Creating robust, reliable, clinically relevant classifiers from spectroscopic data

R. L. Somorjai

Received: 27 October 2009 / Accepted: 31 October 2009 / Published online: 25 November 2009

© International Union for Pure and Applied Biophysics (IUPAB) and Springer 2009

Abstract I describe in detail the intimately connected feature extraction and classifier development stages of the data-driven Statistical Classification Strategy (SCS) and compare them with current practice used in MR spectroscopy. We initially created the SCS for the analysis of MR and IR spectra of biofluids and tissues, and subsequently extended it to analyze biomedical data in general. I focus on explaining how to extract discriminatory spectral features and create robust classifiers that can reliably discriminate diseases and disease states. I discuss our approach to identifying features that retain spectral identity and provisionally relate these features, averaged subregions of the spectra, to specific chemical entities (“metabolites”). Particular emphasis is placed on describing the steps required to help create classifiers whose accuracy doesn’t deteriorate significantly when presented with new, unknown samples. A simple but powerful extension of the discovered features to detect metabolite-metabolite (feature-feature) interactions is also sketched. I contrast the advantages and disadvantages of using either spectral signatures or explicit metabolite concentrations derived from the spectra as sets of discriminatory features. At present, no clear-cut preference is obvious and more objective comparisons will be needed. Finally, I argue that clinical requirements and exigencies strongly suggest adopting a two-phase approach to diagnosis/prognosis. In the first phase the emphasis ought to be on providing as accurate a diagnosis as possible, without any attempt to identify “biomarkers.” That should be the goal of the second, research phase, with a view of providing prognosis on disease progression.

Keywords Statistical classification strategy · Feature selection · Classifier development · Spectral signature · Metabolite mixture · Crossvalidation · Bootstrap · Dissimilarity/distance measure · Genetic algorithm

Abbreviations

AUC	Area under the curve
CV	Cross-validation
EE	Error estimate
FSD	Feature space dimensionality
GA_ORIS	Genetic algorithm-driven optimal feature selector
LDA	Linear discriminant analysis
LOO	Leave-one-out
MM	Metabolite mixture
MON	Monitoring
QDA	Quadratic discriminant analysis
ROC	Receiver operating characteristics
SCS	Statistical Classification Strategy
SFFS	Sequential forward floating selection
SFR	Sample-to-feature ratio
SS	Spectral signature
TR	Training
TS	Test
WCVBTS	Weighted cross-validated bootstrap

Introduction

The need to create noninvasive, robust diagnostic/prognostic clinical procedures has suggested using spectroscopic techniques. Magnetic resonance (MR), infrared (IR), Raman, and mass spectroscopy are the prime candidates

R. L. Somorjai (✉)
Biomedical Informatics Group, Institute for Biodiagnostics,
National Research Council Canada,
435 Ellice Ave.,
Winnipeg, MB R3B 1Y6, Canada
e-mail: Ray.Somorjai@nrc-cnrc.gc.ca

for analyzing biofluids and tissues. MR spectroscopy can also be applied in vivo, a significant advantage in the clinic.

Given that it is feasible to acquire biomedical spectra noninvasively, the next step is to develop analytical procedures that can reliably classify these spectra (e.g., assign them with high confidence to either the disease or the healthy group). Unfortunately, biomedical spectra are plagued by the twin curses of dimensionality (Bellman 1961) and dataset sparsity (Somorjai et al. 2003). The first curse is active because d , the dimensionality of spectral feature space is typically very high, $d=O(1,000)$ to $O(10,000)$; for spectra, the initial features are the various spectral intensity values at the measurement frequencies. The second curse arises because of the difficulty and/or cost of acquiring a statistically meaningful number (N) of biomedical samples; frequently N is only of $O(10)$ to $O(100)$. (The deleterious effect of sparseness is further exacerbated by sample imbalance: the number of samples from diseased patients is typically three- to tenfold smaller than from control cases.) A small N leads to a sample-to-feature ratio (SFR) N/d that is $1/20$ to $1/500$, instead of the preferable 5 to 10 or even larger ratio (Jain and Chandrasekaran 1982). The latter SFR values are needed in order to develop a classifier with high generalization capability, i.e., one that assigns correctly and with high probability samples of unknown provenance to their proper classes.

An appropriately large SFR value seems to be a necessary condition. However, even if this necessary condition is met, sufficiency is not guaranteed for small sample sizes; apart from a few recent exceptions (Simon et al 2003; Dougherty 2001), this latter caveat has not been fully appreciated (Somorjai et al 2003).

How do we arrive at such reliable analytical procedures? “There are no panaceas in data analysis” (Huber 1985), i.e., no single, data-independent, best “black box” classification algorithm could possibly exist for the wide range of biomedical datasets one encounters. As a consequence, the choice of preprocessing methodology, classifier development, etc., is necessarily data-dependent and should be data-driven. This can be achieved by formulating and realizing a flexible classification **strategy**. This is the objective we set for ourselves and tried to meet over the last dozen years (Somorjai et al. 2004a).

Our strategy is called the Statistical Classification Strategy (SCS). It evolved in response to the need to robustly classify biomedical data that suffer from the above twin curses. In particular, the strategy has been formulated with clinical utility in mind: not only should the eventual classifier provide accurate, reliable diagnosis/prognosis, it should also predict class membership with the fewest possible discriminatory features (attributes). Furthermore, these few features should be interpretable in terms of biochemically, medically relevant entities (“biomarkers”).

These two interrelated aspects are not always appreciated and considered for the development of classifiers of clinical relevance. In particular, what is considered interpretability is a contentious issue.

Because of the twin curses, reliable classification of biomedical data, and spectra in particular, is especially difficult and demands a “divide and conquer” approach. Relying on this approach, our strategy evolved gradually and now consists of five stages. In Fig. 1 we show the interrelation among these five stages of the SCS. All these stages are data-driven and only the ultimate goal, data $\rightarrow$ results is immutable; frequent and repeated backtracking between the individual stages is generally necessary and frequently has proved to be beneficial. Although all stages are important, here I shall focus on the two crucial ones, feature extraction and classifier development, and the essential interplay between them. This is dictated by and predicated upon our wrapper-based approach to feature selection/extraction/generation.

Feature selection/extraction/generation

Given the increasing emphasis on interdisciplinary research, I assume that prior to the data analysis, the spectra went through all necessary steps of quality control that the experimentalist could identify and realize (see e.g., Weljie et al. 2006; Goodacre et al. 2007). (Unfortunately, such quality control is not always present

Fig. 1 Flowchart for the stages of the statistical classification strategy

or evident in publications dealing with classification of biomedical data.) Furthermore, for the sake of focus, I assume that the first two stages of the strategy have already been accessed and explored (at least temporarily). Thus, at the visualization stage (stage 1) potential outliers have been identified and removed (Somorjai et al. 2004b), and at stage 2 various preprocessing techniques were investigated and implemented, among them peak alignment, normalization, spectral filtering, as well as various data transformations (linear and nonlinear, e.g., power, logarithmic and exponential, derivative computation, rank ordering, etc.) (Dolenko and Somorjai 1995; Somorjai et al. 1999). The consequences of the stage 2 operations can then be visualized and assessed by backtracking to stage 1. For further details, including the ultimate classifier aggregation stage (if necessary), consult Somorjai et al. (2004a) and Kuncheva (2004).

Thus, we arrive at one of the two crucial stages of the classification strategy, the reduction of feature space dimension. Its primary role is to lift the curse of dimensionality, i.e., increase the SFR to the acceptable 5–10 range. This is both essential and feasible for spectra, for which the majority of the d spectral features is either redundant (correlated) or irrelevant (“noise”).

Feature space dimensionality (FSD) reduction

FSD reduction can be accomplished in two ways. Filter methods do not rely on any knowledge of the eventual classifier algorithm. Wrapper methods, which tie FSD reduction intimately to some prediction/classification algorithm, tend to perform better, but are more biased and are slower than the filter methods, especially since they generally use some form of crossvalidation (Kohavi and John 1997). Furthermore, overtraining is more of a threat for them than for filter methods.

For either method, FSD comes in two major “flavors”. Feature selection finds a subset of the original features. This is the only course of action when feature adjacency (e.g., consecutive spectral data points) has no physical relevance. Feature extraction is the more general approach: it involves obtaining some functional combination of the original features. [The simplest possible and frequently advantageous if suboptimal procedure, used routinely by spectroscopists, is binning, i.e., the reduction of the original d data points into d/D (D an integer ≥ 2) new spectral features, each constructed by averaging D adjacent data points. The optimal D is obtained by trial-and-error.]

Partial least squares discriminant analysis (Sjöström et al. 1986; Eriksson et al. 2001) may be viewed as a particular variant of a wrapper method for feature selection. The N d -point discretized spectra are replaced by d “latent variables”. These new, orthogonal features are linear

combinations of the original d spectral features, but generally bear little resemblance to them. However, unlike principal component analysis (Jackson 1991), they are constructed to maximize the correlation with the matrix $\mathbf{Y}$ of class labels. (For two-class problems, $\mathbf{Y}$ is an $N \times 2$ dummy matrix, with entries 0 and 1. For class 1 samples, column 1 contains ones, for class 2, zeros; in column 2 the situation is reversed.)

Additional, threshold-based processing (using score plots and weight/loading plots) is required to identify which of the original features is relevant. This feature scrambling may hinder direct interpretability.

Such interpretational problems led us to develop a specific feature extraction method (Nikulin et al. 1995, 1998), whose most important advantage is that it retains spectral identity. The new features are functions (typically the averages) of adjacent spectral data points and hence readily interpretable.

Our feature extraction approach uses a genetic-algorithm-based optimal region selection algorithm (GA_ORS) (Nikulin et al. 1998). The genetic algorithm was chosen because it is a global minimizer, working with a string of zeroes and ones. Its inputs are F , the maximum number of features (distinct subregions) required, two random seeds for initialization, and the usual parameters of GA optimization: crossover and mutation rates, number of generations, and number of “elite” chromosomes. The objective function that the GA_ORS optimizes has been devised to simultaneously minimize the misclassification error and the classification “fuzziness” (i.e., increase class assignment probabilities). For spectra, f_k consecutive 1s would be required for the eventual creation of g_k , the k th feature, $k=1, \dots, F$. Unique to our method is that we can specify the type of required mathematical operation on f_k adjacent data points while running GA_ORS. We generally use averaging: a new feature g_k is the average of the spectral intensities of f_k adjacent, original data points. For MR spectra, such averaging is particularly meaningful, since the averaged regions frequently correspond to and estimate specific peak areas. A bonus is that averaging increases the signal-to-noise ratio (i.e., corrects partially for the presence of noise). Clearly, GA_ORS provides an optimized version of binning, because the F g_k s may be interpreted as F bins with different, optimized widths; F , the number of features, is an input parameter, but unlike D in binning, its optimal value is directly tied to the desired classifier optimization.

We may also specify what type of classifier [linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), k -nearest neighbors (k -NN), support vector machine (SVM), etc.] will be used in this wrapper-based feature extraction approach. Because of its low complexity, we usually employ LDA as the classifier, with leave-one-out (LOO) crossvalidation. Using proper feature extraction,

even the simplest classifiers frequently outperform more sophisticated (e.g., nonlinear) variants (Holte 1993; Simon et al. 2003).

We incorporated into the GA_ORs software several algorithmic enhancements and constraints. These turned out to be both necessary and useful for reducing the possibility of bias. Thus, more attention is being paid to avoid selecting spectral subregions that are only accidentally discriminatory due to sample scarcity. The following are among the most useful constraints:

1. Minimize the maximum misclassification error among the classes (for two-class problems this equalizes false positive and false negative rates).
2. Penalize classification errors differently for different classes (relevant when there is considerable imbalance between class sample sizes).
3. Set lower and upper bounds on the allowed spectral subregion widths. (This is important for MR and mass spectra, for which usually only relatively narrow regions, 5–50 adjacent intensity values, are chemically relevant or meaningful.)

When the data can be split into training (TR) and test (TS) sets, we augmented GA_ORs's objective function, such that in addition to the conventional root mean square (RMS) error, the quantity $\delta\text{AUC} = |\text{AUC}(\text{TR}) - \text{AUC}(\text{TS})|$ is also minimized. In particular, we minimize the sum of RMS and δAUC . AUC is the area under the ROC curve, and minimizing δAUC helps equalize the errors for the TR and TS sets while searching for optimal features, thus controlling overfitting.

The GA_ORs approach has been very successful [for its application to the classification of biomedical MR and IR spectra, see Somorjai et al. (1995), and the review of Lean et al. (2002) and references therein].

A good discussion of the conventional feature selection approaches can be found in Jain et al. (2000). The authors assert that the sequential forward floating selection (SFFS) method (Pudil et al. 1994) is particularly powerful and near-optimal. SFFS is a generalization of the “plus- s minus- r ” method, which enlarges the feature subset by s features using forward selection and then deletes r features using backward selection. In SFFS the values of s and r are determined automatically and updated dynamically. An improved, adaptive version was also announced (Somol et al. 1999). However, in our experience, even the adaptive SFFS is less successful than the GA_ORs, especially when the number of original features is large.

In general, no non-exhaustive feature selection method can be guaranteed to find the subset that produces the smallest classification error (optimal subset). In fact, of the possible $K=2^d-1$ feature subsets of the d original features, any ordering of the classification errors may

occur (Cover and van Campenhout 1977). Thus, there is no guarantee that the best L -feature subset consists of the first L “best” features selected and ordered via any univariate method.

When the currently best features are still inadequate to give good classification results, a simple feature generation approach frequently works well. This involves augmenting the current L -term feature set by all $L(L+1)/2$ interaction (i.e., quadratic) terms (Klein and Somorjai 1988). By also including the peak ratios or peak area ratios, a/b , (favored in MR spectroscopy), we gain an additional $L(L-1)$ terms (both a/b and b/a ratios should be included). This produces a total of $F=L(3L+1)/2$ possible features. Even for moderate L , F is fairly large: for $L=3$, $F=15$, for $L=10$, $F=155$. Thus, there are penalties to be paid: the number of parameters (i.e., the classifier coefficients) to be optimized will increase. This is undesirable for biomedical datasets, both because of interpretational difficulties and because they already suffer from sparsity.

Our recommended and generally successful procedure is to select an “optimal” subset from these F features. Since “adjacency” of features no longer makes sense (hence our standard GA_ORs-based feature selection can no longer be used), some form of algorithmic (if possible, exhaustive) search is needed. Exhaustive search for all possible 2^F-1 subsets is rarely feasible, hence we select randomly (with replacement) a large [typically $O(10^5)$] number of feature sets via the wrapper-based approach and assess their optimality. Generally, we keep the number of features in each random set fixed. As an example, we might specify identifying optimal K -feature subsets ($K \ll F$). These would likely contain some of the original L features, but also some quadratic interaction terms or peak ratios. Importantly, we may still use on these K features a linear classifier, hence we would not increase unduly the number of classifier coefficients, i.e., the complexity of the classifier.

Development of reliable classifiers

Having obtained (via GA_ORs or otherwise) K discriminatory features that satisfy the SFR appropriate for the dataset size, the next step is to develop a reliable classifier with high generalization power (i.e., with prediction accuracy for unknown samples comparable to the accuracies obtained on training set samples). To achieve this, one would like (ideally) to partition the dataset into three subsets: a training set, a monitoring (tuning) set, and an independent validation (test) set. However, when the sample size is small, such partitioning is neither feasible nor advisable because the subsets will have large variances. In that case, the simplest approach is to use the entire dataset when developing the classifier. However, this

resubstitution method is known to give an optimistically biased error estimate (EE) of the classification accuracy, i.e., the resubstitution EE would be smaller than the true error.

We may attempt to reduce this optimistic bias by crossvalidation (CV). Amongst the various CV approaches, the leave-one-out (LOO) method trains on $N-1$ of the N samples, and validates the accuracy of this $(N-1)$ sample-based classifier on the left-out sample. By leaving out each of the N samples in turn (hence producing N slightly different classifiers), it provides an essentially unbiased EE; unfortunately, for small N the variance can be unacceptably large. In k -fold CV, the dataset is split into k approximately equal parts, a classifier is trained on $k-1$ parts, validated on the held-out portion, and the process is repeated, cycling through all k folds. The overall EE is the average of the k EEs. Typically, $k=5$ or 10 appears to be a good compromise. The holdout method (i.e., twofold CV) uses approximately half of the data for training and the other half for validation. For the k -fold CV ($k \neq N$, i.e., not LOO), the data can be repeatedly partitioned, preferably in a stratified fashion, i.e., maintaining the original sample proportions in the classes; there is evidence that this improves the EE (Witten and Frank 2005).

The bootstrap method (Efron and Tibshirani 1993), a random resampling with replacement strategy, helps overcome both large bias and large variance. For a chosen feature set and classifier, bootstrapping splits the dataset randomly into training and test sets, according to a user-specified proportion. It trains the classifier on the training set and validates it on the test set. The process is repeated B times on the original dataset [B is large, typically $O(1,000)$]; the B different classifier outcomes are then averaged. Because of one-by-one, sequential replacement of the samples while generating the training-test set pair, in the limit, as $B \rightarrow \infty$, $(1 - 1/B)^B = e^{-1} = 0.368$ ($e = 2.717$, logarithmic base), $\sim 36.8\%$ of the samples will be left out and can be used for validation. However, the replacement strategy implies that in any training set several copies of the same sample would be present. This has implications on what classifier may be used (e.g., LDA will have a singular covariance matrix).

Inspired by the successes of the bootstrap methodology, we have developed a procedure (weighted cross-validated bootstrap, WCVBST) for creating robust classifiers, i.e., classifiers whose accuracy is more reliable for independent validation (test) sets (Somorjai et al. 2002). This methodology evolved and the current, more general and flexible version of WCVBST requires as input the feature set to be used with the classifier of choice. The four required input parameters for the WCVBST are B , α , β_1 , and β_2 ; $0 \leq \beta_1 \leq \beta_2 \leq 1$ (α , β_1 , and β_2 defined vide infra), and any other constraint or penalty parameters that

are deemed necessary for finding the optimal WCVBST classifier.

WCVBST selects randomly about half of the samples (in a stratified manner) to form a training set, develops with this an optimized classifier, and uses the remaining half of the samples to test the efficacy of this classifier (Fig. 2). The process is repeated B times, always starting with the entire dataset; usually $B=5,000-10,000$ (unless N is very small, in which case, with the above B values, there would be several repeated, identical splits). For all B random splits, the B sets of optimized classifier coefficients are saved.

The novelty and power of WCVBST derives from introducing a weighting approach: a weighted average of these B sets of optimized classifier coefficients produces the final, single classifier. The weights determined for classifier j are $Q_j(\alpha; T) = \kappa_j(T)C_j(T)^\alpha$, $T=TR$ or TS , with $0.75 \leq C_j(T) \leq 1$ the crisp fraction of class assignment probabilities, and $\kappa_j(T)$ is Cohen's chance-corrected measure of agreement (Cohen 1968), $\sim 0 \leq \kappa_j(T) \leq 1$, with $\kappa_j(T) = 1$ signifying perfect classification for the data subset T . The parameter α controls the importance we assign to the crispness: $\alpha=0$ means that classification crispness is not considered at all (we found that $\alpha=0.5$ provides a good compromise). The B Q_j values used for the weights are the ones found for the less optimistic test sets, i.e., $Q_j(\alpha; TS)$, not for the training sets. The weighted $Q_j(\alpha; TS)$ average need be based only on a fraction of the B classifiers. This is controlled by the range $\beta_2 - \beta_1$, with input parameter values $0 \leq \beta_1 \leq \beta_2 \leq 1$. We sort the classifiers in descending order according to their $Q_j(\alpha; TS)$ values and "integrate" (sum) the corresponding V classifier coefficients $\{\Omega_{mj}\}$, $m=1,2,\dots,V$ over the range $\beta_2 - \beta_1$:

$$\Omega_m = (1/Q(\alpha; TS)) \sum_{j=r}^s Q_j(\alpha; TS) \Omega_{mj}; \quad r = B\beta_1, s = B\beta_2$$

$$Q(\alpha; TS) = \sum_{j=r}^s Q_j(\alpha; TS);$$

It is more likely that overfitting will be avoided if we do not select the $Q(\alpha; TS)$ obtained for the single, top-ranked classifier ($\beta_1=\beta_2=1$). Coefficients obtained by summing over the interquartile (i.e., $\beta_1=0.25$, $\beta_2=0.75$) range seem to give the most robust results. We report classifier outcomes as class probabilities.

For a reliable and realistic error estimate, especially for small N , it is important to use crossvalidation both at the feature selection/extraction and classifier development and (even aggregation) stages (Ambroise and McLachlan 2002; Simon et al. 2003). This is sometimes referred to as external/internal or double CV. The parameter tuning-feature selection proceeds as follows (for the sake of keeping the description reasonably transparent, we assume that a number of parameters has been fixed a priori and the

Fig. 2 The weighted crossvalidated bootstrap (WCVBST) scheme. Given a predetermined feature set, the complete N -sample dataset ($N=N_1+N_2$, here $N_1=12$, $N_2=8$) is split randomly into training set i (TR_i) and test set i (TS_i), in ~50:50 proportion, $i=1,2,\dots,B$. A classifier is developed on training set i and tested on test set i . $Q_\alpha(TR_i)$ and $Q_\alpha(TS_i)$ are computed. (See text for details)

only relevant parameter to determine is the optimal number of features:

- Step 1: We carry out K_{ext} repetitions, $j=1,\dots,K_{\text{ext}}$ for the external CV.
- Step 2: For each j in the above range, we determine the optimal number of features, say $F(j)$. This is carried out by first randomly splitting, ~50:50, the entire dataset into a training set, $TR(j)$, and a monitoring $MON(j)$ set. The splitting is stratified, i.e., the original relative proportion of the samples in the two classes is maintained. $TR(j)$ is used for the determination of the best feature set $F(j)$; $MON(j)$ attempts to control the extent of overfitting, i.e., the excessive adaptation of the features to the specifics of $TR(j)$. Overfitting is a common and serious problem if the dataset size N is small. The outcomes of this step are two error estimates:

$EE[TR(j)]$ for $TR(j)$, $EE[MON(j)]$ for $MON(j)$. (At this stage, using WCVBST will likely be beneficial.)

- Step 3: The random splitting with replacement in step 2 is carried out $k=1,2,\dots,K_{\text{int}}$ times. (An alternative option is to do K_{int} - fold CV.) K_{int} is the number of internal CVs selected. This leads to K_{int} pairs $[EE(TR(j,k)), MON(j,k)]$ and K_{int} feature sets $F(j,k)$. $F(j)$ is frequently kept constant for each of the K_{int} steps, i.e., $F(j,k) = F(j)$, $\forall k$. (Another option, less commonly used because of the heavier computational burden, is to find the optimal $F(j,k) \neq F(j)$, $\forall k$. In this case, it is useful to construct an occurrence frequency (OF) histogram $H(j)$. In this histogram, features may appear 1,2,... up to K_{int} times. The not unreasonable inference is that the larger the OF for a feature, the more likely that that feature is discriminatory. Frequently, $H(j)$ is cut at decreas-

ing OFs and the number of features for which the CV classification error is minimal is identified.)

- Step 4: For each j , average the N_{int} EEs obtained in the internal CV loop (step 3), and select the smallest average CV error and the corresponding $F(j)$.
- Step 5: Repeat step 2 K_{ext} times.
- Step 6: Compute the classification errors for each j , average over all K_{ext} , and calculate the standard deviations (alternatively, compute the robust equivalents, median and interquartile range).

In principle, this six-step process ought to be carried out for the tuning of all parameters of the problem. In practice, this would be too time-consuming and is rarely done. (Ideally, even K_{ext} and K_{int} should be optimized. Usually, K_{ext} is 10–20, and K_{int} is kept at 5–10.)

Discussion

I did not describe the fifth stage, classifier aggregation/fusion. We invoke this stage if the best classifier found after the first four stages (in particular, stages 3 and 4) is still not accurate enough. There are several options. We may produce T different classifiers by using T different classification methods [e.g., LDA, QDA, SVM, k -nearest neighbors, Breiman's random forests (Breiman 2001), etc.]. By submitting the samples to these T classifiers, T generally different class assignments are produced for each sample. We may then combine these by any of the numerous classifier combining methods. Kuncheva (2004) discusses in extensive detail the various classifier aggregation possibilities and algorithms.

Another option is to select a single classifier but apply it to the T different feature sets and repeat the procedure outlined above.

A very important practical consideration is how to acquire/represent the data on which decisions (e.g., medical diagnoses/prognoses) will be based. The goal is to find those class representations (types of features) that are optimal for the eventual classifier development and application. With clinical applications in mind, we narrow the range of representation possibilities by focusing on noninvasive data acquisition techniques that are based on some form of spectroscopy (MR, IR, Raman, mass, etc.).

For spectra, the two potentially most useful class representations are the following:

1. Explicitly determined spectral signatures (SSs) (peaks, integrated spectral subregions). This is the approach we advocate.
2. Fitting the spectra to the relative concentrations of a mixture of metabolites (MMs).

In analyzing spectra derived from biofluids or tissues, and with increasing emphasis on identifying biomarkers

(e.g., the metabolites that characterize diseases or disease stages), the question naturally arises, which representation is preferable? Should we focus on using MMs or SSs? Our semi-implicit approach, using a wrapper-based feature extraction methodology (exemplified by the GA_ORS component of the SCS) derives an interpretable spectral signature; the interpretability persists and can even be enhanced by explicitly including interaction (quadratic) terms. Bias, due to redundancy, sample imbalance, etc., can be minimized, however, the derived discriminatory spectral regions don't necessarily correspond to known metabolites, i.e., at this stage of the analysis useful biomarkers may not be discovered or discoverable. Thus, "interpretability" only means that the identified discriminatory spectral signature doesn't require additional feature unscrambling, as for PLS-DA-derived secondary features [SIMCA-P's score and weight plots (Eriksson et al. 2001)].

A fully explicit method, more commonly advocated by MR spectroscopists, is to fit the spectra in question to a set of user-chosen metabolite spectra. The in vitro MR spectra of these metabolites are obtained from a previously compiled database. The fitting is usually carried out either by some commercial algorithm, e.g., LCModel (Provencher 1993), MestRe-C (Cobas and Sardina 2003), or an in-house equivalent. The outcome of this fitting procedure is an estimate of the concentration of a number of possible and/or likely metabolites.

Applying any classification methodology (e.g., the SCS) to such fitted spectra simply means that the predetermined features are the actual fitted metabolites and the feature values are metabolite concentrations.

Of course, this approach has its own difficulties and shortcomings. Which, and how many, metabolites to choose for the fitting inevitably introduces bias; this bias is the consequence of an ab initio, not necessarily fully informed decision by the analyst. Assuming the availability of the base spectra to be used for the fitting, the bias may be reduced by simply increasing the number of metabolites included in the database. However, if a hitherto unknown compound is present, it is necessarily missing from the metabolite database; hence it cannot be readily identified.

An important source of error is the generally low accuracy of the concentration values obtained. According to practitioners, 20% uncertainty seems to be common and considered acceptable if determining metabolite concentrations is the only goal of the fitting; however, such large errors in concentration will certainly lead to unacceptable and unusable classification accuracies, no matter how the classification is carried out, because of the large overlap between the individual metabolite concentration ranges of the samples from the two classes.

Using spectral signatures has its own shortcomings. The major one is that some of the discriminatory subregions found will not have known metabolic identity. Furthermore, the number of optimal discriminatory features requested is an input parameter; it has to be validated to reduce the induced bias. When the sample size is small, additional uncertainties arise because of the possibility of non-uniqueness, i.e., several non-overlapping feature sets may produce comparable classification accuracies (this, of course, is also a caveat for MM-based classifiers).

Our very limited comparison of the two feature extraction approaches on a few datasets (we used identical, SCS-based methodology for both) suggests that the spectral-signature-based classification is more accurate. However, this conclusion is unlikely to be always correct, and more extensive comparison studies will be needed. (A combination of the two feature selection approaches is an intriguing possibility for future research.)

Because of our desire to develop classifiers that are immediately relevant in the clinic, the above considerations and caveats persuaded us to change our philosophy regarding feature selection and classifier development. We now advocate a two-stage process, with different goals. At the first stage, we only require that the classifier(s) provide fast, reliable diagnoses for the clinician (by general consensus, this is what our clinical collaborators seem to want first). Thus, it appears that for this diagnostic stage, the clinician doesn't need or require any causal (chemically, biologically identifiable) interpretation of the distinction between the diseased and nondiseased states.

The goal of the second stage is to provide, based on the robust classifiers developed, reliable decision support for prognosis; this would require explicit research effort (not necessarily related directly and immediately to clinical relevance) and identification of the molecular, biological causes of the disease state (discovery of "biomarkers" or panels of biomarkers).

An additional impetus for promoting the above-described two-stage approach is our discovery that after completion of the original wrapper-based selection of interpretable features, the results can be further improved by using a dissimilarity/distance-based classification approach (Somorjai et al. 2007; R.L. Somorjai, B. Dolenko, W. Roberson, N. Thiessen, unpublished data), some aspects of which were also advocated by Duin's group (Pełalska and Duin 2005). This, of course means that the new features, the two distances to two class proximity measures may be complicated, generally non-linear combinations of the original features. (However, at least for the SS representation, interpretability is not lost, only disguised; one may readily backtrack to the original feature selection stage and identify the starting feature set.) The great advantage of this dissimilarity-based

approach is that there is a great deal of flexibility in choosing the class proximity measures (e.g., class centroids, k-nearest neighbors, average distances to members of the two classes, etc.). Additional flexibility is conferred on the approach by being able to select from a great variety of dissimilarity/distance measures [e.g., Minkowski, Anderson-Bahadur/Mahalanobis (Anderson and Bahadur 1962), Chernoff/Bhattacharya (Cover and Thomas 1991), symmetric Kullback-Leibler (Kullback and Leibler 1951)]. An important bonus is that this new, derived feature set is two-dimensional, hence the classification results can be immediately visualized in a class proximity plane, e.g., using stage 1 of the SCS.

In Fig. 3, we show a class proximity mapping to 2 dimensions of the 26-dimensional colon cancer vs. normals dataset (see the Appendix for a shortened description of how we developed the accurate classifiers for this dataset; the 26 dimensions refer to the 26 subregions we arrived at after the initial feature selection, vide infra), using as the class proximity measures the class centroids and as the distance measure the Mahalanobis distance (this is the distance characteristic of the LDA classifier we used). Note that we didn't at all need to carry out the additional feature selection of finding 15 features from the subsequently constructed 377 linear + quadratic features. Both the visualization and classification advantages are evident. Furthermore, the mapping did not worsen the class assignment accuracies.

Fig. 3 Class proximity mapping from 26 dimensions to 2. Normal (red disks) vs. colon cancer patients (blue disks). The proximity measure is the class centroid, the distance measure is the Mahalanobis distance

The two approaches (spectral signatures and mixtures of metabolites) definitely have common aspects. In particular, a spectrum is the manifestation of a mixture of metabolites (and “noise”). However, based on the two-stage approach sketched above, we should, if possible, confine ourselves to the second, research phase — the metabolite mixture identification and interpretation.

Acknowledgments I thank the staff of the Biomedical Informatics Group of the Institute for implementing most of the algorithms discussed and our collaborators who provided the MR spectra on which the SCS could be tested and validated. Incorporating useful suggestions and comments from the reviewer enhanced the clarity and reader-accessibility of the manuscript.

Appendix

Example of an SCS-based classifier development: control vs. colorectal cancer (CRC)

I give an abbreviated version of the two components of the SCS discussed above in general terms, as they apply to distinguishing between controls and patients with CRC. For definitions the reader may consult the main text.

The samples for the ^1H NMR spectra were derived from fecal extracts. There were 412 controls (“normals”) and 111 CRCs. NMR spectral acquisitions were on a Bruker INCA NMR spectrometer (400.13 MHz, at 300°K). An externally introduced TSP peak was used as the chemical shift reference (0 ppm) and also for peak alignment for stage 2. (Additional details of the parameters of the MRS experiments are found in Bezabeh et al. 2009.) The visualization stage of the SCS provided an initial quality control; since for each sample replicate spectra were obtained, displaying them immediately identified “outliers.” Their absence/presence was confirmed by calculating pair-wise Pearson’s correlation coefficients between the replicates.

Preprocessing

We acquired 16,384-data-point spectra, and to reduce phasing errors, only their magnitude was used for the subsequent analyses. From these magnitude spectra, the region (−1.32 to 5.46 ppm) was extracted and the resulting 8,000-point spectra were normalized to unit area. The spectra were then replaced by their first derivatives to obtain more accurate classification results.

We eliminated the “noise” regions of the spectra, i.e., the regions that on visual inspection appeared to lack peaks. (When these regions were included in a preliminary analysis, the classification accuracy invariably decreased, suggesting that they are detrimental.) Only the subregions 2.03–2.45 ppm (500 data points) and

0.80–1.22 ppm (500 data points) survived. These resultant 1,000-point spectra were then rank-ordered (i.e., the actual intensity values were replaced by their ranks, in order to reduce the influence of excessively large resonance intensities.)

Feature extraction

From the above ppm ranges, we extracted a set of maximally discriminatory features, the averaged spectral intensities of the chosen subregions, using GA_ORS driven by LDA/LOOCV as the crossvalidated classifier. We initially specified identifying 12 features, satisfying the minimum 5–10 SFR required for obtaining robust classifiers. With ten different random starting seeds, the GA_ORS produced ~150 different features (averaged subregions) that gave comparable accuracy. In order to obtain classifiers producing balanced sensitivity (SE) and specificity (SP), the feature extraction was constrained by imposing an approximately fourfold penalty for misclassifying the cancer samples. The preliminary sensitivity/specificity-balanced classifiers gave overall accuracies in the unacceptable 70–80% range. (Because of the 412:111 sample imbalance between the classes, misclassifying all the 111 cancer and none of the 412 control samples, the overall accuracy with these sample sizes would still be 78.8%, but with 0% sensitivity and 100% specificity, completely useless both for classification purposes and clinically.)

Hence, we selected 26 distinct, non-overlapping subregions from the initial ~150 as the starting point for additional feature generation: these 26 linear features (averaged subregions) were augmented by all their quadratic (interaction) terms (e.g., from features f_1 and f_2 , f_1^2 , f_2^2 , and f_1f_2 were constructed as additional features). Adding all 351 quadratic terms, we created a total of $26+351=377$ features.

Classifier generation

The relative sample scarcity restricted us to subdividing the dataset into only training and monitoring sets. We split the data randomly, placing 206 normal and 55 cancer samples in the training set, and 206 normal and 56 cancer in the monitoring set. Ten such splits created ten different training/monitoring sets. However, with the larger feature set, we could no longer satisfy the recommended 5–10 SFR. Hence, for each of the ten random splits, we carried out additional feature selection from the new, 377-feature training sets. In the tested range of 8–20 feature subsets, 15 different sets of features (again, obtained using LDA/LOOCV) produced 15 different optimized classifiers.

The excellent, SE-SP-balanced results, based on the ten classifiers are as follows (where SD = standard deviation and AUC = area under the ROC curve):

Training set:

(SE $\pm$ SD) : 87.5 $\pm$ 2.0%,
 (SP $\pm$ SD) : 87.5 $\pm$ 2.0%,
 (AUC $\pm$ SD) : 0.932 $\pm$ 0.01

Monitoring set:

(SE $\pm$ SD) : 85.2 $\pm$ 3.3%,
 (SP $\pm$ SD) : 86.9 $\pm$ 1.4%,
 (AUC $\pm$ SD) : 0.923 $\pm$ 0.02

Considering only the (clinically more relevant) crisp samples (i.e., those with class assignment probabilities ≥ 0.75), we found the following classification results:

Training set:

(SE $\pm$ SD) : 92.5 $\pm$ 3.0%,
 (SP $\pm$ SD) : 92.3 $\pm$ 1.4%,
 (%Crisp $\pm$ SD) : 76.8 $\pm$ 2.9%,

Monitoring set:

(SE $\pm$ SD) : 91.8 $\pm$ 4.5%,
 (SP $\pm$ SD) : 91.9 $\pm$ 2.6%,
 (%Crisp $\pm$ SD) : 78.1 $\pm$ 2.9%

A not unreasonable conclusion is that in the ~77% crisp cases, the clinical decision would likely be reliable.

References

- Ambrose C, McLachlan GJ (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci USA* 99:6562–6566
- Anderson TW, Bahadur RR (1962) Classification into two multivariate normal distributions with different covariance matrices. *Ann Math Statist* 33:420–431
- Bellman R (1961) *Adaptive control processes: a guided tour*. Princeton University Press, Princeton
- Bezabeh T, Somorjai R, Dolenko B, Bryskina N, Levin B, Bernstein CN, Jeyrajah E, Steinhart AH, Rubin D, Smith ICP (2009) Detecting colorectal cancer by ^1H magnetic resonance spectroscopy of fecal extracts. *NMR Med* 22(6):593–600
- Breiman L (2001) Random forests. *Mach Learn* 45:5–32
- Cobas JC, Sardina FJ (2003) Nuclear magnetic resonance processing. MestRe-C: a software package for desktop computers. *Concepts Magn Reson A* 19A:80–96
- Cohen J (1968) Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol Bull* 70:213–220
- Cover TM, Thomas JA (1991) *Elements of information theory*. Wiley, New York
- Cover TM, van Campenhout JM (1977) On the possible orderings in the measurement selection problem. *IEEE Trans Syst Man Cybern SMC* 7(9):657–661
- Dolenko B, Somorjai RL (1995) Time well spent: preprocessing of MR spectra for greater classification accuracy. *Proceedings of Society of Magnetic Resonance, Third Scientific Meeting, Nice, France, 19–25 August, #1936*
- Dougherty ER (2001) Small sample issues for microarray-based classification. *Comp Funct Genom* 2:28–34
- Efron B, Tibshirani RJ (1993) *An introduction to the bootstrap*. Chapman Hill, New York
- Eriksson L, Johansson E, Kettaneh-Wold N, Wold S (2001) *Multi- and megavariate data analysis—principles and applications*. Umetrics AB, Umea
- Goodacre R, Broadhurst D, Smilde A, Kristal BS, Baker JD, Beger R, Bessant C, Connor S, Capuani G, Craig A, Ebbels T, Kell DB, Manetti C, Newton J, Paternostro G, Somorjai R, Sjöström M, Trygg J, Wulfert F (2007) Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomic* 3:231–241
- Holte RC (1993) Very simple classification rules perform very well on most commonly used datasets. *Mach Learn* 11:63–91
- Huber PJ (1985) Projection pursuit. *Ann Stat* 13:435–475
- Jackson JE (1991) *A user's guide to principal components*. Wiley, New York
- Jain AK, Chandrasekaran B (1982) Dimensionality and sample size considerations in pattern recognition practice. North Holland Publishing, Amsterdam
- Jain AK, Duin RPW, Mao J (2000) Statistical pattern recognition: a review. *IEEE Trans Pattern Anal Mach Intell* 22:4–37
- Klein P, Somorjai RL (1988) Nonlinear methods for discrimination and their application to classification of protein structures. *J Theor Biol* 130:461–468
- Kohavi R, John GH (1997) Wrappers for feature subset selection. *Artif Intell* 97:273–324
- Kullback S, Leibler RA (1951) On information and sufficiency. *Ann Math Statist* 22:79–86
- Kuncheva, LI (2004) *Combining pattern classifiers—methods and algorithms*. Wiley-Interscience, New York
- Lean CL, Somorjai RL, Smith ICP, Russell P, Mountford CE (2002) Accurate diagnosis and prognosis of human cancers by proton MRS and a three stage classification strategy. *Annu Rep NMR Spectrosc* 48:71–111
- Nikulin A, Brière KM, Friesen L, Smith ICP, Somorjai RL (1995) Genetic algorithm-guided optimal attribute selection: a novel preprocessor for classifying MR spectra. *Proceedings of the Society of Magnetic Resonance, Third Scientific Meeting, Nice, France, 19–25 August, #1940*
- Nikulin AE, Dolenko B, Bezabeh T, Somorjai RL (1998) Near-optimal region selection for feature space reduction: novel preprocessing methods for classifying MR spectra. *NMR Biomed* 11: 209–216. Our software, GA_ORIS is available on request
- Pękalska E, Duin RPW (2005) *The dissimilarity representation for pattern recognition foundations and applications*. Series in machine perception artificial intelligence, vol. 64. World Scientific, River Edge, NJ
- Provencher SW (1993) Estimation of metabolite concentrations from localized in vivo proton NMR spectra. *Magn Reson Med* 30:672–679
- Pudil P, Novovicová J, Kittler J (1994) Floating search methods in feature selection. *Pattern Recogn Lett* 15:1119–1125

- Simon R, Radmacher MD, Dobbin K, McShane LM (2003) Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Natl Cancer Inst* 95:14–18
- Sjöström M, Eriksson L, Hellberg S, Jonsson J, Skagerberg B, Wold S (1986) PLS discriminant plots. *Proceedings of PARC in Practice*, Amsterdam, June 19–21, 1985. Elsevier, North-Holland
- Somol P, Pudil P, Novovicová J, Paclík P (1999) Adaptive floating search methods in feature selection. *Pattern Recogn Lett* 20:1157–1163
- Somorjai RL, Nikulin AE, Pizzi N, Jackson D, Scarth G, Dolenko B, Gordon H, Russell P, Lean CL, Delbridge L, Mountford CE, Smith ICP (1995) Computerized consensus diagnosis: a classification strategy for the robust analysis of MR spectra. I. Application to ^1H spectra of thyroid neoplasms. *Magn Reson Med* 33:257–263
- Somorjai RL, Dolenko B, Halliday W, Fowler D, Hill N, Ross I, Brière KM, Smith ICP (1999) Accurate discrimination between low- and high-grade human brain astrocytomas: robust multivariate analysis of ^1H MR Spectra. *J Med Biochem* 3:17–24
- Somorjai RL, Dolenko B, Nikulin A, Nickerson P, Rush D, Shaw A, de Glogowski M, Rendell J, Deslauriers R (2002) Distinguishing normal from rejecting renal allografts: application of a three-stage classification strategy to MR and IR spectra of urine. *Vibr Spectrosc* 28:97–102
- Somorjai RL, Dolenko B, Baumgartner R (2003) Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions. *Bioinformatics* 19:1484–1491
- Somorjai RL, Alexander M, Baumgartner R, Booth S, Bowman C, Demko A, Dolenko B, Mandelzweig M, Nikulin AE, Pizzi N, Pranckeviciene E, Summers R, Zhilkin P (2004a) A data-driven, flexible machine learning strategy for the classification of biomedical data. Chapter 5 In: Dubitzky W and Azuaje F (eds) *Artificial intelligence methods and tools for systems biology computational biology series*, vol. 5 Springer, New York, pp 67–85
- Somorjai RL, Demko A, Mandelzweig M, Dolenko B, Nikulin AE, Baumgartner R, Pizzi N (2004b) Mapping high-dimensional data onto a relative distance plane—an exact method for visualizing and characterizing high-dimensional patterns. *J Biomed Inf* 37:366–379
- Somorjai RL, Dolenko B, Mandelzweig M (2007) Direct classification of high-dimensional data in low-dimensional projected feature spaces—comparison of several classification methodologies. *J Biomed Inf* 40:131–138
- Weljie AM, Newton J, Mercier P, Carlson E, Slupsky CM (2006) Target profiling: quantitative analysis of ^1H NMR metabolomics data. *Anal Chem* 78:4430–4442
- Witten IH, Frank E (2005) *Data mining practical machine learning tools and techniques*, 2nd ed. Elsevier, Amsterdam