BIOINFORMATICS

Vol. 19 no. 12 2003, pages 1484–1491
DOI: 10.1093/bioinformatics/btg182

Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions

R. L. Somorjai, B. Dolenko and R. Baumgartner*

Institute for Biodiagnostics, National Research Council Canada, Winnipeg, MB, Canada R3B 1Y6

Received on October 18, 2002; revised on November 20, 2002; accepted on February 16, 2003

ABSTRACT

Motivation: Two practical realities constrain the analysis of microarray data, mass spectra from proteomics, and biomedical infrared or magnetic resonance spectra. One is the '*curse of dimensionality*': the number of features characterizing these data is in the thousands or tens of thousands. The other is the '*curse of dataset sparsity*': the number of samples is limited. The consequences of these two curses are far-reaching when such data are used to classify the presence or absence of disease.

Results: Using very simple classifiers, we show for several publicly available microarray and proteomics datasets how these curses influence classification outcomes. In particular, even if the *sample per feature ratio* is increased to the recommended 5–10 by feature extraction/reduction methods, dataset sparsity can render any classification result statistically suspect. In addition, several 'optimal' feature sets are typically identifiable for sparse datasets, all producing perfect classification results, both for the training and independent validation sets. This non-uniqueness leads to interpretational difficulties and casts doubt on the biological relevance of any of these 'optimal' feature sets. We suggest an approach to assess the relative quality of apparently equally good classifiers.

Contact: Ray.Somorjai@nrc-cnrc.gc.ca

INTRODUCTION

The prospects for effective and reliable disease diagnosis and management have improved significantly with the development of microarray technology in genomics (Eisen *et al.*, 1998; Alon *et al.*, 1999; Brown *et al.*, 2000) and the application of MALDI and SELDI TOF mass spectroscopy in proteomics (Wright *et al.*, 1999; Vlahou *et al.*, 2001; Adam *et al.*, 2002; Li *et al.*, 2002; Petricoin *et al.*, 2002). However, classification of diseases based on either methodology is in its infancy and effective methods for analyzing and interpreting such data have yet to be developed.

Promising results have been reported in the literature, claiming near-perfect classification accuracy for both training and independent validation datasets. However, the number of samples per class is ∼10–30, whereas the original number (N) of features (genes or mass/charge values) is in the thousands. Even after feature selection, the number M of 'optimal' features identified is ∼50–100 for microarrays, 5–10 for mass spectra. Does such combination of feature and sample numbers guarantee statistically significant classification? A seemingly plausible statistical argument often used/implied is that if perfect or near-perfect classification is achievable both for training and *independent validation sets*, then the results must be reliable and hence the classifier is *robust*. Unfortunately, for typical microarray or mass spectral datasets such conclusions are generally unwarranted, and the reliability of the results may be illusory.

Two culprits are responsible for the likely fallacy of such reliability arguments. One is Bellman's '*curse of dimensionality*' (too many *features*); the other is the '*curse of dataset sparsity*' (Somorjai and Nikulin, 1993; too few *samples*). Microarray data, mass spectra from proteomics, and biomedical magnetic resonance (MR), and infrared (IR) spectra are vulnerable to both 'curses': typically, each sample is characterized by several thousand features (e.g. genes or M/Z values or spectral intensities), yet only a few samples are available for analysis. Consequently, although good classifiers are still possible, a large number of features leads to interpretability problems.

METHODS

What do we want from a classifier?

There are two, generally interrelated goals for supervised classifiers. First, we want *robust* classifiers, i.e. classifiers that are insensitive to outliers, and possess high generalization power: *unknown* 'patterns' are classified correctly and reliably by the classifier. Second, with *medical/biological interpretability* in mind, we want to identify the smallest possible subset of maximally

*To whom correspondence should be addressed.

discriminatory features (e.g. genes, proteins, chemical compounds). Eventual disease management and treatment would benefit from having only a few, *biologically relevant* features. Unfortunately, producing robust classifiers is frequently at the expense of achieving this second goal.

How and why do we choose a particular classifier?

For cDNA microarray-based classification, the normalized intensity levels of gene expressions are the features. In proteomics, mass spectroscopy-based classification uses the mass/charge ratios as features. Most studies aim to achieve both classification goals; in particular, special emphasis is placed on the identification of 'marker' genes that are responsible for, or indicative of, the occurrence of different cancers (DeRisi *et al.*, 1996; Golub *et al.*, 1999; Bittner *et al.*, 2000; Furey *et al.*, 2000; Hedenfalk *et al.*, 2001; Khan *et al.*, 2001; Ramaswamy *et al.*, 2001; Su *et al.*, 2001; Yeang *et al.*, 2001; Pomeroy *et al.*, 2002) or other diseases. The same applies for the identification of proteins or protein fragments as 'biomarkers'. The technical complexity of the classifiers used range from the simplest (correlation techniques) to the most sophisticated (non-linear support vector machines, SVMs). (For good descriptions of general classification methodology, see e.g. Duda *et al.*, 2000 and Raudys, 2001). However, the choice of classifier seems not to be dictated by the data at hand, but by 'expert' recommendation (usually based on other types of data), personal experience or preference, or simply software availability. The maxim 'simpler is better' has mostly been ignored. In general, no specific effort has been expended on choosing the classifier most appropriate for a given dataset. This, despite (Raudys, 1998) convincing demonstration that a common classification framework can be created to emulate a wide range of classifiers [from the simplest Euclidean distance classifier through Fisher's LDA, regularized DA (Friedman, 1989), to maximum margin classifiers and SVMs], and that classifier complexity and sample size have to be matched. Instead, the 'best', i.e. most sophisticated classifier is used, whether appropriate or not! [A very simple classifier has recently been proposed for microarrays, (Tibshirani *et al.*, 2002).]

We faced similar challenges, and developed a successful classification strategy for disease profiling, based on MR or IR spectra of tissues or biofluids (Somorjai *et al.*, 2002, overview by Lean *et al.*, 2002). For classification purposes, the similarities (i.e. high dimensionality and small sample size) between microarray data, mass, and other biomedical spectra are striking; hence, our experience will prove useful for analyzing microarray data and proteomics-generated mass spectra. We focus on *supervised* classification problems, i.e. samples with known class labels.

Can we assess classifier reliability?

Classifier design starts by partitioning the dataset into *training* and *test* sets. The chosen classifier is optimized on the training set, and *validated* on the test (validation) set. (To guard against overfitting, one frequently accepts sub-optimal classifiers, if they strike a good balance between training and test set accuracy. Unfortunately, dataset sparsity often prevents the recommended approach of splitting the data into disjoint *training*, *monitoring* and independent *test* sets.) When the dataset is small, (typical for both microarray data and biomedical spectra), *crossvalidation* (CV), frequently the leave-one-out (LOO) method, is applied to the entire dataset. (Other versions, such as 10-fold crossvalidation, are not feasible, because of dataset sparsity.) However, conventional use of LOO (or any other) crossvalidation does not produce a robust classifier, only a reasonably unbiased, if optimistic, estimate of the generalization error. The actual classifier to be used for new samples is not specified. With eventual practical, clinical use in mind, we have developed a bootstrap-inspired method (Somorjai *et al.*, 2002) that produces a single classifier as the weighted average of a large number of individual classifiers, each obtained by a different random partitioning of the dataset into training and test sets. (The weights use the *test* set accuracies.)

Microarray data, mass spectra from proteomics, biomedical MR and IR spectra are all vulnerable to both 'curses'. In the clinical context, this means that even if disease diagnosis and prognosis is possible (goal 1), eventual disease treatment and management (goal 2) is less clear-cut or even problematic.

Caveat 1. The curse of dimensionality: too many features
It has been generally accepted by the machine learning community (Foley, 1972; Jain, 1982) that robust classification requires a *sample per feature ratio* (SFR) of at least 5–10 [although this depends on the data and on the complexity of the classifier (Raudys, 2001)]. However, for microarray data, and mass, IR or MR spectra in a biomedical context, SFR is typically 1/20–1/500. The SFR improves for correlated features, the case for MR and IR spectra, and possibly for genes and M/Z values. However, in general the SFR is still too small to expect robust classifiers.

The conventional solution to this SFR dilemma is to reduce feature space dimensionality, preferably by eliminating redundant and/or irrelevant (noisy) features. For biomedical spectra, we have developed a genetic-algorithm-driven optimal region selector method (GA_ORS; Nikulin *et al.*, 1998). It creates new features from the averages of adjacent original features (spectral data points). For microarray expression profiles, where correlation between the features (genes) is not evident, other feature selection methods are required. The optimal, exhaustive best subset search (ES), is not feasible unless

N is relatively small. (Selecting the best pair from a 15 154-feature proteomics dataset requires evaluating 114 814 281 classifiers. Assuming 0.1 s for developing a single classifier with LOO-CV, this needs 132.9 days of continuous computation!). When ES is not viable, some suboptimal approach, e.g. our dynamic programming-based method (Nikulin *et al.*, 1998), or even a simple random subset sampling, often produces near-optimal discriminatory feature subsets in acceptable execution times.

SVM classifiers (Pontil and Verri, 2001) do not require feature space reduction. Nevertheless, even their proponents now concede that feature selection is beneficial (Yeo and Poggio, 2001; Guyon *et al.*, 2002), not only for helping with the important interpretation problem, but also for providing better classification.

Feature selection approaches fall into two categories. *Filtering* methods rank features by using some *class-independent* index. In contrast, *wrapper* methods use class discrimination requirements and capabilities for optimal feature selection (Kohavi and John, 1997). Univariate versions, which rate features independently, are most common. They ignore interdependence among the features. The 'best' $M << N$ features identified are used to develop the classifiers. Of course, there is no guarantee that a list of the best M features obtained via univariate feature selection methods will contain the best pairs, triplets, etc. There is an extensive literature on feature selection methods, most of them sequential and univariate, with their attendant disadvantages (Siedlecki and Sklansky, 1988; Somol *et al.*, 1999).

To improve single-gene-based ratings, simultaneous analysis of *pairs* of genes ('gene pair ranking') has been suggested (Bø and Jonassen, 2002). Although there is improvement, when there are many features, feature selection methods based on only one or two features may give poor results (Cover, 1974; Kittler, 1978). Hence, even the best pairs found may not provide accurate classification.

Given the too few samples–too many features conundrum, a reasonable approach is to find many random subsets of $K > 2$ (but still small) features of comparable classification power, count the frequency of occurrence of all features, and assume that the most frequently selected features are the most relevant. This is the method implemented recently for microarrays (Li *et al.*, 2001), and the approach we have been applying to biomedical spectra using our GA_ORS method. Both approaches are explicitly and intentionally multivariate, essential if interdependence among features is to be included *ab initio*. However, because of the large number of random trials, many of the best features are discriminatory only by chance, and statistical significance estimates must be corrected (done rarely if at all) for *repeated measurements*. Such correction reduces the optimistic bias in apparent error estimates, and should be carried out, at least by the classical (conservative) Bonferroni correction that assumes independence of the several simultaneous statistical tests (Westfall and Wolfinger, 1997), but preferably by methods based on false discovery rates (Benjamini and Hochberg, 2000) that take into account dependence for the repeated measurements (Benjamini and Yekutieli, 2001).

It appears that no matter what feature selection approach has been used to classify microarray data, generally at least 50 (and frequently more) features have been chosen and used for classification. This suggests a lack of appreciation of the SFR's critical role in robust classifier development. The situation is somewhat better for classifying the mass spectra derived from proteomics, with typically ~5–20 features identified for 50–100 samples per class.

Caveat 2. The curse of dataset sparsity: too few samples This is an insidious curse, because when only a few samples per class are available, it is quite easy to produce seemingly robust classifiers that give excellent results for both training and validation sets. Unfortunately, the reliability of these results is illusory, and conclusions based on them are suspect or perhaps even wrong. One manifestation of this curse is that the good *results are essentially independent* of the type of classifier used: everything seems to work! Furthermore, the results seem immune to the curse of dimensionality: no feature reduction appears necessary. Even if the recommended SFR is achieved (for most currently available microarray datasets, this would limit the maximum allowable features to 2–3, for mass spectra from proteomic studies, to 5–10), we must question the statistical relevance of the high classification accuracy, and the generalizability of the results. In particular, the successful classification of an independent validation set is only meaningful statistically if both training and validation sets are large enough to be representative of their original population distributions. Establishing that training and validation set members were drawn from the same distribution *is essential for sparse datasets*, because they are unlikely to be representative of their classes.

While the importance of achieving the appropriate SFR is increasingly acknowledged in the microarray and proteomics literature, the consequences of dataset sparsity are still not appreciated sufficiently.

Of the two constraints produced by dataset sparsity, one limits the choices for crossvalidation; the second prevents a meaningful splitting of the data into training, monitoring and test sets, necessary for robust classifier development. An important consequence of dataset sparsity is that several sets of 'optimal' features will give identical misclassification errors. This non-uniqueness casts doubt on

the biological relevance of any specific 'optimal' feature (i.e. gene or protein) set. (When using wrapper methods for feature selection, the eventual 'optimal' features are also classifier-dependent, another manifestation of the inevitability of non-uniqueness.) We need methods to discriminate among equally 'good' results.

RESULTS AND DISCUSSION

How do the caveats manifest in practice?

Microarray data We present results obtained for two publicly available microarray datasets: the two-class leukemia set (Golub *et al.*, 1999), and the four-class small, round blue-cell tumor (SRBCT) set (Khan *et al.*, 2001; Yeo and Poggio, 2001). Both datasets were previously analyzed by other investigators, using SVM- or k-NN-based classifiers. We selected a least trimmed squares (LTS) regression method for classification (10% trimming); for two-class problems, LTS is a robust version of the simple linear Fisher discriminant. We used exhaustive search as the feature selection method (with LTS, or LDA and leave-one-out crossvalidation), requesting only two features (genes). This satisfies the SFR requirement for most of the pair classifiers. The datasets were analyzed without additional preprocessing. When the number of classes $K > 2$, both theoretical arguments and our extensive practical experience indicate that developing all $K(K-1)/2$ pair classifiers, and combining their outcomes gives better, more reliable results than an explicit K-class classifier. We use a Bayes-based pair classifier combination (Schürmann, 1996).

Dataset 1: Acute Myeloid Leukemia, AML (47 samples) versus Acute Lymphoblastic Leukemia, ALL (25 samples).

The cDNA microarrays contain 7129 genes, partitioned into a training set (TS, 38 samples, 27 AML + 11 ALL) and a validation set (VS, 34 samples, 20 AML + 14 ALL) (Golub *et al.*, 1999).

The two best feature (gene) pairs, (2300, 4847) and (4211, 4847), give no errors on the TS and one error on the VS. (The single misclassified sample in the VS was misclassified by all other investigators; it appears that this sample was originally assigned to the wrong class.) There is one pair with one TS and one VS error, and nine pairs with no TS and two VS errors. These 10 pairs all share gene 4847, zyxin.

Dataset 2: Small, round blue-cell tumors of childhood (SRBCT) (Khan *et al.*, 2001).

The cDNA microarrays comprise 2308 genes. The TS contains 23 Ewing family tumor (EWS), eight Burkitt lymphoma (BL), 12 neuroblastoma (NB) and 20 rhabdomysarcoma (RMS) samples. The 25 samples in the VS comprise six EWSs, three BLs, six NBs, five RMSs and five non-SRBCTs. For this four-class problem, we developed all six class A versus class B classifiers. Table 1 below summarizes the number of gene pairs that gave no TS and no VS errors for the various pair classifiers.

Table 1. Number of gene pairs that classify both training and test sets without error

Pair classifiers	EWS versus BL	EWS versus NB	EWS versus RMS	BL versus NB	BL versus RMS	NB versus RMS
Number of gene pairs	102	36	14	163	199*	23

*Two of these are single genes: 509 and 1932.

The above analysis strikingly demonstrates how dataset sparsity leads to non-uniqueness of feature sets, and to subsequent interpretational ambiguities (which genes are **really** relevant biologically?).

With only two attributes and very simple linear classifiers, we could surpass or match the best results published for both datasets. However, achieving zero or very few misclassification errors is incidental to this study. Our ultimate message is that sparse datasets induce multiple, non-unique sets of 'best' discriminatory genes, leading to VS classification that is essentially perfect, yet statistically suspect. Hence, attributing definitive statistical or biological significance to any of these 'best' sets of discriminatory genes is problematic.

Mass spectra from proteomics studies We downloaded the ovarian and prostate cancer mass spectroscopy datasets from the NIH/FDA Clinical Proteomics Program Databank (http://clinicalproteomics.steem.com.) to demonstrate some consequences of non-uniqueness. We used simple random sampling (1 million random samples with replacement) to identify the 'best' 2–5 'features' (M/Z values) out of the original 15 154, both for the prostate ('JNCI 7-3-02') and the ovarian ('6-19-02') datasets. For classification, we selected either the robust, least-trimmed-squares (LTS) linear regression method (10% trimming), or the simple LDA with LOO-CV.

Ovarian cancer dataset ('6-19-02') We partitioned randomly the dataset into five different training (TS) and validation (VS) sets, D1–D5, always maintaining the splits at 61 + 61 (TS) and 30 + 101 (VS).

Two to four features classified both TSs and VSs perfectly. In particular, we found several feature sets for many of the dataset partitions that gave 100% accuracy for both TS and VS (e.g. for D3, in the range 567–884 Da, one two-feature-set, sixteen three-sets, and seventy-nine four-sets were identified). Besides interpretational ambiguities, such non-uniqueness implies that many more proteins would require identification than the low (two–four) pattern size suggests. In Table 2 we display the number

Table 2. Ovarian cancer versus healthy: feature sets with perfect classification

Number of features	Dataset partition D_1	D_2	D_3	D_4	D_5	All
2	3	—	1	—	—	1
3	27	1	16	1	—	18
4	142	13	79	3	1	108

Table 3. Prostate cancer versus healthy: average classification accuracy

Number of features	Training sets	Validation sets	All
LTS classifier			
3	98.1%	92.9%	—
5	99.1%	94.6%	—
6	99.8%	95.0%	—
LDA classifier; LOO-CV			
3	97.4%	93.7%	97.7%
4	97.8%	94.9%	97.7%
5	98.6%	96.6%	98.5%
6	98.7%	96.9%	98.5%
7	100.0%	97.1%	99.2%

of perfect solutions for both TSs and VSs, when using two–four features. We also developed classifiers using all samples (All).

Prostate cancer dataset ('JNCI 7-3-02') Each TS had 42 (class 1) + 42 (class 2) samples, each VS, 21 + 27. We used LTS with 10% trimming. Three to six features classified both training and validation sets near-perfectly. In Table 3, first three rows, we display the LTS classifier-based classification results, averaged over the five randomly partitioned datasets, for both the TSs and VSs. In the last five rows, averaged results are shown for three–seven features, when we used LDA with leave-one-out crossvalidation. Under 'All', we also include the results when the dataset wasn't split into training/validation sets. Two five-feature sets classified both TSs and VSs without error.

Comparison of the results in Table 3 shows that for good classification of this dataset, the choice of the particular classifier is not critical. However, the 'optimal' feature sets, obtained by the different classifiers, don't coincide, again emphasizing the non-uniqueness conundrum.

It is unlikely that an 'optimal' feature set, selected using a particular classifier, would retain optimality when it is used with a different classifier. However, if the feature set gives comparable classification results with a variety of conceptually and/or technically different classifiers, then it is reasonable to assume that this set of features is, if not the best possible, at least reasonably robust.

Table 4. Fraction (F) of all reference pairs (S) with perfect classification (N_P) in the RDP

Classifier pairs: A(n1) vs. B(n2)	F (N_P/S)
MD(10) versus MG(10)	0.35 (42/121)
MD(10) versus RH(10)	0.21 (25/121)
MD(10) versus NC(4)	1.00 (55/55)
MD(10) versus PN(8)	0.06 (6/99)
MG(10) versus RH(10)	0.71 (86/121)
MG(10) versus NC(4)	0.66 (36/55)
MG(10) versus PN(8)	0.00 (0/99)
RH(10) versus NC(4)	1.00 (55/55)
RH(10) versus PN(8)	0.45 (45/99)
NC(4) versus PN(8)	0.80 (36/45)

The concept that a set of features (a 'feature profile'), acting in concert produces better classifiers than could be obtained by some 'unique' biomarker is relatively new in the microarray/proteomics fields. However, we have been successfully exploiting this idea for the last decade in analyzing biomedical (IR and MR) spectra (consult the references in Lean *et al.* (2002).

Dataset sparsity and its consequences—a visual assessment

Given the combination of high dimensionality and sparsity, typical of microarray, proteomic or biomedical datasets in general, their visual assessment would be helpful. The difficulty lies in preserving the relative distance relations between high-dimensional samples when projected to two or three dimensions. The conventional approaches, e.g. multidimensional scaling (Borg and Groenen, 1997) or Kohonen's mapping (Kohonen, 2001), preserve the distance relationships only approximately, and require time-consuming optimization of some non-linear target function, with no guarantee that its global minimum is achievable. The most common, principal component analysis (PCA)-based method has its own, inherent difficulties. Most relevant for classification is that the PCs explaining most of the data variance are rarely maximally discriminatory. A good review of earlier attempts is in Siedlecki *et al.* (1988).

Inspired by the philosophy of projection pursuit (Friedman and Tukey, 1974), which advocates searching for 'interesting' directions in the high-dimensional space, we have developed a distance-based mapping method for visualizing high-dimensional patterns and their relative relationships (Somorjai *et al.*, in preparation). It only requires a single computation of a distance matrix. The mapping's most important characteristic is that certain distances in the original space are preserved *exactly*

in a special *relative distance plane* (RDP). All points of the dataset can thus be displayed and visualized, without any distortion of their original distances *to two reference patterns*, say, $\mathbf{R}_1$ and $\mathbf{R}_2$. These reference patterns can be any pair in the dataset, and a line through them defines a possibly 'interesting' direction. For classification, the most useful reference pair should belong to different classes, but need not be the class centroids.

Mapping to a lower dimensional space introduces constraints, i.e. the extent of class separability displayed in the RDP (or a following one-dimensional projection onto the reference axis) is the least favorable of what can be expected of a classifier. Hence, a revealing manifestation of data sparsity is that patterns from two classes separate perfectly, even when the mapping to the RDP is directly from the original high-dimensional space, i.e. without prior feature reduction. If there are n_1 patterns in class 1, n_2 in class 2, with the two centroids included, there are $S = (n_1 + 1)(n_2 + 1)$ possible reference pairs, when the two patterns of a pair belong to different classes. The fraction of the S directions that separate the classes perfectly is a quantitative measure of how 'easy' it is to classify the dataset (due to sparsity or any other reason). A telling demonstration of this is provided by the five-class CNS tumor microarray expression dataset (Pomeroy *et al.*, 2002). The five classes are medulloblastoma (MD, 10 samples), malignant glioma (MG, 10), AT/RT (RH, 10), normal cerebellum (NC, 4), and PNET (PN, 8). This is clearly a very sparse dataset, as shown by the mapping results; mappings were carried out from the original 7129-dimensional feature space. The results for the 10 pair classifiers are collected in Table 4.

Inspection of the table shows that even using the Euclidean distance matrix (corresponding to the Nearest Mean classifier, the simplest linear classifier possible), nine out of the ten pair classifiers had perfect (multiple) solutions in the RDP. (The single exception, MG versus PN, had one pattern misclassified.)

Assume that after feature selection, several classifiers exist with identical TS and VS errors. How do we decide which of these is potentially better, especially for sparse datasets? A possible qualitative assessment is provided by the mapping to the RDP: The classifier (with its feature set) that produces the smallest error for both validation and training sets *after the mapping* will likely be the best. For example, for the prostate cancer dataset 'JNCI 7-3-02', two classifiers with different five-feature sets classified both TSs and VSs without error. When the datasets were mapped from five dimensions to the RDP, one of them produced zero TS and VS errors, the other seven TS and two VS errors. This suggests that the first classifier might also do better with additional unknown data.

CAUTIONS AND CONCLUSIONS

In view of the double goal of classifier development with medical/clinical relevance, classifiers or classifier combinations lacking feature selection/extraction are of more limited value when applied to microarray expression profiles, proteomics mass spectra or biomedical MR and IR spectra. This is because diagnosis and prognosis, although of considerable importance, form only part of the clinical story. The second goal, the identification of a few, biologically relevant discriminatory features (i.e. genes or proteins or chemical species) that could lead to effective disease management and treatment, is probably of even greater significance. Having to satisfy the second goal excludes from consideration classifiers that only use the original, high-dimensional feature space. These include not only the SVM-type methods that map the data into an even higher-dimensional space (Müller *et al.*, 2001), but also approaches that combine several classifiers using different feature sets, such as boosting (Freund, 1995), random subspace methods (Ho, 1998), and stochastic discrimination (Kleinberg, 1990). These latter all have problems with eventual interpretation, i.e. it is impossible to decide conclusively which of the many feature sets are relevant.

The classifiers that endorse and rely on prior feature selection still have to satisfy the appropriate SFR to be robust and reliable. However, although this is generally necessary, it is by no means sufficient: the curse of dataset sparsity also has to be lifted. Unfortunately, the obvious and ideal remedy of increasing the sample size is often not feasible.

We have used our new method of exactly mapping data from the originally high-dimensional feature space to a 2D plane, the RDP. This enables us to visualize directly how 'easy' it is to classify the dataset. Additional useful features of the mapping include the detection of possible outliers, assessing whether the training and validation sets derive from the same distribution, and the evaluation of the efficacy of the particular feature reduction employed. We shall discuss and expand on these aspects in a separate publication.

ACKNOWLEDGEMENTS

We thank those authors who generously made their datasets public. We are grateful to Professor S. Raudys and the reviewers for useful comments.

REFERENCES

Adam,B.-L., Qu,Y., Davis,J.W., Ward,M.D., Clements,M.A., Cazares,L.H., Semmes,O.J., Schellhammer,P.F., Yasui,Y., Feng,Z. *et al.* (2002) Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.*, **62**, 3609–3614.

Alon,U., Barkai,N., Notterman,D.A., Gish,K., Ybarra,S., Mack,D. and Levine,A.J. (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA*, **96**, 6745–6750.

Benjamini,Y. and Hochberg,Y. (2000) On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educational Behavioral Statistics*, **25**, 60–83.

Benjamini,Y. and Yekutieli,D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, **29**, 1165–1188.

Bittner,M., Meltzer,P., Chen,Y., Jiang,Y., Seftor,E., Hendrix,M., Radmacher,M., Simon,R., Yakhini,Z., Ben-Dor,A. *et al.* (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. *Nature*, **406**, 536–540.

Bø,T.H. and Jonassen,I. (2002) New feature subset selection procedure for classification of expression profiles. *Genome Biol.*, **3**, 1–11.

Borg,I. and Groenen,P. (1997) *Modern Multi-Dimensional Scaling*. Springer, Berlin.

Brown,M.P.S., Grundy,W.N., Lin,D., Cristianini,N., Sugnet,C., Furey,T.S., Ares,Jr,M. and Haussler,D. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl Acad. Sci. USA*, **97**, 262–267.

Cover,T.M. (1974) The best two independent measurements are not the two best. *IEEE Trans. Syst. Man. Cybern.*, **4**, 116–117.

DeRisi,J., Penland,L., Brown,P.O., Bittner,M.L., Meltzer,P.S., Ray,M., Chen,Y., Su,Y.A. and Trent,J.M. (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet.*, **14**, 457–460.

Duda,R.O., Hart,P.E. and Stork,D.G. (2000) *Pattern Classification*, 2nd edn, Wiley, New York.

Eisen,M., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14 863–14 868.

Foley,D.H. (1972) Considerations of sample and feature size. *IEEE Trans. Inform. Theor.*, **IT-18**, 618–626.

Freund,Y. (1995) Boosting a weak learning algorithm by majority. *Information and Computation*, **121**, 256–285.

Friedman,J. (1989) Regularized discriminant analysis. *J. Am. Stat. Assoc.*, **84**, 165–175.

Friedman,J. and Tukey,J.W. (1974) A projection pursuit algorithm for exploratory data analysis. *IEEE Trans. Comput.*, **C-23**, 881–889.

Furey,T.S., Cristianini,N., Duffy,N., Bednarski,D.W., Schummer,M. and Haussler,D. (2000) Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*, **16**, 906–914.

Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A. *et al.* (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.

Guyon,I., Weston,J., Barnhill,S. and Vapnik,V. (2002) Gene selection for cancer classification using support vector machines. *Machine Learning*, **46**, 389–422.

Hedenfalk,I., Duggan,D., Chen,Y., Radmacher,M., Bittner,M., Simon,R., Meltzer,P., Gusterson,B., Esteller,M., Kallioniemi,O.-P. *et al.* (2001) Gene-expression profiles in hereditary breast cancer. *New Engl. J. Med.*, **344**, 539–548.

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

Jain,A.K. (1982) *Dimensionality and Sample Size Considerations in Pattern Recognition Practice*. North-Holland, Amsterdam.

Khan,J., Wei,J.S., Ringnér,M., Saal,H., Ladanyi,M., Westermann,F., Berthold,F., Schwab,M., Antonescu,C.R., Peterson,C. *et al.* (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat. Med.*, **7**, 1–10.

Kittler,J. (1978) *Pattern Recognition and Signal Processing*. Chen,C.H. (ed.), Sijthoff Noordhoff, Netherlands, pp. 41–60.

Kleinberg,E.M. (1990) Stochastic discrimination. *Annals of Mathematics and Artificial Intelligence*, **1**, 207–239.

Kohavi,R. and John,G.H. (1997) Wrappers for feature subset selection. *Artificial Intelligence*, **97**, 273–324.

Kohonen,T. (2001) *Self-Organizing Maps*, 3rd edn., Springer, Berlin.

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

Li,L., Weinberg,C.R., Darden,Th.A. and Pedersen,L.G. (2001) Gene selection for sample classification based on gene expression data: study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, **17**, 1131–1142.

Li,J., Zhang,Zh., Rosenzweig,J., Wang,Y.Y. and Chan,D.W. (2002) Proteomics and bioinformatics approaches of serum biomarkers to detect breast cancer. *Clinical Chemistry*, **48**, 1296–1304.

Müller,K.-R., Mika,S., Rätsch,G., Tsuda,K. and Schölkopf,B. (2001) An introduction to kernel-based learning algorithms. *IEEE Transaction on Neural Networks*, **12**, 181–201.

Nikulin,A.E., Dolenko,B., Bezabeh,T. and Somorjai,R.L. (1998) Near-optimal region selection for feature space reduction: novel preprocessing methods for classifying MR spectra. *NMR Biomed.*, **11**, 209–216.

Petricoin,III,E.F., Ardekani,A.M., Hitt,B.A., Levine,P.J., Fusaro,V.A., Steinberg,S.M., Mills,G.B., Simone,Ch., Fishman,D.A., Kohn,E.C. *et al.* (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*, **359**, 572–577.

Pomeroy,S.L., Tamayo,P., Gaasenbeek,M., Sturla,L.M., Angelo,M., McLaughlin,M.E., Kim,J.Y.H., Goumnerova,L.C., Black,P.M., Lau,Ch. *et al.* (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, **415**, 436–442.

Pontil,M. and Verri,A. (2001) Properties of support vector machines. *Neural Computation*, **10**, 955–974.

Ramaswamy,S., Tamayo,P., Rifkin,R., Mukherjee,S., Yeang,Ch.-H., Angelo,M., Ladd,Ch., Reich,M., Latulippe,E., Mesirov,J.P. *et al.* (2001) Multiclass cancer diagnosis using tumor gene expression signatures. *Proc. Natl Acad. Sci. USA*, **98**, 15 149–15 154.

Raudys,J. (1998) Evolution and generalization of a single neurone. I. single-layer perceptron as seven statistical classifiers. *Neural Networks*, **11**, 283–296.

Raudys,S. (2001) *Statistical and Neural Classifiers—An Integrated Approach to Design*. Springer, London.

Schürmann,J. (1996) *Pattern Classification: A Unified View of Statistical and Neural Approaches*. John Wiley & Sons, Inc., New York.

Siedlecki,W. and Sklansky,J. (1988) On automatic feature selection. *Int. J. Pattern Recogn. Artif. Intell.*, **2**, 197–200.

Siedlecki,W., Siedlecka,K. and Sklansky,J. (1988) An overview of mapping techniques for exploratory pattern analysis. *Pattern Recognition*, **21**, 411–429.

Somol,P., Pudil,P., Novovicová,J. and Paclík,P. (1999) Adaptive floating search methods in feature selection. *Pattern Recognition Letters*, **20**, 1157–1163.

Somorjai,R.L. and Nikulin,A. (1993) The curse of small sample sizes in medical diagnosis via MR spectroscopy. *Proc. Soc. Magn. Reson. Med. Twelfth Annual Scientific Meeting*. New York, pp. 685.

Somorjai,R.L., Dolenko,B., Nikulin,A., Nickerson,P., Rush,D., Shaw,A., de Glogowski,M., Rendell,J. and Deslauriers,R. (2002) A 3-stage robust classification strategy for biomedical spectra: application to urine MR and IR spectra to distinguish normal allografts from biopsy proven rejections. *Vibrational Spectroscopy*, **28**, 97–102.

Su,A.I., Welsh,J.B., Sapinoso,L.M., Kern,S.G., Dimitrov,P., Lapp,H., Schultz,P.G., Powell,S.M., Moskaluk,C.A., Frierson,Jr,H.F. *et al.* (2001) Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res.*, **61**, 7388–7393.

Tibshirani,R., Hastie,T., Narasimhan,B. and Chu,G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl Acad. Sci. USA*, **99**, 6567–6572.

Vlahou,A., Schellhammer,P.F., Mendrinos,S., Patel,K., Kondylis,F.I., Gong,L., Nasim,S. and Wright,Jr,G.L. (2001) Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine. *Am. J. Pathol.*, **158**, 1491–1502.

Westfall,P.H. and Wolfinger,R.D. (1997) Multiple tests with discrete distributions. *Am. Statist.*, **51**, 3–8.

Wright,Jr,G.L., Cazares,L.H., Leung,S.-M., Nasim,S., Adam,B.-L., Yip,T.-T., Schellhammer,P.F., Gong,L. and Vlahou,A. (1999) Proteinchip surface enhanced laser desorption/ionization (SELDI) mass spectrometry: a novel protein biochip technology for detection of prostate cancer biomarkers in complex protein mixture. *Prostate Cancer and Prostatic Diseases*, **2**, 264–276.

Yeang,Ch.-H., Ramaswamy,S., Tamayo,P., Mukherjee,S., Rifkin,R.M., Angelo,M., Reich,M., Lander,E., Mesirov,J. and Golub,T. (2001) Molecular classification of multiple tumor types. *Bioinformatics*, **17**, S316–S322.

Yeo,G. and Poggio,T. (2001) *Multiclass classification of SRBCTs*, AI Memo 2001-018; CBL Memo 206, MIT—artificial intelligence laboratory, Cambridge, MA, pp. 1–17.