BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 8 2005, pages 1509–1515
doi:10.1093/bioinformatics/bti171

Gene expression

Optimal number of features as a function of sample size for various classification rules

Jianping Hua[1], Zixiang Xiong[1], James Lowey[2], Edward Suh[2] and Edward R. Dougherty[1,3,*]

[1]Department of Electrical Engineering, Texas A&M University, College Station, TX 77843, USA, [2]Translational Genomics Research Institute, Phoenix, AZ 85004, USA and [3]Department of Pathology, University of Texas M.D. Anderson Cancer Center, Houston, TX 77030, USA

Received on September 16, 2004; revised on November 17, 2004; accepted on November 19, 2004
Advance Access publication November 30, 2004

ABSTRACT

Motivation: Given the joint feature-label distribution, increasing the number of features always results in decreased classification error; however, this is not the case when a classifier is designed via a classification rule from sample data. Typically (but not always), for fixed sample size, the error of a designed classifier decreases and then increases as the number of features grows. The potential downside of using too many features is most critical for small samples, which are commonplace for gene-expression-based classifiers for phenotype discrimination. For fixed sample size and feature-label distribution, the issue is to find an optimal number of features.

Results: Since only in rare cases is there a known distribution of the error as a function of the number of features and sample size, this study employs simulation for various feature-label distributions and classification rules, and across a wide range of sample and feature-set sizes. To achieve the desired end, finding the optimal number of features as a function of sample size, it employs massively parallel computation. Seven classifiers are treated: 3-nearest-neighbor, Gaussian kernel, linear support vector machine, polynomial support vector machine, perceptron, regular histogram and linear discriminant analysis. Three Gaussian-based models are considered: linear, nonlinear and bimodal. In addition, real patient data from a large breast-cancer study is considered. To mitigate the combinatorial search for finding optimal feature sets, and to model the situation in which subsets of genes are co-regulated and correlation is internal to these subsets, we assume that the covariance matrix of the features is blocked, with each block corresponding to a group of correlated features. Altogether there are a large number of error surfaces for the many cases. These are provided in full on a companion website, which is meant to serve as resource for those working with small-sample classification.

Availability: For the companion website, please visit http://public.tgen.org/tamu/ofs/

Contact: e-dougherty@ee.tamu.edu

1 INTRODUCTION

Given the joint feature-label distribution, increasing the number of features always results in decreased classification error; however, this is not the case when a classifier is designed via a classification rule from sample data. Typically (but not always), for fixed sample size, the error of a designed classifier decreases and then increases as the number of features grows. This peaking phenomenon was first rigorously demonstrated for discrete classification (Hughes, 1968), but it can affect all classifiers, the manner depending on the feature-label distribution. The potential downside of using too many features is most critical for small samples, which are commonplace for gene-expression-based classifiers for phenotype discrimination. For fixed sample size and feature-label distribution, the issue is to find an optimal number of features. A seemingly straightforward approach would be to find the distribution of the error as a function of the feature-label distribution, number of features and sample size. In fact, this has rarely been achieved. This leaves open the simulation route, and this approach has been taken in the past for quadratic and linear discriminant analysis (Van Ness and Simpson, 1976; El-Sheikh and Wacker, 1980; Jain and Chandrasekaran, 1982). To apply simulation for various feature-label distributions and classification rules, and across a wide range of sample and feature-set sizes requires enormous computation. This study employs contemporary massively parallel computation. There are a large number of resulting three-dimensional graphs. A website is provided for comparison of the many results.

Determining the optimal number of features is complicated by the fact that if we have D potential features, then there are $C(D,d)$ feature sets of size d, and all of these must be considered to assure the optimal feature set among them (Cover and Van Campenhout, 1977). Owing to the combinatorial intractability of checking all feature sets, many algorithms have been proposed to find good suboptimal feature sets; nevertheless, feature selection remains problematic. Evaluation of methods is generally comparative and based on simulations (Jain and Zongker, 1997; Kudo and Sklansky, 2000). One method used to avoid the confounding effect of feature selection is to make a distributional assumption so that any d features possess the same marginal distribution. To model the situation in which subsets of genes are co-regulated and correlation is internal to these subsets, we assume that the covariance matrix of the features is blocked, with each block corresponding to a group of correlated features, and that feature selection follows the order of the features in the covariance matrix (details below). While this does not necessarily produce the optimal feature set for each size, it does provide comparison of the classification rules relative to a global selection procedure that takes

*To whom correspondence should be addressed.

into account correlation—as opposed to the less realistic assumption of equal marginal distributions.

Feature-set size was first addressed for multinomial discrimination via the histogram rule by finding an expression for the expected error $E[\varepsilon_d(S_n)]$ over all probability models, assuming equally likely probability models (Hughes, 1968). The drawback of this approach is twofold. First, in practice there is only one probability model and the behavior of the error can vary substantially for different models. Second, all models are not equally likely in realistic settings. Recently, an analytic representation of the distribution of the error $\varepsilon_d(S_n)$ for multinomial discrimination has been discovered that is distribution-specific (as is our simulation procedure here), and from which a distribution-specific expected error $E[\varepsilon_d(S_n)]$ is derived (Braga-Neto and Dougherty, 2004a). The optimal feature number is at once determined by this expectation.

The other case historically addressed is classification with two Gaussian class-conditional distributions. The Bayes classifier Ψ_d is determined by a discriminant $Q_d(\boldsymbol{x})$, with $\Psi_d(\boldsymbol{x}) = 1$ if and only if $Q_d(\boldsymbol{x}) > 0$. In practice, the discriminant is estimated from sample data. The standard plug-in rule to design a classifier from a feature-label sample of size n is to obtain an estimate $Q_{d,n}$ of Q_d by replacing the means and covariance matrices in the discriminant by their respective sample means and sample covariance matrices. The designed classifier $\Psi_{d,n}$ is determined by the estimated discriminant according to $\Psi_{d,n}(\boldsymbol{x}) = 1$ if and only if $Q_{d,n}(\boldsymbol{x}) > 0$. Since the discriminant yields a quadratic decision boundary, the method is known as *quadratic discriminant analysis* (QDA). For equal covariance matrices, the discriminant is a linear function, and the method is called *linear discriminant analysis* (LDA). For LDA, representation of the distribution of $Q_{d,n}$ goes back more than 40 years (Bowker, 1961), as does the discovery of an analytic expression for the expected error $E[\varepsilon_d(S_n)]$ under the assumption that the sample is evenly split between the two classes (Sitgreaves, 1961). Recently, a representation of the distribution for $Q_{d,n}$ has been discovered for QDA (McFarland and Richards, 2002). This representation has been used in an analytic approach to the expected error $E[\varepsilon_d(S_n)]$ in the following way: the mean and variance of $Q_{d,n}$ are derived exactly from the McFarland–Richards representation; with these a normal approximation to the distribution of $Q_{d,n}$ is constructed; and $E[\varepsilon_d(S_n)]$ is found based on the normal approximation (Hua *et al*., 2004).

It is reasonable to expect that error representations will be difficult to obtain for most classification rules. Moreover, as the current simulation study shows, one must be wary of generalizing the behavior of LDA classifiers. Even for LDA there are significant differences in error behavior depending on the feature-label distribution, for instance, between correlated and uncorrelated features. The combination of the mathematical difficulty of deriving error representations and large differences in error makes it important to conduct large simulation studies under different conditions to gain an understanding of the kind of feature-set sizes that should be employed.

One might conjecture that a practical way to proceed would be to try feature sets of varying sizes and then choose the designed classifier having the least error; however, this approach is fraught with danger. When a classifier is designed from a small sample, its error must be estimated using the sample data by a method such as cross-validation, but such methods are very inaccurate in the sense that the expected absolute deviation between the estimated error and the true error is often very large, the situation being worse for complex classification rules and with increasing numbers of features (Braga-Neto and Dougherty, 2004b). Thus, trying numerous feature sets and selecting the one with the lowest estimated error presents a multiple-comparison type problem in which it is likely that some feature set will have an estimated error far below its true error, and therefore appear to provide excellent classification. Since variation is worse for large feature sets, this could create a bias in favor of large feature sets, which goes directly into the teeth of the peaking phenomenon. Hence, gaining an idea of feature-set sizes that provide good classification in various circumstances is of substantial benefit.

2 SIMULATION WITH SYNTHETIC DATA

Seven classifiers are considered in our study: 3-nearest-neighbor (3NN), Gaussian kernel, linear support vector machine (linear SVM), polynomial support vector machine (polynomial SVM), perceptron, regular histogram and linear discriminant analysis (LDA). For linear SVM and polynomial SVM, we use the codes provided by LIBSVM 2.4 (Chang and Lin, 2000) with the default setting, except that for polynomial SVM the degree in the kernel function is set to 6. For the Gaussian kernel, the smoothing factor h has been set to 0.2 after various trials. For the regular histogram classifier, the cell number along each dimension is set to 2 or 3 and evaluated separately, after which the optimal value between the two is selected. We consider three two-class distribution models:

Linear model: The class-conditional distributions are Gaussian, $N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, with identical covariance matrices, $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$. The Bayes classifier is linear and the Bayes decision boundary is a hyperplane. Without loss of generality, we assume that $\boldsymbol{\mu}_0 = (0, 0, \ldots, 0)$ and $\boldsymbol{\mu}_1 = (1, 1, \ldots, 1)$.

Non-linear model: The class-conditional distributions are Gaussian with covariance matrices differing by a scaling factor, namely, $\lambda\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}$. Throughout the study, $\lambda = 2$. The Bayes classifier is non-linear and the Bayes decision boundary is quadratic. Again we assume that $\boldsymbol{\mu}_0 = (0, 0, \ldots, 0)$ and $\boldsymbol{\mu}_1 = (1, 1, \ldots, 1)$.

Bimodal model: The class-conditional distribution of class S_0 is Gaussian, centered at $\boldsymbol{\mu}_0 = (0, 0, \ldots, 0)$, and the class-conditional distribution of class S_1 is a mixture of two equiprobable Gaussians, centered at $\boldsymbol{\mu}_{10} = (1, 1, \ldots, 1)$ and $\boldsymbol{\mu}_{11} = (-1, -1, \ldots, -1)$. The covariance matrices of the classes are identical. The Bayes decision boundaries are two parallel hyperplanes. Owing to the extreme non-linear nature of this model, the perceptron, linear SVM and LDA classifiers are omitted from our study in this model.

Throughout, we assume that the two classes have equal prior probability. The maximum dimension is $D = 30$. Hence, the number of features available is $\leq$30 and the peaking phenomenon will only show up in the graphs for which peaking occurs with $<$30 features.

As noted in the Introduction, to avoid the confounding effects of feature selection, we employ a covariance-matrix structure. We let all features have common variance, so that the 30 diagonal elements in $\boldsymbol{\Sigma}$ have the identical value σ^2. To set the correlations between features, the 30 features are equally divided into G groups, with each group having $K = 30/G$ features. To divide the features equally, G cannot be arbitrarily chosen. The features from different groups are uncorrelated, and the features from the same group possess the same correlation ρ among each other. If $G = 30$, then all features are uncorrelated. We denote a particular feature with the label $F_{i,j}$, where i, $1 \leq i \leq G$, denotes the group to which the feature belongs and j, $1 \leq j \leq K$, denotes its position in that group. The full feature set takes the form $\boldsymbol{F} = \{F_{1,1}, F_{2,1}, \ldots, F_{G,1}, F_{1,2}, \ldots, F_{G,K}\}$. For

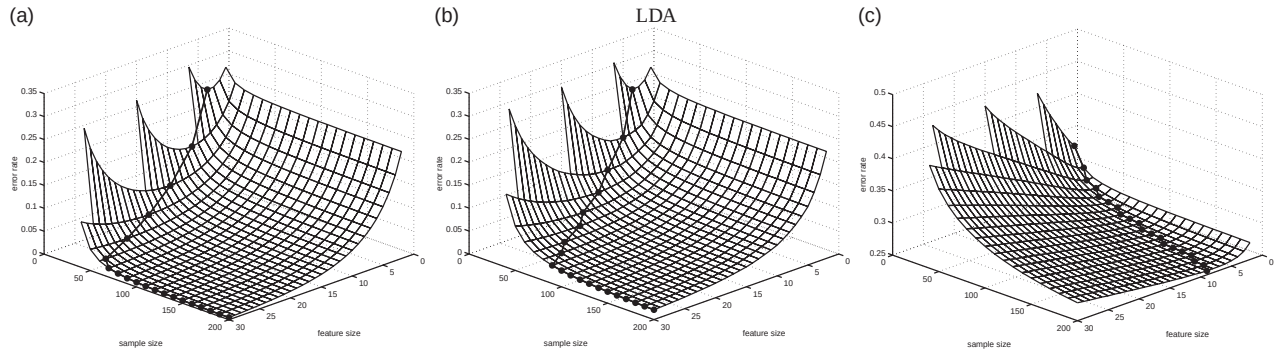

Fig. 1. Optimal feature size versus sample size for LDA classifier. Linear model is tested. σ^2 is set to let Bayes error be 0.05. **(a)** Uncorrelated features. **(b)** Slightly correlated features, $G = 5$, $\rho = 0.125$. **(c)** Highly correlated features, $G = 1$, $\rho = 0.5$.

any simulation based on a feature subset of d features, the first d features in $\boldsymbol{F}$ are picked. For example, if $G = 10$, then each group has $K = 3$ features, and the covariance matrix, with features ordered as $F_{1,1}, F_{1,2}, F_{1,3}, F_{2,1}, \ldots, F_{10,1}, F_{10,2}, F_{10,3}$, is

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & \rho & \rho & & & & & & & & & \\ \rho & 1 & \rho & & 0 & & \cdot & \cdot & \cdot & & 0 & \\ \rho & \rho & 1 & & & & & & & & & \\ & & & 1 & \rho & \rho & & & & & & \\ & 0 & & \rho & 1 & \rho & \cdot & \cdot & \cdot & & 0 & \\ & & & \rho & \rho & 1 & & & & & & \\ & \cdot & & & \cdot & & \cdot & & & & \cdot & \\ & \cdot & & & \cdot & & & \cdot & & & \cdot & \\ & \cdot & & & \cdot & & & & \cdot & & \cdot & \\ & & & & & & & & & 1 & \rho & \rho \\ & 0 & & & 0 & & \cdot & \cdot & \cdot & \rho & 1 & \rho \\ & & & & & & & & & \rho & \rho & 1 \end{bmatrix}.$$

In our study, all seven classifiers are applied to the three distribution models (except for the perceptron, linear SVM and LDA for the bimodal model, as already explained). For each model, altogether 30 different cases are considered according to different covariance-matrix structures and variances:

Variance (σ^2): Three different variances σ^2 are chosen for each model. They correspond to Bayes errors 0.05, 0.10 and 0.15, under the assumption of 10 uncorrelated features.

Covariance matrix: Four basic covariance-matrix structures are studied by dividing the 30 features into $G = 1$, 5, 10 and 30 groups. For $G = 1$, 5 and 10, three different correlation coefficients, $\rho = 0.125$, 0.25 and 0.5, are considered. Thus, the total number of covariance-matrix structures studied is 10. For each variance σ^2, different covariance-matrix structures will have different Bayes errors. The increase in correlation among features, either by decreasing G or increasing ρ, will increase the Bayes error for a fixed feature size.

For each case, performances, i.e. error rates, of various classifiers are estimated at various feature sizes and sample sizes based on Monte Carlo simulations:

Feature size (d): Except for the regular histogram, all classifiers are tested on 29 different feature sizes from 2 to 30. For the regular histogram, owing to the exponentially increasing cell numbers, feature sizes are limited from 1 to 10.

Sample size (n): Sample sizes run from 10 to 200, increased by steps of 10, for a total of 20 sample sizes.

For each feature size d and sample size n, the simulation generates n training points according to the distribution model, variance and covariance matrix being tested. The trained classifier is applied to 200 independently generated test points from the identical distribution. This procedure is repeated 25 000 times for all classifiers except LDA with 100 000 repetitions and linear SVM and polynomial SVM with 5000 repetitions, the latter being extremely computationally intensive. Error rates are then averaged. The results are presented by a 2-D mesh plot like that in Figure 1. The black lines with circular markers are those with the lowest error rates, and hence the ones showing the optimal feature sizes. There are 540 mesh plots on the website. In the next section we discuss some representative results.

3 RESULTS ON SYNTHETIC DATA

Figure 1 shows the effect of correlation for LDA with the linear model. Note that the sample size must exceed the number of features to avoid degeneracy. For uncorrelated features in Figure 1(a), the optimal feature size is around $n - 1$, which matches well with previously reported LDA results (Jain and Waller, 1978). As feature correlation increases, the optimal feature size decreases, becoming very small when correlation is high. This also matches results in the same paper, which claims that the optimal feature size is proportional to $\sqrt{n}$ for highly correlated features. In all three parts, uncorrelated, slightly correlated and highly correlated, we see the peaking phenomenon and observe that the optimal number of features increases with increasing sample size. For microarray-based studies, where sample sizes of <50 and feature correlation are commonplace, one should note that with slight correlation, the optimal number of features for $n = 30$ and $n = 50$ is $d = 12$ and $d = 19$, respectively, and with high correlation, the optimal number of features for $n = 30$ and $n = 50$ is $d = 3$ and $d = 4$, respectively. Similar results have been obtained for the non-linear model.

Figure 2 provides some results for the regular histogram classifier on the three models. The cell number increases exponentially with feature size and the optimal number of features is quite small in all three models. The curve of the optimal number of features as a function of the sample size shows the common increasing monotonicity. The optimal feature size for the bimodal model is larger, indicating

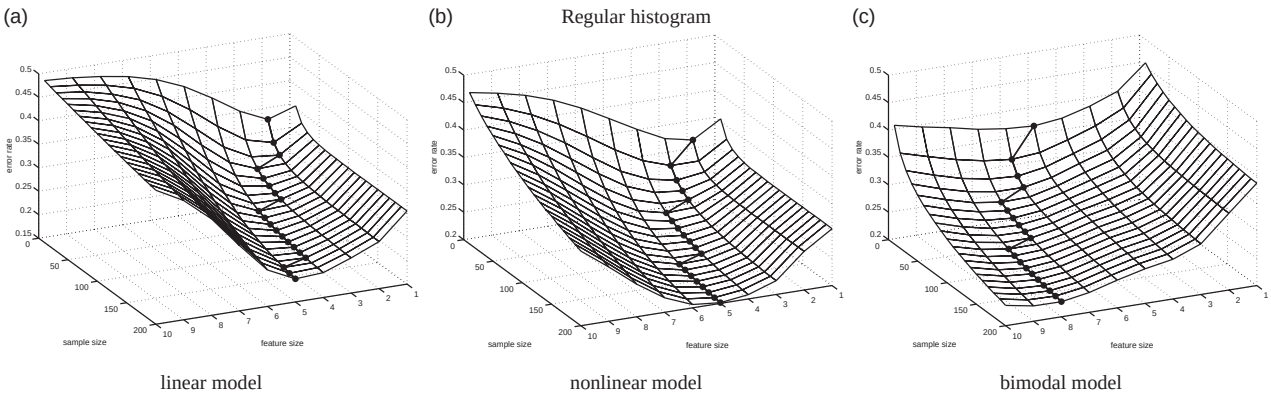

Fig. 2. Optimal feature size versus sample size for regular histogram classifier. Uncorrelated features. σ^2 is set to let Bayes error be 0.05.

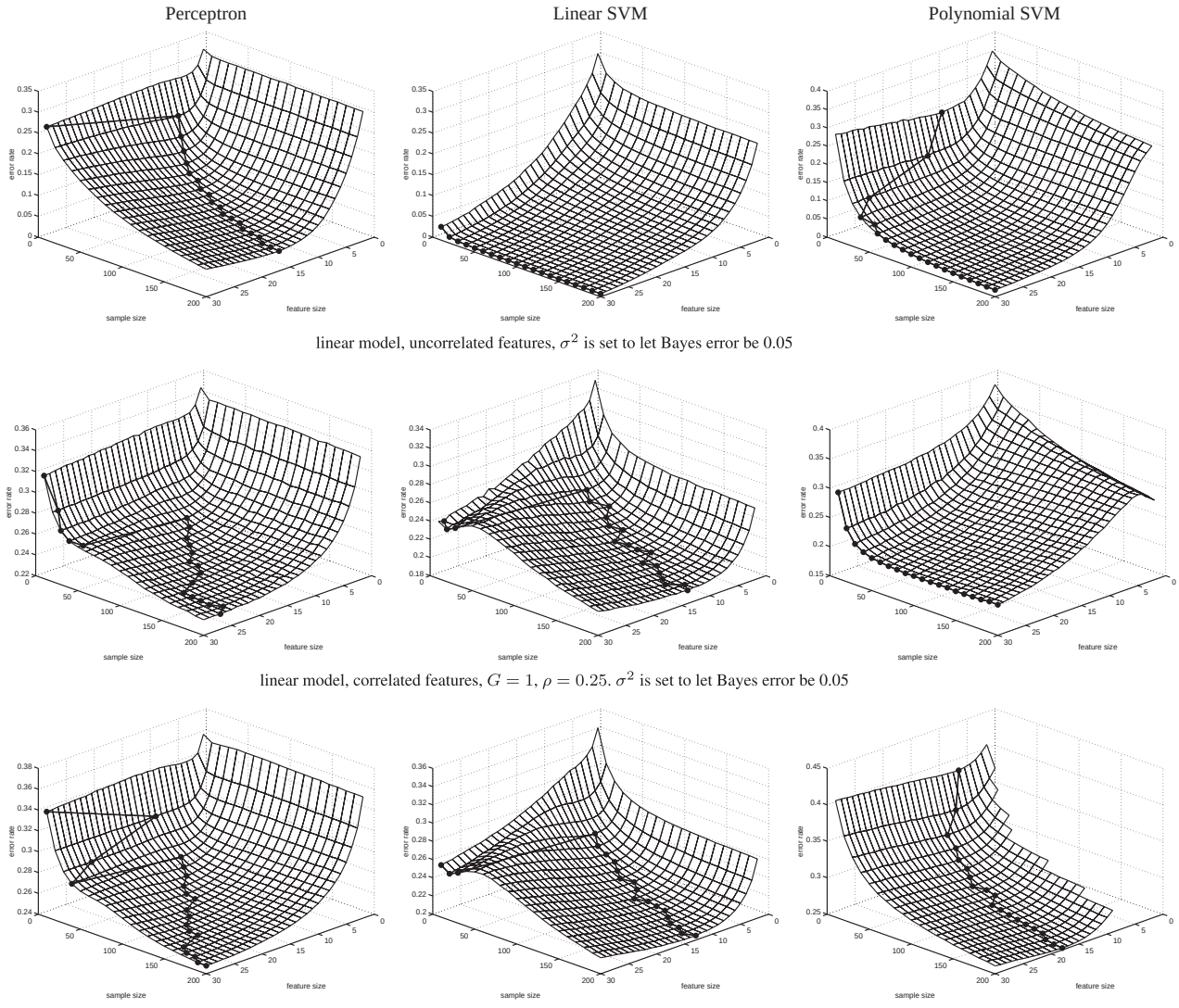

Fig. 3. Optimal feature size versus sample size for perceptron and SVM classifiers.

the need for more features to separate three concentrations of mass as opposed to two.

In Figure 3, we compare the perceptron, linear SVM and polynomial SVM classifiers. Of practical importance, the linear SVM shows no peaking phenomenon for up to 30 features, the polynomial SVM peaks at under 30 only for quite small samples on the uncorrelated linear model and the polynomial SVM shows no peaking at up to 30 features for the correlated linear model. When there is no peaking, one can safely use a large number of features even for small samples. The optimal-feature-size curves for the perceptron and linear SVM for the correlated linear and non-linear models are quite similar, whereas they are very different for the uncorrelated linear model. Note also that the error rate drops much faster relative to sample size for the polynomial SVM in comparison to the linear SVM for the correlated model.

Perhaps the most interesting aspect of Figure 3 is that there are cases in which the optimal number of features is not monotonically increasing with the sample size (and here we are not referring to the slight wobble owing to a flat surface). When it applies, monotonicity follows from the peaking point as the sample size increases. Two phenomena are observed here. With an extremely small sample size ($n = 10$), we observe no peaking for the perceptron and linear SVM except for the perceptron in the non-linear model, and the peaking is extremely slight. More striking is that, for the perceptron in all cases and the linear SVM in the correlated cases, in a range of sample sizes we do not observe the typical concave behavior of the error as a function of the number of features. On the contrary, in some feature size range, the classification error will increase and then decrease with the feature size, thereby forming a ridge across the error surface. A zoomed plot for the perceptron in the uncorrelated case in Figure 4(a) shows the ridge.

This phenomenon can be appreciated by decomposing the error of the designed classifier into the sum of the error, ε_d, of the optimal classifier for the classification rule relative to the feature-label distribution and the cost, $\Delta_d(S_n)$, of designing a classifier from the sample S_n. Then, taking expectation with respect to the distribution of the samples yields

$$E[\varepsilon_d(S_n)] = E[\Delta_d(S_n)] + \varepsilon_d.$$

Considering the expected error as a function of the feature number d, the common interpretation is that $E[\varepsilon_d(S_n)]$ decreases to a minimum at d_0 and thereafter increases with increasing d. This means that for $d < d_0$, the optimal error ε_d is falling faster than the design cost $E[\Delta_d(S_n)]$ is rising, and that for $d > d_0$, the optimal error ε_d is falling slower than the design cost $E[\Delta_d(S_n)]$ is rising. The feature sets for $d < d_0$ are said to *underfit* the data because there is insufficient classifier complexity to take full advantage of the data to separate the classes, whereas feature sets for $d > d_0$ are said to *overfit* the data because the complexity of the classifier allows it to produce decision regions that too closely follow the sample points. Under this interpretation, $E[\varepsilon_d(S_n)]$ decreasing to a minimum at d_0 and thereafter increasing mean there is decreasing underfitting and then increasing overfitting. The situation may not be so simple. For example, in Figure 4, we observe the following phenomenon: there are feature numbers $d_0 < d_1$ such that for $d < d_0$, ε_d is falling faster than $E[\Delta_d(S_n)]$ is rising; for $d_0 < d < d_1$, ε_d is falling slower than $E[\Delta_d(S_n)]$ is rising; and for $d > d_1$, ε_d is falling faster than $E[\Delta_d(S_n)]$ is rising. For sample size $n = 10$, simulations

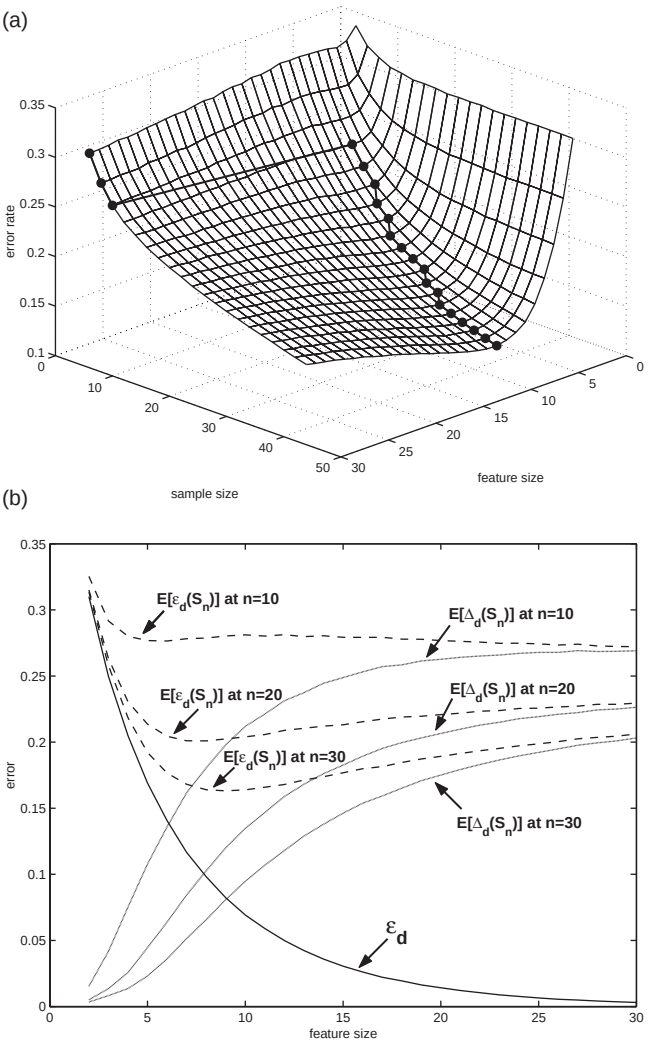

Fig. 4. A case of perceptron classifier: linear model, uncorrelated features, σ^2 is set to let Bayes error be 0.05. **(a)** Optimal feature size versus sample size. **(b)** Relationship among $E[\varepsilon_d(S_n)]$, $E[\Delta_d(S_n)]$ and ε_d for $n = 10$, 20 and 30.

have been run up to 400 features and ε_d still falls no slower than $E[\Delta_d(S_n)]$ rises. Similar phenomena can be observed for other cases of perceptron and some of the SVM classifiers on the complementary website.

Figure 5 shows results for the 3NN classifier on all three models. The surfaces and optimal-feature-size curves for the Gaussian-kernel classifier have almost identical shapes (these being provided only on the website owing to space limitations). Since for the Gaussian kernel the distance between sample points increases with feature size, the posterior probability of the test sample points will be largely determined by the nearest neighbors. Thus, the Gaussian kernel should have similar properties to the 3NN classifier regarding optimal feature size, and this is confirmed by our simulation. A key observation is that for the linear and bimodal models, in which the optimal decision boundaries are flat, there is no peaking up to 30 features. Peaking has been observed in some cases at up to 250 features with sample size $n = 10$, which should have little impact in practical applications.

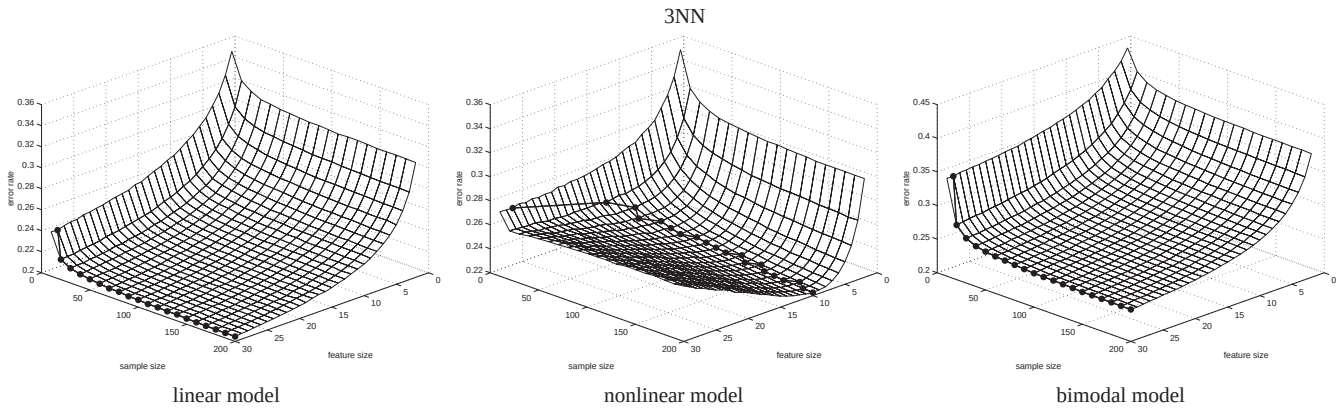

Fig. 5. Optimal feature size versus sample size for the 3NN classifier. Correlated features, $G = 1$, $\rho = 0.25$. σ^2 is set to let Bayes error be 0.05.

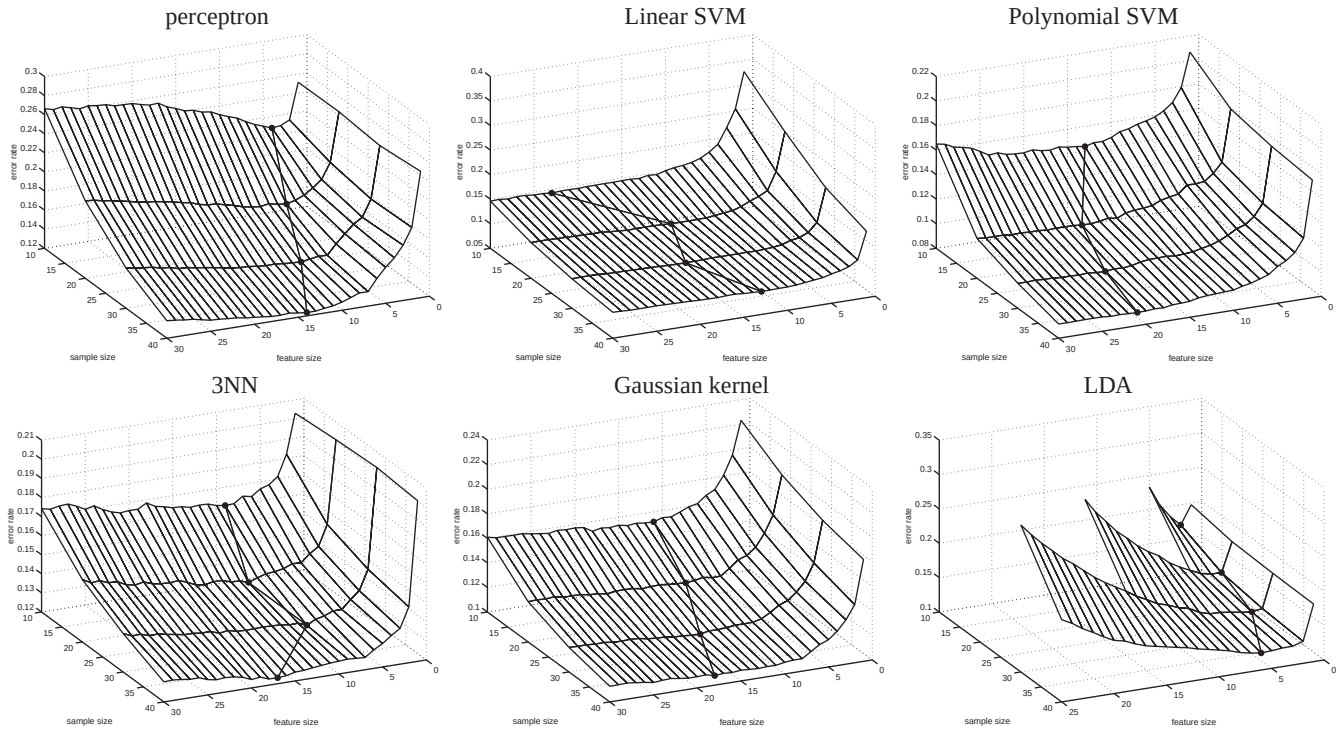

Fig. 6. Optimal feature size versus sample size for various classifiers on real patient data. Sample size $n = 40$.

Once again the optimal-feature-number curve is not increasing as a function of sample size—this being observed in the non-linear model for both classifiers. The optimal feature size is larger at very small sample sizes, rapidly decreases, and then stabilizes as the sample size increases. To check this stabilization, we have tested the 3NN classifier on the non-linear model case in Figure 5 for sample size up to 5000. The result shows that the optimal feature size increases very slowly with sample size. In particular, for $n = 100$ and $n = 5000$, the optimal feature sizes are 9 and 10, respectively. This suggests a useful property of kNN and Gaussian-kernel classifiers: once we find an optimal feature size for a very modest sized sample, we can use the same number of features for much larger samples without sacrificing optimality. Based on our simulations, using more than $d = 10$ features is counterproductive for the models considered.

4 PATIENT DATA

In addition to the synthetic data, we have conducted experiments based on real patient data. With real data it is not possible to perform the kind of systematic study performed on synthetic data arising from parameterized distributions. Our purpose in considering a particular set of real microarray data is to demonstrate that the behavior of optimal feature-set sizes for such data can bear similarity to that arising from synthetic data, in particular, with correlated synthetic data.

The patient data come from a microarray-based cancer-classification study (van de Vijver *et al*., 2002) that analyzes a large number of microarrays prepared with RNA from breast tumor samples of 295 patients. Using a previously established 70-gene prognosis profile (van't Veer *et al*., 2002), a prognosis signature based on gene-expression is proposed in van de Vijver *et al*. (2002) that correlates well with patient survival data and other existing clinical measures. Of the 295 microarrays, 115 belong to the 'good-prognosis' class, whereas the remaining 180 belong to the 'poor-prognosis' class.

All classifiers are tested on various feature sizes from 1 to 30, except the regular histogram, which is omitted for the patient data because its error surface is too rough with the limited number of replications used. To mitigate the confounding effects of feature selection, for each feature-set size, floating forward selection (Pudil *et al*., 1994) is used to find a (hopefully) close-to-optimal feature subset based on all 295 data points. This will provide 'population-based' feature sets whose sample-based performances can then be evaluated. To evaluate the performance of each feature subset, we approximate the classification error with a hold-out estimator. For a sample size of n, 1000 samples of size n are drawn independently from the 295 data points, and for each observation the different classifiers trained on the n points are tested on the $295 - n$ points not drawn. The 1000 error rates are averaged to obtain an estimate of the sample-based classification error. Since the observations are actually not independent, a large n will induce inaccuracy in the estimation. Hence, we limit n under 40 to reduce the impact of observation correlation. The results are shown in Figure 6, where all classifiers show some degree of overfitting beginning at feature size from 10 to 20, some significant and some insignificant. Owing to only 1000 samples, there is some wobble in the flat regions of the graphs. Ignoring this, there is decent concordance with the correlated synthetic data—one should not expect complete concordance. Note that the flatness of the SVM graphs, especially in the polynomial case, again indicates the robustness of SVM classification relative to using large feature sets with small samples. Compare this to the lack of feature-size robustness for LDA classification. As with the model cases, there is similarity in the optimal-feature-size performance of the 3NN and Gaussian-kernel classifiers; however, with the patient data, there is earlier peaking for sample sizes below 20, but this is fairly slight.

5 CONCLUSION

Two conclusions can safely be drawn from this study. First, the behavior of the optimal-feature-size relative to the sample size depends strongly on the classifier and the feature-label distribution. An immediate corollary is that one should be wary of rules-of-thumb generalized from specific cases. Second, the performance of a designed classifier can be greatly influenced by the number of features and therefore one should attempt to use a number close to the optimal number. This means that it can be useful to refer to a database of optimal-feature-size curves to choose a feature size, even if this means making a necessarily very coarse approximation of the distribution model from the data—even perhaps just a visual assessment of the data. Owing to the roughness of these kinds of approximations, a classifier like the polynomial SVM, which shows strong robustness with respect to large feature sets, has inherent advantages over a classifier like LDA, which does not show robustness.

REFERENCES

Bowker,A.H. (1961) A representation of Hotelling's T^2 and Anderson's classification statistics. In Solomon,H., ed. *Studies in Item Analysis and Prediction*. Stanford University Press, Stanford, pp. 285–292.

Braga-Neto,U. and Dougherty,E.R. (2004a) Exact performance of error estimators for discrete classifiers, submitted.

Braga-Neto,U. and Dougherty,E.R. (2004b) Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, **20**, 374–380.

Chang,C.-C. and Lin,C.-J. (2000) LIBSVM: Introduction and benchmarks. Technical Report, Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan.

Cover,T.M. and Van Campenhout,J.M. (1977) On the possible orderings in the measurement selection problem. *IEEE Trans. System. Man. Cybernetics*, **7**, 657–661.

El-Sheikh,T.S. and Wacker,A.G. (1980) Effect of dimensionality and estimation on the performance of gaussian classifiers. *Pattern Recog.*, **12**, 115–126.

Hua,J., Xiong,Z. and Dougherty,E.R. (2004) Determination of the optimal number of features for quadratic discriminant analysis via the normal approximation to the discriminant distribution. *Pattern Recog.*, **38**, 403–421.

Hughes,G.F. (1968) On the mean accuracy of statistical pattern recognizers. *IEEE Trans. Information Theory*, **14**, 55–63.

Jain,A.K. and Chandrasekaran,B. (1982) Dimensionality and sample size considerations in pattern recognition practice. In Krishnaiah,P.R., Kanal,L.N., eds. *Handbook of Statistics, Vol. II* North-Holland, Amsterdam, pp. 835–855.

Jain,A.K. and Waller,W.G. (1978) On the optimal number of features in the classification of multivariate gaussian data. *Pattern Recog.*, **10**, 365–374.

Jain,A.K. and Zongker,D. (1997) Feature selection—evaluation, application, and small sample performance. *IEEE Trans. Pattern Anal. Machine Intell.*, **19**, 153–158.

Kudo,M. and Sklansky,J. (2000) Comparison of algorithms that select features for pattern classifiers. *Pattern Recog.*, **33**, 25–41.

McFarland,H.R. and Richards,D.St.P. (2002) Exact misclassification probabilities for plug-in normal quadratic discriminant functions II. The heterogeneous case. *J. Multivariate Anal.*, **82**, 299–330.

Pudil,P., Novovič ová,J. and Kittler,J. (1994) Floating search methods in feature selection. *Pattern Recog. Lett.*, **15**, 1119–1125.

Sitgreaves,R. (1961) Some results on the distribution of the W-classification. In Solomon,H., ed. *Studies in Item Analysis and Prediction*. 241–251, Stanford University Press, Stanford, pp. 241–251.

van de Vijver,M.J., He,Y.D., van't Veer,L.J., Dai,H., Hart,A.A.M., Voskuil,D.W., Schreiber,G.J., Peterse,J.L., Roberts,C., Marton,M.J. *et al*. (2002) A gene-expression signature as a predictor of survival in breast cancer. *New Engl. J. Med.*, **347**, 1999–2009.

Van Ness,J.W. and Simpson,C. (1976) On the effects of dimension in discriminant Analysis. *Technometrics*, **18**, 175–187.

van't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A.A.M., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.