Feature Selection with Single-Layer Perceptrons for a Multicentre ^{1}H-MRS Brain Tumour Database

Enrique Romero[1], Alfredo Vellido[1], and Josep María Sopena[2]

[1] Departament de Llenguatges i Sistemes Informàtics, Universitat Politècnica de Catalunya
{eromero,avellido}@lsi.upc.edu
[2] Laboratori de Neurocomputació, Universitat de Barcelona
jsopena@ub.edu

Abstract. A Feature Selection process with Single-Layer Perceptrons is shown to provide optimum discrimination of an international, multi-centre ^{1}H-MRS database of brain tumors at reasonable computational cost. Results are both intuitively interpretable and very accurate. The method remains simple enough as to allow its easy integration in existing medical decision support systems.

1 Introduction

Diagnostic decision making in brain oncology is, for rather obvious reasons, an extremely sensitive matter. Much of the responsibility of brain tumour diagnostic decisions ultimately rest on the expert clinician's shoulders. Taking into account that most diagnostic techniques have to be non-invasive in this domain, clinicians might benefit from the use of an at least partially automated computer-based medical decision support system (MDSS) that embedded data mining processes. The reluctance of a clinician to seek the support of a computer-based MDSS should not be underestimated, though, as exemplified by the few products of this kind reaching mainstream medical practice [1]. This makes simplicity and robustness compulsory requirements for the employed data analysis methods.

In this study, we analyze Magnetic Resonance Spectroscopy (MRS) brain tumor data from the INTERPRET European research project [2]. The original database contains records corresponding to many brain tumour pathologies (many of which are represented by a very small sample of cases) and even to healthy brain tissue. This makes their computer-based automated classification a non-trivial undertaking that must be carefully designed. Most importantly, the high dimensionality of the data also precludes the straightforward interpretation of the obtained results, limiting their usability in a practical medical context, in which interpretability is paramount.

In the case of MRS data mining, one way to comply simultaneously with the aforementioned simplicity and interpretability requirements is through dimensionality reduction and, more specifically, through feature selection (FS). In this paper a FS procedure with Single-Layer Perceptron (SLP) classifiers that yields very accurate results with a parsimonious subset of interpretable spectral MRS frequencies is described. The method is based on the hypothesis that irrelevant features produce smaller variations in the output values than relevant ones. Hence, a natural way of comparing the

J. Cabestany et al. (Eds.): IWANN 2009, Part I, LNCS 5517, pp. 1013–1020, 2009.

relevance of two features is by comparing the absolute values of the derivatives of the output function with respect to the corresponding input units in the trained model. For SLP, the variation (in absolute value) of the output function is smaller for input features with smaller weights (in absolute value). Therefore, the magnitude of the weight can be considered as an indicator of its importance.

A backward selection technique was used as search procedure: starting from the complete set of available features, several features are removed at every step. The number of features removed at every step is a parameter of the system that controls the granularity of the selection and its computational cost. 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 data set), an aggressive strategy structured in three phases was designed to save computational time: in a first phase, 50% of the features were removed at every step of the backward selection procedure. In a second phase, a 20% of the remaining features were removed. Finally, in a third phase, features were removed one by one.

The experiments with ^{1}H-MRS data reported in the following sections validate the usefulness of the described method, since the results presented are, to the best of our knowledge, the best reported up to date using this database. In addition, most of the selected features have a direct interpretation in terms of metabolites and molecules often mentioned in the MRS literature as descriptors of brain tumor pathologies.

2 Description of the INTERPRET ^{1}H-MRS Brain Tumour Data

The echo time is a determinant parameter in ^{1}H-MRS data acquisition. In short-echo time (SET) spectra some metabolites are better resolved (e.g. lipids, myo-inositol, glutamine and glutamate). However, there may be numerous overlapping resonances making the spectra difficult to interpret [3]. The use of a long-echo time (LET) in the acquisition of spectra yields less clearly resolved metabolites but also less baseline distortion, resulting in a more readable spectrum. LET data also allow a more reliable analysis and testing of classification methods [4].

In this study we consider three different data sets of single voxel ^{1}H-MR spectra, acquired *in vivo* from brain tumor patients. The first was acquired at SET, the second at LET and, finally, the third is a fusion of the two previous ones, where the spectra acquired at two echo times obtained for the same patient (when both echo times were available) are combined through straight concatenation. The clinically-relevant regions of both the SET and LET spectra were sampled to obtain 195 frequency intensity values.

Class (tumour pathology type) 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 experiments, spectra were bundled into three groups, namely: *G1*: low grade gliomas (astrocytomas grade II, oligoastrocytomas grade II and oligodendrogliomas grade II); *G2*: high grade malignant tumors (metastases and glioblastomas); and *G3*: meningiomas. In summary, three data sets were analyzed: SET (217 cases and 195 features), LET (195 cases and 195 features) and SET+LET (195 cases and 390 features).

3 Some Precedents of Feature Selection and Classification with INTERPRET ^{1}H-MRS Brain Tumour Database

Previous published work analyzing similar ^{1}H-MRS INTERPRET data used feature extraction (PCA, specifically) 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 6 principal components [5]. 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 6 principal components. Independent Component Analysis (ICA), an alternative feature extraction method, has also been applied to analyze an earlier version of the INTERPRET data in [6].

This type of binary classification, in different combinations, has been carried out with different versions of INTERPRET ^{1}H-MRS data and with varying degrees of success (see, for instance, the combination of spectrometric and imaging data in [7]). This is a somehow easier problem than the multi-class one that we deal with in this paper. In [8], this time using exactly the same three groups of tumours that we have analyzed in this study, a basic linear model (LDA) with 6 spectral frequencies (3.72, 3.04, 2.31, 2.14, 1.51 and 1.20 ppm) achieved a 83% of correct classification on an independent test set for LET, and a 89% correct classification for SET, using 5 frequencies (namely 3.76, 3.57, 3.02, 2.35 and 1.28 ppm). Similar results were found for LET data in [4] for a combination of PCA and LDA and for different versions of support vector machines.

Only a few recent works have addressed the problem of multi-class classification of ^{1}H-MRS data by combination of echo times. In [3], SET and LET data were combined in a classification problem that involved four groups of tumours: *high grade malignant tumours* and *meningiomas*, as in the current study, and astrocytomas grade II (part of our *low grade gliomas*) and anaplastic astrocytomas. Using LDA, the correct diagnosis was suggested by at least one of the echo times in 90% of all cases. A far more similar approach to the one followed in the current study (combining the SET and LET data through concatenation) was used in [9], also to classify *G1*, *G2* and *G3*: feature selection followed by LDA achieved 88.71% test accuracy, while PCA followed by LDA achieved a maximum of just over 90% test accuracy for 8 principal components.

4 Feature Selection Process

The problem of FS can be defined as follows: given a set of d features, select a subset that performs the best under certain evaluation measure. From a computational point of view, the definition of FS usually leads to a search problem in a space of 2^d elements. In this case, two components must be specified: the feature subsets evaluation measure and the search procedure through the space of feature subsets. If any of these two componentes depends on an external model, it must also be specified.

In the rest of the section, the constituent elements of the FS process for the ^{1}H-MRS brain tumour database are described.

4.1 Model

SLP Artificial Neural Networks (ANN) with sigmoidal output units were used both for the feature subsets evaluation measure (within the FS process) and to obtain the

test accuracy (within the learning process). The number of output units was set to the number of classes of the problem. Therefore, the activation y_j of the output unit j for a d-dimensional input vector x is computed as

$$y_j = g\left(\sum_{k=1}^{d} x_k \cdot \omega_{jk} + b_j\right), \tag{1}$$

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

There are several reasons for using SLP 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 ^{1}H-MRS brain tumour data set [10]. In addition, MLP parameters are more difficult to adjust. Alternatively, FS with linear Support Vector Machines (SVM) [11] 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 SVM. Therefore, the saliency of every feature is likely be more independent for SLP than for linear SVM. In addition, linear models had shown quite good performance with these data in previous studies [8].

4.2 Feature Subsets Evaluation Measure

The evaluation measure (the relevance) of a feature subset was computed as the sum of the individual saliencies of its features. The saliency s_i of a feature i over O outputs was computed as: $s_i = \sum_{j=1}^{O} |\hat{\omega}_{ji}|$, 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 absolutes 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

$$\frac{\partial y_j}{\partial x_i} = g'\left(\sum_{k=1}^{d} x_k \cdot \hat{\omega}_{jk} + b_j\right) \cdot \hat{\omega}_{ji},$$

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 SLP, the magnitude of the weights corresponding to a feature is considered as an indicator of its importance. Similar ideas can be found elsewhere (see, for example, [11] or [12]).

4.3 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 removed at every step of the algorithm according to the evaluation measure. Since the evaluation measure of a feature subset is computed as the sum of the saliencies of its features, the features to be removed at every step are those with the smallest saliency. The number of features removed at every step is a parameter of the system, that controls the granularity of the selection and the computational cost. The main reason for choosing a backward procedure instead of a forward one is twofold: First, a backward procedure allows, at the onset, to take into account all the interactions among variables. Second, the parameters of the SLP are easier to adjust.

4.4 Algorithm

The FS algorithm finally applied in this study consisted of three general phases:

1. Perform a backward selection procedure (see section 4.3), starting with the whole set of features, until only one feature remain. At every step:
   (a) Train a SLP with the remaining features (see section 4.1).
   (b) Compute the saliency of every feature (see section 4.2).
   (c) Remove the 50% of the remaining features with the lowest saliency.
   For every feature subset obtained, estimate its generalization performance. Out of all the results, keep the previous to the best one for the next phase (to avoid missing a possible generalization maximum in intermediate, not analyzed, subsets).
2. The second phase is similar to the first one, except for:
   (a) The initial feature subset is the one obtained in the first phase.
   (b) At every step, 20% of the remaining features are removed.
3. The third phase is similar to the second one, except for:
   (a) The initial feature subset is the one obtained in the second phase.
   (b) At every step, one feature is removed.

5 Experiments

5.1 Experimental Setting

No preprocessing of the data was done, since all the features were in the same range of values. The target values were generated with a 1-of-C coding scheme (a value of 1 for the correct class, an 0 for all the others).

The SLP models were trained with the Delta rule in on-line mode for $10,000$ epochs. The logistic function $g(z) = \frac{1}{1+e^{-z}}$ was used for the output units. Initial weights and momentum were set to 0. Learning rates were heuristically adjusted.

The saliencies were calculated using the whole data set, and five runs of a 5-fold stratified Cross-Validation (CV) were performed to estimate the generalization performance. Prior to every CV, the data were randomly shuffled. The complete experiment took around 7 hours in an Intel Xeon CPU at 2,000 MHz.

Table 1. Classification and FS results

Data Set	Test	NF	Features Selected (ppm)
SET	95.72%	29	3.13 1.51 4.15 1.65 3.74 3.60 3.51 1.30 1.82 3.45
			2.31 2.22 2.27 3.32 3.77 1.87 0.94 3.81 3.24 2.29
			2.03 1.97 1.47 1.63 3.56 3.93 2.23 1.59 3.34
SET	90.51%	18	1.51 3.13 2.22 2.27 3.56 3.60 3.45 3.51 1.87 2.29
			2.03 1.47 1.59 1.97 3.74 4.15 3.77 1.30

Data Set	Test	NF	Features Selected (ppm)
LET	95.79%	50	1.23 2.16 2.27 2.33 3.36 0.87 3.72 1.42 2.99 2.39
			3.09 3.05 1.76 0.54 2.88 1.53 1.44 2.52 2.54 2.56
			3.70 3.64 3.32 1.32 3.91 2.94 1.06 3.20 3.55 3.85
			1.59 3.53 3.18 3.79 1.04 0.64 3.26 2.48 0.73 1.27
			1.51 3.37 3.45 3.03 1.84 4.19 1.91 3.39 3.94 0.89
LET	90.26%	8	2.27 2.16 1.23 2.99 2.88 0.87 3.72 1.76

Data Set	Test	NF	Features Selected (ppm)
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

5.2 Results

Classification and FS results of the experiments for the three data sets studied are summarized in Table 1 (SET: top, LET: middle, and SET+LET: bottom). For the combination of echo times, SET+LET, up to a 98.46% average test accuracy (an average of only 3 misclassifications out of 195 spectra) was achieved using a parsimonious subset of only 18 spectral frequencies, 8 of them belonging to SET and 10 to LET. The results reported in the second rows of Table 1 illustrate how the test accuracy deteriorates as we detract features from the selected subset. Nevertheless, for the SET+LET data set, a selection of 9 spectral frequencies (a mere 2.3% of the original 390) was still able to retain over 91% of the average test classification accuracy in the multi-class problem involving *low grade gliomas*, *high grade malignant tumors*, and *meningiomas*.

For illustration, the 18 selected spectral frequencies are displayed, together with the mean spectra of the three tumour classes, in Figure 1, both for SET (top) and LET (bottom) data separately. Many of them have a clear interpretation as resonances of metabolites and molecules often reported in the MRS literature as descriptors of brain tumor pathologies. They include, at SET: Alanine (^{2}CH-group) at 3.77 ppm, Glycine ($^2CH_2$-group) or Myo-Inositol at 3.58, possibly Taurine at 3.45, Choline and other trimethylamine-containing compounds at 3.18, Glutamate and Glutamine ($^3CH_2$- and $^4CH_2$- groups) at 2.12, Lactate and Lipids at 1.32 and 1.25, and also lipids at 0.89. They also include, at LET: Glutamate/Glutamine-containing compounds (^{2}CH-groups) at 3.79 and, possibly, 3.70, Glutamate and Glutamine metabolites (this time $^4CH_2$-groups) at 2.54 and 2.27, and N-acetylaspartate and other N-acetyl-containing compounds at 2.03. All this metabolic information should provide medical experts with intuitive insights on the diagnoses of the analyzed brain tumour pathologies.

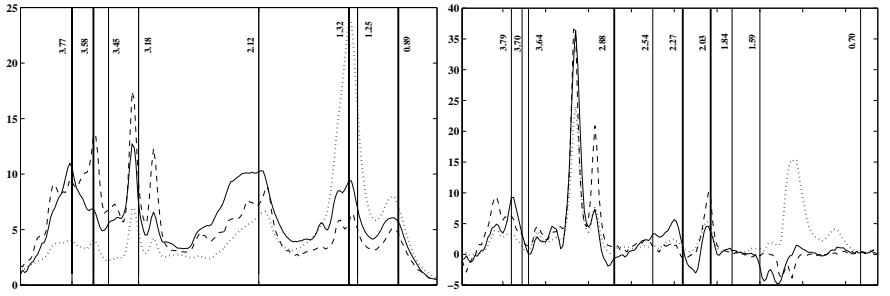

Fig. 1. Representation of the 18 selected spectral frequencies from the SET+LET data set as vertical lines, with their value in ppm tagged by their side. Left plot: SET results; right plot: LET results. The 3 most relevant frequencies of each echo time are represented with thicker lines. Mean spectra of each class are represented as dashed lines (low grade gliomas), dotted lines (high grade malignant tumors), and solid lines (meningiomas).

The use of a single echo time yields a 95.72% average test accuracy (an average of between 8 and 9 misclassifications out of 195 spectra) for SET using 29 spectral frequencies and a very similar 95.79% for LET, this time using a less compact subset of 50 selected frequencies. According to these results, a classification based on SET only might be preferred to one based on LET only (on the grounds of easier interpretation), although LET reaches a rather good compromise between accuracy and interpretability with a 90.26% result using only 8 frequencies. In any case, according to the results reported in Table 1, models using SET+LET instead of a single echo time should always be the choice in order to achieve optimal generalization.

Furthermore, our results using SET+LET to discriminate between *low grade gliomas*, *high grade malignant tumors*, and *meningiomas* with a combination of FS and SLP improve on those reported in [9] for the same problem (using stepwise FS with LDA), by almost a 10% average test accuracy. Our results using only SET and LET also improve on those previously reported for the same problem in [8,4,3].

6 Conclusions and Future Work

A simple SLP with FS has achieved a near perfect classification (measured by test accuracy) of an international, multi-centre ^{1}H-MRS data set by combining data acquired at two different echo times. This performance is better than that achieved using data obtained at any of the echo times separately, reinforcing previous results [9]. In future research, out-of-sample data should be acquired in order to fully ensure the generalization capabilities of the model. Importantly, these good results are obtained while retaining model simplicity and interpretability, as they only require less than 5% of the original frequencies (most of them identifiable as known descriptors of brain tumor pathologies).

The FS procedure was designed with an aggressive strategy that allowed to save computational time, making it feasible for data sets with a large number of variables. However, for highly correlated features the values of the weights (and therefore the

saliency) may be distributed in the trained SLP, leading to unsuitable FS. In this case, the FS process would benefit from a previous filter procedure where highly correlated features were considered as a single one.

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, Dr. M. Julià-Sapé, Ms. T. Delgado, Ms. J. Martín, Mr. I. Olier and Mr. A. Pérez (all from GABRMN-UAB). GABRMN-UAB coordinator: Prof. C. Arús. The authors acknowledge funding from M.E.C. research project TIN2006-08114. We also acknowledge the use of the UPC MA1 Department computing cluster system (http://www.ma1.upc.edu/eixam/index.html).

References

1. Vellido, A., Lisboa, P.J.G.: Neural networks and other machine learning methods in cancer research. In: Sandoval, F., Prieto, A.G., Cabestany, J., Graña, M. (eds.) IWANN 2007. LNCS, vol. 4507, pp. 964–971. Springer, Heidelberg (2007)
2. INTERPRET: International Network for Pattern Recognition of Tumours Using Magnetic Resonance project, `http://azizu.uab.es/INTERPRET`
3. Majós, C., Julià-Sapé, M., Alonso, J., Serrallonga, M., Aguilera, C., Acebes, J.J., Arús, C., Gili, J.: Brain tumor classification by proton MR spectroscopy: Comparison of diagnostic accuracy at short and long TE. American Journal of Neuroradiology 25, 1696–1704 (2004)
4. Lukas, L., et al.: Brain tumor classification based on long echo proton MRS signals. Artificial Intelligence in Medicine 31, 73–89 (2004)
5. Devos, A.: Quantification and Classification of MRS Data and Applications to Brain Tumour Recognition, PhD thesis, Katholieke Univ., Leuven, Belgium (2005)
6. Huang, Y., Lisboa, P.J.G., El-Deredy, W.: Tumour grading from Magnetic Resonance Spectroscopy: A comparison of feature extraction with variable selection. Statistics in Medicine 22, 147–164 (2003)
7. Luts, J., Heerschap, A., Suykens, J.A.K., Van Huffel, S.: A combined MRI and MRSI based multiclass system for brain tumour recognition using LS-SVMs with class probabilities and feature selection. Artificial Intelligence in Medicine 40, 87–102 (2007)
8. Tate, A., 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)
9. García-Gómez, J.M., Tortajada, S., Vidal, C., Julià-Sapé, M., Luts, J., Moreno-Torres, A., Van Huffel, S., Arús, C., Robles, M.: The influence of combining two echo times in automatic brain tumor classification by Magnetic Resonance Spectroscopy. NMR in Biomedicine (2008) (accepted for publication)
10. Romero, E., Sopena, J.M.: Performing feature selection with Multi-Layer Perceptrons. IEEE Transactions on Neural Networks 19, 431–441 (2008)
11. Guyon, I., Weston, J., Barnhill, S., Vapnik, V.N.: Gene selection for cancer classification using support vector machines. Machine Learning 46, 389–422 (2002)
12. Steppe, J.M., Bauer, K.W.: Feature saliency measures. Computer & Mathematics with Applications 33, 109–126 (1997)