12

Complexity of Magnetic Resonance Spectrum Classification

Richard Baumgartner, Tin Kam Ho, Ray Somorjai, Uwe Himmelreich, Tania Sorrell

Summary. We use several data complexity measures to explain the differences in classification accuracy using various sets of features selected from samples of magnetic resonance spectra for two-class discrimination. Results suggest that for this typical problem with sparse samples in a high-dimensional space, even robust classifiers like random decision forests can benefit from sophisticated feature selection procedures, and the improvement can be explained by the more favorable characteristics in the class geometry given by the resultant feature sets.

12.1 Introduction

Biomedical spectra obtained by magnetic resonance (MR) spectroscopy are characterized by (a) high dimensionality and (b) a typically small number of available samples. A statistically meaningful analysis of a limited number of high-dimensional data points presents a serious challenge, due to the extreme sparsity of samples in high-dimensional spaces [13]. Dimensionality reduction techniques using feature selection and extraction provide a natural way to address this problem [10, 11, 14]. Interpretable feature selection is especially desirable in disease profiling applications when using biomedical data such as spectra or gene microarrays [8, 9, 14], because they provide hypotheses for the domain experts for further corroboration. The main goal of this chapter is to explore the utility of data complexity measures [6] in assessing several feature selection and extraction procedures in a real-world two-class discrimination problem using MR spectra.

Robust classifiers are needed to generalize the class boundary from severely limited training samples in a high-dimensional problem. Questions remain on how such classifiers interact with dimensionality reduction techniques. Recently, in the application of handwritten word recognition it has been demonstrated that feature selection and extraction can be beneficial for the random subspace method (RSM) [2]. In addition, a successful, standard application of RSM has been reported in disease profiling applications using high-dimensional gene microarray data, where the combination of feature selection and extraction with RSM was proposed as a topic for further research [1]. Motivated by the observations in [1] and [2], we use RSM classification accuracy as a guide for comparing features extracted by our algorithms.

12.2 Materials, Methods, and Data

Conventional biochemical techniques frequently have difficulty identifying closely related species or subspecies of fungi or yeasts. At best, the procedures are time-consuming. In contrast, MR spectroscopy, combined with multivariate classification methods, has proven to be very powerful. As a typical application of the methodology, we have used MR spectra of isolates of two pathogenic yeast species, *Candida albicans* and *Candida parapsilosis* [3].

The yeast colonies were suspended in phosphate-buffered saline made up with deuterated water. The suspension was immediately transferred to a 5-mm NMR tube (Wilmad Glass Co., Buena, NJ). The ^{1}H MR spectra were acquired at 37degC on a Bruker Avance 360-MHz MR spectrometer using a 5-mm ^{1}H, ^{1}3C inverse-detection dual-frequency probe. The following acquisition parameters were used: frequency 360.13 MHz, pulse angle 90deg, repetition time 2.3 s, spectral width 3600 Hz. For a more specific description of the technical details of the acquisition procedure, see [3]. Spectra were processed using the Xprep software (IBD, Winnipeg, Manitoba). The feature extraction and classification methods were carried out on the magnitude spectra. The dimensionality of the spectra was 1500, corresponding to intensity values in the range of 0.35 to 4.00 ppm. A typical spectrum is shown in Figure 12.1. The training set contained 124 spectra (62 in each class). The independent test set not used for classifier development contained 73 spectra (35 in class 1, *Candida albicans*, and 38 in class 2, *Candida parapsilosis*).

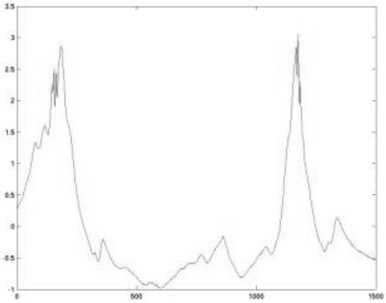

Fig. 12.1. A example of a magnetic resonance spectrum.

12.3 Dimensionality Reduction Techniques

Components of the raw spectral feature vectors are ordered by the channel frequencies. Because many spectral peaks and valleys have a naturally occurring width, intensity values in neighboring channels are expected to be correlated to some extent. To quantify this, we computed the 1500×1500 correlation coefficient matrix using the training set. Figure 12.2 shows the heat-map representation of this correlation matrix. We can clearly recognize the bands or clusters of highly correlated intervals of adjacent-neighboring features formed along the main diagonal. The motivation for the dimensionality reduction techniques under investigation is to take advantage of this structure in both an unsupervised and a supervised manner.

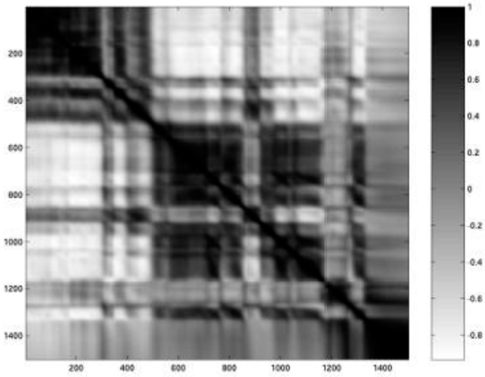

Fig. 12.2. Heat-map representation of the correlation matrix computed from the training set. The bands of highly correlated neighboring features are clearly recognizable.

In particular, let a sample in the original 1500-dim ($p_{dim} = 1500$) feature space be represented by a $1 \times p_{dim}$ row vector $s_{original}$. The dimensionality reduction is achieved using a $p_{dim} \times n_{feat}$ matrix B so that the sample in the reduced n_{feat}-dimensional space is given by the row vector $s_{reduced}$, where

$$s_{reduced} = s_{original} B.$$

The columns of the matrix B represent the basis functions onto which the data are projected. In the two feature extraction procedures we used, the new features are averages over intervals of neighboring features in the original spectrum. Thus, each column in the matrix B represents a basis function with nonzero values in some of the p_{dim} positions where the orignial features are to be averaged.

Averaging highly redundant features also has a smoothing effect and improves the signal-to-noise ratio. An example of a column of the matrix B is shown in Figure 12.3, where the feature interval 200 to 300 is averaged.

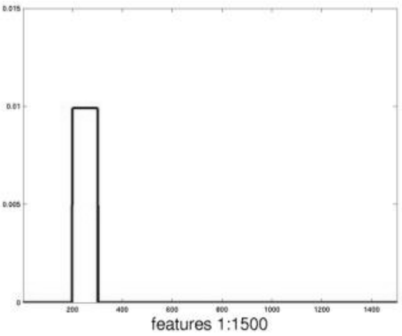

Fig. 12.3. An example of a basis function that averages the intensity values in the spectral interval 200 to 300.

12.3.1 Unsupervised Feature Extraction

In unsupervised feature selection, we determine the set of features to be averaged from the training set using a sequential clustering algorithm. Clusters of features are defined as intervals of neighboring frequency channels for which the pairwise correlation coefficient is not lower than a specific threshold. We start with the feature at position 1 and keep adding the neighboring features to the current feature cluster, as long as their correlation coefficients with the first feature are at or above the chosen threshold. If the criterion is violated, the current feature cluster is assumed to be complete. The (last) feature that caused the violation of the cluster criterion is declared the first feature of a new cluster and the procedure is repeated. The procedure ends with the last feature (at position p_{dim}). The number of feature clusters identified gives the dimensionality of the reduced space.

12.3.2 Supervised, Genetic-Algorithm-Driven Feature Extraction

Alternatively, features can be selected with some regard to their discriminating power for the two classes. The feature selection algorithm we have used for supervised feature extraction is the near-optimal region selector (ORS) [8, 10]. ORS searches for intervals of neighboring features that are maximally discriminatory. ORS is guided by a genetic algorithm (GA), explicitly optimized for preprocessing spectra. GA is particularly appropriate for spectra, since the latter are naturally representable as "chromosomes," vectors of length p_{dim}, with 1's indicating the presence and 0's the absence of features. The GA's input includes (1) n_{feat}, the maximum number of features, which is the number of distinct spectral subregions required in the type of dimensionality reduction operation/transformation to be carried out (averaging of the spectral windows); (2) the population size; (3) the number of generations; and (4) two random seeds. The operations comprise the standard GA options: mutation and crossover. To achieve robust classification, the number of desired features is typically kept much smaller than the sample size. GA_ORS begins by searching the entire feature space, i.e., the complete spectrum. The output is the set of (averaged) feature intervals that optimally separate the classes. GA_ORS is applied as a wrapper, for which the search of the feature space is guided by the leave-one-out accuracy of a linear discriminant analysis (LDA). Once n_{feat} ($\ll p_{dim}$) good features have been found, the results are validated using an independent test set that was not used in the feature extraction procedure.

12.4 Random Subspace Method and Decision Forests

The random subspace method (RSM)[4] is known to produce robust classifiers for many high-dimensional problems. The method combines the decision of a large number of classifiers with sufficient differences in the generalization power [7]. Each classifier is trained to perfection, but uses only a randomly selected subset of features. If the classes are unambiguous in the chosen subspace, the classifier is perfect and at the same time is insensitive to differences in the unselected features. As a result, it has some built-in generalization power to avoid overtraining.

RSM often performs better than any individual classifier in its collection. The individual classifiers are best taken to be decision trees, but success has been reported on nearest neighbor classifiers [5], pseudo-Fisher linear discriminants [12], and support vector machines [1]. The built-in defense against overtraining has made RSM useful in problems involving a large number of features where some may be redundant. Thus RSM appears to be promising in classifying MR spectra that typify such problems.

In RSM the degree of improvement over individuals varies with the specifics of the individual classifiers. There are variations between different fractions of features chosen to use, or between different realizations of the random projections. In the case of decision trees, variations are also observed between different types of splitting hyperplanes used in the trees. Such variations have not been thoroughly analyzed, but a large range usually means that the discriminating power is concentrated in a small number of features, so that their presence in the chosen set is important.

In our experiments we use RSM with decision trees that use oblique or axis-parallel splits. In oblique trees, at each internal node the tree splits the data set using a linear function of the features obtained by a simplified Fisher's procedure, central axis projection [4]. In axis-parallel trees, each node splits the data set using the feature that maximizes information gain. The classifier that uses RSM on decision trees is also called a random decision forest.

In Table 12.1 we report the test set accuracies when the random decision forest is applied to several feature sets resulting from variations of the two feature extraction procedures. For comparison, we also show the accuracies of a nearest-neighbor classifier using Euclidean distance, and the two types of decision trees applied individually without participating in an RSM ensemble. The feature sets are

- original: the 1500 dimensional vector containing the raw spectrum;
- GA averaged: the three features that are average intensities in three spectral regions selected by the genetic algorithm;
- GA regions: the concatenation of the intensity values in the three spectral regions selected by the genetic algorithm (channel 82–96, 908–933, and 1080–1242);
- corr 0.90: averages of 90 spectral windows selected by correlation coefficient clustering with a threshold at 0.9;
- corr 0.99: averages of 330 spectral windows selected by correlation coefficient clustering with a threshold at 0.99;
- corr 0.998: averages of 849 spectral windows selected by correlation coefficient clustering with a threshold at 0.998.

Table 12.1. Nearest neighbor, single decision trees, and random decision forest accuracies (% correct on the test set) using different sets of selected features.

Feature set	Original	GA averaged	GA regions	corr 0.90	corr 0.99	corr 0.998
Dimensionality	1500	3	204	90	330	849
1-nearest neighbor	91.78	91.78	79.45	86.30	87.67	91.78
1 oblique tree	82.19	87.67	75.34	78.08	84.93	82.19
1 axis-pl. tree	73.97	86.30	83.56	78.08	83.56	75.34
Random decision forest	94.52	95.89	90.41	90.41	94.52	91.78

This problem demonstrates an extreme case where only a few of a large number of features are relevant for discrimination, and the number of available training samples is very small compared to the feature space dimensionality. In such a space, many single classifiers would suffer from overtraining, as we can see from single tree accuracies in Table 12.1. But the RSM ensembles are robust and are able to take advantage of the large number of features. Nevertheless, applying sophisticated feature selection techniques is still important, as evidenced by the accuracy improvement achieved by the feature set GA-averaged on the RSM ensemble, as well as on the single tree classifiers. Moreover, with better features to train on, RSM also shows less performance variation among different training options.

12.5 Evaluation of the Feature Set Complexity

To find explanations for the differences in classification accuracies using these feature sets, we computed the values of several measures of classification complexity as described in [6]. More details about these measures can be found in other chapters in this volume. The measures used in this experiment are

1. `boundary`: length of the class boundary estimated by the fraction of points on a class-crossing edge in a minimal spanning tree
2. `intra-inter`: ratio of average intra-class nearest-neighbor distance to average inter-class nearest-neighbor distance
3. `pretop`: fraction of points with the maximal within-class covering ball not fully contained in other balls
4. `overlap`: overlap volume of the class bounding boxes
5. `maxfeaeff`: maximum feature efficiency, or the largest fraction of points classified by a single feature
6. `nonlin-NN`: nonlinearity of nearest neighbor classifier
7. `nonlin-LP`: nonlinearity of linear classifier minimizing error distance

Table 12.2 lists the values of these measures computed from the training samples (TR) and the test samples (TE) represented by each feature set. There are some indications that the GA-averaged feature set makes the classification problem easier in at least three ways: (1) it puts fewer points on boundary, (2) the classes have less spread (lower `intra-inter` ratio), and (3) the classes are more spherical (smaller `pretop` value). Two of the metrics, volume of `overlap` and maximum feature efficiency, are heavily affected by the orientation angle of the class gap, and are thus not too revealing in this problem. The nonlinearity of the nearest neighbor classifier is relatively low for the GA-averaged feature set, suggesting that with this feature set, the nearest neighbor boundary can largely avoid cutting off part of the convex hulls of the two classes to the wrong side of the decision surface. On the other hand, the nonlinearity of the linear classifier is actually higher for the GA-averaged feature set while being zero for all others. This suggests that linear separability by itself does not necessarily give an easily learnable problem; sparse samples in a high-dimensional space have a higher chance to be linearly separable, but severe overtraining may prevent the learning algorithms from yielding a satisfactory classifier.

12.6 Conclusion

We described a study of MR spectra classification where two feature selection and extraction procedures were used to derive several feature sets representing the problem. We applied decision forests constructed with the random subspace method to each feature set, and observed that the averaged intensity values in three frequency windows selected by a genetic algorithm yielded the most accurate classifier. We further attempted to explain the superiority of this feature set using several measures of data complexity, and observed that its higher utility in classification is consistent with favorable values with three measures most relevant to the class geometry.

We expect that more studies along this line will lead to a way of using the data complexity measures to guide feature selection for classification. In this methodology, feature selection procedures may be designed for minimizing classification complexity, as measured by the most useful descriptors of the class geometry.

Table 12.2. Complexity measures of the feature sets on the training sample (TR) or test sample (TE).

		Original	GA averaged	GA regions	corr 0.90	corr 0.99	corr 0.998
boundary	TR	0.25	0.16	0.40	0.26	0.27	0.27
	TE	0.27	0.22	0.40	0.32	0.37	0.33
intra-inter	TR	0.59	0.41	0.69	0.64	0.62	0.61
	TE	0.58	0.46	0.66	0.64	0.63	0.60
pretop	TR	0.98	0.91	1.00	1.00	1.00	0.99
	TE	1.00	0.93	0.99	0.99	1.00	1.00
overlap	TR	0.00	0.38	0.00	0.00	0.00	0.00
	TE	0.00	0.29	0.00	0.00	0.00	0.00
maxfeaeff	TR	0.27	0.14	0.19	0.23	0.26	0.27
	TE	0.52	0.23	0.40	0.49	0.52	0.52
nonlin-NN	TR	0.05	0.02	0.11	0.07	0.06	0.06
	TE	0.05	0.08	0.14	0.07	0.06	0.05
nonlin-LP	TR	0.00	0.01	0.00	0.00	0.00	0.00
	TE	0.00	0.03	0.00	0.00	0.00	0.00

References

[1] A. Bertoni, R. Folgieri, G. Valentini. Bio-molecular cancer prediction with random subspace ensembles of support vector machines. *Neurocomputing*, 63C, 535–539, 2005.

[2] S. Gunter, H. Bunke. Feature selection algorithms for the generation of multiple classifier systems and their application to handwritten word recognition. *Pattern Recognition Letters*, 25(11), 1323–1336, 2004.

[3] U. Himmelreich, R.L. Somorjai, B. Dolenko B, et al. Rapid identification of Candida species by using nuclear magnetic resonance spectroscopy and a statistical classification strategy. *Applied and Environmental Microbiology* 69(8), 4566–4574, 2003.

[4] T.K. Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8), 832–844, 1998.

[5] T.K. Ho. Nearest neighbors in random subspaces. In *Proceedings of the 2nd International Workshop on Statistical Techniques in Pattern Recognition*, Sydney, Australia, August 11–13, 1998, pages 640–648.

[6] T.K. Ho, M. Basu. Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5), 289–300, 2002.

[7] E.M. Kleinberg. Stochastic Discrimination. *Annals of Mathematics and Artificial Intelligence*, 1, 207–239, 1990.

[8] C.L. Lean, R.L. Somorjai, I.C.P. Smith, P. Russell, C.E. Mountford. Accurate diagnosis and prognosis of human cancers by proton MRS and a three stage classification strategy. *Annual Reports on NMR Spectroscopy* 48, 71–111, 2002.

[9] C. Mountford, R. Somorjai, P. Malycha, et al. Diagnosis and prognosis of breast cancer by magnetic resonance spectroscopy of fine-needle aspirates analyzed using a statistical classification strategy. *British Journal of Surgery*, 88(9), 1234–1240, 2001.

[10] A. Nikulin, B. Dolenko, T. Bezabeh, R. Somorjai. Near optimal region selection for feature space reduction: novel preprocessing methods for classifying MR spectra. *NMR in Biomedicine*, 11, 209–216, 1998.

[11] S. Raudys, A. Jain. Small sample size effects in statistical pattern recognition: recommendation for practitioners. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 13(3), 252–264, 1998.

[12] M. Skurichina, R.P.W. Duin. Bagging, boosting, and the random subspace method for linear classifiers. *Pattern Analysis and Applications*, 5(2), 121–135, 2002.

[13] R. Somorjai, B. Dolenko, R. Baumgartner. Class prediction and discovery using gene expression and proteomics mass spectroscopy data. Curses, caveats, cautions. *Bioinformatics*, 19, 1484–1491, 2003.

[14] R. Somorjai, B. Dolenko, A. Nikulin, et al. Distinguishing normal from rejecting renal allografts: application of a three-stage classification strategy to MR and IR spectra of urine. *Vibrational Spectroscopy*, 28, 97–102, 2002.