Available online at www.sciencedirect.com
SCIENCE DIRECT®
Pattern Recognition Letters 27 (2006) 1383–1389
Pattern Recognition Letters
www.elsevier.com/locate/patrec

Data complexity assessment in undersampled classification of high-dimensional biomedical data

R. Baumgartner *, R.L. Somorjai

Institute for Biodiagnostics, National Research Council Canada, 435 Ellice Avenue, Winnipeg, Man., Canada R3B 1Y6

Received 27 June 2005; received in revised form 18 December 2005
Available online 20 March 2006

Communicated by Prof. F. Roli

Abstract

Regularized linear classifiers have been successfully applied in undersampled, i.e. small sample size/high dimensionality biomedical classification problems. Additionally, a design of data complexity measures was proposed in order to assess the competence of a classifier in a particular context. Our work was motivated by the analysis of ill-posed regression problems by Elden and the interpretation of linear discriminant analysis as a mean square error classifier. Using Singular Value Decomposition analysis, we define a discriminatory power spectrum and show that it provides useful means of data complexity assessment for undersampled classification problems.

In five real-life biomedical data sets of increasing difficulty we demonstrate how the data complexity of a classification problem can be related to the performance of regularized linear classifiers. We show that the concentration of the discriminatory power manifested in the discriminatory power spectrum is a deciding factor for the success of the regularized linear classifiers in undersampled classification problems. As a practical outcome of our work, the proposed data complexity assessment may facilitate the choice of a classifier for a given undersampled problem.

Keywords: Classification; Data complexity; Regularization; Undersampled biomedical problems

1. Introduction

Regularized linear discriminants have been successfully applied in undersampled, (i.e. small sample size combined with high dimensionality), biomedical classification problems e.g., for gene microarrays or biomedical spectra. They are often competitive with other (nonlinear) state-of-art algorithms (Tibshirani et al., 2002). In the small sample size–high dimensionality scenario, overfitting is a major issue (Somorjai et al., 2003a; Simon et al., 2004; Ambroise and McLachlan, 2002). Due to their limited capacity, the use of linear classifiers in undersampled biomedical problems has been advocated (Simon et al., 2004). There are numerous successful applications of regularized linear classifiers in undersampled, and in particular, biomedical problems; they include partial least squares (PLS) (Nguyen and Rocke, 2002), support vector machines (SVM) (Guyon et al., 2002), pseudoinverse linear discriminant analysis (LDA) (Ye et al., 2004), shrunken centroids (Tibshirani et al., 2002) etc.

Data complexity measures for classification have been proposed by Ho and Basu (2002) in order to discern the behavior of the classification algorithms in a given context.

Motivated by the analysis of (Elden, 2004) linear regression problems, we developed an approach to estimate data complexity in the small sample size/high dimensionality classification problems, using singular value decomposition (SVD) analysis and the minimum square error (MSE) formulation of the linear discriminant analysis. Using five real-life biomedical datasets of increasing difficulty, we

* Corresponding author. Fax: +1 204 984 5472.
E-mail address: richard.baumgartner@nrc-cnrc.gc.ca (R. Baumgartner).

0167-8655/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2006.01.006

show how the data complexity measures of a given classification problem can be related to the performance of linear classifiers.

2. Theoretical background

2.1. Linear regression and the singular value decomposition

Let X be a centered data matrix with dimensions $(n, p_{\dim})$. Consider the singular value decomposition (SVD) of X:

$$X = USV^{\mathrm{T}} = TV^{\mathrm{T}}, \tag{1}$$

where U is an $n \times n$ matrix of left singular vectors as columns, V is a $p_{\dim} \times p_{\dim}$ matrix of right singular vectors as columns, S is an $n \times p_{\dim}$ matrix, T is an $n \times p_{\dim}$ matrix.

The columns of U (or T), corresponding to the nonzero singular values of X, form a new set of eigenfeatures, which are usually referred to as principal components (Elden, 2004; Phatak and De Jong, 1997; Jollife, 1986). In the more specific context of microarray data analysis, they are also called eigengenes (Bair et al., 2005). They represent new coordinates in the space spanned by right singular vector of the data matrix (columns of the matrix V), or equivalently, to the eigenvectors of the matrix $X^{\mathrm{T}}X$ and the covariance matrix $\frac{1}{n-1}(X^{\mathrm{T}}X)$. Traditionally, the principal components are ordered according to the magnitude of the singular values of the matrix X (also the eigenvalues of matrix $X^{\mathrm{T}}X$ or the eigenvalues of the sample covariance matrix $\frac{1}{n-1}(X^{\mathrm{T}}X)$). The directions of right singular vectors of the matrix X (or the eigenvectors of the covariance matrix) correspond geometrically to the directions of the maximum variance of the data (Jollife, 1986). Note, that the ordering based on the magnitude of the singular values of X is given solely by the properties of the data matrix X. Hence, selecting the principal components according to this criterion is done in an unsupervised manner and corresponds to traditional principal component regression.

In particular, if we define $S_r = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$, where

$\sigma_1 \geqslant \sigma_2 \geqslant \cdots \sigma_r > 0$, are the singular values of X, and

$r = \mathrm{rank}(X) \leqslant \min(n, p_{\dim})$ is the rank of X

(in exact arithmetic),

then the matrix S can be written as

$$S = \begin{pmatrix} S_r & 0 \\ 0 & 0 \end{pmatrix}.$$

Furthermore, let $\lambda_i = \sigma_i^2$ be the ith eigenvalue of the matrix $X^{\mathrm{T}}X$; then the amount of variance explained by the ith principal component is given by λ_i, so that the normalized amount of variance explained by ith principal component is given by

$$\lambda_i \Big/ \sum_{j=1}^{r} \lambda j. \tag{2}$$

Consider now the least-squares fitting problem:

$$y \approx Xw,$$

$$w = \arg\min R^2 = \arg\min \|y - Xw\|_2, \tag{3}$$

where y is a $n \times 1$ vector of target values, $\| \ \|_2$ denotes L_2 norm; R^2 is also referred to as the residual of the least square problem (Elden, 2004).

The solution of (3) is given by the solution of the normal equations:

$$X^{\mathrm{T}}Xw = X^{\mathrm{T}}y. \tag{4}$$

As shown, e.g., in (Elden, 2004), given k singular values with their corresponding singular vectors, the objective function of (2) (i.e. the residual R_k^2), can be expressed using the SVD of X as follows:

$$R_k^2 = \sum_{i=k+1}^{n} (u_i^{\mathrm{T}}y)^2. \tag{5}$$

Thus, to obtain the minimum R_k^2, one should order the principal components according to the magnitudes of the components $U^{\mathrm{T}}y$, as has been proposed for principal component regression (PCR) (Jollife, 1986). This ordering was also suggested for the modified truncated SVD in (Elden, 2004). Note, that in contrast to the ordering of the principal components by the magnitudes of the singular values of the data matrix X, this ordering is based not only on the properties of the data, but it also takes into account the vector of target values (y). Therefore, this ordering is done in a supervised manner and it can be understood as a supervised version of the PCR (Jollife, 1986; Elden, 2004).

If $p_{\dim} > n(\mathrm{rank}(X) = \mathrm{rank}(X^{\mathrm{T}}X) \leqslant n)$, then the problem (3) is undersampled or ill-posed, and some form of regularization is desirable to obtain a unique and stable solution (Hansen, 1998; Vapnik, 1999).

2.2. Linear discriminant analysis (LDA) as a mean square error classifier

Consider the formulation of the LDA as a MSE classifier. In particular, if we consider a 2-class problem, with the n_1 samples of class 1 as rows of the matrix X_1 and the n_2 samples of class 2 as rows of the matrix X_2, then we can define the data matrix X as

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \text{ and the target vector } y \text{ as } y = \begin{pmatrix} \frac{n}{n_1} e_{n1} \\ -\frac{n}{n_2} e_{n2} \end{pmatrix},$$

where e_{n_i} is the $n_i \times 1$ vector with elements 1.

Then we can formulate the Fisher linear discriminant as a solution of the problem given by Eq. (3). This formulation is also referred to as a minimum square error (MSE) formulation of the LDA (Duda et al., 2000).

Thus, the mass of the vector y along the ith left singular vector u_i gives the degree of discriminatory power of ith principal component. Therefore, if we keep the commonly used ordering of the components, based on the singular

values of the matrix X as given above, but we attribute to each component u_i the magnitude of $u_i^{\mathrm{T}}y$, we obtain the "discriminatory" power spectrum for the dataset under study. As it is the case for the normalized amount of variance explained by the ith principal component in the unsupervised case given by Eq. (2), we define the normalized amount of the discriminatory information carried by the ith principal component by

$$|u_i^{\mathrm{T}}y| \Big/ \sum_{j=1}^{r} |u_j^{\mathrm{T}}y|. \tag{6}$$

3. Data

The first four datasets we used comprised samples of magnetic resonance (MR) spectra. An MR spectrum is a collection of intensity peaks and valleys that carry discriminatory information due to the physical/chemical basis of the class separation (Lean et al., 2002). A typical MR spectrum (taken from dataset 2) is shown in Fig. 1a. Datasets 1–3 derive from the fast identification of pathogenic yeasts, using MR spectra (Himmelreich et al., 2003). Dataset 4 comprises MR spectra of biofluids obtained from normal subjects and cancer patients (Somorjai et al., 2003b).

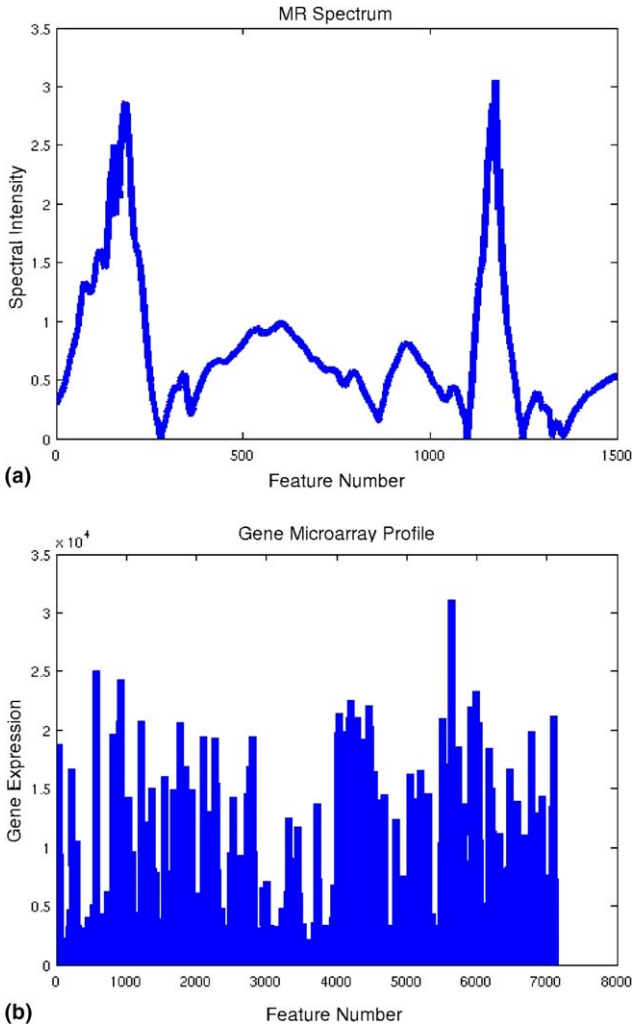

Fig. 1. (a) Representative MR spectrum. (b) Representative sample of a gene microarray profile.

The fifth dataset we used is a well-known benchmark of DNA microarray leukemia profiling. Samples in this dataset contain a vector of gene expressions from a number of subjects with two different types of leukemia (Golub et al., 1999). A typical sample from this dataset is shown in Fig. 1b.

The dimensionality and the sample size of each dataset are given in Table 1.

Table 1
Description of the datasets (p_{dim}—dimensionality of the data, n_1, n_2—number of samples in class 1 and class 2, respectively)

	$p_{\mathrm{dim}}/n_1 + n_2$
Dataset 1	1500/104 + 75
Dataset 2	1500/104 + 93
Dataset 3	1500/175 + 129
Dataset 4	300/61 + 79
Dataset 5	7129/47 + 25

4. Classifiers and evaluation

We used two representatives of regularized linear classifiers in our experiments. As a realization of the regularized MSE LDA, we used PLS. We employed the NIPALS (Wold et al., 1984) algorithm as implemented in Matlab toobox PLSTOOLS (Eigenvector) for calculating the linear discriminant solution. We chose PLS, because it has been successfully applied in undersampled biomedical problems (Bennett and Embrechts, 2003) and also because of its relationship to the SVD analysis in regression (Elden, 2004). Furthermore, PLS-related Krylov subspace methods are successfully used for solution of ill-posed inverse problems in other fields, such as numerical analysis and engineering. For examples see e.g. (Bjorck, 2004).

We also employed in our experiments linear SVM, currently considered a powerful state-of-art classifier (Vapnik, 1999). We used the SVM as implemented in Matlab toolbox PRTOOLS (Van der Heijden et al., 2004).

We split the data k times ($k = 10$) into a training set and an independent test set. We trained the classifier of choice on the training set. To avoid overoptimistic assessment of the classification error, we evaluated the classifier on the independent test set (Simon et al., 2004; Ambroise and McLachlan, 2002). As an estimate of the classification error, we used the average and standard deviation over all data splits. The data split proportions were 2/3 for the training set and 1/3 for test set. We split the data in a stratified way.

5. Results

The classification errors for the SVM and PLS classifiers are given in Table 2. The PLS and SVM classifiers show

Table 2
Classification error for the five datasets (averaged over 10 test sets) using SVM and PLS

	SVM mean (std. deviation)	PLS mean (std. deviation)
Dataset 1	0.08 (0.06)	0.08 (0.03)
Dataset 2	0.04 (0.02)	0.06 (0.02)
Dataset 3	0.21 (0.05)	0.18 (0.03)
Dataset 4	0.30 (0.05)	0.29 (0.05)
Dataset 5	0.03 (0.03)	0.04 (0.03)

comparable performance for all five datasets. The lowest classification error was achieved in Dataset 5 and the highest in Dataset 4, whereas moderate classification error was found for datasets 1, 2 and 3. Therefore, according to the classification error obtained by the two classifiers under consideration, Dataset 5 is the "easiest" and Dataset 4 is the "most difficult".

The means of discriminatory power spectra (over all data splits) from the five datasets under investigation are given in Fig. 2a–e. For all datasets, there is clearly a dominant component, which carries most of the discriminatory information. From the discriminatory power spectrum one can also see that the ranking of the discriminatory components does not necessarily coincide with the ranking based on the variance explained. In all datasets however, the discriminatory components are located at the beginning of the spectrum, corresponding to higher singular values of X (eigenvalues of the sample covariance matrix).

We were interested in how the dominant peak influences the classification error. From the mean discriminatory power spectrum we calculated the normalized discriminatory power of the dominant component according to Eq. (6). We take this number as a useful, rough characteristic of a given classification problem. The graph of the normalized discriminatory power of the first dominant discriminatory component vs. the mean classification error for the PLS classifier is displayed in Fig. 3. (For the SVM we obtained essentially the same graph.) The greater the discriminatory power of the dominant component, the lower the (mean) classification error. Thus, the "easier" classification problems (e.g. Dataset 5) and the "more difficult" classification problems (e.g. Dataset 4) reside in the lower, the

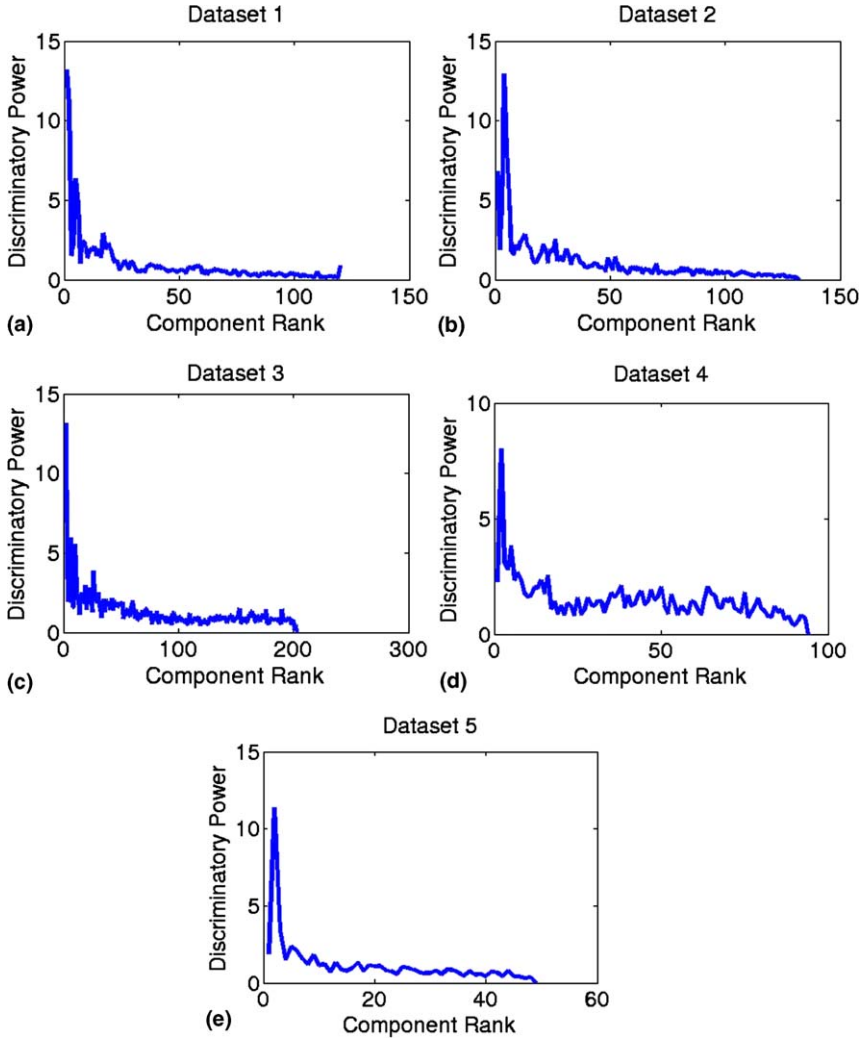

Fig. 2. Discriminatory power spectra for the five datasets.

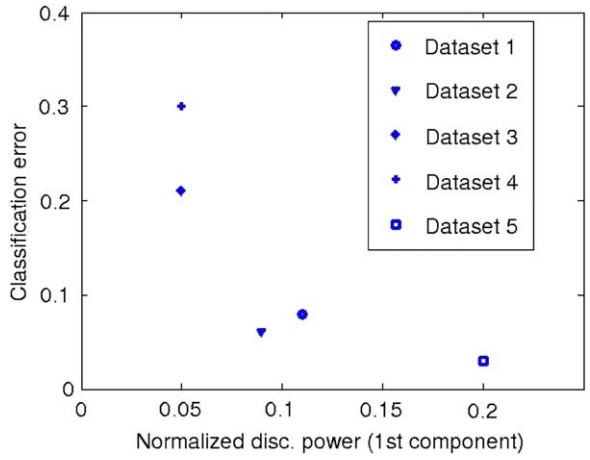

Fig. 3. Normalized discriminatory power of the dominant discriminatory component vs. (mean) classification error for the PLS classifier. Note, that the larger the normalized discriminatory power of the dominant component, the lower the classification error.

right corner and upper left corner of this display, respectively. Note, that the discriminatory power of the first discriminatory component for Datasets 3 and 4 is approximately the same. The superior classification performance obtained for Dataset 3 is due to the contribution of additional, strong discriminatory components in its discriminatory spectrum.

Using the interpretation of the linear decomposition: $X = TV^{\mathrm{T}} = \sum_{i=1}^{r} t_i v_i^{\mathrm{T}}$ in Eq. (1), as a sum of outer products (Phatak and De Jong, 1997; Jollife, 1986), one can remove (project out, deflate) any of the components and obtain a new matrix X_{proj}:

$$X_{\mathrm{proj}} = X - t_i v_i^{\mathrm{T}},$$

where t_i, v_i are the columns of the matrices T and V, respectively.

Next, we removed the dominant discriminatory component in the training set and then assessed how the classification error changed for the test set. We have performed this procedure for all datasets over all splits. The mean increases (with standard deviations) in the classification error for the five datasets and for the PLS and SVM classifiers are given in Table 3. For Dataset 5, with a clearly dominant peak, its removal was detrimental and resulted in a dramatic increase of the classification error. The increase of the classification error for Datasets 1–4 was large, but it was not detrimental. Looking at the corresponding discriminatory spectra, for all four data sets the next most discriminatory components carry considerable discriminatory power and therefore, even after removal of the most dominant component, they prevent a larger increase in the classification error.

Table 3
Classification accuracies for the five datasets (averaged over 10 test sets) using SVM and PLS after the removal of the most discriminatory feature

	SVM mean (std. deviation)	PLS mean (std. deviation)
Dataset 1	0.21 (0.10)	0.16 (0.04)
Dataset 2	0.21 (0.14)	0.18 (0.05)
Dataset 3	0.32 (0.05)	0.27 (0.05)
Dataset 4	0.37 (0.05)	0.39 (0.03)
Dataset 5	0.38 (0.08)	0.28 (0.08)

To investigate the influence of removing additional discriminatory components with high discriminatory power, we successively eliminated the first 10 discriminatory components and measured the change in classification error over the 10 data splits. The graphs for all datasets and for the PLS classifier are shown in Fig. 4. As expected, the more discriminatory components were removed, the larger was the increase of the classification error in all five datasets. Depending on the magnitude of the successive

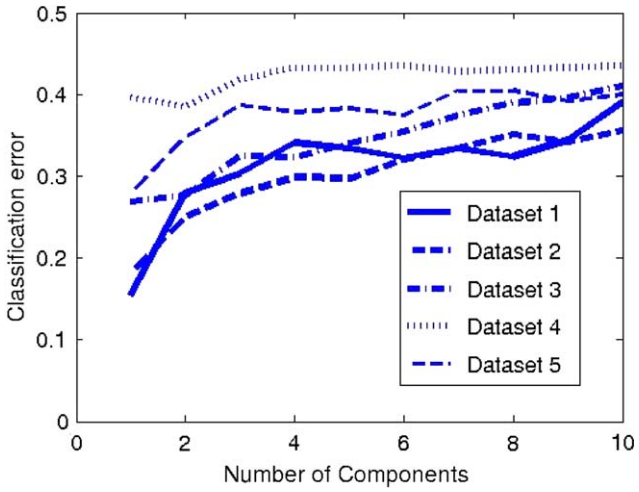

Fig. 4. The number of removed discriminatory components (with highest discriminatory power) vs. (mean) classification error for the PLS classifier. Note, that the classification error exhibits the same trend for all datasets. It generally increases with the number of removed discriminatory components.

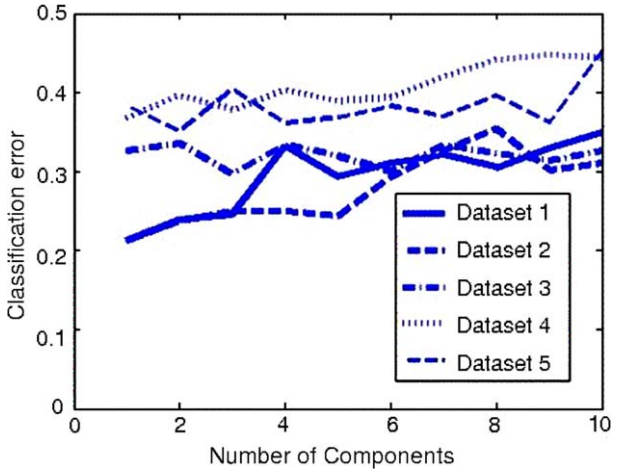

Fig. 5. The number of removed discriminatory components (with highest discriminatory power) vs. (mean) classification error for the SVM classifier. Note that again, the classification error generally increases with the number of discriminatory components removed.

discriminatory components, the increase of the classification error was more or less abrupt for different datasets. For example, for Dataset 5, as an extreme case, the increase of classification error was abrupt, as it could have already been seen from the removal of the first dominant discriminatory component. The corresponding graphs for all datasets and for the SVM classifier are shown in Fig. 5. They exhibit a trend similar to those obtained for the PLS classifier, although the SVM does not explicitly solve the least-squares problem as given by Eq. (3). However, for datasets 3 and 5, the classification error stabilizes after removal of the first discriminatory component.

6. Discussion

The PLS classifier is competitive with the SVM for all datasets. This performance of the PLS is in agreement with similar results obtained, e.g. by (Bennett and Embrechts, 2003; Elden, 2004). However, the main goal of our study was not the quest for the best classifier, but a better understanding of classifier behavior in specific real-life settings.

For all five datasets, the concentration of the discriminatory information is quite high, as manifested by dominating discriminatory peaks in the discriminatory power spectrum; they determine classifier performance. In fact, for all datasets, we created non-informative versions by randomly permuting the class labels (Ambroise and McLachlan, 2002). We obtained the discriminatory power spectra in the same way as described in the Results section. The discriminatory power spectra of these datasets were flat, with no clearly dominating peaks. This further supports the importance of the concentration of discriminatory information as expressed in the SVD representation.

Ranking the discriminatory components based on the amount of discriminatory information, one can carry out supervised principal component analysis using the spaces generated by the SVD of the data matrix (Jollife, 1986; Elden, 2004). An interesting alternative for supervised principal components was proposed in (Bair et al., 2005). There, the data are screened in the data space for "interesting" features (in this case gene expressions) and then submitted to the PCA. A criterion similar to Eq. (6) (with the ith column of matrix X instead of the left singular vector u_i) is used to identify the relevant features. Provided such features exist in the data (as it is likely in gene expression profiling) this is a reasonable approach and it can be interpreted using our notion of discriminatory spectrum. Excluding the original features, i.e. those with low covariance with the target y, one enhances the amount of discriminatory information of the features with high covariance with the target. The corresponding peak in the discriminatory spectrum will be more pronounced and the new data configuration will be more suited for the success of a linear classifier.

Interpreting the columns of the matrix T as coordinates of the data in the basis given by the right singular vectors of the matrix X, (or the eigenvectors of the sample covariance matrix), concentration of the discriminatory information in a small number of components suggests that the high-dimensional data for biomedical classification problems often lies near a low-dimensional linear manifold, well approximated by a subset of the right singular vectors of the matrix X (i.e. a subset of the eigenbasis of the sample covariance matrix).

In conclusion, we have shown that the MSE formulation of the LDA and the related notion of discriminatory power spectrum provide valuable insight into the properties of undersampled (biomedical) classification problems. In all datasets in our study, the discriminatory components identified using the MSE-LDA-derived criterion were also relevant for the state-of-art Support Vector Machine classifier. Using the discriminatory spectrum, it was possible to link classifier performance data configuration for a given classification problem. Moreover, data complexity assessment supports previous findings of the high degree of competence of regularized linear classifiers for undersampled biomedical classification problems (Simon et al., 2004). It also provides means for deciding the appropriateness of a regularized linear classifier for a particular undersampled classification problem.

We are currently investigating the applicability of our approach to the elucidation of data complexity for regularized linear classifiers in nonlinear (kernel) feature spaces.

Acknowledgements

We thank Drs. T. Sorrell and U. Himmelreich for providing us with some of the datasets used in our experiments. We also thank an anonymous reviewer for his comments, which greatly improved the presentation.

References

Ambroise, C., McLachlan, G., 2002. Selection bias in gene extraction on the basis of microarray gene expression data. PNAS 99, 6562–6566.

Bair, E. et al., 2005. Prediction by Supervised Principal Components. Technical Report. Stanford University.

Bennett, K., Embrechts, M., 2003. An optimization perspective on partial least squaresAdvances in Learning Theory: Methods, Models and Applications. NATO Science Series II: Computer and Systems Sciences. IOS Press, Amsterdam, pp. 227–250.

Bjorck, A., 2004. The calculation of linear least squares problems. Acta Numer. 13, 1–53.

Duda, R. et al., 2000. Pattern Classification. John Wiley & Sons.

Elden, L., 2004. Partial least squares vs. Lanczos bidiagonalization—I: Analysis of a projection method for multiple regression. Comput. Statist. Data Anal. 46, 11–31.

Golub, T. et al., 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286, 531–537.

Guyon, I. et al., 2002. Gene selection for cancer classification using Support Vector Machines. Mach. Learn. 46, 389–422.

Hansen, P., 1998. Rank-Deficient and Discrete Ill-posed Problems. SIAM, Philadelphia.

Himmelreich, U. et al., 2003. Rapid identification of Candida species by using nuclear magnetic resonance spectroscopy and a statistical classification strategy. Appl. Environ. Microbiol. 69, 4566–4574.

Ho, T.K., Basu, M., 2002. Complexity measures of supervised classification problems. IEEE Trans. Pattern Anal. Machine Intell. 24, 289–300.

Jollife, I., 1986. Principal Component Analysis. Springer-Verlag, New York, NY.

Lean, C. et al., 2002. Accurate diagnosis and prognosis of human cancers by proton MRS and a three stage classification strategy. Annual Rep. NMR Spectrosc. 48, 71–111.

Nguyen, D.V., Rocke, D.M., 2002. Multiclass cancer classification via partial least squares with gene expression profiles. Bioinformatics 18, 1216–1226.

Phatak, A., De Jong, S., 1997. The geometry of partial least squares. J. Chemometr., 311–338.

PLS Toolbox©, 1997–1998. Eigenvector Research, Inc., Manson, WA, USA.

Simon, R.M. et al., 2004. Design and Analysis of DNA Microarray Investigations. Springer-Verlag, New York.

Somorjai, R. et al., 2003a. Class prediction and discovery using microarray and proteomics mass spectroscopy data: Curses, caveats, cautions. Bioinformatics 19, 1484–1491.

Somorjai, R. et al., 2003b. Comparison of Two Classification Methodologies on a Real World Biomedical Problem. Lecture Notes in Computer Science, vol. 2396. Springer-Verlag, New York, pp. 433–441.

Tibshirani, R. et al., 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS 99, 6567–6572.

Van der Heijden, F. et al., 2004. Classification, Parameter Estimation and State Estimation—An Engineering Approach using Matlab. John Wiley & Sons. Available from: <http://www.prtools.org/>.

Vapnik, V., 1999. The Nature of Statistical Learning Theory. Springer-Verlag, New York.

Wold, S. et al., 1984. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comput. 5, 735–743.

Ye, J. et al., 2004. An optimization criterion for generalized discriminant analysis on undersampled problems. IEEE Trans. Pattern Anal. Machine Intell. 26, 982–994.