ELSEVIER

Magnetic Resonance Imaging 22 (2004) 251–256

MAGNETIC RESONANCE IMAGING

Unsupervised feature dimension reduction for classification of MR spectra

R. Baumgartner[a], R. Somorjai[a,*], C. Bowman[a], T.C. Sorrell[b], C.E. Mountford[b], U. Himmelreich[b]

[a]*Institute for Biodiagnostics, National Research Council Canada, Winnipeg, Canada*
[b]*Institute for Magnetic Resonance Research and Department of Magnetic Resonance in Medicine, University of Sydney, Sydney, Australia*

Received 19 February 2003; received in revised form 27 August 2003

Abstract

We present an unsupervised feature dimension reduction method for the classification of magnetic resonance spectra. The technique preserves spectral information, important for disease profiling. We propose to use this technique as a preprocessing step for computationally demanding wrapper-based feature subset selection. We show that the classification accuracy on an independent test set can be sustained while achieving considerable feature reduction. Our method is applicable to other classification techniques, such as neural networks, support vector machines, etc. © 2004 Elsevier Inc. All rights reserved.

Keywords: Feature selection; Extraction; MR spectroscopy; Classification

1. Introduction

Biomedical spectra obtained by MR spectroscopy are characterized by 1) high dimensionality and 2) scarcity of available samples. A statistically meaningful analysis of a limited number of high-dimensional data points presents a serious challenge due to the extreme sparseness of high-dimensional spaces. It is generally accepted by the pattern recognition community that robust classifier development requires 5–10 samples per feature [1,2]. Hence, some form of feature selection/extraction provides a natural way to address this problem [3,4] . Feature selection/extraction is especially desirable in disease profiling applications when using biomedical spectra [5,6], for which the main interest lies in identifying discriminatory spectral regions (adjacent spectral intensities).

Fortunately, MR spectral features are highly redundant, suggesting that the data do not span the entire (original) high-dimensional space; instead, they lie on (or close to) some low-dimensional manifold. Neighboring spectral features of MR spectra are highly correlated; in fact, they are almost identical and therefore form natural clusters [3]. We present a feature reduction method, using unsupervised clustering that exploits the highly correlated characteristics of neighboring spectral intensities. We propose this technique as a preprocessing step for wrapper-based feature extraction procedures.

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*.

2. Materials and methods

The yeast colonies were suspended in phosphate-buffered saline (PBS). The suspension was immediately transferred to a 5-mm NMR tube (Wilmad Glass Co, Inc, Buena, NJ, USA).

The ^{1}H MR spectra were acquired at 37C on a Bruker Avance 360 MHz MR spectrometer using a 5-mm inverse-detection dual frequency probe. Spectra were processed using XWINMR 2.6 software. The feature extraction-classification was carried out on the magnitude spectra.

For the mathematical description of a two-class classifi-

* Corresponding author. Tel: +204-984-4538; fax: +204-984-5472.
E-mail address: ray.somorjai@nrc-cnrc.gc.ca (R. Somorjai).

0730-725X/04/$ – see front matter
doi:10.1016/j.mri.2003.08.033

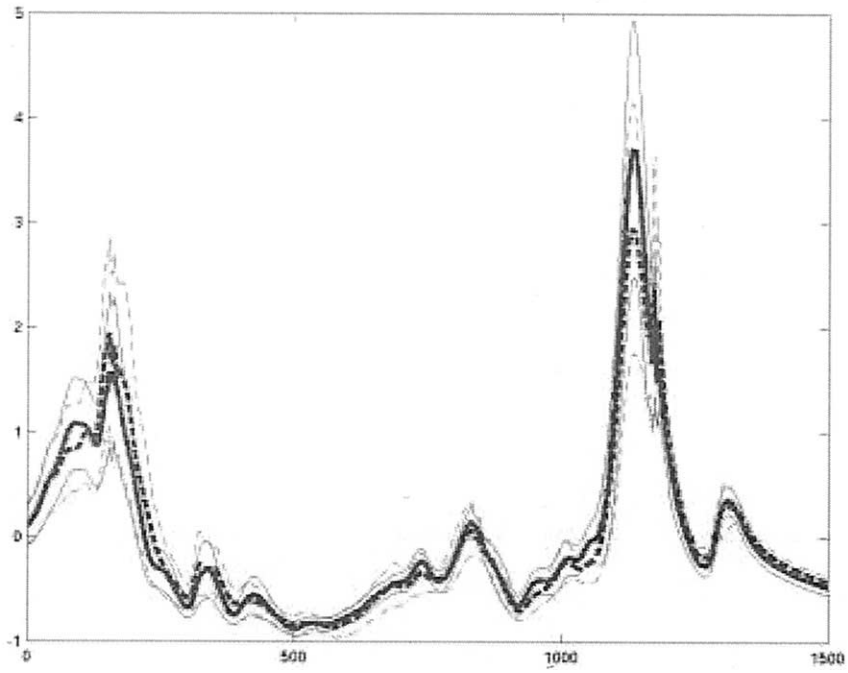

Fig. 1. Thick lines represent the centroids of the two classes as given by a training set (– centroid(A_m), — centroid(B_m)). Thin lines (–, —) represent the variation of the features across the spectra (samples); –, centroid(A_m) ± stdev(A_m); and —, centroid(B_m) ± stdev(B_m). Note the high overlap between the classes.

cation problem, see the Appendix I. The dimensionality of the spectra (number of spectral features) was 1500 (p_{dim} = 1500). The training set contained 124 (ntr_1 = 62, ntr_2 = 62), the test set 73 ($ntest_1$ = 35, $ntest_2$ = 38) spectra (samples).

Our algorithm is based on previous work on feature extraction/selection [7–9]. The goal of the algorithm is to identify clusters of highly correlated neighboring features. We achieve this by assigning neighboring features to a cluster and assessing redundancy in the currently identified cluster. As a criterion of redundancy, we use the minimum correlation coefficient of the correlation matrix of the cluster. Thus, we require that all pairwise correlations between the features in a cluster be very high, above a preselected threshold.

The pseudocode for identifying feature clusters is given below. Consider M_{train}:

Algorithm:

```
Select a correlation threshold;
Let the current cluster be the empty set;
Create a new current cluster containing only the 1st
    feature M_train(:,1);
for i = 2 to p_dim

add feature i to the current cluster
calculate the correlation matrix (CC) of the current
    cluster
if minimum(CC) < threshold then create new current
    cluster containing only the ith feature
end if
end for
```

Note that the dimensionality of the correlation matrix of the current cluster is equal to the number of elements (features) in the cluster.

After clusters of features are identified, each cluster is represented by its feature centroid (mean). Thus the dimensionality of the data (number of spectral features) reduces to the number of clusters identified n_{clus}. Finally, we calculate the correlation matrix of the whole reduced data set. If min(CC) < threshold, it indicates that further clustering, now among non neighboring features may be necessary. This rarely happens in practice.

At the next step, we used the reduced data as input to a supervised wrapper-based feature selection procedure, based on dynamic programming and least-trimmed-square classification [3]. In a wrapper-based approach [10], a classification algorithm is used to optimize subsets of features, generally using crossvalidation on the training set. Once the optimal subset of features is selected, the same classification algorithm is used to classify unseen samples, from a test set. Thus the classifier is "wrapped" around both the training and the test set.

3. Results

Fig. 1 demonstrates the difficulty of the classification problem. The two class centroids and the standard devia-

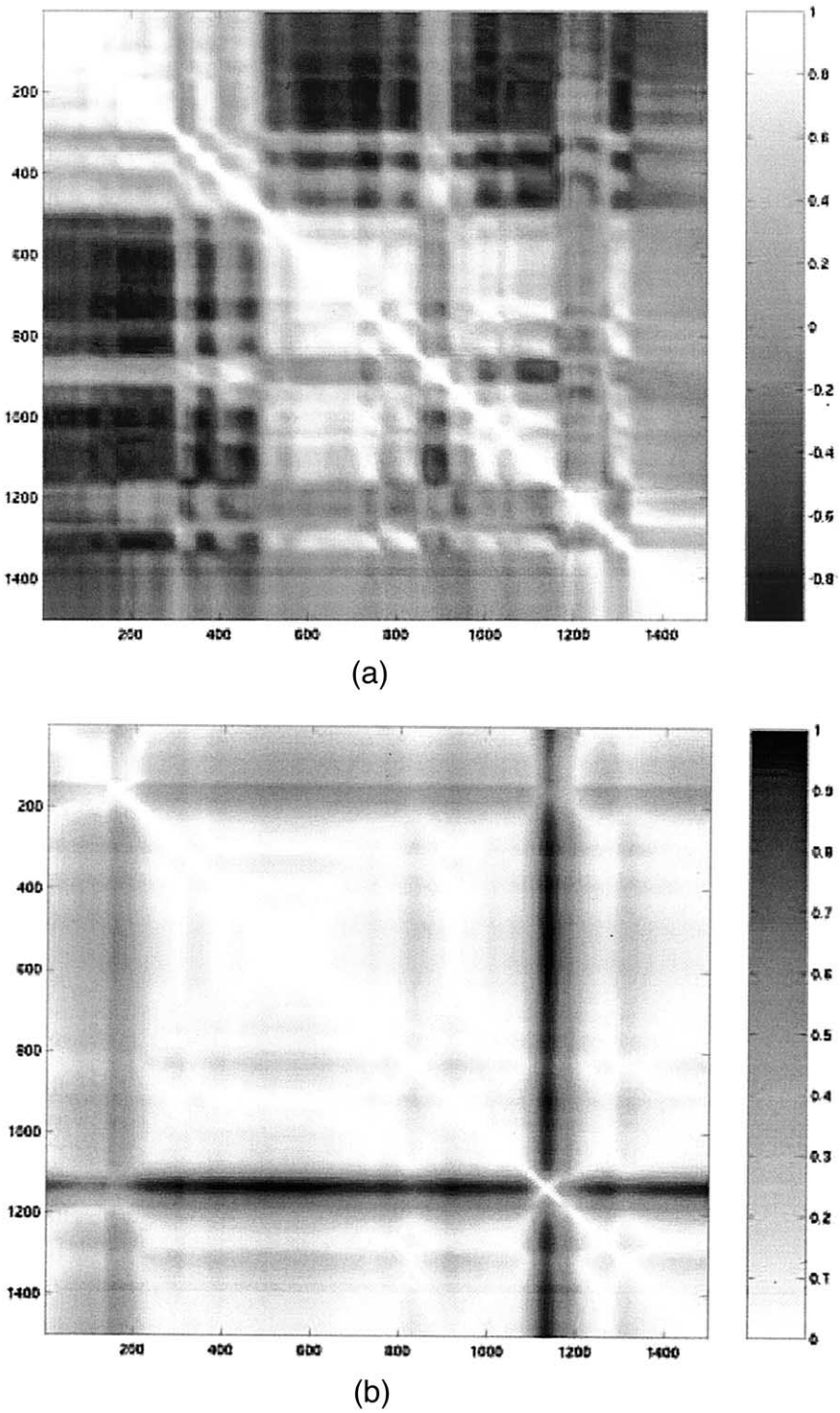

Fig. 2. (a) Correlation matrix of the spectral features. (b) Distance matrix of spectral features.

tions of the two data sets (light solid and dashed lines) overlap strongly.

Fig. 2a displays the correlation matrix, and Fig. 2b, the distance matrix of the features for the original feature space, respectively (the dimensionality of the matrix is 1500 × 1500, the main diagonal of the correlation matrix contains ones). Note the strong correlation between neighboring features (spectral intensities), manifested by the high-intensity bands formed along the main diagonal of the correlation matrix. Note similar bands in the distance matrix, indicating that the neighboring spectral features are in fact almost identical.

Figs. 3 and 4 show the influence of the correlation coefficient threshold on dimensionality (Fig. 3) and on clas-

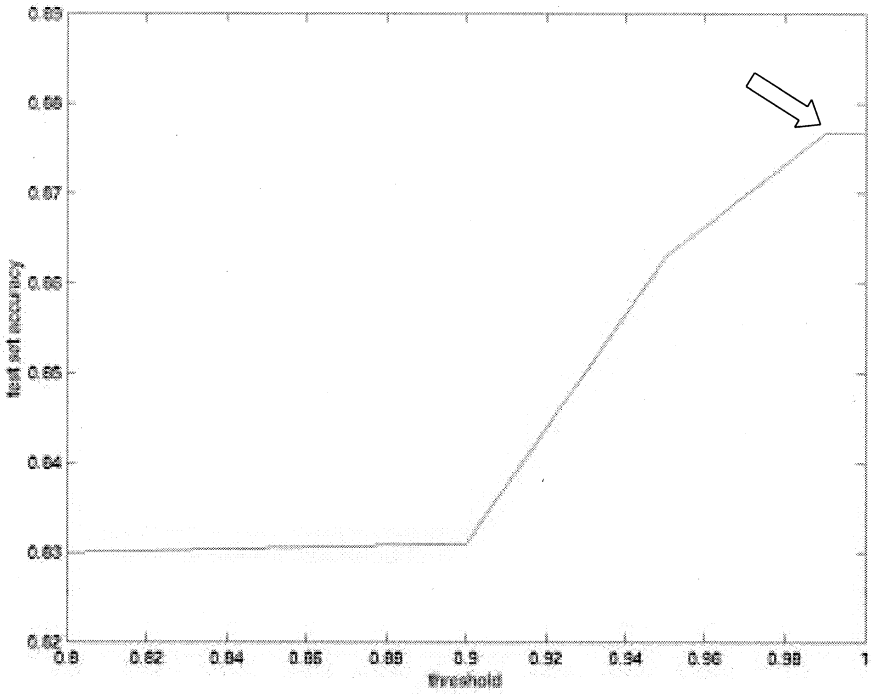

Fig. 3. Accuracy of independent test set vs. threshold. The "optimum" threshold, identified by a decrease in classification accuracy, is denoted by the arrow.

sification accuracy for the independent test set (Fig. 4). The correlation coefficient threshold ranges between 0.8-1. High accuracy on the independent test set is maintained throughout this parameter range, although there is a clear trade-off between accuracy and dimensionality reduction. Note the decrease in accuracy at the "optimum" threshold value (threshold = 0.99). The feature space dimensionality at this threshold is 330, 22% of the original 1500-dimensional space. Fig. 5 shows the reduction of the dimensionality of the original feature space; displayed along the horizontal

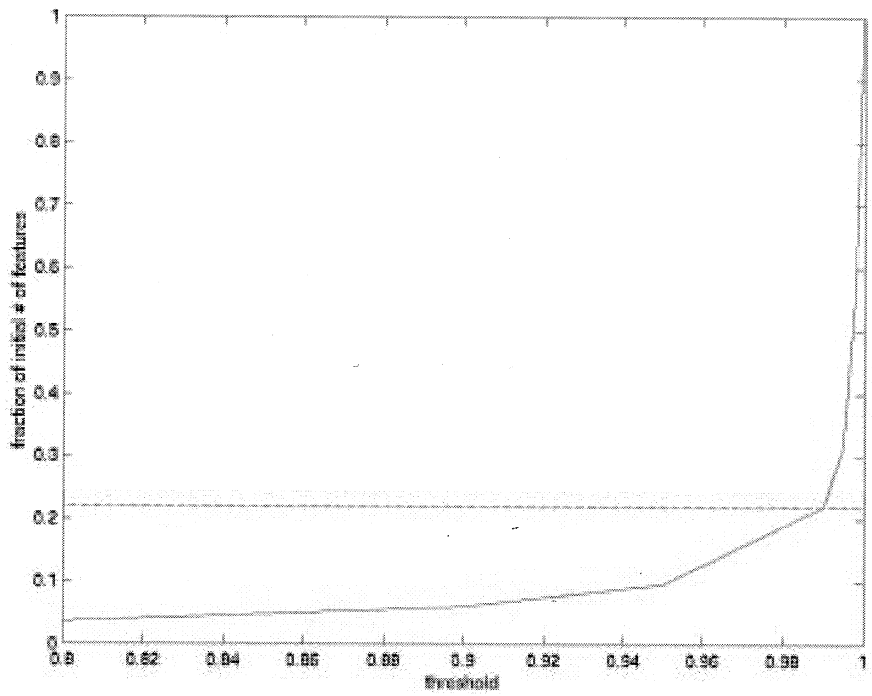

Fig. 4. Feature dimension reduction vs. threshold. At the "optimum" threshold, the feature dimensionality is 22% of the original feature dimension.

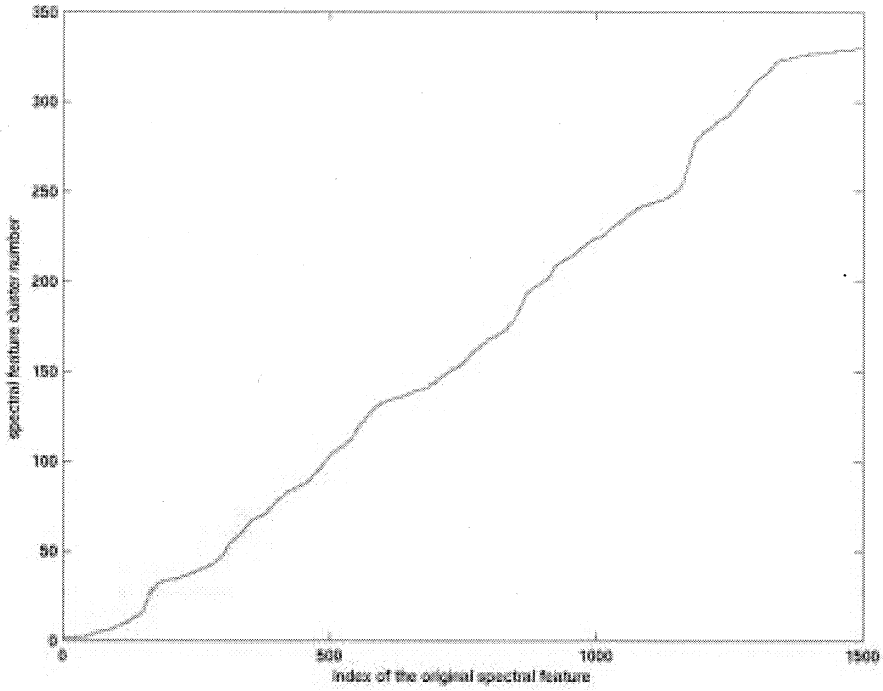

Fig. 5. Reduction of the original feature space to 330 features (number of clusters on the vertical axes).

axis is the index of the feature in the original space, on the vertical axes we show the cluster number to which the spectral feature was assigned. (Note that the maximum number on the vertical axis is 330.) When lowering the threshold below the "optimum," the accuracy decreases.

4. Discussion

We have demonstrated the real-life utility of a clustering-based feature reduction technique by applying it to the classification of MR spectra. The method uses a specific

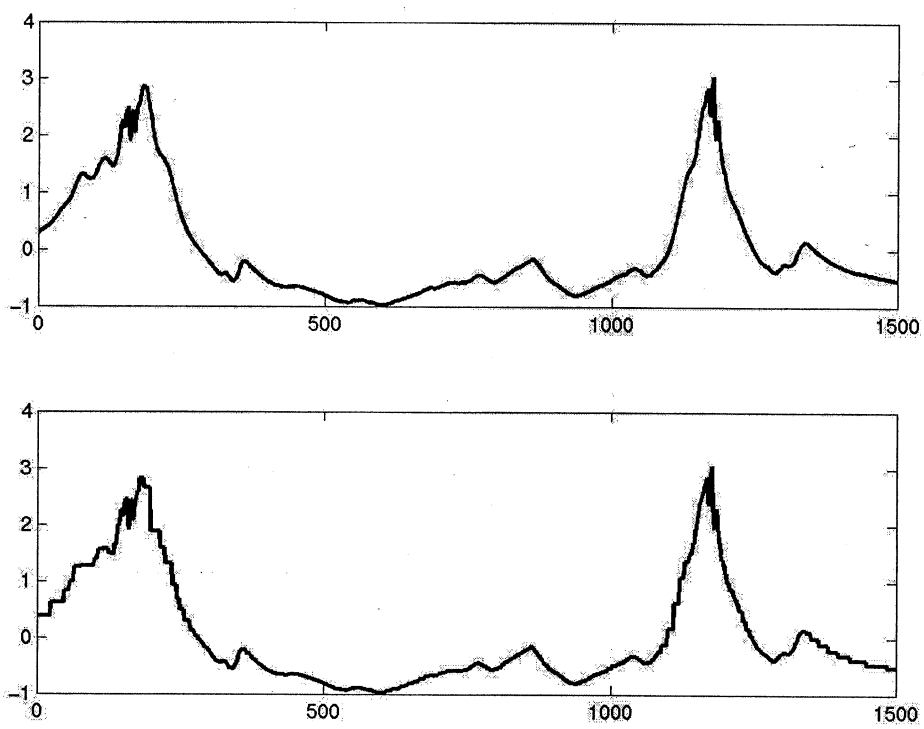

Fig. 6. Spectrum for sample 1. The top plot shows the original spectrum, the bottom after its reduction to 330 features (99% correlation threshold).

characteristic of biomedical spectra, i.e., the high correlation between neighboring spectral features. Because averaging neighboring spectral features forms the feature clusters, this technique retains spectral identity. The dimensionality of the spectra was significantly reduced while maintaining high accuracy on the independent test set. Consequently, the combinatorial explosion due to the large number of features was effectively prevented when using the dynamic programming-based feature extraction. Note that data processed by our technique may be used in connection with any other optimization technique such as a genetic algorithm for feature extraction or other classification techniques, including neural networks [11], support vector machines etc. It may be also used as a preprocessing step for more sophisticated dimension reduction techniques [12].

Note that our method is not confined to MR spectra; initial experiments show similar success for ovarian cancer profiling using mass spectroscopy [6].

Appendix. Notation and mathematical description of the problem

In real-life classification settings, one usually works with two data sets, the training and the test set. The labels of the samples (spectra) in the training set are used in developing the classifier. The classifier is then applied to predict the labels of the spectra in an independent test set. The known labels in the test set are used only in evaluating the prediction accuracy of the classifier.

Let us consider a two-class classification problem, with p_{dim} spectral peaks (features). Define:

$A_m = (a_{m\ ij})$ is a $ntr_1 \times p_{dim}$ matrix of the training samples from the first class, with ntr_1 samples (spectra) in the first class,

$B_m = (b_{m\ ij})$ is a $ntr_2 \times p_{dim}$ matrix of the training samples from the second class, with ntr_2 samples (spectra) in the second class,

$A_t = (a_{t\ ij})$ is a $ntest_1 \times p_{dim}$ matrix of the test samples from the first class, with $ntest_1$ samples (spectra) in the first class,

$B_t = (b_{t\ ij})$ is a $ntest_2 \times p_{dim}$ matrix of the test samples from the second class, with $ntest_2$ samples (spectra) in the second class.

Then the training set can be represented as a $(ntr_1 + ntr_2) \times p_{dim}$ matrix $M_{train} = [A_m', B_m']'$, where ' denotes the matrix transpose.

The test set can be represented as a $(ntest_1 + ntest_2) \times p_{dim}$ matrix $M_{test} = [A_t', B_m']'$.

The centroid of a group of spectra (i.e., summation is carried out across samples) represented by a $n_A \times p_{dim}$ matrix A is a p_{dim}-dimensional row vector, whose ith element is given by: $\text{sample\_centroidA}(i) = (1/n_A)\ \Sigma A(:,i)$, where (:,i) stands for ith column of matrix A; The corresponding $p_{dim} \times p_{dim}$ correlation matrix is $CC = (cc_{ij})$, where cc_{ij} is the correlation coefficient between ith and jth feature vector (column of matrix A).

The corresponding distance matrix is $DD = (dd_{ij})$, where dd_{ij} is the Euclidean distance between ith and jth feature vector (column of matrix A).

The centroid of a group (cluster) of features (i.e., summation is carried out across features) represented by a $n_A \times n_{feat}$ matrix A ($n_{feat} \leq p_{dim}$), where feature_centroid is a n_A-dimensional column vector, whose ith element is given by:

$$\text{feature\_centroidA}(i) = (1/n_{feat})\ \Sigma A(i,:),$$

where (i,:) stands for ith row of matrix A.

The standard deviation of a sample centroid is given by a p_{dim}-dimensional vector, whose i-th element is given by the standard deviation of the i-th column of matrix A, $\text{stdevA}(i) = \text{stdev}(A(:,i))$.

References

[1] Raudys S, Jain A. Small sample size effects in statistical pattern recognition: recommendation for practitioners, IEEE Trans Pattern Analysis Machine Intell 1991;13:252–64.

[2] Somorjai R, Janeliunas A, Baumgartner R, Raudys S. Comparison of two classification methodologies on a real-world biomedical problem. In: Caelli T, Amin A, Duin R, Kamel M, de Ridder D, editors. Springer notes on computer science. Berlin: Springer-Verlag, 2002. p. 433-42.

[3] Nikulin A, Dolenko B, Bezabeh T, Somorjai R. Near optimal region selection for feature space reduction: novel preprocessing methods for classifying MR. SpectraNMR Biomed 1998;11:209–16.

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

[5] Mountford C, Somorjai R, Malycha P, et al. Diagnosis and prognosis of breast cancer by magnetic resonance spectroscopy of fine-needle aspirates analyzed using a statistical classification strategy. Br J Surg 2001;88:1234–40.

[6] Petricoin E, Ardekani A, Hitt B, et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359:572–7.

[7] Das SK. Feature selection with a linear dependence measure. IEEE Trans Computers 1971;20:1106-9.

[8] Jain AK, Dubes R. Feature definition in pattern recognition with small sample size. Pattern Recognition 1978;10:85–97.

[9] Mitra P, Murthy CA, Pal SK. Unsupervised feature selection using feature similarity. IEEE Trans Pattern Recognition Machine Intell 2002;24:301–12.

[10] Kohavi R, John G. The wrapper approach. In: Huan L, Motada H, editors. Feature extraction, construction and selection. Dordrecht: Kluwer Academic Publishers, 1998. p. 33-51.

[11] Axelson D, Bakken I, Gribbestad I, Ehrnholm B, Nilsen G, Aasly J. Applications of neural network analyses to in vivo ^{1}H magnetic resonance spectroscopy of Parkinson disease patients. J Magn Reson Imaging 2002;16:13–20.

[12] Bowman C, Baumgartner R, Somorjai R. Dimensionality reduction for biomedical spectra. IEEE Canadian Conference on Computer and Electrical Engineering Proceedings, vol. 2, 2002. p. 1073-6.