Near-optimal region selection for feature space reduction: novel preprocessing methods for classifying MR spectra

Alexander E. Nikulin,1,2 Brion Dolenko,1 Tedros Bezabeh1 and Ray L. Somorjai1*

1Institute for Biodiagnostics, National Research Council, 435 Ellice Ave, Winnipeg, Manitoba, Canada R3B 1Y6
2Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, Manitoba, Canada R3T 5V6

Received 28 July 1997; revised 4 November 1997; accepted 1 December 1997

ABSTRACT: We introduce a global feature extraction method specifically designed to preprocess magnetic resonance spectra of biomedical origin. Such preprocessing is essential for the accurate and reliable classification of diseases or disease stages manifest in the spectra. The new method is genetic algorithm-guided. It is compared with our enhanced version of the standard forward selection algorithm. Both seek and select optimal spectral subregions. These subregions necessarily retain spectral information, thus aiding the eventual identification of the biochemistry of disease presence and progression. The power of the methods is demonstrated on two biomedical examples: the discrimination between meningioma and astrocytoma in brain tissue biopsies, and the classification of colorectal biopsies into normal and tumour classes. Both preprocessing methods lead to classification accuracies over 97% for the two examples. © 1998 John Wiley & Sons, Ltd.

KEYWORDS: enhanced forward selection; genetic algorithm; principal component; preprocessing; optimal region selector

INTRODUCTION

1H MR spectroscopy promises to be a powerful non-invasive tool in detecting and classifying diseases and disease stages. However, to fulfil this promise it must be combined with reliable and robust classification strategies. These will have to deal with the complications that may prevent accurate classification of MR spectra of biomedical origin, whether ex or in vivo. The first complication arises because such spectra are complex, noisy, with overlapping peaks, all of which conspire against the identification of their metabolic constituents. The second is that usually only a limited number of spectra are available for analysis. Third, the scarcity of samples is contrasted with the overabundance of spectral features: typically each spectrum consists of hundreds or thousands of attributes (the frequencies), with the majority redundant or at best correlated.

In principle, the first two complications could be rectified. In practice, improving experimental conditions in a clinical environment is generally not feasible, and acquiring sufficient number of samples is either not possible or may take too much time. However, it is possible to deal with the third complication by the appropriate feature space reduction method. In fact, it is imperative to reverse the large ratio of number-of-attributes to number-of-spectra if reliable classification is required, whatever classifier might be used. Both theoretical and empirical evidence1 suggests that this sample-to-attribute ratio must be greater than 10 to 1, rather than the more typical 1 to 10 or less for spectra. Our experience with different classification strategies for spectra of various tissue biopsies2–4 suggests that most of the effort should be directed toward preprocessing. Accurate classification will then follow even with simple classifiers such as linear discriminant analysis (LDA).

In conventional preprocessing of MR spectra of biomedical origin, principal component analysis (PCA)5,6 is applied to the data to reduce the dimensionality of the feature space. PCA involves diagonalising the L×L sample covariance matrix S, i.e. rotating the L-dimensional coordinate system of the original feature vectors (the L-point spectra) into another orthonormal system. Its coordinate axes, the principal components (PCs), when sorted in order of decreasing eigenvalues of S, define variance-maximizing directions in the data set. PCA is a legitimate feature reduction method. However, two of its characteristics argue against using it to extract features from spectra. The first is that the PCs, being linear combinations of the original spectral features, tend to lose or at best mask useful original spectral information. This motivated the development of our subset selection algorithms, which are designed to retain
situations the number of training samples is invariably finite. Under these circumstances the so-called ‘peaking phenomenon’ \(^7,8\) will likely occur, i.e. beyond a certain number of features, classifier performance will begin to degrade. Thus, there is an optimal number of features, and this number depends on the classifier, on the characteristics of the features, and on sample size. For the LDA classifier, which uses the inverse of the pooled covariance matrix \(S\), there is a maximum number of theoretically allowed attributes \(M: M \leq N - 1\) (where \(N\) is the number of samples). If \(M\) is larger than \(N - 1\), then \(S\) becomes singular and its inverse does not exist. In practice, even for \(M\) less than but close to \(N - 1\), matrix inversion is often unstable numerically, leading to unreliable results. This is true even if the generalized inverse (SVD) is used, as we do in our implementation of LDA. In fact, in practice it is advisable to restrict the number of attributes to be less than \(\min\{N_k\}, k = 1, 2, \ldots, C\), where \(N_k\) is the number of spectra in the \(k\)th class.

Before we introduce our subset selection procedures, it is worthwhile stating the criteria used to assess the quality of any particular subset. The goal is to create from the spectra, (the objects to be classified) a limited number of features that will ensure reliable, robust classification. The additional constraint we impose is that the features must retain their spectral identity: hence the subsets chosen must be (or at least have an obvious one-to-one correspondence with) spectral subregions. Since feature space reduction is essential for robust classification, we find that some preprocessing/transformation in any subregion is invariably beneficial. Such transformation defines the feature derivable form the original spectral intensities in the subregion. The most common form of reducing transformation is averaging, i.e. replacing the \(K\) intensities in the subregion by their mean value. This single number then is the new feature. Of course, in certain circumstances other transformations might be better; on occasion we have successfully used variances or ratios of averaged intensities.

Having defined our features, we now need an objective (fitness) function \(F\) to rank any subset of the original complete set. Since our primary goal is improved classification, a natural choice for \(F\) is the mean square error between the training set classification results (probability \(p_{jq}\) that spectrum \(j\) is in class \(q\)) and the a priori class indicator \(I_{jq}\) (=1 if in class \(q\), 0 otherwise) for that spectrum:

\[
F = \frac{1}{N.C} \sum_{q=1}^{C} \sum_{j=1}^{N} (p_{jq} - I_{jq})^2
\]

Such an objective function will simultaneously maximize the overall accuracy for the training set and produce crisper (less fuzzy) class assignments. The choice of the classification method to obtain the \(p_{jq}\) is also natural. It has been shown\(^9\) that LDA is the best choice for a small training set, even if conditions, essential for optimal LDA classification (normality, common class covariance matrices) are not met. Furthermore, the leave-one-out (LOO) cross-validation can be carried out with minimal additional cost. This is because one can use a standard matrix identity for a direct update of the matrix inverse \(S^{-1}\) of the pooled covariance matrix \(S\). The complete \(S\) need be inverted only once, at the beginning, and from \(S^{-1}\) all \(S_{(k)}^{-1}\) (the subscript \((k)\) indicates that the \(k\)th spectrum was left out) are obtained by fast vector and matrix multiplications. Our computational experience suggests that a complete LDA/LOO for 100–200 samples is only marginally slower (~5%–10%) than a single LDA/resubstitution calculation. Of course, any other classifier can be used to optimise \(F\). Naturally, fast and robust classifiers are preferable.

It is important to appreciate the difference between using all the data and cross-validating the classifier with e.g. LOO, and partitioning the data into training and test sets, the latter being only used to validate the training set-based classifier’s accuracy and robustness. Using the entire data set, especially when it is of limited size, may be the only recourse. Cross-validation merely reduces the optimistic bias that a classifier might have, and provides a more realistic assessment of the expected classification error. Contrary to standard practice, where the actual classifier used to assign new samples to classes is a resubstitution-based one, we use the average classifier obtained from the cross-validation process (e.g. with \(N\) samples, \(M\) features and LOO, there are \(N\) (slightly) different classifiers produced. We simply average the \(M+1\) LDA coefficients). This turned out to give realistic results when additional samples became available for validation.

When there are enough samples to partition the data, the ever-present danger is overfitting, i.e. producing excellent classification results on the training set and...
Our method is a modified, improved version of the sequential forward subset selection algorithm,\(^{10,11}\) found in commercial software, such as SAS or Statistica. For example, SAS uses two methods of subset selection—forward selection and backward elimination. Backward elimination is simply not feasible computationally, due to the large number of data points comprising typical MR spectra. Thus the only choice left is forward selection (FS). However, the standard version of FS has a potentially serious drawback, since it cannot guarantee that combining the best \(K\) single attributes will work better than the combination of \(K\) individually inferior ones. (We have already encountered such a situation when the individual attributes were PCs.) By always choosing the best current attribute, there is the danger of eliminating potentially ‘good’ combinations. This danger is especially great on the very first step. We reduce this risk by not confining the process to the best single attribute, but using the FS scheme on all single attributes.

Feature selection based on the branch-and-bound algorithm\(^2\) cannot be usefully implemented because it requires that the objective function be ‘monotonic’ on the subsets of the original feature set, i.e. the function must decrease if we add new features. This condition is met with the Bayesian classifier. Unfortunately, our classifiers are only approximations to a true Bayesian one and additional ‘nuisance’ features can easily deteriorate the classifier’s performance. Because we can only approximate the true Bayesian classifier, the objective function is generally not monotonic, a necessary condition to guarantee that the optimal subset be found.

Assume a data set consisting of \(N\) samples with known class labels from \(C\) classes. Let \(L\) be the number of attributes characterizing these samples (i.e. \(L\)-point MR spectra) in the original data set. Our goal is to select a subset \(K < L\) of attributes which optimizes some objective function. For the prespecified number of \(K\) attributes we have \(L! / K! (L-K)!\) possible combinations. Such a dependence on \(K\) does not allow exhaustive search for large or even moderate \(K\). Therefore, we have chosen a faster way of selecting a (sub)optimal subset of attributes.

The data set is a collection of \(L\)-point spectra \((L \sim O(1000), i.e. of order 1000)\). First, we reduce the dimensionality of the feature space, since even EFS cannot deal with 1000 attributes in realistic computation times. The reduction may be achieved by subsampling: \(M < L\) spectral subregions are formed by selecting \(L/M\) consecutive data points. These \(L/M\) points in each subregion are then processed to form \(M\) features. The EFS algorithm can deal with \(M \sim 100 - 200\).

The EFS algorithm is then used to further reduce the \(M\) subregions to a \(K\)-member subset. \(K\) is an input parameter that in practice we limit to \(\sim 10\) or less. The subregions retain their spectral (hence biochemical) identities, one of the reasons this preprocessing approach was developed.

On the first step of the algorithm, \(M\) sets of attributes are initialized by placing the individual attributes in corresponding sets \((\Omega_1 = \{1\}, \ldots, \Omega_M = \{M\})\). Now for \(j = 2, \ldots, K\) we repeat the same process, each \(\Omega_i\), \(i = 1, \ldots, M\) is augmented by the attribute that gives the best classification result in combination with those already included in \(\Omega_i\). After the selection process is complete, we choose the best out of the various \(\Omega_i\), \(i = 1, \ldots, M\). This process is illustrated by the following diagram:

\[
\begin{align*}
\Omega_0 &= \{1\}, \quad \Omega_0 = \{1\} \\
\Omega_1 &= \{1, 2\}, \quad \Omega_1 = \{1, 2\} \\
\vdots \\
\Omega_M &= \{1, 2, \ldots, M\}, \quad \Omega_M = \{1, 2, \ldots, M\}
\end{align*}
\]

For instance, the sequence \(\Omega_1 = \{1\}, \{1, z_1^{(2)}\}, \ldots, \{1, z_1^{(2)} \ldots z_1^{(K)}\}\) means that the first feature \(\Omega_1 = \{1\}\) is updated by, say, the feature \(z_1^{(2)} = j \neq 1\) that gives the best two-feature classification involving feature 1. At the next stage the procedure is repeated for the best third feature not equal to 1 and \(j\), and this process is continued until the prespecified number \(K\) of features is reached. The entire process is then repeated for the other \(\Omega_i\). This is in contrast to standard stepwise (sequential) forward selection methods, which do not \textit{tract} all the attributes. In fact, the standard FS corresponds to the single process in the above diagram that starts with the \textit{best} individual attribute, say, \(\Omega_j = \{J\}\). EFS does not guarantee finding the optimal subset of attributes because it does not allow backtracking. However, the proposed enhancement helps prevent the premature elimination of important attributes and always produces a good suboptimal subset.

We have first used EFS to select an improved subset of principal components in a principal component analysis, with promising results.\(^3\) Of course, any attribute set can be used; however, EFS’s major application is to spectra. Despite its successes, EFS has some limitations: (1) It is a suboptimal algorithm; there is no guarantee that the best set of subregions will be found. Since full backtracking is computationally prohibitive, the algorithm may continue along a suboptimal branch and never find the global optimum. (2) EFS operates on subregions of equal size and the algorithm lacks the flexibility to merge or overlap these. (3) EFS is relatively slow (although not excessively so). Choosing a near-optimal subset of 10 features out of 100 takes approximately 30 min on an SGI workstation, with performance equivalent to that of a 200 MHz PC.)
The main drawback of EFS is its inflexibility, i.e. the boundaries of the averaging window are fixed, and only the size of the window can be varied. Trying to rectify this defect led to the idea of using a genetic algorithm (GA)-based approach, with the freedom to select interesting subregions of arbitrary size.

GA-guided optimal region selection (ORS) for spectra

Because our problem is special, we did not use the standard GA\textsuperscript{14–16} but designed and implemented a problem-specific version. There are two parts to our GA implementation: (1) Mapping the original attribute space onto a bit string set; (2) designing an overall scheme to create the population and allowing its evolution with subsequent generations. (The objective function \( F \) that drives the algorithm has already been defined.) Before we describe these in detail, we present a simple pseudo code for the overall operation of the algorithm:

1. Select \( M \), the maximum number of desired attributes/subregions, \( G \), the number of generations, \( P \), the size of the population.
2. Create \( P \) binary strings of length \( L \) (\( P \) chromosomes), each containing \( M \) subregions, i.e. different sets of contiguous but not overlapping ones; the remaining chromosome locations are filled with zeroes.
3. Process the \( M \) subregions to derive the \( M \) features.
4. For each of the \( P \) strings evaluate the previously defined fitness function \( F \), applying an \( M \)-feature LDA/LOO classifier to the training set. Sort the \( P \) fitness values in ascending order (because of our definition of \( F \), the lower its value the fitter the chromosome). In this sorted list the 10 best are referred to as the elite.
5. ‘Breed’ the population by mutation and/or crossover, keeping the elite. Steps 3–5 constitute one generation.
6. Go to step 3 and repeat until the number of generations equals the preset \( G \).

Mapping

The problem of mapping the original feature space onto a bit string (zeroes and ones) has not received much attention in the literature. A possible reason is that the typical feature set for the standard GA application is a collection of numerical values, each represented by its own bit string in the computer. We have a somewhat different problem. A spectrum is a set of \( L \) different intensity values, one at each of \( L \) frequencies. The natural and simplest mapping onto a bit string is to put \( L \) ‘zeroes’ into a logical array (a ‘chromosome’) of \( L \) dimensions. Selecting subregions from the spectrum translates to converting some of the ‘zeroes’ into ‘ones’.\textsuperscript{17} However, because we plan to do more than simply eliminate a subset of the frequencies, we have extended the above simple representation. Thus, further dimension reduction is possible if in each subregion we replace the individual intensities by fewer attributes via additional processing. Limiting cases of such processing would create and associate a single attribute, e.g. the average value, or the variance, with each subregion.

To achieve such flexibility, we treat a spectrum as a set of segments, each comprised of a pair of adjacent data points. Hence, an \( L \)-point spectrum becomes an \((L-1)\)-point ‘chromosome’. Now a ‘one’ in the \( i \)th position means that points \( i \) and \( i+1 \) are connected. A ‘zero’ means the converse. A set of consecutive ‘ones’ corresponds to an attribute range. Therefore, any given chromosome is a combination of a set of connected spectral subregions, and unconnected spectral points. This is illustrated below:

\[
\begin{array}{ccccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\
1 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 \\
\end{array}
\]

\[
\begin{array}{ccccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 \\
\end{array}
\]

With the above construction we have two opportunities to reduce dimensionality: by selecting subregions, and by further processing (by some transformation) the data points within the subregions. If this transformation is averaging, then the influence of noise is also decreased. We have used averaging.

The GA Architecture

We now introduce the genetic operators we use, and the details of how the population is created for any given generation. There are two canonical genetic operators, mutation and crossover. Our implementation of crossover is the same as that described in the literature. We select two parents randomly from the current population, and choose a random crossover point at which the chromosomes are to be cleaved. Exchanging the parts of the parent chromosomes creates two new chromosomes. This operation makes GA very flexible and enables the search to move far away from the initial locations in the high dimensional feature space. The other genetic operator, mutation, is specifically tailored to our problem. Its distinguishing characteristics are that it is not a single-point operator, and that its size evolves with evolving generations. Extensive experimentation showed that in early generations we must change more than one bit in our chromosomes, hence the introduction of a \( k \)-block mutation. A large mutation size (large \( k \)) allows a rapid but coarse-grained initial exploration of the feature space, but as the process evolves, we need finer tuning, hence
the size $k$ of the $k$-block mutation decreases with increasing generations. The initial size of the $k$-block mutation is $1/64$th of the full spectral range, i.e. $k = L/64$, and $k$ is decreased gradually as the optimisation proceeds.

The input parameter set consists of the size $P$ of the population, the maximum number of subregions allowed, the number of generations $G$, and the mutation $p_M$ and crossover $p_C$ probabilities. The former is set to 0.001, the latter to 0.66.

To help create an initial population that is representative of the data, we build a separability profile using the $r^2$ statistics. For two-class problems this computed statistic is stored for each spectral position. For many-class problems all pairwise statistics are computed and the profile is produced as their sum. Then the initial subregions are chosen randomly but with relative probabilities proportional to the profile.

The algorithm starts by generating a population of $P$ random bit strings according to the above initialization. $P$ is typically 200–600. A generation consists of (1) changing members of the population by the above outlined mutation and crossover rules (‘breeding’), (2) sorting the entire population according to their fitness ($F$) values. The best $P_1$ strings (the elite, $P_1 \sim 10$) are kept intact, the rest, $P - P_1$, are produced randomly by a process governed by the assigned mutation and crossover probabilities, and by the current fitness ranking. Two random chromosomes are chosen, with probabilities depending linearly on the rank of the chromosomes in the ordered generation list. They are first mutated with probability $p_M$, and either mated via crossover with probability $p_C$, or added directly to the new generation. This process is continued until a new generation is formed. We stop when either the size of point mutation is zero or the pre-set number of generations $G$ has been reached. $G$ is obviously problem-dependent, but for most applications $G \sim 50–100$ was sufficient.

Our mapping allows a huge number of potential features for classification, up to half the size of the full spectrum. However, any classifier that uses many attributes will be unstable and slow. Therefore, in our GA implementation we restrict the maximum number of features allowed, treating it as an input parameter ($M$) under user control.

GA-ORS is respectably fast. With parameters $P$ (population size) = 200, $L$ (number of points in the spectra) = 1024, $G$ (number of generations) = 30, and $M$ (number of subregions required) = 10, the processing time on the SGI workstation referred to earlier is $\sim 25$ min.

DATA AND MATERIALS

Tissue specimens

Multiple surgical biopsies (0.02–0.1 g) were obtained during resection from brain tumours. The colorectal biopsies (0.004–0.1 g) were obtained from both surgery and colonoscopy. Each biopsy was individually placed in phosphate-buffered saline in $D_2O$ and frozen in liquid nitrogen within 1 min of excision. After thawing, between one and four samples of suitable size (0.02 to 0.06 g) were cut from each brain biopsy. Similarly, one or two samples (0.004–0.05 g) were cut from each colon biopsy. Each sample was placed in a capillary (i.d. of 2.5 or 3.2 mm) as described previously. 18

Magnetic resonance spectroscopy

$^1$H MR spectra were obtained on a Bruker AM360 spectrometer at $37^\circ$C without spinning of the sample and with low power presaturation of the signal from water (20 dB of 0.2 W). The spectra were acquired with the following parameters: 256–640 acquisitions, spectral width 5 KHz, recycle delay 2.41 s, time domain 4 K.

$^1$H MR spectra of tissue biopsies of meningiomas (95 spectra) and astrocytomas (74 spectra) of human brain neoplasms 19 and of normal (50 spectra) and cancerous colorectal tissue (48 spectra) formed our data sets. Both sets of spectra were normalized to the area, and when necessary, aligned to an internal reference peak. In both cases the magnitude spectra were used. (Our extensive experience suggests that phasing is superfluous, and even deleterious for reliable classification.) The regions of interest were 450 data points ($\sim$0.5–3.6 ppm) for the spectra of colorectal biopsies, condensed into 90 subregions, by averaging five adjacent points, and 550 points ($\sim$0.27–4.0 ppm) for the astrocytic spectra, collapsed into 110 subregions. These were our attributes for a ‘baseline’ LDA classification, the results made more robust by cross-validating with the leave-one out (LOO) method. The EFS algorithm started with these 90 (colon) and 110 (brain) attributes, and was required to find the nine and ten best subregions, respectively, by minimizing the objective function $F$. For the sake of a fair comparison with the EFS algorithm, the GA method was also required to find the nine and ten best subregions, but used as input attributes the entire spectral range, without subregion condensation.

RESULTS AND DISCUSSION

The classification results on the colorectal biopsies are collected in Table 1. We did not partition the data set into training and test sets because of the relatively few spectra available per class. The 450-point spectra were reduced to 90 equal subregions (designated by 90 (Raw) in Table 1), by averaging every 5 adjacent points. (For this data set of 98 samples, 90 is near the maximum number of patterns that can be classified with LDA without numerical instabilities due to near-colinearity). The cross-validated classification accuracies for the mini-

mally preprocessed 90 (Raw) classification are (normal, cancerous) (69%, 85%), with an overall average of 77.4%. These are improved to (98%, 100%), (average 98.9%) after EFS-ORS preprocessing (Table 1). Only 45 of the 450 data points were utilized, forming the nine final subregions (attributes). These are (in ppm): 0.80–0.84, 1.82–1.85, 1.85–1.89, 2.02–2.06, 2.23–2.26, 2.53–2.57, 2.73–2.77, 3.28–3.31 and 3.31–3.34. The GA-ORS preprocessor produced an improvement essentially equivalent to the EFS results. The only noticeable difference between the two is that GA-ORS classified crisply 96.9% of the original samples, whereas EFS-ORS achieved a slightly worse 91.8%. GA-ORS used 48 of the 450 data points; the subregions ranged from 2 to 9 points in width, compared to the uniform 5-point subregions of EFS-ORS. The nine subregions selected are (in ppm, number of data points in parentheses): 0.72–0.73 (2), 1.00–1.05 (9), 1.48–1.50 (4), 1.68–1.72 (7), 1.86–1.89 (6), 2.74–2.77 (5), 2.87–2.92 (8), 3.28–3.30 (4), 3.31–3.32 (3). There is considerable overlap between subregions selected by EFS and GA. This is gratifying, since there is no a priori reason to expect this from a purely algorithmic viewpoint. Several comparably good solutions are found routinely both by EFS and GA. Such agreement implies that biochemical understanding of differences between the normal and diseased state is possible.

The classification results on meningiomas vs astrocytomas are collected in Table 2. The classification approach mimics the one used for the normal vs colon cancer samples. The original 550-point spectra were reduced by averaging to 110 (Raw) for the baseline LDA classification, and to serve as starting subregions for the EFS-ORS. GA-ORS does not need such averaging and starts with the full 550-point spectral region. The 100 (Raw) results are considerably better than for the colon classification: (M, As) (85%, 90%), 87.6% overall. These are again improved by both EFS (97%, 100%), 98.8% overall, and GA (98.8%, 100%), 99.4% overall. Only 50 of the 550 data points were utilised by EFS, forming the 10 final subregions (features). These are (in ppm): 1.19–1.22, 1.73–1.76, 2.41–2.44, 3.12–3.15, 3.32–3.36, 3.36–3.39, 3.46–3.49, 3.63–3.66, 3.70–3.73, 3.76–3.80. GA-ORS used 75 of the 550 data points; the subregions ranged from 3 to 17 points in width, compared to the uniform 5-point subregions of EFS-ORS. The 10 subregions selected are (in ppm, number of data points in parentheses): 0.38–0.44 (9), 1.12–1.13 (3), 1.71–1.74 (6), 2.77–2.84 (11), 2.99–3.07 (13), 3.13–3.14 (3), 3.32–3.34 (3), 3.45–3.48 (6), 3.60–3.71 (17), 3.74–3.76 (4). Again, there is considerable overlap between subregions selected by EFS and GA.

In Table 3 we display the classification outcome for the meningioma vs astrocytoma problem when the data are partitioned into independent training and test sets. We have assigned randomly 111 of the 169 spectra to the training set and 58 to the test set. We wanted to see whether comparable accuracies are achievable on test sets that the classifiers were not trained on, whether the optimal regions selected are related to those obtained for the full data set, and whether the relatively good results obtained with 110 (Raw) using all samples were reliable. Inspection of Table 3 reveals that the results are much poorer (~53% for the training and ~49% for the test set overall) than the corresponding outcome for the full 110 (Raw). This is to be expected, given the near-equality of sample and attribute numbers, but such practical confirmation speaks more eloquently. What we observe

<table>
<thead>
<tr>
<th>Class</th>
<th>N</th>
<th>A</th>
<th>C</th>
<th>A</th>
<th>C</th>
<th>A</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>50</td>
<td>69.4</td>
<td></td>
<td></td>
<td></td>
<td>98.0</td>
<td>92.0</td>
</tr>
<tr>
<td>Cancer</td>
<td>48</td>
<td>85.4</td>
<td>97.9</td>
<td>100.</td>
<td>100.</td>
<td>91.7</td>
<td>97.9</td>
</tr>
<tr>
<td>Overall</td>
<td>98</td>
<td>77.4</td>
<td>99.0</td>
<td></td>
<td></td>
<td>98.9</td>
<td>91.8</td>
</tr>
</tbody>
</table>

90 (Raw) stands for 90 subregions without any feature selection, 9 (EFS) indicates preprocessing with the enhanced forward selection-based ORS, using nine subregions, 9 (GA) refers to the nine-subregion genetic algorithm-based ORS. N denotes the number of spectra in the class; A stands for percent accuracy, C for the percentage of the samples classified crisply, i.e. with probability greater than 75%. All classifications were carried out with LOO cross-validation. Only the clinically relevant crisp i.e. unambiguous results are displayed.

<table>
<thead>
<tr>
<th>Class</th>
<th>N</th>
<th>A</th>
<th>C</th>
<th>A</th>
<th>C</th>
<th>A</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>95</td>
<td>85.1</td>
<td>98.9</td>
<td></td>
<td></td>
<td>97.5</td>
<td>85.3</td>
</tr>
<tr>
<td>As</td>
<td>74</td>
<td>90.1</td>
<td>95.9</td>
<td>100.</td>
<td>78.4</td>
<td>100.</td>
<td>85.1</td>
</tr>
<tr>
<td>Overall</td>
<td>169</td>
<td>87.6</td>
<td>97.6</td>
<td></td>
<td></td>
<td>98.8</td>
<td>82.2</td>
</tr>
</tbody>
</table>

See legend for Table 1.

for $N = 111$ and $M = 110$ is that at this theoretical limit for sample size vs attribute number even the generalized matrix inversion is unreliable. This numerical instability is further exacerbated by the LOO procedure, because updates of the generalized inverse depend heavily on correctly estimating the rank of the nearly-singular $S$, which is done $N$ times with LOO. In contrast, both the EFS- and GA-preprocessed data produced impressive results. For EFS, the overall accuracy for training and test sets was (97.8% and 95.0%), whereas for GA this was an even better (99.1% and 100%)! EFS still used 50 of the 550 data points, with the following 10 subregions (in ppm): 1.12–1.15, 1.42–1.46, 1.70–1.73, 2.85–2.88, 3.29–3.32, 3.39–3.42, 3.49–3.53, 3.63–3.66, 3.70–3.73, 3.83–3.86. The more accurate GA utilized 85 of the 550 points, with the following 10 subregions (in ppm, number of data points in the subregion in parentheses): 0.44–0.48 (6), 0.81–0.86 (9), 1.72–1.75 (5), 2.25–2.27 (3), 3.11–3.19 (14), 3.32–3.36 (6), 3.36–3.41 (8), 3.45–3.51 (9), 3.61–3.70 (15), 3.74–3.80 (10). Again, there is considerable overlap between the EFS and GA subregions; even more interestingly, there is also excellent overall agreement between EFS (all samples) and EFS (train & test), as well as between GA (all samples) and GA (train & test).

Note that all classification results displayed in the Tables are based on the crisply classified spectra. As an example, in Table 2 in the two columns beneath 10(EFS), $A (%\text{ accuracy}) = 97.5$ for class $M$. This classification accuracy applies to the 85.3% (C) of the 95 spectra that the classifier could assign with greater than 75% probability to class $M$. We maintain that only such unambiguous class assignments would be useful and relevant in a clinical environment, where uncertain class identification must lead to additional tests, experiments.

**CONCLUSIONS**

Comparison of the GA-based ORS method with its EFS-based counterpart reveals both similarities and differences. Both produced reduced feature sets that enabled the simple LDA/LOO classifier to produce highly accurate results. GA does have several advantages over EFS. The most striking is its ability to select attributes (subregions) of varying sizes. This may become crucial for difficult classification problems, particularly involving *in vivo* spectra. This is because the ability to create features built from spectral subregions of varying widths helps in restricting the optimal features to a reasonably low number, yet a larger fraction of the full spectral range could be 'sampled'. For instance, when classifying meningiomas vs astrocytomas, EFS used 50 of the original data points to produce 10 features, whereas GA, in selecting 10 features, recruited 75 data points and produced a classifier that was better in every respect than the EFS-based classifier. Furthermore, region condensation, an essential pre-preprocessing step for EFS, is done dynamically and automatically by GA-ORS. Last but not least, GA-ORS is generally faster, although this depends strongly on the data.

---

**REFERENCES**

12. Narendra P. M. and Fukunaga K. A branch and bound algorithm for


