ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Feature Selection in Proton Magnetic Resonance Spectroscopy for Brain Tumor Classification

Félix F. González Navarro and Lluís A. Belanche Muñoz. *
{fgonzalez,belanche}@lsi.upc.edu
Universitat Politècnica de Catalunya, Ω-Building. Barcelona, Spain

Abstract. ^{1}H-MRS is a technique that uses response of protons under certain magnetic conditions to reveal the biochemical structure of human tissue. An important application is found in brain tumor diagnosis, due to the known complications of physical exploration and as a help to other kind of non invasive methods. It is possible to analize spectral data with machine learning methods to classify tumor classes in an automated fashion. One important characteristic of these data is their high dimensionality. In this work we present a contribution to lighten this situation with an algorithm based on entropic measures of subsets of spectral data. Experimental results show that the approach used has a good classification performance, both in terms of prediction accuracy and number of involved spectral frequencies.

1 Introduction

Proton magnetic resonance spectroscopy (^{1}H-MRS) is a non-invasive technique that provides information about the biochemical profile (metabolites and lipids) of brain tissue. Originally used for *in vitro* chemical analysis of small samples, it has been used in the diagnosis of adult brain tumors [1]. Previous existing work shows that it is possible to classify brain tumors using the values of data points of ^{1}H-MRS using machine learning techniques [2]. Some of these efforts perform dimensionality reduction with algorithms for feature extraction or pairwise feature selection techniques [3]. In this work, an Entropic Filtering Algorithm (EFA) for feature selection is described as a method to generate a relevant subset of spectral frequencies. This is a feature selection method based on finding feature subsets evaluated as a whole with respect to their ability to classify tumors, rather than on ranking individual contribution (that implicitly denies interaction between features). The EFA is tested on real ^{1}H-MRS data of 6 tumor classes (grouped in three super-classes). Several machine learning algorithms are used to test the reliability of the obtained subsets in the classification of tumors, within an appropriate experimental framework.

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

2 Entropic Filtering Algorithm for Cancer Classification

Mutual Information (MI) measures the mutual dependence of two random variables. It has been used with success as a criterion for feature selection in machine learning tasks. In this work we use this concept by deriving a fast algorithm that computes MI between a set of variables and the class variable by generating first a “super-feature”, obtained considering the concatenation of each combination of possible values of its forming features. In symbols, let $X = \{X_1, ..., X_n\}$ be the original feature set, consider a subset $\tau \subseteq X$ and define the operator $\biguplus$ as *concatenation*. Then define $\mathcal{V}_\tau$ as:

$$\mathcal{V}_\tau = \biguplus_{X_i \in \tau} X_i \tag{1}$$

Given $\tau = \{\tau_1, \cdots, \tau_k\}$, a single feature $\mathcal{V}_\tau$ is obtained uniquely, whose possible values are the concatenations of all possible values of the features in τ. The conditional entropy between $\mathcal{V}\tau$ and the class feature Y is then:

$$H(Y|\tau_1, \cdots, \tau_k) = H(Y|\mathcal{V}_\tau) = -\sum_{v \in \mathcal{V}_\tau} \sum_{y \in Y} p(v, y)\, log \frac{p(v, y)}{p(y)} \tag{2}$$

Proceeding in this way, the MI can be determined as a simple bivariate case: $I(\mathcal{V}_\tau; Y) = H(Y) - H(Y|\mathcal{V}_\tau)$. An *index of relevance* of the feature $X_i \in X$ to a class Y with respect to a subset $\tau \subset X$, inspired on [4], is given by:

$$R(X_i; Y|\tau) = \frac{I(X_i; Y|\mathcal{V}_\tau)}{H(Y|\mathcal{V}_\tau)} = \frac{H(Y|\mathcal{V}_\tau) - H(Y|X_i; \mathcal{V}_\tau)}{H(Y|\mathcal{V}_\tau)} \tag{3}$$

This index of relevance of a feature subset is to be maximized (it has a maximum value of 1). This measure is used next to evaluate subsets of spectral frequencies, embedded into a fast filter forward-search strategy, conforming the *Entropic Filtering Algorithm* (EFA). Let $D_{p\times(n+1)}$ be a discrete data matrix described by n variables $X = \{X_1, \ldots, X_n\}$ (plus the class variable Y, in column $n+1$). The matrix D is first sorted using lexicographical order, which accelerates future computations. Then a new matrix $T_{p\times 2}$ is generated formed by the super-variable $\mathcal{V}_\tau$ and the class Y. The pseudo-code of the algorithms is detailed below.

3 Experimental work

The analyzed ^{1}H-MRS dataset corresponds to 266 single voxel long echo time spectra acquired *in vivo* from brain tumour patients, out of which 195 are used in this study, including: meningiomas (55 cases), glioblastomas (78), metastases (31), astrocytomas Grade II (20), oligoastrocytomas Grade II (6), and oligodendrogliomas Grade II (5).

Algorithm 1: Conditional Multivariated Entropy

```
Function H $(Y, \tau \subset X)$
$v^- \leftarrow \uplus\{d_{1,j} \,|\, X_j \in \tau\}$   ;   $y^- \leftarrow d_{1,n+1}$
$cv \leftarrow cy \leftarrow 1$   ;   $H \leftarrow 0$
for $i \leftarrow 2$ to $p$ do
    $v \leftarrow \uplus\{d_{i,j} \,|\, X_j \in \tau\}$   ;   $y \leftarrow d_{i,n+1}$
    if $v^- = v$ and $y^- = y$ then
        $cv \leftarrow cv + 1$   ;   $cy \leftarrow cy + 1$
    else if $v^- \neq v$ then
        $H \leftarrow H + \frac{cy}{p} \log \frac{cy}{cv}$
        $cv \leftarrow cy \leftarrow 1$
    else
        $t \leftarrow \#\{v \,|\, v = T_{i,1}, i = 1, \ldots, p\}$        /* recall $T$ is sorted */
        $H \leftarrow H + \frac{cy}{p} \log \frac{cy}{t}$
        $cv \leftarrow cv + 1$   ;   $cy \leftarrow 1$
    $v^- \leftarrow v$   ;   $y^- \leftarrow y$
returns $-(H + \frac{cy}{p} \log \frac{cy}{cv})$
```

Algorithm 2: Index of relevance R

```
Function R $(X_i \in X \setminus \tau, Y, \tau \subseteq X)$
returns $\frac{\mathbf{H}(Y,\tau) - \mathbf{H}(Y,\tau \cup \{X_i\})}{\mathbf{H}(Y,\tau)}$
```

Algorithm 3: Entropic Filtering Algorithm

```
$\Phi \leftarrow \emptyset$                                /* Best Spectral Subset BSS */
repeat
    $x' \leftarrow \underset{x \notin \Phi}{\operatorname{argmax}}\{\mathbf{R}(x, Y, \Phi)\}$
    $\Phi \leftarrow \Phi \cup \{x'\}$
until $R(Y, \Phi) = 1$ or $\Phi = X$ ;
```

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 analysis in this study, spectra were grouped into three superclasses: high-grade malignant tumours (metastases and glioblastomas), low-grade gliomas (astrocytomas, oligodendrogliomas and oligoastrocytomas) and meningiomas. The analyzed spectra consist of 195 frequency intensity values, from 4.21 ppm down to 0.51 ppm.

3.1 Experimental setup

The main purpose of this research was achieving good generalization results with a model as simple as possible for better interpretation. To this end, the experimental conditions enforce several aspects: a well-balanced class representation, a feature selection process independent of the final classifier, and the selection of a small number of features.

The ^{1}H-MRS data set was randomly split into two parts: 70% for the feature selection process itself and *a posteriori* classifier induction and model selection by means of 3-fold cross validation (hereafter called the *training* set) and the remaining 30% that will be used to ascertain the generalization ability of the classifiers (the *test* set). This division was done keeping the relative proportion of classes in the whole data.

We were first interested in ascertaining whether a feature selection process could deliver similar results than those obtained using the full set of frequencies. To this end, we began by designing four different classifiers using the training set and the full set of frequencies. The classifiers are the nearest-neighbour technique with Euclidean metric (*NN*) with parameter k (number of neighbours), the *Naïve Bayes classifier* (NB), a *C4.5* decision tree with parameter cp (complexity parameter) and a Random Forest (RF) [5] with parameter nt (number of trees). In order to apply the EFA, a discretization process is needed. Many dimensionality reduction studies use discretization schemes as a way to favor classification tasks (such as [6], [7]). This change of representation does not often result in a significant loss of accuracy (sometimes significantly improves it); it also offers large reductions in learning time. In this sense, and keeping in mind our predictive objective, ^{1}H-MRS data were discretized using the CAIM algorithm [8]. This method was selected for two reasons. It is designed to work with supervised data and does not require the user to define a specific number of intervals for each feature. The EFA is applied to the discretized ^{1}H-MRS data (the training part) to obtain what will be called Best Spectral Subset (BSS). Note that the EFA does not need an inducer. Given the obtained BSS, the four classifiers can then be built in the training set using the original continuous frequencies and evaluated in the test set. In addition, the filtering algorithm RELIEF [9] is used as a comparative reference. This is a feature-weighting algorithm that takes feature interactions into account and yields a set of feature weights that can be sorted in descending order. A cut-point can be established to obtain a feature subset. The Pareto principle states that, for many events, 80% of the effects (viz. classification ability) comes from 20% of the causes (viz. frequencies). Loosely following this principle, we first linearly renormalize the obtained weights so that the smallest weight equals zero and their sum equals one. Then we select the top m features such that their accumulated weight is closest to 0.2.

A feature selection process was also developed *ex novo* in the training set using Forward Selection in wrapper mode using the same classifiers referenced above as evaluation function. The goodness of each subset was again evaluated for each classifier by means of 3-fold cross-validation in the training set. The BSS was the subset that obtained the maximum evaluation.

3.2 Experimental results and discussion

For every feature selection experiment, the size of the corresponding BSS, the accuracy (Acc), the macro-averaged F_1-measure on the test set and its corresponding parameter found in model selection are reported. The results for the

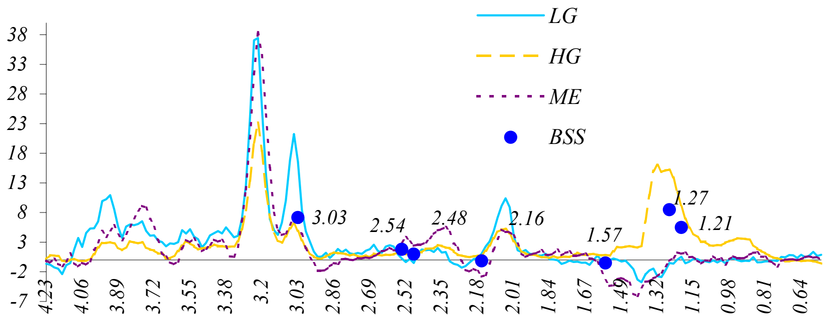

Fig. 1: The 7 spectral frequencies obtained by EFA.

four classifiers using the full set of frequencies are displayed in the first row of Table 1; the results using EFA are displayed in the second row while the results obtained using RELIEF algorithm are presented in the third row. The results using Forward Selection in wrapper mode are presented in Table 2 (left).

Reduction	BSS	NN			NB		C4.5			RF		
method	size	Acc	F_1	k	Acc	F_1	Acc	F_1	cp	Acc	F_1	nt
NR	195	85.0%	81.0%	9	**90.5%**	**88.1%**	**83.0%**	**77.2%**	**0.25**	83.0%	75.4%	9
EFA	7	**94.3%**	**93.1%**	**13**	86.8%	85.1%	79.2%	73.2%	0.25	**85.0%**	**82.6%**	**8**
RELIEF	15	71.7%	61.0%	10	60.4%	52.7%	62.2%	54.6%	0.25	71.7%	64.4%	5

Table 1: Classification performance on the test set (NR = no reduction). k = number of neighbours. cp = complexity parameter. nt = number of trees.

The best result is obtained for the EFA using NN with 7 spectral frequencies: **3.03**, **2.54**, **2.48**, **2.16**, **1.57**, **1.27** and **1.21**, expressed in ppm. The positions in the spectrum of these frequencies are depicted in Fig. 1, shown against average spectra per class. Adding to the interpretability of the results, some of the selected frequencies can be related to known metabolites: 3.03 corresponds to the Creatine peak, a measure of energy status; 2.16, 2.48 and 2.54 are roughly in the area of glutamine-glutamate and lipid/macromolecule summed peaks; 1.57 is located nearby the Alanine peak; and, finally, 1.21 and 1.27 are within the lactate/lipids peak area, which specifically characterizes high-grade malignant tumours. Performance is similar or much better using this subset than using the full set of frequencies, thus providing evidence in favor of a feature selection process. Among the classifiers, NN seems the best alternative in general. In order to ascertain which super-classes are the most difficult to predict, the full confusion matrix of EFA using NN is shown in Table 2 (right).

There exists previous work analizing this ^{1}H-MRS data in similar settings. PCA followed by LDA was used in [10] to distinguish between high-grade malignant tumours and meningiomas, obtaining a mean AUC (area under the ROC curve) of 0.94, using 6 principal components. The same method was used to distinguish between high-grade malignant tumors and astrocytomas Grade II (part of the low-grade gliomas super-class), obtaining a mean AUC of 0.92, also

wrapper	size	Acc	F_1	k/cp
FSS-NN	12	88.7%	87.0%	4
FSS-NB	11	81.1%	78.0%	-
FSS-C4.5	7	73.6%	67.0%	0.25

True	**EFA+13NN**		
Class	**LG**	**HG**	**ME**
LG	7	1	0
HG	1	29	0
ME	0	1	14

Table 2: Left: Test set results using Forward Selection in wrapper mode. Right: Test set confusion matrix for EFA+NN.

using 6 principal components. Note that these are both more limited and less interpretable settings that the one in this paper. Similarly, in [2], LDA with 6 spectral frequencies (3.72, 3.04, 2.31, 2.14, 1.51 and 1.20 ppm: note that some of them are similar to the ones selected by our algorithm) achieved a 83% of correct classification on an independent test set, this time using exactly the same three super-classes that we have analyzed in this study.

4 Conclusions

Several feature selection methods have been applied to a high-dimensional ^{1}H-MRS data set. An entropic algorithm has shown to be able to provide a drastic dimensionality reduction of the problem, while improving on the performance of the full dataset. An added advantage of this method is its simplicity and the absence of any parameter tuning. Comparative results with similar studies show that our solution is competitive both in terms of the prediction accuracy and the parsimony of the selected spectral frequencies, therefore opening a promising research avenue in the problem of brain tumor classification using MRS data.

References

[1] N. A. Sibtain et. al. The clinical value of proton magnetic resonance spectroscopy in adult brain tumours. *Clinical Radiology*, 62:109–119, 2007.

[2] A. R. Tate et. al. Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra. *NMR in Biomedicine*, 19:411–434, 2006.

[3] C. Ladroue. *Pattern Recognition Techniques for the Study of Magnetic Resonance Spectra of Brain Tumours.* PhD thesis, St. George's Hospital Medical School, 2003.

[4] H. Wang. *Towards a unified framework of relevance.* PhD thesis, U. of Ulster, 1996.

[5] B. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[6] M. Ng and L. Chan. Informative gene discovery for cancer classification from microarray expression data. In *IEEE Workshop on Machine Learning for Signal Processing*, pages 393–398. IEEE, 2005.

[7] D. Le and S. Satoh. Robust object detection using fast feature selection from hugh feature sets. In *13th International Conference on Image Processing*, pages 961–964. IEEE, 2006.

[8] L. Kurgan and K. Cios. Caim discretization algorithm. *IEEE Transactions on Knowledge and Data Engineering*, 16(2):145–153, 2004.

[9] K. Kira and L. Rendell. The feature selection problem: Traditional methods and a new algorithm. In *Procs of the Natl. Conf on Artificial Intelligence*, pages 129–134, 1992.

[10] A. Devos. *Quantification and classification of magnetic resonance spectroscopy data and applications to brain tumour recognition.* PhD thesis, Katholieke Univ. Leuven, 2005.