria. *Journal of Machine Learning Research*, 3:1357–1370, 2003.

[16] E. Romero and J. M. Sopena. Performing Feature Selection with Multi-Layer Perceptrons. *IEEE Transactions on Neural Networks*, 19(3):431–441, 2008.

[17] D. W. Ruck, S. K. Rogers, and M. Kabrisky. Feature Selection using a Multilayer Perceptron. *Journal of Neural Network Computing*, 2(2):40–48, 1990.

[18] J. W. Sammon. A Non-linear Mapping for Data Structure Analysis. *IEEE Transactions on Computers*, C-18:401–408, 1969.

[19] J. M. Steppe and K. W. Bauer. Feature Saliency Measures. *Computer & Mathematics with Applications*, 33(8):109–126, 1997.

[20] G. L. Tarr. *Multi-layered Feedforward Neural Networks for Image Segmentation*. PhD thesis, Air Force Institute of Technology, 1991.

[21] A. R. Tate, C. Majós, A. Moreno, F. A. Howe, J. R. Griffiths, and C. Arús. Automated Classification of Short Echo Time in Vivo 1H Brain Tumor Spectra: a Multicenter Study. *Magnetic Resonance in Medicine*, 49:29–36, 2003.

[22] A. R. Tate, J. Underwood, D. M. Acosta, M. Julià-Sapé, C. Majós, A. 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, and C. Arús. Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra. *NMR in Biomedicine*, 19(4):411–434, 2006.

[23] A. Vellido and P. J. G. Lisboa. Handling outliers in brain tumour MRS data analysis through robust topographic mapping. *Computers in Biology and Medicine*, 36(10):1049–1063, 2006.

[24] J. H. Wikel and E. R. Dow. The Use of Neural Networks for Variable Selection in QSAR. *Bioorganic & Medicinal Chemistry Letters*, 3(4):645–651, 1993.

[25] P. M. Wong, T. D. Gedeon, and I. J. Taggart. An Improved Technique in Porosity Prediction: A Neural Network Approach. *IEEE Transactions on Geoscience and Remote Sensing*, 33(4):971–980, 1995.