Deriving biomedical diagnostics from NMR spectroscopic data

Ian C. P. Smith · Ray L. Somorjai

Received: 17 August 2010 / Accepted: 14 February 2011 / Published online: 8 March 2011
© Springer-Verlag (outside the USA) 2011

Abstract Biomedical spectroscopic experiments generate large volumes of data. For accurate, robust diagnostic tools the data must be analyzed for only a few characteristic observations per subject, and a large number of subjects must be studied. We describe here two of the current data analytic approaches applied to this problem: SIMCA (principal component analysis, partial least squares), and the statistical classification strategy (SCS). We demonstrate the application of the SCS by three examples of its use in analyzing ^1H NMR spectra: screening for colon cancer, characterization of thyroid cancer, and distinguishing cancer from cholangitis in the biliary tract.

Keywords Biomedical spectroscopy · Statistical classification strategy (SCS) · Soft independent modelling of class analogies (SIMCA) · Cancer screening · ^1H NMR spectra

List of Abbreviations

FLD	Fisher's linear discriminant
FOBT	Fecal occult blood test
NMR	Nuclear magnetic resonance
PC	Principal component
PCA	Principal component analysis
PCR	Principal component regression
PLS	Partial least squares
PSC	Primary sclerosing cholangitis
SCS	Statistical classification strategy

SIMCA Soft independent modelling of class analogies
WCVBST Weighted cross validated bootstrap

Introduction

Modern spectroscopic methods are being increasingly used to develop medical diagnostics. The methods lead to a plethora of data that defy analysis by conventional approaches. Over the past 10–20 years, a number of methods have been reported to simplify the data and analyze them accurately. The most commonly used approach, favored by many spectroscopists and chemometricians, is soft independent modelling of class analogies (SIMCA), exemplified by the software SIMCA-P (Eriksson et al. 2001). The strategy of SIMCA-P is to rely almost exclusively on principal component analysis (PCA), and for classification, on its supervised versions, partial least squares (PLS) or principal component regression (PCR). For short, we shall denote by “SIMCA” the entire corpus of the PCA/PLS/PCR-based methodology, detailed by Eriksson et al. (2001). The other approach is the statistical classification strategy (SCS), developed at the Institute for Biodiagnostics (IBD). A detailed description of the SCS is found in Somorjai et al. (2004a) and Nikulin et al. (1998). A simpler outline of some aspects of its methodology may be found in Somorjai (2009). The basic philosophy of the SCS is based on the realization that “there is no panacea in data analysis” (Huber 1985), hence the methods and algorithms one needs to use are dependent on the nature of the data. We shall introduce both approaches and compare their relative utility for the classification of nuclear magnetic resonance (NMR) spectra. Note that there are numerous different types of classifiers, as well as many feature selection and classifier aggregation methods and algorithms in the machine learning literature (consult, e.g., Hastie et al. 2009; Duda et al. 2001;

I. C. P. Smith (✉) · R. L. Somorjai
Institute for Biodiagnostics, National Research Council,
Winnipeg R3B 1Y6, Canada
e-mail: ian.smith@nrc-cnrc.gc.ca

Kuncheva 2004; Witten and Frank 2005). However, our focus here is to compare two *strategies*, not individual, mostly unrelated methods/algorithms.

The problems

Medical specimens or subjects are difficult to obtain in large numbers. Often, studies involving small numbers of subjects are reported, suggesting great potential for ultimate medical use, but regrettably, very few proceed further to actually demonstrate this and produce medically useful classifiers. This is the first problem.

Most biomedical spectra involve a large number of data points (“features”), many of which are not useful or may be even detrimental (“noise”), and only a few of which may actually be diagnostic. The second problem therefore is to extract the clinically meaningful features and limit their number (reduce the size of the feature space). A large number of features lead to overfitting and a non-robust classifier (one that fails on challenge by data not used to develop the classifier).

Problem three involves identifying the appropriate mathematical method to develop the classifier. The conventional “scatter plots,” used formerly in biomedical reports, are at best based on univariate statistical tests (e.g., *t*-tests) and do little more than suggest possible potential for class separation. Actual confirmation of such potential requires appropriate multivariate methods and classifiers.

The top of Fig. 1 demonstrates the problem of insufficient number of samples. With very few samples, an apparently accurate separation is not difficult to achieve. Introduction of more samples will often produce a poorer separation because of the likelihood of filling the initially empty overlap between classes (bottom of Fig. 1). Finally, an adequate number of specimens may produce both a realistic, acceptable class separation and a potentially robust classifier. A rule of thumb is that the number of specimens must be roughly 10 times the number of discriminating features. Unfortunately, the converse is often reported.

Spectra must be prepared for analysis (preprocessing). This involves referencing to a standard frequency (chemical shift in NMR), scaling the features, and normalizing to the area under the entire spectrum. Absorption or derivative spectra may be used. With absorption, the use of magnitude spectra simplifies analysis. Rank ordering of intensities has also proven useful, as did adding nonlinear terms to the feature space.

Determination of the best discriminating features in a spectrum is a critical requirement for classifier reliability and robustness. Generally, the full dataset is first split into training and validation (monitoring) sets for classification. The optimal feature set is found using only the training set. The validation set serves as a control; it helps prevent

Fig. 1 Representation of the risks of reaching conclusions with a sparse data set. Increasing the number of subjects generally lowers the accuracy, but this is much closer to the true accuracy. The lower accuracy solution will also be more robust: challenging the resultant classifier with new specimens will yield accuracy similar to that found by a reliable classifier

excessive training and hence overfitting. Ideally, an independent, blinded test set—one that didn’t participate in the classifier development—should be challenged by the classifier to assess future performance realistically. When the data set size is too small, even the training-validation set split is unfeasible and some version of cross-validation, e.g., leave-one-out (LOO) or *k*-fold method is commonly used, with attendant caveats (Somorjai et al. 2004a).

Soft independent modelling of class analysis (SIMCA)

The strategy of SIMCA is simple: it uses PCA/PLS/PCR for all data analysis. SIMCA’s feature selection is in fact a data compression procedure. It reduces the original, high-dimensional feature space to a low number of new features, the uncorrelated principal components (PCs), derived via principal component analysis (PCA). PCA is an unsupervised approach, i.e., it doesn’t use class information. The PCs are linear combinations of the original features. They are eigenvectors of the data matrix, arranged so that the first PC explains most of the data variance, the second the majority of what remains, etc. The relative magnitudes of the data matrices’ eigenvalues determine the number of PCs necessary to explain most of the variability (variance) in the data. Two important quantities are the *scores* (orthogonal

projections of the data values onto the PCs) and *loadings* (whose magnitudes reflect the relative importance of the features for the PCs). PCA rotates the original coordinate system such that the PCs, the new orthogonal coordinate axes, point along the maximal variance. These help assess the classification relevance of the original features comprising the PCs for classification.

Typically, the first two or three PCs provide the two- to three-dimensional reduced coordinate system in which the individual instances (samples) of the dataset may be displayed. However, the major disadvantage of PCA when the PCs are used as new features for classification is that the PCs' directions, the directions along which the dataset variance is maximal, are not necessarily the best directions for classification.

An important facet of SIMCA is that it allows computation of class-dependent PCAs. These are more flexible than the conventional PCs that don't use class-specific information. In particular, the relevance of different features and measures of class separation can be assessed more readily. The class-specific PCs are orthogonal within their class; they are not so between classes. This is not necessarily a disadvantage. However, in the SIMCA-based classification literature there doesn't seem to be any assessment of dependencies/correlations between PCs belonging to different classes. Such dependencies may be detrimental for robust, reliable classification.

Partial least squares (PLS) and principal component regression (PCR) are two components of the SIMCA corpus that may be and are used for classification. (For two-class problems, linear classification and regression are equivalent.) They are more appropriate for classification than PCA because the PLS/PCR PCs are determined by maximizing the covariance between the independent variables (the derived PCs) and the dependent ones (class membership). However, the caveats raised in connection with PCA still hold.

What is missing in SIMCA for classification? In principle, nothing. However, because the PCs are linearly "scrambled" versions of the original data points, direct identification and validation of the discriminating features ("biomarkers") are more difficult. Furthermore, for spectra the original features are single data points (resonances), hence they cannot be readily and meaningfully interpreted as biomarkers. This lack of interpretability provided us the first impetus to develop the feature selection stage of the statistical classification strategy (SCS). Furthermore, the general philosophy of the SIMCA approach, relying almost exclusively on PCA to reduce feature space dimensionality, independently of the data to be classified, led us to develop a *data-driven* classification strategy, without prior, preset assumptions. A flexible strategy is essential for biomedical spectroscopic data.

The statistical classification strategy (SCS)

This approach was developed in our institute. Its major difference from SIMCA is that the components of its stages are selected according to the properties of the data to be analyzed, i.e., the SCS is data-driven at all of its stages. It now consists of five stages. Of these, feature selection is critical. The particular method we developed and advocate was designed to produce features that retain spectral identity. In particular, for NMR spectra the new features are averages of spectral intensities of adjacent data points of varying ranges (subregions), thus better representing specific peak locations and areas. Our genetic algorithm-based optimal feature selector (Nikulin et al. 1998) is used; it optimizes both the number of discriminating subregions and their widths. The feature selection is "driven" by some preselected classifier that will eventually be employed for the actual classification (wrapper method). This approach leads to an optimal classifier-feature set pair. Because of its robustness, we frequently use Fisher's linear discriminant (FLD). FLD has the advantage of not only assigning samples to one of the classes, but also yielding a measure of the probability that they belong to that class. A more detailed mathematical description of the methods is given in Nikulin et al. (1998).

When the sample size is small, in addition to or instead of the standard cross-validation methods (e.g., LOO), we generally use our bootstrapping-inspired approach, *weighted cross-validated bootstrap* (WCVBST) (Somorjai 2009). The steps of this process are shown in Fig. 2. WCVBST was designed to create more robust classifiers, i.e., classifiers whose accuracy is more reliable for independent test sets. WCVBST selects randomly about half of the samples 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. The process is repeated B times (typically, $B=5,000$ – $10,000$), always starting with the entire dataset. For all B random splits, the B sets of optimized classifier coefficients are saved. WCVBST is powerful because of the weighting we introduced: a *weighted average* of these B sets of coefficients produces the final, *single* classifier. The B weights are the ones found *not* for the training sets, but for the less optimistic *test sets*. We report classifier outcome as *class probability*.

When we are not satisfied with the accuracy of our classifier, we resort to classifier aggregation, a consensus technique (Kuncheva 2004). One approach is to use different representations of the spectra, such as absorption, first derivative, rank order, and develop full classifier-feature set pairs for each. Thus, we use a specific classifier, but on different types of feature sets. The classification outcomes are generally different and will emphasize different aspects

WEIGHTED CROSSVALIDATED BOOTSTRAP: WCVBST

TR = Training set VL = Validation set Q = Classifier Quality

Fig. 2 Representation of the bootstrapping procedure. The data for the normals and the cancer patients are divided randomly into two groups. A classifier is calculated from one and tested by the other (TR₁ and VL₁). Two different random sets (TR₂ and VL₂) are then extracted from the entire data set and the process used for the TR₁/VL₁ split repeated. This procedure is carried out many times (thousands) yielding many classifiers from which the VL-weighted, highest quality combined classifier is determined

of the spectra. Alternatively, we may select a specific feature set, perhaps based on prior information, but now develop, for this particular feature set, different classifiers, such as linear or quadratic discriminant analysis, some nearest neighbor classifier, neural nets, support vector machines. In either case, we combine the classification outcomes to yield a consensus result (Somorjai et al. 2004a, b).

We emphasize that the SCS doesn't necessarily exclude PCs as possible features. In fact, PCs were also tried, with less success, in early versions of the SCS. However, unlike SIMCA, the SCS, because of its data-driven philosophy and a built-in flexibility, doesn't confine itself to PCs as the only possible features. The strategy allows for and encourages experimentation with different combinations of preprocessing, feature selection, classifier choice, and classifier aggregation. As an example, we generated class-specific feature sets (Somorjai et al. 2004b) that didn't require the use of class-dependent PCAs and PCs. Instead, in each class, lines that passed through pairs of samples were constructed and a sample point was assigned to the class whose line was closer to it. These lines are general-

izations of PCs. They are more flexible, because neither orthogonality nor confinement to the direction of maximal variability was required and imposed.

Examples

Screening for colon cancer

In most countries, colon cancer is the second worst killer among the cancers. If detected early, the prognosis is excellent (92%): detected late and 5-year survival drops to 6%. Clearly the way to prevent deaths is to find a method to test asymptomatic subjects and thus diagnose early stages of colon cancer, i.e., subject screening. The current gold standard for screening is colonoscopy, where a flexible probe is inserted into the large intestine via the rectum and a visual assessment is made. The method is expensive, objectionable to many, and there is a risk of puncturing the intestinal wall and causing serious abdominal infection. A less expensive, less invasive method with a high accuracy is necessary. A start was made with the fecal occult blood test (FOBT), where a smear of stool is placed on a supportive strip that is subsequently tested chemically for the presence of hemoglobin. Its accuracy ranges widely, from 40–80%, often due to the consumption of meat or red wine before the test. A risk of false positives leads to unnecessary

Fig. 3 NMR spectra (360 MHz, 27°C) of fecal water from healthy controls and from patients with advanced colon cancer (Bezabeh et al. 2009)

colonoscopy, whereas a false negative leads to undetected advancement of the stage of the colon cancer. Subject compliance, actually doing the test, is low.

In our institute, we have perfected a method to detect early colon cancer by performing NMR spectroscopy on fecal water, the supernate from a suspension of several grams of feces (Bezabeh et al. 2009). Typical ^1H spectra from unseparated suspension are shown in Fig. 3. The spectra from the supernates show much greater resolution.

While differences between the spectra can be seen, it is necessary to determine the spectral regions most valuable for classification, step three of the SCS method. This is done by use of the genetic algorithm (see Examples), which selects a few (e.g., four or five) discriminating regions. The classification results are based on 523 individual cases (412 normal, 111 with colorectal cancer). An overall accuracy of 92% was obtained. Because we used FLD, we also obtained the probability that specimens belong in one of the classes. This probability may be viewed as the credibility (“crispness”) of the class assignment. The high accuracy was obtained on samples having $\geq 75\%$ crispness (405 specimens).

We are now in the process of moving this test into hospital use, which is expected to be complete in 2011. This study was performed in partnership with the MD Anderson Hospital, University of Texas, and the Health Sciences Centre, University of Manitoba.

Thyroid cancer

Another example demonstrates the method of consensus analysis (Somorjai et al. 1995, 2004a). This involves analysis of the ^1H spectra of thyroid tissue biopsies by use of different classifier approaches and then combining the resultant classifier outcomes to form a consensus result. These are the methods used in step five of the SCS (Somorjai et al. 2004a). The data were split into training and test sets. The three classifiers used are linear discriminant analysis, genetic programming, and neural networks. In this case we selected as the discriminating features the first ten PCs found by PCA. Cross-validation on the training sets was done by the standard LOO method. For N samples, N slightly different classifiers are obtained; samples from the test set are assigned N times to one of the classes. The consensus result is the median of the N assignments. This early method is a precursor of the classifier aggregation step of the SCS; it produced an accuracy of 99% in a group of 107 thyroid biopsies from normal and various malignant cancers.

Biliary cancer

In a recent publication (Albiin et al. 2008), the SCS method was applied to diagnose cancer in a cohort with primary sclerosing cholangitis (PSC)±cancer by analysis of bile ex

vivo. This is necessary because strictures of the biliary system are difficult to classify. The only successful treatment for PSC is liver transplant. If cancer is present in the system, the transplant will fail. Applying the above methods to a cohort of 45 specimens, overall accuracy of 89% was achieved. This study was conducted in collaboration with the Karolinska Hospital and Stockholm University.

The future

It is clear from these, and many other published studies, that in combination with robust, accurate analytical methods, high resolution NMR is a powerful tool for the classification of specimens. In the case of readily available specimens such as blood or urine, this approach suffices. However, when tissue biopsies are used, or fluids such as bile are obtained by endoscopic techniques, a degree of invasion is involved. In the future, for the latter cases we must seek to obtain useful spectra by studies in vivo. The MRI instruments currently available for such purposes involve magnetic fields of 1.5 to 4.0 Tesla, compared to the high resolution instruments reaching as high as 22 Tesla. The separation between the component resonances in NMR spectra is linearly dependent on the strength of the magnetic field, as is approximately the signal-to-noise ratio of a given NMR spectrum. One is tempted to measure in vivo NMR spectra at the highest field possible, but there are several impediments: ultra high field instruments of adequate bore size for human studies are prohibitively expensive; there is risk involved at high fields due to neural stimulation and deposition of radiofrequency power. Hence, these high field instruments are not likely to appear in hospitals due to cost concerns, installation problems (the magnets are huge), and safety concerns. We must optimize the resolution of spectra by careful deconvolution and design the various instrument components for optimum performance. Useful spectra have been obtained at reasonable fields (1.5–3.0 Tesla) and the quality of the spectra is improving rapidly. Thus, we believe that 3.0 or 4.0 Tesla instruments will be able to yield spectra that will be adequate for screening. An example is a ^1H NMR spectrum of bile in gall bladder in vivo in a commercial 3 Tesla instrument obtained in several minutes (Mohajeri, unpublished results).

The future looks bright!

References

- Albiin N, Smith ICP, Arnelo V, Lindberg B, Berquist A, Dolenko B, Bryksina N, Bezabeh T (2008) A detection of cholangiocarcinoma with magnetic resonance spectroscopy of bile in patients with and without primary sclerosing cholangitis. *Acta Radiol* 49:855–862
- Bezabeh T, Somorjai RL, Dolenko B, Bryksina 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 Biomed* 22:593–600
- Duda RO, Hart PE, Stork DG (2001) Pattern classification, 2nd edn. Wiley-Interscience, New York
- Eriksson L, Johansson E, Kettaneh-Wold N, Wold S (2001) Multi- and megavariate data analysis—principles and applications. Umetrics, Umea
- Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference and prediction, 2nd edn. Springer, Berlin
- Huber PJ (1985) Projection pursuit. *Ann Stat* 13:435–475
- Kuncheva LI (2004) Combining instance classifiers—methods and algorithms. Wiley-Interscience, New York
- 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* 11:209–217
- Somorjai RL (2009) Creating robust, reliable, clinically relevant classifiers from spectroscopic data. *Biophys Rev* 1:201–211
- 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 NMR spectra. I. Application to thyroid neoplasms. *Magn Res Med* 33:257–263
- Somorjai RL, Alexander M, Baumgartner R, Booth S, Bowman C, Demko A, Dolenko B, Mandelzweig M, Nikulin AE, Pizzi N, Prankeviciene E, Summers R, Zhilkin P (2004a) A data-driven, flexible machine learning strategy for the classification of biomedical data. In: Dubitzky W, 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—a novel, exact method for visualizing and characterizing high-dimensional instances. *J Biomed Inform* 37:366–379
- Witten IH, Frank E (2005) Data mining: practical machine learning tools and techniques, 2nd edn. Morgan Kaufmann, San Francisco