A combined MRI and MRSI based Multiclass System for Brain Tumour Recognition using LS-SVMs with Class Probabilities and Feature Selection

Jan Luts [a,*], Arend Heerschap [b], Johan A.K. Suykens [a], Sabine Van Huffel [a]

[a] *Katholieke Universiteit Leuven, Department of Electrical Engineering, ESAT-SCD (SISTA), Kasteelpark Arenberg 10, B-3001 Leuven-Heverlee, Belgium*

[b] *University of Nijmegen, University Medical Center Sint Radboud, Department of Radiology, Geert Grooteplein Z18, PO Box 9101, 6500 HB Nijmegen, The Netherlands*

Abstract

Objective: This study investigates the use of automated pattern recognition methods on magnetic resonance data with the ultimate goal to assist clinicians in the diagnosis of brain tumours. Recently, the combined use of magnetic resonance imaging (MRI) and magnetic resonance spectroscopic imaging (MRSI) has demonstrated to improve the accuracy of classifiers. In this paper we extend previous work that only uses binary classifiers to assess the type and grade of a tumour to a multiclass classification system obtaining class probabilities. The important problem of input feature selection is also addressed.

Methods and Material: Least squares support vector machines (LS-SVMs) with radial basis function kernel are applied and compared with linear discriminant analysis (LDA). Both a Bayesian framework and cross-validation are used to infer the parameters of the LS-SVM classifiers. Four different techniques to obtain multiclass probabilities as a measure of accuracy are compared. Four variable selection methods are explored. MRI and MRSI data are selected from the INTERPRET project database.

Results: The results illustrate the significantly better performance of automatic relevance determination (ARD), in combination with LS-SVMs in a Bayesian framework and coupling of class probabilities, compared to classical LDA.

Conclusion: It is demonstrated that binary LS-SVMs can be extended to a multiclass classifier system obtaining class probabilities by Bayesian techniques and pairwise coupling. Feature selection based on ARD further improves the results. This classifier system can be of great help in the diagnosis of brain tumours.

Key words: brain tumours, multiclass classification, class probabilities, feature selection, magnetic resonance imaging (MRI), magnetic resonance spectroscopic imaging (MRSI)

---

* Corresponding author. Tel.: +32 16 321065; fax: +32 16 321970
Email addresses: jan.luts@esat.kuleuven.be (Jan Luts),
a.heerschap@rad.umcn.nl (Arend Heerschap),
johan.suykens@esat.kuleuven.be (Johan A.K. Suykens),
sabine.vanhuffel@esat.kuleuven.be (Sabine Van Huffel).

1 Introduction

Contrast-enhanced magnetic resonance imaging (MRI) is a major tool for the anatomical assessment of tumours in the brain. However, several diagnostic questions, such as the type and grade of the tumour, are difficult to address using MRI. The histopathology of a tissue specimen remains the gold standard, despite the associated risks of surgery to obtain a biopsy. In recent years, the use of magnetic resonance spectroscopy (MRS), which provides metabolic information, has gained a lot of interest for a more detailed and specific non-invasive evaluation of brain tumours. In particular magnetic resonance spectroscopic imaging (MRSI), which can provide quantitative metabolite maps of the brain, is attractive as this may also enable to view the heterogenous spatial extent of tumours, both in- and outside the MRI detectable lesion.

As individual viewing and analysis of the multiple spectral patterns, obtained by an MRSI exam, is time-consuming and often requires specific spectroscopic expertise, it is not practical in a clinical environment. Automatic processing and evaluation of the data and easy and rapid display of the results as images or maps is needed for routine clinical interpretation of an exam. At this point, machine learning techniques and pattern recognition systems come up. It is known that different (pathological) tissue types contain specific metabolic patterns [1]. If particular pattern recognition techniques can be automated and integrated into a clinical decision support system (DSS), MRI and MRS can actually become part of clinical practice. Several studies have presented progress in this direction. For example, Preul *et al.* [2] and Szabo de Edelenyi *et al.* [3] conducted some early work. In addition, in the EU framework 6 project INTERPRET [4] a DSS was developed using mainly single-voxel and multivoxel MR spectra combined with MRI [5].

In the past, many researchers explored the use of pattern recognition to build classifiers for different tissue types based on MRI or MRS. First, people have only been using MRI data to distinguish different tissues. It was illustrated that MRI has only limited potential to specify the type and grade of a tumour [6,7]. Later on, one started to construct classifiers using MRS data based on artificial neural networks, linear discriminant analysis (LDA), fuzzy techniques, support vector machines (SVMs) and least squares support vector machines (LS-SVMs) [8–13]. However, only few researchers achieved to combine the information that is present in MRI and MRS. In [3] one specific contrast from MRI was combined with spectroscopic information. [14] added extra image variables for fusion with metabolic information and used distribution plots for classification. In [15], the authors explored the use of LDA and LS-SVMs to binary classify different tissues. These studies all agreed that the use of image intensities and spectroscopic information can improve the accuracy of brain tumour classifiers.

In this paper, we extend the work of Devos *et al.* that was presented in [15]. Devos *et al.* demonstrated that LS-SVMs with a radial basis function (RBF) kernel often achieve a significantly higher performance than LDA and LS-SVMs with linear kernel. In addition, it is known that dealing with unbalanced data sets or small data sets, which is often the case, is problematic if one uses LDA. The linear decision boundaries might also strongly correlate with the training cases. Further, all classifiers presented in [15] are binary ones and are just illustrating the combined use of MRI and MRS. However, if DSSs have to be implemented, the development of multiclass classifier systems is of very high interest. Moreover, clinicians are also interested in a measure of uncertainty when using a DSS. Obviously, it is not enough to output a single tumour type for the case to be classified, without a measure of its confidence.

This paper is organized as follows. First, Section 2 gives an overview of the methods and data set. In the next section we introduce the four methods that are used to handle the feature selection problem. In addition, we describe the four different methods used to combine pairwise class probabilities. Afterwards the results are described in Section 5. Finally, the discussion and conclusion are formulated.

2 Methods and material

In this study, image intensities and spectroscopic information are used to build multiclass classifier systems. In order to obtain a measure of uncertainty, class probabilities are calculated. The output of the classifier system for a specific case are class probabilities for each possible tissue type. This means that instead of binary output scores (0, for 'no tumour of this class' and 1, for 'tumour of this class') we get probability values for each type of tumour. Based on the results of Devos *et al.*, we decide to use LS-SVMs with an RBF kernel in our study. Both binary LS-SVMs and full Bayesian binary LS-SVMs with RBF kernel can output pairwise class probabilities [16–18]. For the purpose of training, testing and (hyper-)parameter estimation we use the KULeuven's LS-SVMlab MATLAB/C Toolbox[1] . To obtain class probabilities instead of binary outputs, the *softmax* function is used for the LS-SVMs; posterior class probabilities for the Bayesian LS-SVMs are computed as explained in [18]. Four different methods that combine pairwise probabilities are compared.

Although kernel-based techniques are known to be less sensitive to the high dimensionality of the input space, reduction may further improve the accuracy of the classifiers as is demonstrated in [19] on some benchmark data sets. To handle the important feature selection problem, four methods to separate

[1] http://www.esat.kuleuven.ac.be/sista/lssvmlab/ (Accessed: 3 December 2006)

irrelevant features are explored.

An important improvement of our classifier system is that more tissue types can be used, compared to the approach in [15]. Thus we are able to classify not only the main type of a tissue but also the grade and subtype of a tumour. In this study, tissue classification follows the pathway that is summarized in Figure 1.

The data are selected from the INTERPRET project database [4]. The clinical information was acquired in the University Medical Center Nijmegen (UMCN) and data from 25 patients with a brain tumour and 4 volunteers are used. This study has been approved by the ethical committee of the UMCN and followed the rules of the World Health Organization. Each case passed a strict quality control and the tumour type was determined by a consensus on a histopathological study. Only patient data where at least two of the three pathologists agreed about the diagnosis was included. For one patient there was no consensus and that patient is not included in our study. To obtain a sufficiently large data set, several voxels, situated in the tumour area, were selected from each patient as described in [20]. The selection of voxels was based on the spectral information and the MRI data. The four high resolution images were plotted together with a segmented image in which voxels are clustered by a model-based algorithm [21]. Since the clustering provided an objective segmentation, this was considered to be helpful for voxel selection. Next, an expert in spectroscopy selected voxels for each class of pathology only if the considered spectra were found to be typical for that pathology and if they were clearly within the affected brain region. Although this method is subjective, it is chosen because tumours are known to be heterogeneous. We think this procedure is appropriate since there is no "ground truth" in the diagnosis of brain tumours at the voxel level and the number of patients is often limited. Further, cerebrospinal fluid (CSF) and normal tissue from volunteers and patients are selected. The data set includes ten classes of pathologies: normal brain tissue from volunteers and apparently normal tissue from the contralateral half of the brain of patients (218 voxels from 8 persons), CSF from patients (100 voxels from 8 patients), grade II diffuse astrocytomas (90 voxels from 5 patients), grade II oligoastrocytomas (45 voxels from 2 patients), grade II oligodendrogliomas (22 voxels from 2 patients), grade III astrocytomas (16 voxels from 2 patients), grade III oligoastrocytomas (28 voxels from 1 patient), grade III oligodendrogliomas (25 voxels from 2 patients), meningiomas (48 voxels from 3 patients) and grade IV gliomas (70 voxels from 7 patients).

The data set, containing both MR images and MR spectra, is acquired and preprocessed as described in [14]. The MR data are acquired on a 1.5 T Siemens Vision Scanner with CP-head coil. Four different image contrasts are acquired: T1-weighted image (TE/TR = 15/644ms), T2-weighted image (TE/TR = 16/3100ms), proton density-weighted image (TE/TR = 98/3100ms), gadolin-

ium enhanced T1-weighted image (15 ml 0.5 M Gd-DTPA). Both water suppressed and unsuppressed proton MR Spectroscopic Images are acquired. The MRSI data is acquired using a 2D STEAM sequence with the STEAM box positioned totally in the brain (TR/TE/TM = 2000 or 2500/20/30 ms, slice thickness = 12.5 or 15 mm, FOV = 200 mm, spectral width = 1000Hz and NS = 2). Disturbing signals arising from the fat tissue surrounding the skull are avoided. The location of the STEAM box is determined using the gadolinium enhanced T1-weighted image showing the largest tumour area. The MRSI slice is centered around an MRI slice of 5 mm. Since there is 1.5 mm of space between the MRI slices, only one MRI slice is used.

The images are co-aligned and all data are semi-automatically preprocessed as in [14]. The images are registered with respect to the proton density-weighted image by shifting and maximizing the spatial correlation. It is assumed that the MRSI data are registered with the proton density-weighted image since they are acquired in a consecutive manner. Further, only pixels within the boundary of the STEAM box are included. Preprocessing of MRSI included filtering of k-space data by a Hanning filter of 50 % using the LUISE software package (Siemens, Erlangen, Germany), zero filtering to 32x32, spatial 2D Fourier transformation to obtain time domain signals for each voxel, correction for eddy current effects by a technique which prevents occasional occurrence of eddy current correction induced artifacts [22,23], water removal using HLSVD from 4.3 ppm to 5.5 ppm [24], frequency alignment and a simple baseline correction using an exponential filter with a width of 5 ms followed by subtraction of the residual of the original signal. All first order phases are corrected by first manually optimizing the mean spectrum which is calculated from all spectra in the STEAM box of each patients MRSI data. Next, this correction is applied to each seperate signal of the patients MRSI data. Finally, the spectra are normalized using the water signal [25]. Hereafter, all spectra are quantified using peak integration. Ten different features are extracted from each spectrum [26]: L2 (0.835-0.965 ppm), L1 (1.2 ppm) + Lac + Ala (1.265-1.395 ppm), NAA (1.955-2.085 ppm), Glx (2.135-2.265 ppm), Cr (2.955-3.095 ppm), Cho (3.135-3.265 ppm), Tau (3.375-3.505 ppm), mI + Gly (3.495-3.625 ppm), Glx + Ala (3.685-3.815 ppm) and Cr (3.885-4.015 ppm). The resolution of four MRI images is lowered to the one of the MRSI grid by averaging pixel intensities within each voxel. The final data set containing 14 variables has also been used in earlier studies [14,15].

In the remainder of this work, existing algorithms and techniques most relevant for this study are briefly described. For further details about the techniques used, an extended overview of the literature and more detailed results, the interested reader is referred to [27].

3 Feature selection

Today, one of the main problems in machine learning and statistics is keeping track of the most relevant information. For this purpose, feature selection techniques are addressed. The major aims of feature selection for classification are finding a subset of variables that result in more accurate classifiers and constructing more compact models. Therefore, feature selection will filter out those variables that are irrelevant for the specific model. The selection should only capture the relevant features while not overfitting the data. Also there is a reduction in the sample size needed for good generalization [28]. In this work we mainly focus on feature weighting and feature selection mechanisms. Techniques like principal component analysis are also able to reduce the dimension of the input space and can extract features, too. However, in this study we prefer methods that provide features with a direct biological meaning.

3.1 Feature selection methods

As feature selection is one of the most important topics in pattern recognition, many attempts have been made to develop feature extraction algorithms. An extensive overview can be found in [29–31]. Basically, three major types of methods are distinguished [32]. The first category is the filter model [30]. The feature filter model filters the variables independently of the classifying algorithm. In this way, an initial analysis is performed on the training data and afterwards the selected features are fed to the classifier. A simple filtering technique ranks or scores each variable based on some measure like the information gain criterion, mutual information, cross-entropy measure, Fisher discriminant criterion or the Kruskal-Wallis test. Apart from these simple ranking methods, more advanced methods like FOCUS or Relief exist [33,34].

Because the learning algorithm (e.g. the classifier) is never used, the main advantage of the filter model is its low computational cost. On the contrary, the weakest point of the filter method is that it completely ignores the impact of the learning algorithm. The performance of a specific feature subset is not tested with the classification technique. Therefore, in [35] it is claimed that the selection procedure should take the learning algorithm into account. This leads us to the second category of selection methods: feature selection techniques using the wrapper model [30]. Different subsets of features are tried on the classification algorithm to estimate the performance of each set, after which the best set is kept. As an exhaustive search through the input space is not feasible, heuristic search methods using backward, forward or stepwise variable selection are often used [36]. In addition, more sophisticated methods like best-first search are also able to traverse the space of subsets [30,37]. To evaluate

each subset, n-fold cross-validation or leave-one-out cross-validation can be used. In [30] it is concluded that the wrapper models result in an increased accuracy because the interaction between the algorithm and the training set interaction is considered. The disadvantage of wrapper methods is the high computational cost of the search.

Apart from the filter and wrapper methods, there also exist some embedded methods. These methods aim to immediately integrate the variable selection or weighting procedure into the learning algorithm. This study does not cover these techniques any further, however, an overview of integrated techniques can be found in [31].

In our application, we build a classifier system that aims at discriminating among 10 different classes. To handle the multiclass problem, we decide to build classifiers between every pair of classes in the data set. This implies that 45 classifiers have to be tuned, trained and tested using cross-validation or similar techniques. This strategy immediately excludes the use of an exhaustive search using for instance stepwise variable selection. In order to avoid tuning the parameters of each LS-SVM classifier a huge number of times, simple methods are preferred. Hence, in this study, an efficient filter technique seems to be an appropriate approach. As there is no overall best variable selection method for LS-SVM classifiers, different filter methods need to be compared before a multiclass classifier system can be constructed. We decide to use a filter model using the Kruskal-Wallis test, the Fisher discriminant criterion and the Relief-F algorithm. Relief-F is an improved version of the original Relief algorithm [38]. Relief-F can be used for multiclass problems, it is more robust and it can handle noisy data. In the next paragraphs, the algorithm is described. To perform variable selection for Bayesian LS-SVMs, an automatic relevance determination (ARD) mechanism is proposed in [39]. In total, these four methods are used in our study for variable selection. In the following part we will discuss these techniques, the experimental setup and the evaluation in more detail.

3.2 Fisher discriminant criterion

Fisher's criterion takes the mean and the within class scatter of the groups into account to compare the correlation between variables and the class label [40]. For all variables in the training/validation set, a score is obtained and the features are ranked according to these scores. Hereafter, different models are built by backwards removing the feature with the smallest Fisher discriminant criterion score. In this way, different models containing the most relevant variables are constructed. Using again 10 times stratified random sampling on the original 2/3 of the data set, the performance of the models on valida-

tion data is checked. Finally the model with the highest average performance on validation data is selected and used on the independent test set. In this way, we use a filter model for selection and check its performance like in the wrapper approach without having to perform an exhaustive search.

3.3 Kruskal-Wallis test

The Kruskal-Wallis test [41] is a non-parametric alternative to the well-known one-way independent-samples analysis of variance [42]. The null hypothesis of the test is that the samples come from populations with equal medians. Given n_C groups, the Kruskal-Wallis test statistic should be compared with the chi-square statistic with $n_C - 1$ degrees of freedom if the sample size within each group is large enough (e.g., > 5). This score is derived for all the features so they can be ranked according to their chi-square value. The same procedure as in the Fisher criterion approach is used: different models are built by removing the variables with the smallest chi-square value. In the end, the variables that are included in the model best performing on validation data, using stratified random sampling, are selected for use on test data. This procedure selects optimal variables in a relatively fast way without causing a massive search process.

3.4 Relief-F

Relief-F is an extended and more robust version of the original Relief algorithm [38]. In contrast to many heuristic measures for feature selection, Relief-F does not assume conditional independence of the variables. The main idea of Relief-F is to estimate the quality of features based on how good their values discriminate between samples that are close. Consecutively random samples are drawn from the data set. Each time the k (e.g. 10) nearest neighbors of the same class and the opposite class are determined. Based on these neighboring cases the weights of the attributes are adjusted. As within the two previous algorithms the variables are ranked and different models are built by dropping the variable with the smallest weight. The remaining part of the selection procedure is completely analogous to the one followed in the two previous methods. Although the Relief-F algorithm is computationally more expensive and complex than the previous techniques, the cost of an exhaustive search is still much higher.

3.5 ARD for LS-SVMs

In [18] the Bayesian evidence framework has been applied to LS-SVMs. Additionally the automatic detection of relevant features in the Bayesian framework has been developed in [39]. To illustrate this and because this work concentrates on LS-SVMs, we start with the model formulation of the LS-SVM classifier

$$\min_{w,b,e} \mathcal{J} = \mu E_W + \zeta E_D, \tag{1}$$

$$y_i(w^T \varphi(x_i) + b) = 1 - e_i, i = 1, ..., N \tag{2}$$

with

$$E_W = \frac{1}{2} w^T w, \tag{3}$$

$$E_D = \frac{1}{2} \sum_{i=1}^{N} {e_i}^2, \tag{4}$$

where x_i is a vector containing the input features, y_i the matching class label (i.e. -1 or $+1$), e_i the error variable, w a weighting vector and b a bias term. In the dual space the LS-SVM classifier is then built as follows

$$y(x) = \text{sign}\left(\sum_{i=1}^{N} \alpha_i y_i K(x, x_i) + b\right), \tag{5}$$

where x is the case to be classified, α_i are Lagrange multipliers and $K(\cdot,\cdot)$ is a positive definite kernel.

In general, the Bayesian LS-SVM framework makes use of three different levels of inferences. On the first level of inference, the bias b and weight w of the LS-SVM are determined. The hyperparameters for regularisation (μ, ζ) are calculated on the second level and the third level performs model comparison to infer the kernel parameters (e.g. σ, the bandwidth of an RBF kernel). The strategy of the ARD procedure is to assign a weight to every input feature by introducing a diagonal weighting matrix U into the kernel function [43]. In this study, an RBF kernel is used and this implies that the kernel has the form

$$K(x_i, x_j) = \exp\left(-\frac{(x_i - x_j)^T U (x_i - x_j)}{\sigma^2}\right). \quad (6)$$

Now, U is inferred by maximizing the model evidence on the third level of inference. As before, the relevant features will have large weights and the less important features will have smaller weights. Instead of doing a backwards variable selection procedure based on ARD, we only reweight the original features according to the weights computed by one iteration of the ARD algorithm. The reason for this approach is that a backwards search would be too time-consuming in this study.

3.6 Experimental setup and evaluation

For each pair of classes in the total data set, the four selection methods are compared using stratified random sampling. The data set is 50 times randomly split in a set used for training, validation and one for testing purposes. One third of the data is used for the test set, 2/3 is used for training and validation. The random splitting is done in a stratified way. Model selection and training happens on the training and validation set while the test set is only used to check the performance of the obtained classifier. To test statistically the performance of the feature selection techniques, each performance measure is averaged over the 50 runs for each single method and every pairwise classifier. Next, we use the Friedman test [41] over all pairwise classifiers (i.e. 45) since the performance of the feature selection methods is correlated for each pair of classes. Further, to study the behaviour of the methods for a specific pairwise classifier in detail, the Friedman test is used since for every of the 50 runs the performance of the different methods is correlated as they are used on the same training and test set.

As performance measure, we use the accuracy (percentage of correctly classified cases), the sensitivity (the ratio of true positives and the sum of true positives and false negatives) and the specificity (the ratio of true negatives and the sum of false positives and true negatives) at a cutoff of 0.50. As some classes might be unbalanced, it is often more appropriate to use the sensitivity and specificity. The cutoff of 0.50 is chosen because it is intuitively a very suitable one. Theoretically it is possible to add a value to the bias term in the LS-SVM classifier and choose another cutoff to correct for unbalance. However, in practice, because of the high number of different pairwise combinations (i.e. 45) and the repeated stratified sampling procedure (i.e. 50) the tuning of an extra correction value becomes a massive task. Therefore we will restrict ourselves to the value of 0.5. Also, the performance of Bayesian LS-SVMs without ARD and the well-known classical technique LDA is provided.

4 Multiclass classification

Until now our discussion focussed on binary classifiers. As mentioned before, if DSSs need to be developed, the study of multiclass classifiers is essential. However, the upgrade of binary LS-SVMs to multiclass LS-SVMs is not straightforward since SVM-based methods employ direct decision functions [19]. The typical procedure is to break down the multiclass problem into a number of smaller binary problems. The procedure to combine these binary classifiers into a multiclass system can be performed in many ways and overall there is no single best performing method for all kinds of classification problems. In the next part, we briefly overview some of the standard methods from the literature and motivate our decisions.

4.1 Combination schemes

In minimal output coding, each class is represented by a unique binary codeword using k bits or k classifiers to encode $n_C = 2^k$ classes [44]. Error correcting output codes use more than the minimal number of bits for encoding to enhance the generalization of the multiclass classifier system [45]. One-versus-all is a method that constructs n_C binary classifiers for the n_C class problem by separating each class from the combination of all others [46]. A disadvantage of the latter method is that the data set is often very asymmetric after grouping together $n_C - 1$ classes. When using one-versus-one coding the unbalance in the data set is often less extreme [47]. For the n_C class problem $n_C(n_C - 1)/2$ binary classifiers need to be built. If the number of classes increases to a very large number, this method seems to become cumbersome. However, when the number of classes is not too abundant, each binary classifier needs to be trained on a smaller number of data so the training and tuning of the classifier can actually become faster. To decide the final class for the one-versus-one approach, a simple voting scheme or max-wins criterion is used.

In this study we decide to use a one-versus-one combination scheme. Apart from the fact that the data are more balanced and that the training and tuning problems are most of the time less computationally intensive, there is also another good reason to use one-versus-one coding. In practice, medical doctors often have a clue about the diagnosis for a specific patient. Frequently, the medical doctors only doubt between two types of tissue such that a binary classification method is sufficient for diagnosing these patients. In fact, one can see the binary classifiers as very powerful stand-alone entities that can also be combined when multiclass classification is needed. Furthermore, clinicians also want a measure of uncertainty when performing classification; it would

be interesting to provide class probabilities for every tissue class. All these issues are addressed in the next part of this section. We cover four different methods that can combine one-versus-one pairwise class probabilities in order to retrieve final class probabilities.

4.2 Pairwise combination of probabilities

In the literature, a few authors provide algorithms to obtain class probabilities based on pairwise combination. In this study, we compare the methods of Price *et al.* [48], Hastie and Tibshirani [49] and two algorithms of Wu *et al.* [50]. The method of Refregier and Vallet [51] and voting [52] are not considered in this work. The reason to omit the algorithm of Refregier and Vallet is that some arbitrary choices about the selection of pairwise probabilities have to be made. It has been pointed out by Price *et al.* and Wu *et al.* that the results are very sensitive to this choice and that finding the optimal selection is often very expensive. Voting is a very simplistic method and it is illustrated in [50] that the errors are high compared to the other methods. Before overviewing the methods and explaining the experimental setup, we state the problem more mathematically. Given a data set x and a corresponding set of class labels y, the pairwise probabilities r_{ij} are denoted as estimates of $\mu_{ij} = P(y_k = i | y_k = i \text{ or } j, x_k)$. As such, the pairwise probabilities r_{ij}, which are the probabilities to predict class i, are retrieved from the binary (i.e. pairwise) classifier that is only trained on data coming from group i and group j. The main goal of coupling probabilities is to obtain the probability $p_i = P(y_k = i | x_k)$ based on the r_{ij} values.

4.3 Price et al.

Price *et al.* develop a method that combines pairwise neural network classifiers with probabilistic outputs for a handwriting recognition system [48]. Although originally intended for classification between a limited number of classes, Price *et al.* show that the approach is also applicable for problems with more than ten classes. The final class probabilities are obtained by

$$p_i = \frac{1}{\sum_{j:j \neq i} \frac{1}{r_{ij}} - (n_C - 2)}. \tag{7}$$

Afterwards the probabilities are normalized such that the sum is exactly one. From the implementation point of view, this method is very simple. On the other hand, the method does not take into account the number of cases for each class.

4.4 Hastie and Tibshirani

In [49], an algorithm which is a special case of the Bradley-Terry model for paired comparisons is presented. In order to obtain $p = (p_1; ...; p_{n_C})$, the algorithm minimizes the Kullback-Leibler distance criterion such that r_{ij} approximates $p_i/(p_i + p_j)$,

$$l(p) = \sum_{i<j} n_{ij} \left(r_{ij} \log \frac{r_{ij}}{\mu_{ij}} + (1 - r_{ij}) \log \frac{1 - r_{ij}}{1 - \mu_{ij}} \right). \tag{8}$$

In this equation, the n_{ij} variable denotes the sum of the number of data points in class i and class j, $r_{ji} = 1 - r_{ij}$ and the model is $\mu_{ij} = p_i/(p_i + p_j)$. One needs to estimate the p_i such that μ_{ij} is close to r_{ij}. Hastie and Tibshirani establish the iterative procedure, depicted below.

Algorithm 1. Coupling approach by Hastie and Tibshirani
1: Start with some initial guess for p_i and corresponding μ_{ij}
2: Repeat $(i = 1, 2, ..., n_C, 1, ...)$ 3 and 4 until convergence
3: $p_i \leftarrow p_i \frac{\sum_{j \neq i} n_{ij} r_{ij}}{\sum_{j \neq i} n_{ij} \mu_{ij}}$
4: renormalize p_i and recompute μ_{ij}
5: $p \leftarrow p / \sum p_i$

Remark that the method takes the number of cases for each class into account.

4.5 Wu et al. - method 1

The first method proposed by Wu *et al.* makes use of an approximate solution to an identity [50]. The existence of this solution is proven based on finite Markov chains. More specifically, Wu *et al.* propose to solve the equations

$$p_i = \sum_{j:j \neq i} \left(\frac{p_i + p_j}{n_C - 1} \right) r_{ij}, \quad \sum_{i=1}^{n_C} p_i = 1, \quad p_i \geq 0. \tag{9}$$

This can be re-expressed as

$$Qp = p, \quad \sum_{i=1}^{n_C} p_i = 1, \quad p_i \geq 0 \quad \text{with} \quad Q_{ij} = \begin{cases} \sum_{s:s \neq i} r_{is}/(n_C - 1), & \text{if i = j} \\ r_{ij}/(n_C - 1), & \text{otherwise.} \end{cases} \tag{10}$$

The main advantage of this method is that only a linear system needs to be solved, no iterative procedure is needed. However, in contrast with the previous algorithm, this method assumes equal weighting (i.e. equal n_{ij}).

4.6 Wu et al. - method 2

The second approach by Wu *et al.* [50] is an improved version of the method of Refregier and Vallet [51]. Wu *et al.* hypothesize the minimum problem

$$\min_p \frac{1}{2} \sum_{i=1}^{n_C} \sum_{j:j \neq i} (r_{ji} p_i - r_{ij} p_j)^2 \quad \text{with} \quad \sum_{i=1}^{n_C} p_i = 1, \quad p_i \geq 0. \tag{11}$$

It is proven that there is a unique solution for p and it can be solved using the simple linear system

$$\begin{pmatrix} Q & 1_{n_C \text{x} 1} \\ 1^T_{n_C \text{x} 1} & 0 \end{pmatrix} \begin{pmatrix} p \\ b \end{pmatrix} = \begin{pmatrix} 0_{n_C \text{x} 1} \\ 1 \end{pmatrix} \quad \text{with} \quad Q_{ij} = \begin{cases} \sum_{s:s \neq i} r_{si}^2, \text{ if i = j} \\ -r_{ji} r_{ij}, \quad \text{otherwise.} \end{cases} \tag{12}$$

$1_{n_C \text{x} 1}$ and $0_{n_C \text{x} 1}$ are column vectors with respectively n_C ones and n_C zeros.

4.7 Experimental setup and evaluation

The pairwise combination methods are compared using stratified random sampling. Like in the feature selection analysis, the data set is repeatedly (i.e. 115) randomly split into a training, validation set and a test set. The pairwise classifiers are built using the training and validation set. Afterwards, we verify the performance of each multiclass combination scheme on the test set. Again, the Friedman test [41] and Tukey's honestly significant difference criterion [53] are used to check whether differences in performance are statistically significant.

The performance measures to compare the different methods are the accuracy, the Brier score [54] and the confusion matrices. The Brier score is related to the mean square error

$$\frac{1}{N} \sum_{j=1}^{N} \frac{1}{n_C} \sum_{i=1}^{n_C} (p_{ij} - t_{ij})^2 \tag{13}$$

where t_{ij} is set to 1 if case j is coming from class i and 0 otherwise. N denotes the number of cases in the test set, n_C is the number of classes and p_{ij} is the predicted posterior probability of class i for case j. This score takes the amount of uncertainty about the predictions into account. The accuracy measure is calculated by assigning each case to the class with the highest posterior probability. Confusion matrices [55] are used to have a clear view on the discriminative power of the classifier for the different classes. The results of the multiclass classification system are summarized in a matrix structure, having on the horizontal axis the actual classification and on the vertical axis the predicted classification. Percentages are calculated so that the total sum for each actual class outcome becomes 100 %.

5 Results

In this section the results of the feature selection methods and pairwise class probability coupling methods are summarized.

5.1 Feature selection

First, to illustrate the importance of good feature selection techniques, we compare the effect of feeding only a selected number of features and feeding all available features to a LS-SVM classifier in Table 1. We choose to make a binary classifier for grade II oligoastrocytomas versus meningiomas and a binary classifier for grade II oligodendrogliomas versus grade III oligodendrogliomas because these data sets are almost balanced. For the first classifier, three features are selected, for the latter five variables are chosen. The choice of the variables is based on prior knowledge. A stratified random sampling procedure is used to calculate the mean percentage of correctly classified cases over 50 runs. Based on the Wilcoxon signed rank test [41], the medians are significantly different. Although there are only results presented for two classifiers in Table 1, one can generalize the observed trend that feature selection can improve the accuracy of a classifier in this study. Therefore, it is important to address this topic before building multiclass classifier systems.

The results for the comparison of the four feature selection techniques are summarized in Figures 2-4. The abbreviations used are ARD for ARD with Bayesian LS-SVMs, FC for Fisher discriminant criterion with LS-SVMs, K-W for the Kruskal-Wallis test with LS-SVMs and R-F for Relief-F with LS-SVMs. The Friedman test and Tukey's honestly significant difference criterion [53] for multiple comparison are used to check for significant differences between the different feature selection methods. Each figure contains a comparison interval

for the mean rank of the averaged performance measure for every method. There are significant differences if the intervals are disjoint. It is observed that the combination of Bayesian LS-SVMs and ARD variable selection generally performs better than the other three approaches. In Figure 2 one can see that the accuracy for ARD is significantly higher than the one of the other methods. Similar results are obtained for the specificity in Figure 3. Concerning the sensitivity, no significant difference is observed between ARD and Relief-F in Figure 4. However, there is a significant difference between ARD and the other two approaches. The differences between the Fisher discriminant criterion, the Kruskal-Wallis test and Relief-F are statistically not significant.

To have a more detailed look, the averaged accuracies for a number of pairwise problems are listed in Table 2. The corresponding class number for a tissue type is 1 for normal tissue, 2 for CSF, 3 for grade II diffuse astrocytomas, 4 for grade II oligoastrocytomas, 5 for grade II oligodendrogliomas, 6 for grade III astrocytomas, 7 for grade III oligoastrocytomas, 8 for grade III oligodendrogliomas, 9 for meningiomas and 10 for grade IV gliomas. Each element represents the mean accuracy over 50 times of stratified random sampling on the test data. The Friedman test and Tukey's honestly significant difference criterion [53] for multiple comparison are used to check for significant differences between the four different feature selection methods for each of the pairwise classifiers. If there is any significant difference between the four methods, the techniques that are not significantly different from the best performing method are printed in boldface, otherwise, no method's performance is printed in boldface. The best performing method is underlined. Further in Table 2, we added the results of Bayesian LS-SVMs without any feature weighting (BL). Often, the performance of Bayesian LS-SVMs without feature selection is already good. However, for certain specific problems (e.g. class 3 versus class 4) the importance of ARD is clear. The performance of classical LDA is also listed in Table 2. The global trend, observed over all pairwise classifiers, is that LDA classifies well between healthy tissue and tumour tissue, but, when discriminating between different tumour types or grades, LS-SVM-based methods often perform better. This is further illustrated in Figure 5 where the accuracy of LDA for each of the 45 pairwise classifiers is plotted. The first nine classifiers distinguish healthy tissue from all other types. As can be observed, these accuracies are generally higher than the ones of all other pairwise problems. According to the Friedman test and Tukey's honestly significant difference criterion, LS-SVM-based methods perform significantly better than LDA for all the specific problems summarized in Table 2. For the other pairwise problems no significant differences are observed.

In the remainder of this section we merely focus on combining pairwise class probabilities into global class probabilities. We restrict the feature selection to the Bayesian methods with ARD, if it improves the results (e.g. class 3 versus class 4), to avoid exhaustive training and tuning times.

The results for the averaged accuracy and the averaged Brier score on test set after 115 times of stratified random sampling are shown in Table 3. Concerning the accuracy, one can observe that the results for the first method of Wu *et al.* and the technique by Hastie and Tibshirani are not significantly different. According to the test statistic, their performances are significantly better than the one of the method by Price *et al.* and the second approach by Wu *et al.*. By looking at the averaged Brier scores, one can compare the prediction uncertainties of each method. The results of the approach by Hastie and Tibshirani seem to degrade if the stopping condition is too loose. If the convergence criterion is taken too high, the iterative procedure produces higher Brier scores than the other methods. When this stopping condition is small enough, the method of Hastie and Tibshirani performs equally well as the first method of Wu *et al.*. Although the Brier scores are relatively close together, statistical differences are found according to the Friedman test and Tukey's honestly significant difference criterion. Further, the method of Price *et al.* seems to produce smaller Brier scores than the approach by Hastie and Tibshirani for some stopping criteria, while its accuracy is smaller. By looking at the results, we observe that the algorithm of Price *et al.* predicts more extreme class probabilities. As such, when predicting correct probabilities, these extreme predictions cause smaller Brier scores.

Confusion matrices for each approach are provided in Figures 6-9. The corresponding class number for each tissue type is equivalent to the one introduced in the previous section. Because of the method's performance and computational simplicity, we will focus on the confusion matrix, produced by the first technique of Wu *et al.*. It is observed that all normal tissue cases are classified correctly, no normal tissue is assigned to a tumour class. For CSF an accuracy of 99.66 % is obtained. The accuracy for grade II diffuse astrocytomas is 96.78 %. Grade II oligoastrocytomas attain an accuracy of 93.97 %. The performance of grade II oligodendrogliomas (92.05 %) is somehow downgraded. Grade III astrocytomas obtain an accuracy of 96 %. The accuracies for grade III oligoastrocytomas and grade III oligodendrogliomas are respectively 99.90 % and 98.80 %. Meningiomas achieve an accuracy of 95.82 %, Grade IV gliomas obtain 98.53 %. In total, 98.24 % of the cases are classified correctly. The first method of Wu was also used in combination with classical LDA instead of the LS-SVM-based approach. The same experimental setup resulted in an accuracy of 96.31 % for classification using LDA. According to

the Wilcoxon rank sum test [41] this performance is significantly lower than the one of the LS-SVM-based approach.

6 Discussion and Conclusion

In this study MRSI and MRI are used to construct a multiclass classifier system for brain tumours. Before discussing the results, we make a remark about the database used. As explained, the same data set has already been used in earlier studies [14,15,56]. In these studies, the authors used six classes of tissue types: normal tissue (8 persons), CSF (8 persons), grade II gliomas (9 persons), grade III gliomas (5 persons), grade IV gliomas (7 persons) and meningiomas (3 persons). These studies constructed test sets that contain voxels coming from patients from which also other voxels were selected for training. Strictly speaking, the test sets were not totally independent. In our work we are also confronted with this issue. Like in the previous studies, one has to keep this in mind when interpreting the results. Additionally, in this work we decided to split the grade II and grade III gliomas further into three different subtypes (astrocytomas, oligoastrocytomas, oligodendrogliomas). The authors are aware that the number of cases in each class decreases in this way, also the number of patients decreases per tissue class. However, the goal of this study is not to stress on the global performance of the classifier. The aim is to compare different methodologies and show their importance for brain tumour classification. Additionally, it is widely known that SVM techniques can handle higher dimensional input spaces and smaller data sets. Moreover, [50] points out the fact that the differences between the pairwise combination schemes become more pronounced with an increasing number of classes. Therefore, it is important to do an analysis with a reasonable amount of classes (e.g. 10). In a later phase it becomes interesting to verify our findings and the ones of [14,15,56] in a multi-center study when more data become available via acquisition through international projects [57,58]. This can possibly result in prospective studies. Furthermore, we plan to integrate the various techniques discussed in this study in the DSSs that are being developed by the eTUMOUR consortium [57] and the HealthAgents project [58].

6.1 Feature selection

First, it should be stressed that the omission of relevant features can improve a classifier as suggested in [30]. If a feature is relevant, this does not automatically mean that it is included in the optimal set. Moreover, if a variable is irrelevant it can sometimes be used in an optimal variable subset. Therefore, the selected variables after the feature selection procedure are not discussed.

This fact also illustrates that it is important to not simply use the pure filter approach. It is of great importance to use the classifier for model selection.

A possible explanation for the results is that, although weights can become zero, reweighting the input features via ARD might increase the performance compared to techniques that are only selecting features. Selecting features is just a 'black or white' decision, while weighting techniques can specifically rescale a variable according to its importance. In addition, in pure selection methods the number of input features has to be determined via a cross-validation analysis or stratified random sampling procedure. In our work, we fixed the number of stratified random sampling runs to ten for determining the size of the feature set. Increasing this number might improve the results for the the non-weighting methods. However, this will also result in longer training and tuning times. As such, from the practical point of view, using ARD with Bayesian LS-SVMs is less computationally intensive than an extra cross-validation analysis or a stratified random sampling procedure to find the optimal number of features. Finally, in contrast to the softmax function for the LS-SVMs, when making predictions using Bayesian LS-SVMs and ARD the unbalance of the data set is taken into account by specifying prior class probabilities. Technically, it is possible to correct for unbalance in non-Bayesian LS-SVM methods, too. However, this comes down to tuning an extra parameter that is a correction on the original bias term of the LS-SVM. As stated above, this extra tuning procedure makes the development of a classifier a massive task. Further, although no statistical differences are observed between the Fisher discriminant criterion, the Kruskal-Wallis test or Relief-F, it seems that the performance of the latter is partly superior. Though this effect is minimal, it can be explained by the fact that Relief-F is not assuming conditional independence of the features.

It is observed that the performance of Bayesian LS-SVMs without feature weighting is sometimes fairly good. This is important because leaving out ARD saves training and tuning time. Depending on the problem, one can decide to use ARD or to omit it. Further, one can argue to use simple and fast methods like LDA for discriminating between tumour tissue and healthy tissue and to apply more advanced methods for determining the specific type and grade of a tumour. Remark that the performance of LDA sometimes seems to be better than the one of LS-SVMs with Kruskal-Wallis test, Fisher criterion or Relief-F. However, as mentioned before, one has to keep in mind that a correction on the bias of the LS-SVM methods was not tuned, causing a downgraded performance.

Finally, the facts, discussed above, and the results illustrate the usefulness of the Bayesian LS-SVMs with ARD in the context of applications with strict time and hardware limits. Dynamic DSSs, containing self-learning classifiers, require methods that can relatively fastly train and tune parameters and per-

form feature selection. Bayesian LS-SVMs with ARD fulfill these requirements and obtain good results. On the contrary, it can be meaningful to use LS-SVMs and feature selection methods based on cross-validation or repeated stratified sampling in static (not self-learning) DSSs. Also a correction on the bias term can be calculated to handle unbalanced problems since there are no direct time constraints.

6.2 Multiclass classification

The use of class probabilities obtained via LS-SVM classifiers and pairwise class probability combination schemes for multiclass classification is illustrated. Four different methods that combine pairwise class probabilities into global class probabilities are compared.

In general, the trends agree with the observations in [50]. Wu *et al.* argue that the differences between the algorithms increase when the number of classes raises (e.g. 10). In particular, this has a stronger impact on the performance of the method by Hastie and Tibshirani. In our 10 group study we also notice this tendency when the stopping conditions for the Hastie and Tibshirani method are not strict enough. In these cases, a downgraded performance is observed for this algorithm. Like in [50], it is noticed that the results of the approach by Hastie and Tibshirani are dependent on the stopping condition. In our analysis, modifying the stopping criterion led to an improvement in the performance of this method. But, since the choice of the condition is application dependent, it is not clear how to choose a suitable stopping criterion in advance, while avoiding an extensive amount of computationally intensive iterations. This is an important drawback of the method of Hastie and Tibshirani. As such, the first non-iterative procedure by Wu *et al.* is preferred. This method can be implemented in a very straightforward way.

As mentioned before, the main aim of this study is to introduce new methodologies for the diagnosis of brain tumours. Although, due to the nature of the data set and the retrospective character of this study, one has to be careful when drawing medical conclusions, certain trends are evident when looking at the confusion matrix obtained by the method of Wu *et al.*. Normal tissue is clearly recognized by the classifier. Some tumour classes are mixed with CSF voxels. This can be clarified by the fact that all the CSF voxels are coming from patients; this may have an influence. Grade II diffuse astrocytomas are classified rather well. Most of the time, this tumour class is mixed up with CSF and grade II oligoastrocytomas. A possible explanation for the relatively poor performance of grade II oligodendrogliomas is the small number of cases in this class. Often, the same voxel is repeatedly misclassified when doing repeated stratified sampling. This problem is more persistent in small data

sets. Grade III astrocytomas are sometimes mixed with the lower grade diffuse astrocytomas. Since the number of cases for this class is small, more data should be acquired. The accuracy for grade III oligoastrocytomas and grade III oligodendrogliomas is fairly good, however, also more data should be acquired. Meningiomas tend to be mixed with grade II diffuse astrocytomas and grade IV gliomas. Conversely, grade IV gliomas are sometimes confused with grade II diffuse astrocytomas and meningiomas.

Compared to classical LDA, the LS-SVM-based approach achieves a significantly higher performance. This is also noted in [15] and can be explained by the fact that certain subtypes of tumours are hard to distinguish with a linear method. This is also supported by the results of the feature selection analysis where LDA mainly attains a high performance for the classification between healthy tissue and tumour.

Acknowledgements

Research supported by Research Council KUL: GOA-AMBioRICS, Centers-of-excellence optimisation, several PhD/postdoc & fellow grants; The Biomedical Magnetic Resonance Research Group Radboud University Nijmegen Medical Center for providing the data; Institute for Molecules and Materials, Analytical Chemistry, Chemometrics Research Department of the Radboud University Nijmegen for preprocessing the data; Flemish Government: FWO: PhD/postdoc grants, projects, G.0360.05 (Advanced EEG analysis techniques for epilepsy monitoring), G.0321.06 (Numerical tensor techniques for spectral analysis), G.0302.07 (Support vector machines and kernel methods), research communities (ICCoS, ANMMM); IWT: PhD Grants; Belgian Federal Government: DWTC (IUAP IV-02 (1996-2001) and IUAP V-22 (2002-2006): Dynamical Systems and Control: Computation, Identification & Modelling); EU: BIOPATTERN (contract no. FP6-2002-IST 508803), eTUMOUR (contract no. FP6-2002-LIFESCIHEALTH 503094), HealthAgents (contract no. FP6-2005-IST 027213).

References

[1] R.E. Danielsen, B. Ross, Magnetic resonance spectroscopy diagnosis of neurological diseases, Marcel Dekker Inc, New York, 1999.

[2] M.C. Preul, Z. Caramanos, D.L. Collins, J.G. Villemure, R. Leblanc, A. Olivier, R. Pokrupa, D.L. Arnold, Accurate, noninvasive diagnosis of human brain tumors by using proton magnetic resonance spectroscopy, Nature Medicine 2 (1996) 323–325.

[3] F. Szabo de Edelenyi, C. Rubin, F. Esteve, S. Grand, M. Decorps, V. Lefournier, J.F. Le Bas, C. Remy, A new approach for analyzing proton magnetic resonance spectroscopic images of brain tumors: nosologic images, Nature Medicine 6 (2000) 1287–1289.

[4] International network for Pattern Recognition of Tumours Using Magnetic Resonance.
URL `http://azizu.uab.es/INTERPRET/ (Accessed: 3 December 2006)`

[5] A.R. Tate, J. Underwood, D.M. Acosta, M. Julia-Sape, C. Majos, 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. Cabanas, A.W. Simonetti, W. Gajewicz, J. Calvar, A. Capdevila, P.R. Wilkins, B.A. Bell, C. Remy, A. Heerschap, D. Watson, J.R. Griffiths, C. Arus, Development of a decision support system for diagnosis and grading of brain tumours using in vivo magnetic resonance single voxel spectra, Nuclear Magnetic Resonance in Biomedicine 19 (4) (2006) 411–434.

[6] F. Earnest, P.J. Kelly, B.W. Scheithauer, B.A. Kall, T.L. Cascino, R.L. Ehman, G.S. Forbes, P.L. Axley, Cerebral astrocytomas: histopathologic correlation of MR and CT contrast enhancement with stereotactic biopsy, Radiology 166 (1988) 823–827.

[7] B.L. Dean, B.P. Drayer, C.R. Bird, R.A. Flom, J.A. Hodak, S.W. Coons, R.G. Carey, Gliomas: classification with MR imaging, Radiology 174 (1990) 411–415.

[8] M.C. Preul, Z. Caramanos, R. Leblanc, J.G. Villemure, D.L. Arnold, Using pattern analysis of in vivo proton MRSI data to improve the diagnosis and surgical management of patients with brain tumors, Nuclear Magnetic Resonance in Biomedicine 11 (1998) 192–200.

[9] H. Poptani, J. Kaartinen, R.K. Gupta, M. Niemitz, Y. Hiltunen, R.A. Kauppinen, Diagnostic assessment of brain tumours and non-neoplastic brain disorders in vivo using proton nuclear magnetic resonance spectroscopy and artificial neural networks, Journal of Cancer Research and Clinical Oncology 125 (1999) 343–349.

[10] J.C. Lindon, E. Holmes, J.K. Nicholson, Pattern recognition methods and applications in biomedical magnetic resonance, Progress in Nuclear Magnetic Resonance Spectroscopy 39 (2001) 1–40.

[11] C.Z. Ye, J. Yang, D.Y. Geng, Y. Zhou, N.Y. Chen, Fuzzy rules to predict degree of malignancy in brain glioma, Medical Biological Engineering Computing 40 (2002) 145–152.

[12] A.R. Tate, C. Majos, A. Moreno, F.A. Howe, J.R. Griffiths, C. Arus, Automated classification of short echo time in in vivo 1H brain tumor spectra: a multicenter study, Magnetic Resonance in Medicine 49 (2003) 29–36.

[13] A. Devos, L. Lukas, J.A.K. Suykens, L. Vanhamme, A.R. Tate, F.A. Howe, C. Majos, A. Moreno-Torres, M. van der Graaf, C. Arus, S. Van Huffel, Classification of brain tumours using short echo time 1H MR spectra, Journal of Magnetic Resonance 170 (2004) 164–175.

[14] A.W. Simonetti, W.J. Melssen, M. van der Graaf, A. Heerschap, L.M.C. Buydens, A new chemometric approach for brain tumor classification using magnetic resonance imaging and spectroscopy, Analytical Chemistry 75 (2003) 5352–5361.

[15] A. Devos, A.W. Simonetti, M. van der Graaf, L. Lukas, J.A.K. Suykens, L. Vanhamme, L.M.C Buydens, A. Heerschap, S. Van Huffel, The use of multivariate MR imaging intensities versus metabolic data from MR spectroscopic imaging for brain tumour classification, Journal of Magnetic Resonance 173 (2005) 218–228.

[16] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Processing Letters 9 (3) (1999) 293–300.

[17] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, Least Squares Support Vector Machines, World Scientific, Singapore, 2002.

[18] T. Van Gestel, J.A.K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, J. Vandewalle, Bayesian framework for least squares support vector machine classifiers, Gaussian processes and kernel Fisher discriminant analysis, Neural Computation 14 (2002) 1115–1147.

[19] S. Abe, Support Vector Machines for Pattern Classification, Springer-Verlag, New York, 2005.

[20] A.W. Simonetti, W.J. Melssen, F. Szabo de Edelenyi, J.J.A. van Asten, A. Heerschap, L.M.C. Buydens, Combination of feature-reduced MR spectroscopic and MR imaging data for improved brain tumor classification., Nuclear Magnetic Resonance in Biomedicine 18 (2005) 34–43.

[21] R. Wehrens, A.W. Simonetti, L.M.C. Buydens, Mixture modelling of medical magnetic resonance data, Journal of Chemometrics 16 (2002) 274–282.

[22] U. Klose, In vivo proton spectroscopy in presence of eddy currents, Magnetic Resonance in Medicine 14 (1990) 26–30.

[23] A.W. Simonetti, W.J. Melssen, M. van der Graaf, A. Heerschap, L.M.C. Buydens, Automated correction of unwanted phase jumps in reference signals which corrupt MRSI spectra after eddy current correction, Journal of Magnetic Resonance 159 (2002) 151–157.

[24] W.W.F. Pijnappel, A. van den Boogaart, R. de Beer, D. van Ormondt, SVD-based quantification of magnetic resonance signals, Journal of Magnetic Resonance 97 (1992) 122–134.

[25] Z. Tong, T. Yamaki, K. Harada, K. Houkin, In vivo quantification of the metabolites in normal brain and brain tumors by proton MR spectroscopy using water as an internal standard, Magnetic Resonance Imaging 22 (2004) 735–742.

[26] V. Govindaraju, K. Young, A.A. Maudsley, Proton NMR chemical shifts and coupling constants for brain metabolites, Nuclear Magnetic Resonance in Biomedicine 13 (2000) 129–153.

[27] J. Luts, A. Heerschap, J. Suykens, S. Van Huffel, A combined MRI and MRSI based multiclass system for brain tumour recognition using LS-SVMs with class probabilities and feature selection, Internal Report 06-143, ESAT-SISTA, K.U.Leuven (Leuven, Belgium).

[28] A. Blumer, A. Ehrenfeucht, D. Haussler, M.K. Marmuth, Occam's razor, Information Processing Letters 24 (1987) 377–380.

[29] D. Wettschereck, D.W. Aha, T. Mohri, A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms, Artificial Intelligence Review 11 (1997) 273–314.

[30] R. Kohavi, G.H. John, Wrappers for feature subset selection, Artificial Intelligence 97 (1997) 273–324.

[31] A. Blum, P. Langley, Selection of relevant features and examples in machine learning, Artificial Intelligence 97 (1997) 245–271.

[32] T.G. Dietterich, Machine learning research: Four current directions, The Artificial Intelligence Magazine 18 (1998) 97–136.

[33] H. Almuallim, T.G. Dietterich, Learning with many irrelevant features, in: Proceedings of the Ninth National Conference on Artificial Intelligence (AAAI-91), Vol. 2, AAAI Press, Anaheim, California, 1991, pp. 547–552.

[34] K. Kira, L.A. Rendell, A practical approach to feature selection, in: D.H. Sleeman, P. Edwards (Eds.), The Ninth International Workshop on Machine Learning, Morgan Kaufmann, Palo Alto, 1992, pp. 249–256.

[35] G.H. John, R. Kohavi, K. Pfleger, Irrelevant features and the subset selection problem, in: International Conference on Machine Learning, Morgan Kaufmann, Palo Alto, 1994, pp. 121–129.

[36] J. Neter, M.H. Kutner, W. Wasserman, C.J. Nachtsheim, Applied Linear Statistical Models, McGraw-Hill/Irwin, Boston, 1996.

[37] M.L. Ginsberg, Essentials of Artificial Intelligence, Morgan Kaufmann, Palo Alto, 1993.

[38] M. Robnik-Sikonja, I. Kononenko, Theoretical and empirical analysis of ReliefF and RReliefF, Machine Learning 53 (2003) 23–69.

[39] T. Van Gestel, J.A.K. Suykens, B. De Moor, J. Vandewalle, Automatic relevance determination for least squares support vector machine classifiers, in: European Symposium on Artificial Neural Networks, D-facto publications, Evere, 2001, pp. 13–18.

[40] R.A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics 7 (1936) 179–188.

[41] M. Hollander, D.A. Wolfe, Nonparametric Statistical Methods, Wiley, New York, 1973.

[42] R.V. Hogg and J. Ledolter, Engineering Statistics, MacMillan, New York, 1987.

[43] D.J.C. MacKay, Introduction to gaussian processes, in C.M. Bishop (Eds.), Neural Networks and Machine Learning, NATO Advanced Study Institute, Vol. 168, Springer-Verlag, Berlin, 1998.

[44] J.A.K. Suykens, J. Vandewalle, Multiclass least squares support vector machines, in: Proceedings International Joint Conference on Neural Networks (IJCNN'99), Washington DC, 1999.

[45] T.G. Dietterich, G. Bakiri, Solving multiclass learning problems via error-correcting output codes, Journal of Artificial Intelligence Research 2 (1995) 263-286.

[46] E. Allwein, R. Schapire, Y. Singer, Reducing multiclass to binary: A unifying approach for margin classifiers, Journal of Machine Learning Research 1 (2000) 113-141.

[47] U.H.-G. Kressel, Pairwise classification and support vector machines, in: B. Scholkopf, C.J.C Burges, A.J. Smola (Eds.), Advances in kernel methods: support vector learning, MIT Press, Cambridge, 1999.

[48] D. Price, S. Knerr, L. Personnaz, G. Dreyfus, Pairwise neural network classifiers with probabilistic outputs, Neural Information Processing Systems 7 (1994) 1109–1116.

[49] T. Hastie, R. Tibshirani, Classification by pairwise coupling, The Annals of Statistics 26 (1998) 451–471.

[50] T. Wu, C. Lin, R. Weng, Probability estimates for multi-class classification by pairwise coupling, Journal of Machine Learning Research 5 (2004) 975–1005.

[51] P. Refregier, F. Vallet, Probabilistic approach for multiclass classification with neural networks, in: Proceedings of International Conference on Artificial Networks, North-Holland, Amsterdam, 1991, pp. 1003–1007.

[52] J. Friedman, Another approach to polychotomous classification, Technical report (1996), Stanford University.

[53] Y. Hochberg, A.C. Tamhane, Multiple Comparison Procedures, Wiley, New York, 1987.

[54] G.W. Brier, Verification of forecasts expressed in probabilities, Monthly Weather Review 78 (1950) 1–3.

[55] G.J. McLachlan, Discriminant Analysis and Statistical Pattern Recognition, Wiley, New York, 1992.

[56] A.W. Simonetti, Investigation of brain tumor classification and its reliability using chemometrics on MR spectroscopy and MR imaging data, PhD thesis, University of Nijmegen.

[57] The eTUMOUR consortium.
URL `http://www.etumour.net/ (Accessed: 3 December 2006)`

[58] The HealthAgents project.
URL `http://www.healthagents.net/ (Accessed: 3 December 2006)`

Captions of tables

Table 1: Average performance on test sets over 50 runs of stratified random sampling for a LS-SVM classifier with and without feature selection.

Table 2: Averaged accuracy on test sets over 50 runs of stratified random sampling. The best performing method and not significantly different approaches are printed in boldface. The score of the best performing technique is underlined. LS-SVM-based methods perform significantly better than LDA according to the Friedman test.

Table 3: Mean accuracy and Brier score on test set over 115 runs of stratified random sampling. The accuracies of the technique by Hastie and Tibshirani (H-T) and the first method of Wu *et al.* are significantly higher compared to all others. In contrast to the other approaches, the Brier score of the technique by Hastie and Tibshirani (if the stopping criterion is sufficiently small, e.g. $\leq 10^{-3}$) and the first method of Wu *et al.* are not significantly different.

Table 1

	all features	selection of features
grade II oligoastrocytomas vs meningiomas	0.9826	0.9955
grade II oligodendrogliomas vs grade III oligodendrogliomas	0.9880	0.9973

Table 2

	ARD	FC	K-W	R-F	BL	LDA
1 vs 3	**0.9984**	0.9922	0.9946	0.9942	0.9926	0.9913
2 vs 3	**0.9873**	0.9822	0.9822	**0.9848**	0.9914	0.9784
3 vs 4	**0.9853**	0.9618	0.9649	0.9591	0.9613	0.9209
3 vs 8	**1.0000**	0.9858	0.9895	**0.9932**	0.9958	0.9926
4 vs 9	**0.9935**	0.9781	**0.9806**	**0.9942**	0.9987	0.9903
5 vs 8	0.9920	0.9867	0.9867	0.9867	0.9947	0.9760
6 vs 8	0.9738	0.9800	0.9754	0.9892	0.9969	0.9738
7 vs 10	0.9938	0.9906	0.9850	0.9956	0.9956	0.9806

Table 3

	H-T (10^{-1})	H-T (10^{-2})	H-T (10^{-3})	H-T (10^{-4})	H-T (10^{-5})	Price *et al.*	Wu *et al.* 1	Wu *et al.* 2
Accuracy	**0.9826**	**0.9829**	**0.9829**	**0.9829**	**0.9829**	0.9817	**0.9824**	0.9817
Brier score	24.784 10^{-3}	3.7268 10^{-3}	**2.8853** 10^{-3}	**2.8923** 10^{-3}	**2.8941** 10^{-3}	3.1379 10^{-3}	**2.9179** 10^{-3}	2.9640 10^{-3}

Captions of figures

Figure 1: Scheme denoting the various steps to perform tissue classification. First, the MRSI and MRI data are acquired using an MR scanner. The spectra are preprocessed, peak integrated and image intensities are averaged within each voxel. Prior to one-versus-one classification, relevant features are extracted. Class probabilities are generated by pairwise coupling.

Figure 2: The comparison intervals for the mean rank of the averaged accuracy on test set. The accuracy of ARD is significantly higher compared to Relief-F (R-F), Fisher discriminant criterion (FC) or the Kruskal-Wallis test (K-W).

Figure 3: Comparison intervals for the mean rank of the averaged specificity on the test set. The specificity of ARD is significantly higher compared to the one of Relief-F (R-F), Fisher discriminant criterion (FC) or the Kruskal-Wallis test (K-W).

Figure 4: Comparison intervals for the mean rank of the averaged sensitivity on the test data. The sensitivity of Relief-F (R-F) is not significantly different from the sensitivity of ARD. The latter is significantly different from Fisher discriminant criterion (FC) and the Kruskal-Wallis test (K-W).

Figure 5: The averaged accuracies for LDA on test set for each of the 45 pairwise classifiers. The first nine pairwise classifiers distinguish healthy tissue from all other tissue types. In general, the accuracy of LDA for these classification problems tend to be higher.

Figure 6: Confusion matrix of the method by Hastie and Tibshirani (with convergence criterion 10^{-3}) over 115 runs of stratified random sampling on test set. On the horizontal axis the true classes are indicated, the vertical axis represents the test set predictions.

Figure 7: Confusion matrix of the method by Price *et al.* over 115 runs of stratified random sampling on test set. On the horizontal axis the true classes are indicated, the vertical axis represents the test set predictions.

Figure 8: Confusion matrix of the first method by Wu *et al.* over 115 runs of stratified random sampling on test set. On the horizontal axis the true classes are indicated, the vertical axis represents the test set predictions.

Figure 9: Confusion matrix of the second method by Wu *et al.* over 115 runs of stratified random sampling on test set. On the horizontal axis the true classes are indicated, the vertical axis represents the test set predictions.

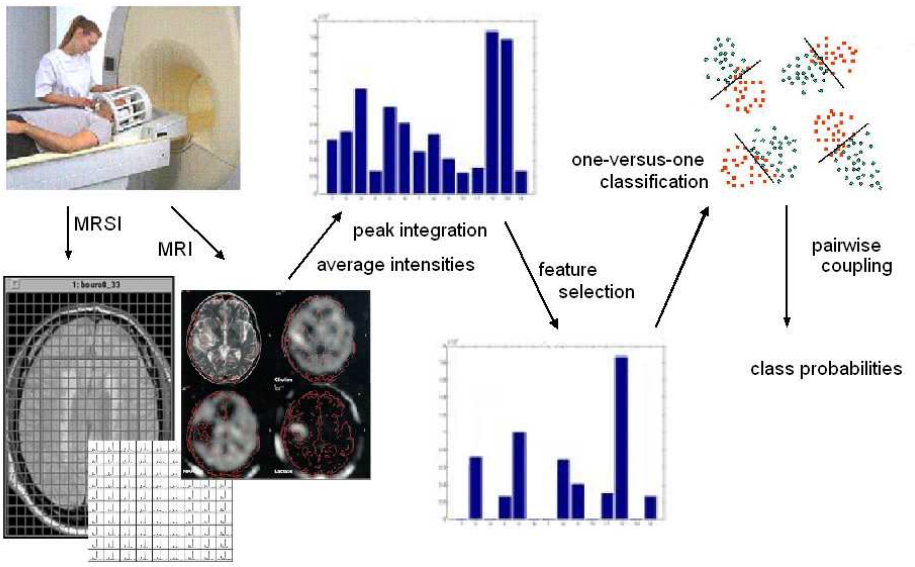

Fig. 1. Luts *et al.*

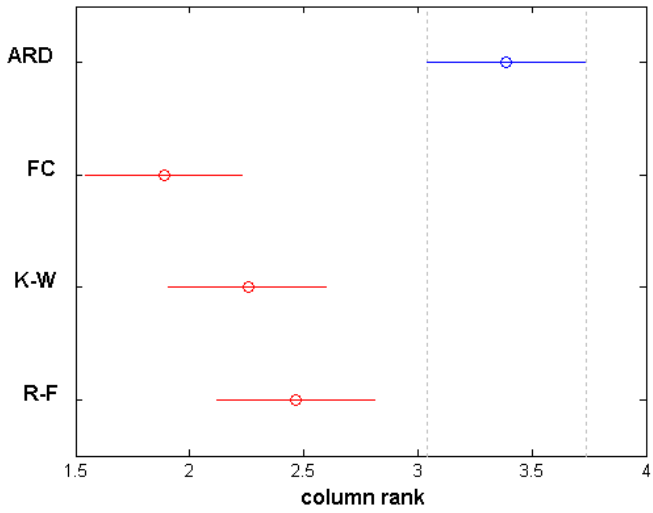

Fig. 2. Luts *et al.*

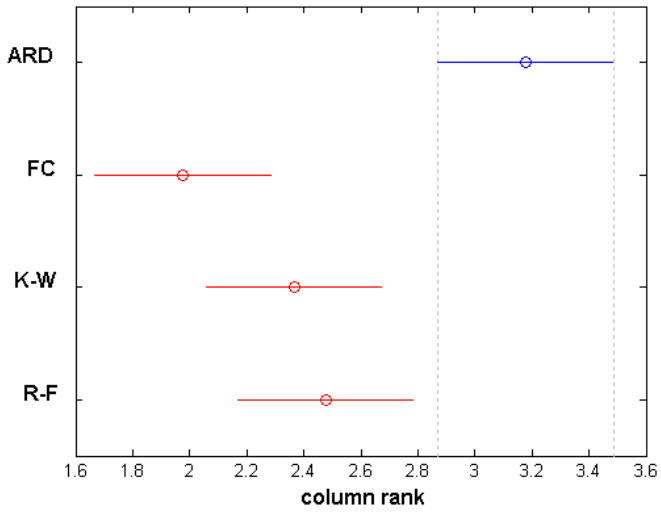

Fig. 3. Luts *et al.*

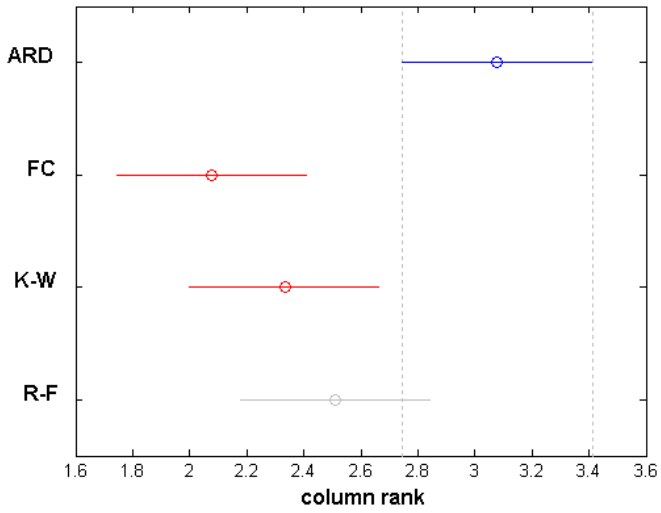

Fig. 4. Luts *et al.*

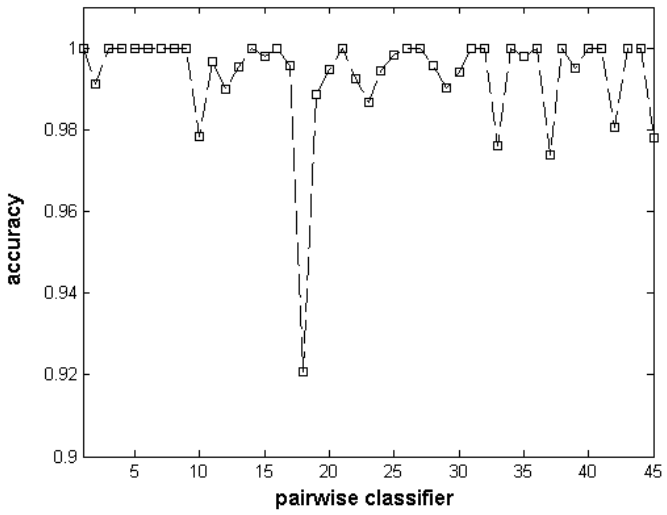

Fig. 5. Luts *et al.*

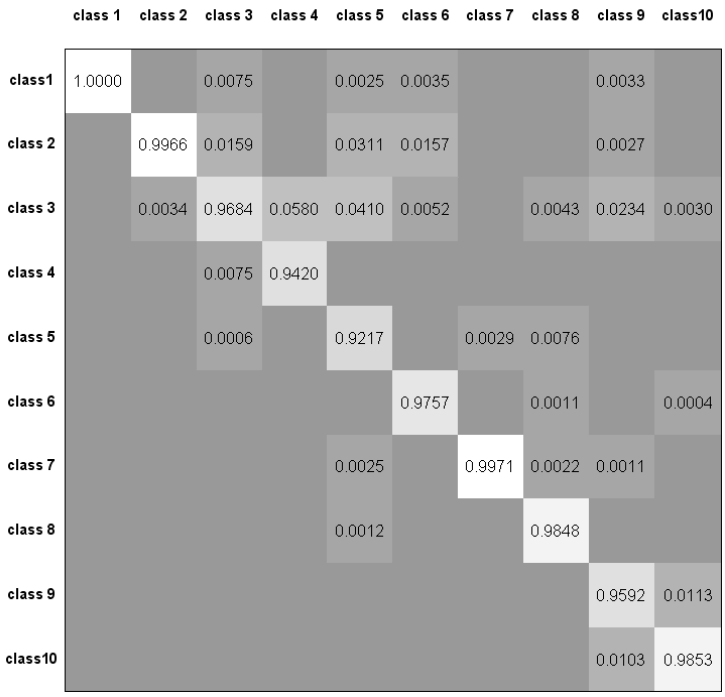

Fig. 6. Luts *et al.*

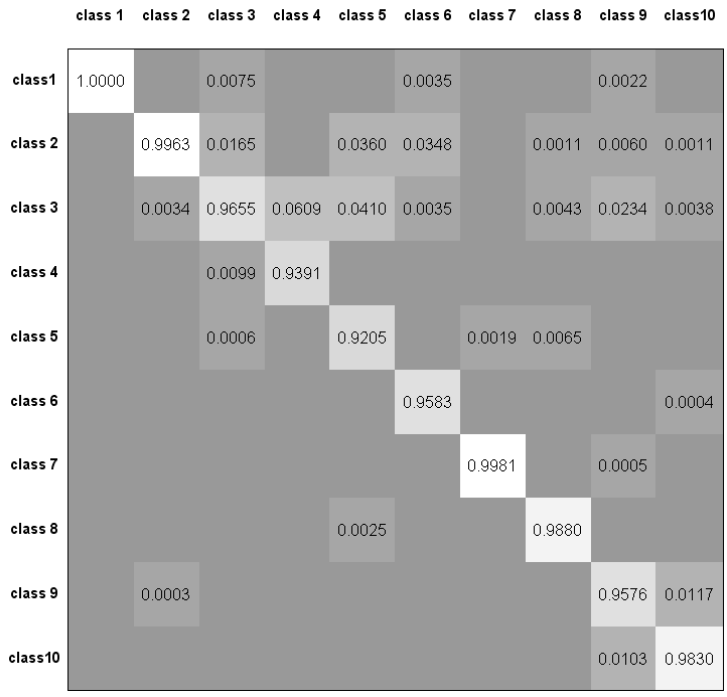

Fig. 7. Luts *et al.*

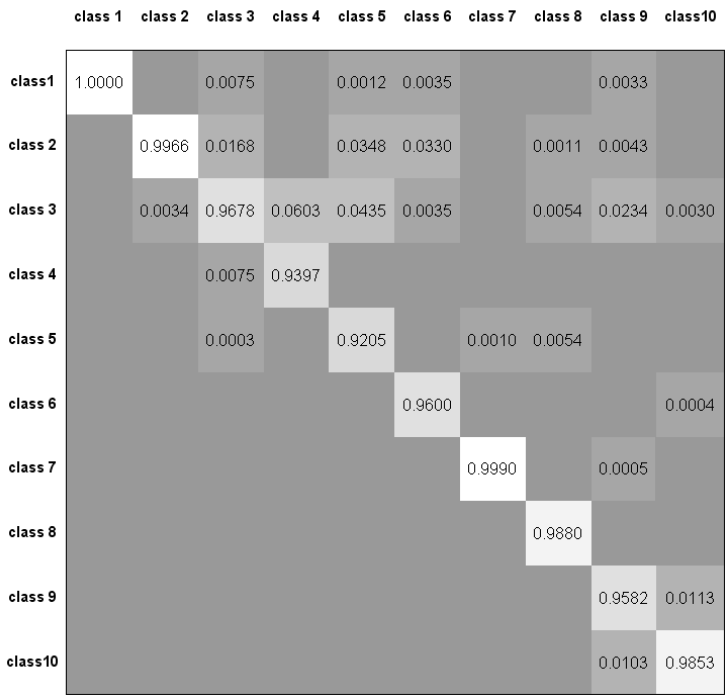

Fig. 8. Luts *et al.*

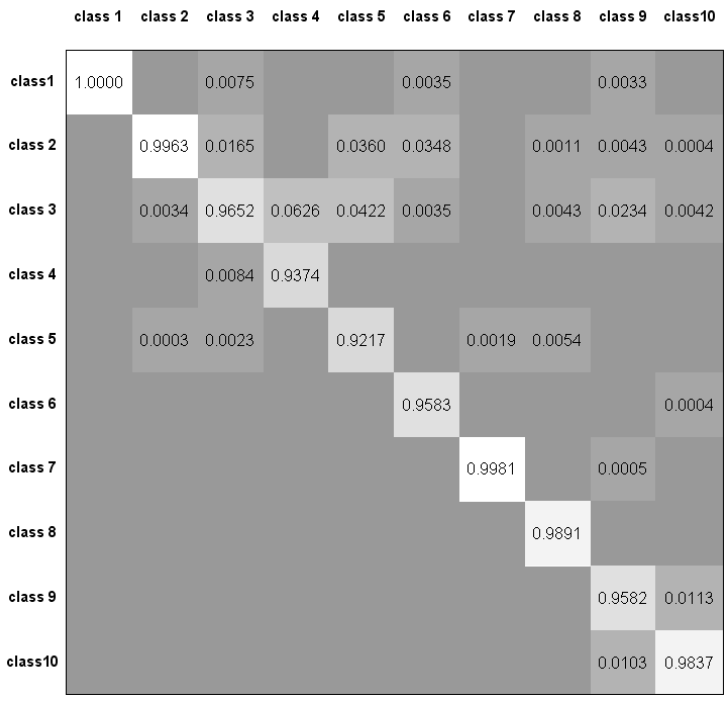

Fig. 9. Luts *et al.*