Classifier Combination for *In Vivo* Magnetic Resonance Spectra of Brain Tumours

Julià Minguillón[1], Anne Rosemary Tate[2,4], Carles Arús[3], and John R. Griffiths[2]

[1] Unitat de Combinatòria i Comunicació Digital, Escola Tècnica i Superior d'Enginyeries, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain
jminguillon@ccd.uab.es
[2] CRC Biomedical MR Research Group, St George's Hospital Medical School, University of London, Cranmer Terrace, London, SW17 0RE, UK
[3] Departament de Bioquímica i Biologia Molecular, Universitat Autònoma de Barcelona, 08193 Bellaterra, Spain
[4] School of Cognitive and Computing Sciences, University of Sussex, Falmer, Brighton, UK.

Abstract. In this paper we present a multi-stage classifier for magnetic resonance spectra of human brain tumours which is being developed as part of a decision support system for radiologists. The basic idea is to decompose a complex classification scheme into a sequence of classifiers, each specialising in different classes of tumours and trying to reproduce part of the WHO classification hierarchy. Each stage uses a particular set of classification features, which are selected using a combination of classical statistical analysis, splitting performance and previous knowledge. Classifiers with different behaviour are combined using a simple voting scheme in order to extract different error patterns: LDA, decision trees and the k-NN classifier. A special label named "unknown" is used when the outcomes of the different classifiers disagree. Cascading is also used to incorporate class distances computed using LDA into decision trees. Both cascading and voting are effective tools to improve classification accuracy. Experiments also show that it is possible to extract useful information from the classification process itself in order to help users (clinicians and radiologists) to make more accurate predictions and reduce the number of possible classification mistakes.

1 Introduction

^{1}H Magnetic Resonance Spectroscopy (MRS) [1] is attracting much attention for non-invasive diagnosis of brain tumours. These tumours currently present a difficult clinical problem: the oncologist needs to know the type of cell from which the cancer originates, as well as the "grade", or degree of malignancy, before choosing appropriate therapy. Some benign tumours respond well to surgery, whereas more aggressive types that are essentially incurable may respond temporarily to palliative treatment. Radiological examination by MRI does not usually give

F. Roli and J. Kittler (Eds.): MCS 2002, LNCS 2364, pp. 282–292, 2002.

conclusive diagnoses, and an incorrect diagnosis could result in the patient failing to receive life-saving treatment. Consequently, the current "gold standard" is stereotactic biopsy followed by histopathology. Biopsy of the brain is expensive, unpleasant for the patient and sometimes has severe side effects, with even occasional deaths. There is thus much interest in MRS, which is a totally non-invasive method - nothing is injected or biopsied. ^{1}H MRS of brain tumours can be performed with many hospital MRI instruments, after slight modification. It gives a spectrum in which the peaks represent signals from hydrogen atoms in chemicals within the tumour. Different tumour types contain characteristic patterns of chemicals, and there are also patterns associated with greater or lesser degrees of malignancy. However, visual interpretation of these spectra is difficult, with many ambiguous cases, and few doctors are trained in it. Consequently, clinical MRS is little used at present, and there have been many attempts to develop automated classification procedures. Hitherto, these have only worked with artificial datasets in which the spectra are drawn from a few well-characterised tumour types [2, 3, 4]. Our study is the first, to our knowledge, to tackle the "real world" problem in which an unknown brain tumour can represent any possible tumour type or grade, and to make classifications according to the standard WHO categories.

There are some fundamental problems when developing an automatic procedure for classifying brain tumour spectra. There is a long list of tumour types [5], some of which are very rare. In addition, some diagnostic criteria are of fundamental importance (e.g. "is it benign or malignant?") whereas others are of merely academic significance. Furthermore, some spectra are less satisfactory than others, either for technical reasons or because the tumour itself contains areas of cyst or haemorrhage. Developing a classifier that can take a spectrum from any undiagnosed tumour and assign it unequivocally to the appropriate class, may therefore not be possible. But this is not necessarily an important goal as in most cases there will be much useful evidence from factors such as the clinical presentation or the anatomical MRI that narrow down the diagnostic possibilities. We have therefore approached this as a multi-stage problem. Ideally, the system would: reproduce the WHO classification grading structure: perform well when the number of samples is low; use previous knowledge about the problem; be robust when the training data might contain errors, since the "gold standard" pathology classification is not always 100% accurate; and finally, help the users to extract relevant information from the classifier, rather than provide (possibly more accurate) "black box" classifiers that they cannot understand. In addition we need a method for selecting the best points or regions of the spectra for classification, since an MR spectrum is a vector of between 512 and 4096 spectral intensities.

Decision trees [6] allow us to build classifiers that partially fulfil all these requirements. Previous experiments [7] with MR spectra show that different classifiers make different mistakes. This can be exploited using a simple voting scheme which labels as "unknown" those samples where different classifiers disagree. An advantage of combining several classifiers in a multi-stage scheme

is that different features (i.e. points or regions of the spectra) may be used at different levels. Several classifiers may be used to establish a minimum threshold to ensure class, or "unknown" and, when two classifiers disagree, this fact may be used to find (or indicate the possibility of) mixtures of two or more classes.

This paper is organised as follows: Section 2 describes the structure of the tumour classification problem and the available data sets and pre-processing. Section 3 describes the classifiers used and the multi-stage scheme. Section 4 describes the experiments and Section 5 summarises the conclusions and proposes future work.

2 Classification of MR Spectra

2.1 Data

The spectra were acquired at three clinical centres: Institut de Diagnòstic per la Imatge (IDI) Bellvitge, Spain, Centre Diagnostic Pedralbes (CDP) Barcelona, Spain, and St. George's Hospital (SGH), London, U.K. One short echo ^{1}H (20 or 30 ms) spectrum was acquired for each patient. Prior to entering the spectra into the analysis, strict quality control and validation procedures were applied to all the data. Following biopsy, the pathology slides for each case was examined by a panel of neuro-pathologists to provide a consensus diagnosis (see http://carbon.uab.es/INTERPRET/cdap.html). Only those tumour classes for which we had at least 4 representatives were used. This resulted in the following classes of spectra: 81 astrocytomas (18 grade II, 6 grade III, and 57 grade IV), 32 metastases, 37 meningiomas, 6 oligodendrogliomas, 6 lymphomas, 5 primitive neuroectodermal tumour (pnets), 4 schwannomas, 4 haemangioblastomas and 14 samples from normal volunteers. Figure 1 shows a plot of a typical spectrum.

2.2 Pre-processing

All spectral processing (from raw signal to processed spectrum) was carried out automatically using a set of stand-alone programs developed (in C) for the decision support tool. The intensities in frequency region known to represent the major peaks was extracted from each spectrum and the resulting vector (of 512 intensity values) was then normalised to have norm $L_2 = 1$. We do not use all 512 points, because only those in range [0.5, 4.2] ppm[1] are considered to have relevant information for tumour classification, giving a total of 195 variables, from v_{151} (4.2 ppm) to v_{345} (0.5 ppm).

3 Combining Classifiers

Combining classifiers with different bias-variance decomposition behaviour, can reduce both bias and variance and thus improve classification error [8]. In this

[1] the ppm scale defines the positions of the peaks with respect to a predefined reference.

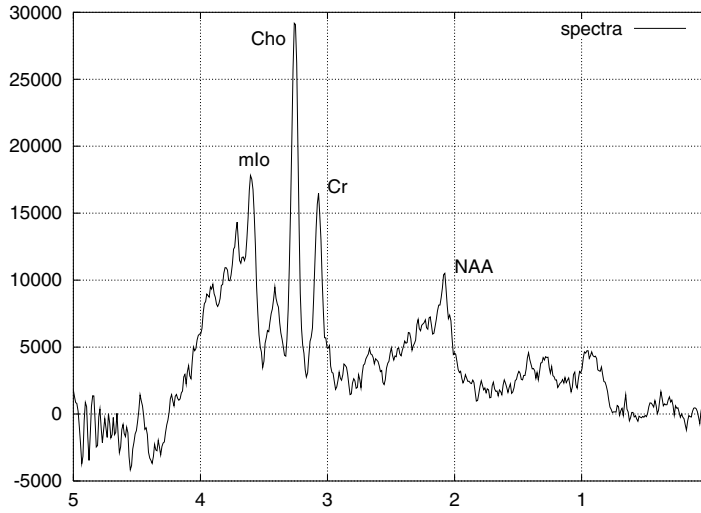

Fig. 1. A typical spectrum from a grade II astrocytoma, showing the position of the peaks representing myoinositol (mIo), choline (Cho), creatine (Cr) and N-acetylaspartate (NAA).

paper we describe how we build a sequence of classifiers each specialising in a concrete problem, and combine them in a way that allows the extraction of different information about the samples being classified. We decided to develop a multi-stage classification system that reproduces the hierarchical structure of the WHO tumour classification that is used by clinicians. Three different types of classifiers are used: linear discriminant analysis (LDA) [9], decision trees [6] and the nearest neighbour classifier (k-NN) [10]. These classifiers were chosen because LDA is simple and interpretable (it is easy to show results and reasons), and good for small sample set sizes. Nevertheless, it may not be very accurate depending on the complexity of the boundaries defined by the different tumour classes and whether or not these overlap, or are mixtures. Decision trees are also simple and interpretable (when using orthogonal splits). They have a good generalisation performance and may include *a priori* knowledge about the problem being solved. Finally, the k-NN classifier is very simple and fast when the number of samples and data dimensionality are small. No training is needed, and as the number of samples is small, no special techniques for reducing nearest neighbour cost are required. Experiments show that the optimal value for k is very sensitive to the number of samples available for each class, so we try several values for k choosing the smallest k yielding a good generalisation error. Other methods (support vector machines or neural networks, for example) generally require more samples than we have available or do not allow the users to interpret results easily.

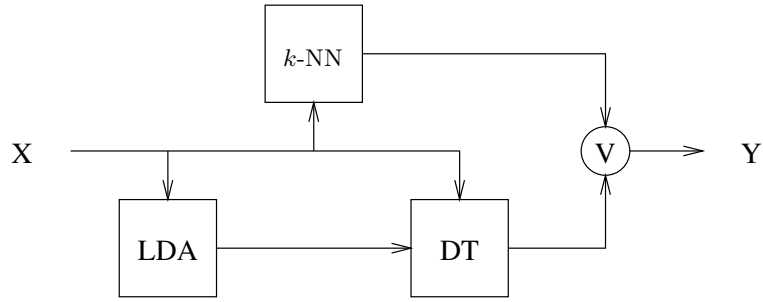

Fig. 2. Cascading and voting architecture for each classification stage. The LDA classifier and the decision tree (DT) are cascaded and then combined with the k-NN classifier using a simple voting scheme (V).

At each stage, two classifiers are built: a k-NN classifier and a decision tree. The results of both classifiers are combined using a voting scheme with a very simple rule: if both outcomes agree, the label will be the outcome of the voting scheme, otherwise, generate "unknown". The basic idea of combining k-NN and decision trees is to exploit their different behaviour to improve classification performance. Since they try to solve the same problem using different approaches, they may make different mistakes. When enough data is available, LDA is used with decision trees as the first stage of a cascading ensemble [11]; the latter uses the class distances and transformed points computed by LDA. Figure 2 shows the combined cascading and voting scheme used at each stage of the classification system. Cascading can be thought of as the process of asking a sequence of experts to give a decision. The cheapest (simplest) expert is consulted first and then the information that it provides is passed to the next expert in the sequence, and so on. The second expert decides whether to use this additional information or not.

3.1 Feature Selection

It is known that different points from the spectra provide plausible biochemical explanations for discriminating between tumour classes. We use this "prior" knowledge to select the points used as classification features. Trying to build a classification system using LDA or the k-NN classifier using the 195 input variables is pointless: data dimensionality is too high and the number of samples for several classes is too small. It is important to include only those variables which are relevant to the classification problem. A classical correlation analysis may be used to find the points with higher discriminating properties. A decision tree can also be used to find other classification variables that remain hidden using correlation analysis: at each stage we take the data set containing the samples from the classes being classified, and a limited depth decision tree (three or four levels is enough for such small data sets) is built without any pruning. For each

split we rank classification features according to its splitting performance, and (depending on the stage and data set being considered), the first or the two first classification features are selected. Then, all these variables that have statistical significance for classification purposes are checked with an expert spectroscopist and those that cannot represent known metabolites are discarded. This approach has been successfully used previously [4, 7].

3.2 The "Unknown" Class

As described in the previous section, when the two classifiers of the same stage disagree, the outcome of the classifier is "unknown". This value may also be used to label those predictions made by a decision tree with a small margin. This margin is defined for each leaf as the probability of making a right prediction minus the probability of making a mistake. Therefore, the new labelling rule for a leaf i is

$$l'_i(t) = \begin{cases} l_i(t) & \text{if } P\{t_i(x) = y\} - P\{t_i(x) \neq y\} > \epsilon \\ \text{"unknown"} & \text{otherwise.} \end{cases}$$

where $t_i(x)$ is the computed label using majority voting. This allows us to discard those samples that fall in leaves which contain elements from several classes. The value for ϵ depends on the number of classes and it is determined empirically. A similar approach has been successfully used in [12].

4 Experimental Results

The classification system was split into four stages: 1) discriminates between tumour and normal samples, 2) tries to classify tumour samples as benign or malignant, 3) tries to separate malignant tumours according to their malignancy grade and in 4) several classifiers are built for each malignancy grade in order to discriminate between WHO tumour classes. Each experiment was carried as follows: at each stage, the original data set containing all samples related to such stage is split using N-fold cross-validation (NFCV) with $N = 10$ following the recommendations of [9] for small size data sets. This process is repeated five times and all results are averaged, resulting in a total of 50 experiments at each stage. Decision trees were built using entropy as the splitting criterion and pruned back using tree size and misclassification error as complexity measures.

4.1 Stage 1: Tumour Vs. Normal

Separating tumours from normal samples is the easiest stage. It is well known that, unless the voxel (volume from which the spectra is acquired) has been placed in a region where an aggressive infiltrating tumour is mixed with brain parenquima, tumour samples have little or no N-Acetylaspartate (NAA) which is shown by a peak at 2.0 ppm. In addition, tumours have much higher choline

levels (shown by a peak at 3.22 ppm). When a decision tree was used for feature selection, two regions were detected as the most important for separating the two groups: the NAA peak and the Creatine peak (around 3.03 ppm). Thus, we selected variables v_{200} (Choline), v_{210} (Creatine), and v_{263} (NAA). Using a k-NN classifier with $k = 1$, a 99.3% classification accuracy is achieved, showing that this first stage can be fully automated with almost no intervention by the radiologist. Two of the samples were misclassified: S1 and S2. When the spectra were visually inspected it appeared that this may have been due to the fact that the voxel from which the spectra were obtained included a large proportion of normal brain parenquima (due to voxel mispositioning). A decision tree was also built using the same points. In this case, classification accuracy is 99.2%, and two samples are misclassified: S1 again, and a normal sample misclassified as tumor (S3). Decision trees always use v_{263} to separate tumors from normal samples. If we combine the outcomes of the two classifiers, only one sample remains unclassified, S1. For the other two samples, the generated label is "unknown", as both outcomes disagree. This fact can be used to alert the spectroscopist to place the voxel in another position.

4.2 Stage 2: Malignant Vs. Benign Tumours

All samples labelled as tumour by the previous classifier are used to build a new classifier which tries to determine whether a tumour is malignant or not. Meningiomas and schwannomas are benign tumors, the rest are malignant or have malignant potential. We also have several tumors which are in the borderline (haemangioblastomas), and it would be useful to identify them. However, there are only four samples of these. Since haemangioblastomas are considered benign tumors, but with uncertain malignant potential, we decided to treat them as benign and delay final classification to an optional third stage. Six points were used as classification features: v_{244} (2.36 ppm, glutamate, glutamine and macromolecules), v_{172} (3.74 ppm, glutamate, glutamine, alanine), v_{210} (2.98 ppm, creatine), v_{191} (3.38 ppm), v_{159} (3.99 ppm) and v_{304} (1.22 ppm, lipid/lactate). We tested several values for k, and $k = 3$ produced the best classification performance. LDA and decision trees were tested alone but also in a cascading ensemble. Table 1 shows the results for this stage. Notice that cascading reduces decision tree misclassification error noticeably.

4.3 Stage 3: Malignancy Grade

The third stage tries to establish the malignancy grade for those samples classified as malignant tumours by the previous stage. We do not try to separate benign tumours into meningiomas, schwannomas and haemangioblastomas because we only have 4 samples of the two last tumour classes as compared with the 37 meningiomas, and the results we obtain are completely biased towards accurately classifying meningiomas, even if we force equal *a priori* probabilities. This will be accomplished when new samples are available. Determining the degree of malignancy is important, because it gives an indication of the

Table 1. Left table: classification accuracy for each classifier, the cascading ensemble and the resulting voting scheme. P is the percentage of samples not classified as "unknown", and E is the misclassification error. Right table: confusion matrix for the voting scheme. β is the percentage of tumours classified as a class that really belong to such class.

classifier	P	E
LDA	—	0.1682
k-NN ($k = 3$)	—	0.1235
Tree	—	0.0894
LDA + Tree	—	0.0788
Voting scheme	89.9%	0.0563

class	benign	malignant	unknown	*total*
benign	139	31	48	218
malignant	12	582	38	632
total	151	163	86	850
β	92.1%	94.9%	—	—

patient outcome, and may determine the treatment prescribed. We use three malignancy grades commonly accepted, labelled low, medium and high. Low grade (WHO grade II) consists of low-grade astrocytomas, oligo-astrocytomas and oligo-dendrogliomas. Medium grade (WHO grade III) consists of astrocytomas (and anaplastic oligoastrocytoma, but we do not have enough samples to include them in our experiments). Finally, high grade (WHO grade IV) includes metastasis, glioblastomas, pnets and lymphomas. Results are shown in Table 2. The variables selected for this stage are v_{181}, v_{201}, v_{309}, v_{317}, v_{264} and v_{197}.

Table 2. Left table: classification accuracy for each classifier, the cascading ensemble and the resulting voting scheme (no cascading). Right table: confusion matrix for the voting scheme (no cascading).

classifier	P	E
LDA	—	0.2154
k-NN ($k = 5$)	—	0.1815
Tree	—	0.1108
LDA + Tree	—	0.1231
Voting scheme	83.8%	0.0679

class	low	medium	high	unknown	*total*
low	57	0	8	55	120
medium	4	0	17	9	30
high	8	0	451	41	500
total	69	0	476	105	650
β	82.6%	—	94.7%	—	—

Notice that in this case cascading does not improve tree performance. The reason is that we only have 6 samples for medium grade tumours and LDA performs poorly. Furthermore, those samples do not form a cluster, so k-NN makes also a lot of mistakes, and therefore, not any sample is labeled as medium grade. Because grade III is an intermediate stage, classifying medium grade tumours is often a problem even for the pathologists.

4.4 Stage 4: Tumour Class

The last stage of our classifier consists of two different classifiers. The first one tries to separate low-grade tumours into oligos (both oligo-astrocytomas and oligo-dendrogliomas) and astrocytomas. The second classifier tries to separate

high-grade tumours in primary tumours and metastasis, and then primary tumours in glioblastomas, lymphomas and pnets.

Low grade malignant tumours A very simple classifier for separating low-grade astrocytomas and oligodendrogliomas was built using v_{227}, v_{172} and v_{181}. Table 3 shows the results for this classifier. Using only these three points, results are very good. Cascading drops misclassification error of the decision tree from 0.12 to 0.08, even with a high misclassification error of the LDA classifier.

Table 3. Left table: Classification accuracy for each classifier, the cascading ensemble and the resulting voting scheme. Right table: confusion matrix for the voting scheme.

classifier	*P*	*E*
LDA	—	0.21
k-NN ($k = 3$)	—	0.16
Tree	—	0.12
LDA + Tree	—	0.08
Voting scheme	84.0%	0.0476

class	astro	oligo	unknown	*total*
astro	64	0	8	72
oligo	4	16	8	28
total	68	16	16	100
β	94.1%	100.0%	—	—

High grade malignant tumours This is probably the most difficult question nowadays related to tumour classification, since this is the most common tumour group, and the different tumour types within it are those the radiologists most easily confuse when using MRI. This classifier is in fact a two-stage classifier: the first one tries to separate primary tumours from the rest (metastasis). The second one tries to identify each one of the primary tumour classes (glioblastomas, lymphomas and pnets). Due to the lack of space, we only show results for the first classifier, which is in fact the hardest problem to solve. Furthermore, we only have a few lymphomas and pnets, so our results are biased towards glioblastomas. We used v_{317}, v_{304}, v_{242}, v_{236}, v_{215} and v_{220}. Table 4 shows the results for this classifier. We decided to include the cascading ensemble into the voting scheme results because β values are more balanced. Notice that these results are the worst, as they correspond to the hardest problem we try to solve in this paper.

Table 4. Left table: Classification accuracy for each classifier, the cascading ensemble and the resulting voting scheme. Right table: confusion matrix for the voting scheme.

classifier	*P*	*E*
LDA	—	0.392
k-NN ($k = 5$)	—	0.326
Tree	—	0.188
LDA + Tree	—	0.218
Voting scheme	75.0%	0.1733

class	primary	secondary	unknown	*total*
primary	278	3	59	340
secondary	62	32	66	160
total	340	35	125	500
β	81.8%	91.4%	—	—

5 Conclusions and Future Work

In this paper we have presented a multi-stage classifier for classification of ^{1}H MR spectra from brain samples. Our goal was to build a classification system which may help clinicians to take decisions and learn from the classification process itself. Several conclusions may be drawn:

- The inherent hierarchical structure of the tumour classification problem is well described using a sequential combination of classifiers.
- Each stage uses its own set of classification features reducing classification cost and learning algorithm resilience.
- When the number of samples for each class is large enough, cascading improves decision tree performance, using LDA as a first classifier.
- Combining several classifiers with different bias-variance behaviour under a voting scheme allows us to have partial classification and a more robust classification system.

Further work is in progress to improve the classification results, but also to learn more about the classification process itself: which tumor classes are misclassified more often, which stage is more critical, system response to rare tumors, and so. Cascading not only at each stage but also between stages is also an interesting subject of study. A completely independent test set is being prepared to test the performance of the classification path developed, so we will be able to check our classification system in a real scenario.

Acknowledgements

This paper was partially supported by Spanish government grant TIC2000-0739-C04-01, and EU grant INTERPRET IST-199-10310. We thank the IDI, CDP and SGH centres for providing the validated data set used in this paper.

References

[1] Danielsen, E.R., Ross, B.: MRS Diagnosis of Neurological Diseases. Marcel Dekker, Inc, NY (1999)
[2] Preul, M.C., et al.: Accurate, noninvasive diagnosis of human brain tumors by using proton magnetic resonance spectroscopy. Nature Medicine **2** (1996) 323–325
[3] Hagberg, G., et al.: *In vivo* proton MR spectroscopy of human gliomas: Definition of metabolic coordinates for multi-dimensional classification. Magnetic Resonance in Medicine **34** (1995) 242–252
[4] Tate, A.R., et al.: Towards a method for automated classification of 1H MRS spectra from brain tumours. NMR in Biomedicine **11** (1998) 177–191
[5] Kleihues, P., Sobin, L.H.: WHO classification of tumors. Cancer **88** (2000) 2887
[6] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth International Group (1984)

[7] Tate, A.R., Ladroue, C., Minguillón, J.: Developing classifiers for single-voxel ^{1}H brain tumour *in vivo* spectra for the INTERPRET decision support tool. Technical Report CSRP543, U. of Sussex, Cognitive and Comp. Sciences (2002)
[8] Domingos, P.: A unified bias-variance decomposition and its applications. In: Proc. of the 17th Int. Conf. on ML, Stanford, CA, USA (2000) 231–238
[9] Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning. Data mining, inference and prediction. Springer series in statistics. Springer (2001)
[10] Dasarathy, B.: Nearest Neighbor Pattern Classification Techniques. IEEE Computer Society Press, Los Alamitos, CA, USA (1991)
[11] Gama, J., Brazdil, P.: Cascade generalization. ML **41** (2000) 315–343
[12] Minguillón, J., Pujol, J., Zeger, K.: Progressive classification scheme for document layout recognition. In: SPIE Proc., Volume 3816, Denver, CO (1999) 241–250