Computer Methods and Programs in Biomedicine (2004) **73**, 91—99

Computer Methods and
Programs in Biomedicine

www.intl.elsevierhealth.com/journals/cmpb

Many are called, but few are chosen. Feature selection and error estimation in high dimensional spaces

Helene Schulerud[a,*], **Fritz Albregtsen**[b]

[a] *SINTEF, PB 124 Blindern, 0314 Oslo, Norway*
[b] *Department of Informatics, University of Oslo, PB 1080 Blindern, 0316 Oslo, Norway*

Received 15 November 2001; received in revised form 27 November 2002; accepted 1 February 2003

KEYWORDS
Feature selection;
Error estimation;
Selection bias;
Discriminant analysis;
Sample size effects;
Feature candidates;
Training set

Summary We address the problems of feature selection and error estimation when the number of possible feature candidates is large and the number of training samples is limited. A Monte Carlo study has been performed to illustrate the problems when using stepwise feature selection and discriminant analysis. The simulations demonstrate that in order to find the correct features, the necessary ratio of number of training samples to feature candidates is not a constant. It depends on the number of feature candidates, training samples and the Mahalanobis distance between the classes. Moreover, the leave-one-out error estimate may be a highly biased error estimate when feature selection is performed on the same data as the error estimation. It may even indicate complete separation of the classes, while no real difference between the classes exists. However, if feature selection and leave-one-out error estimation are performed in one process, an unbiased error estimate is achieved, but with high variance. The holdout error estimate gives a reliable estimate with low variance, depending on the size of the test set.

© 2003 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

In many applications of pattern recognition, the designer finds that the number of possible features, which could be included in the analysis is surprisingly high. Normally, a substantial part of the feature candidates are only noise or are correlated to the other features, meaning that they contain no additional information about the classes. If such features are included in a classifier the classification performance can easily be degraded.

In medical applications the data set is often limited. Evaluating many features on a small set of data is a challenging problem which has not yet been solved. In this paper some pitfalls in feature selection and error estimation in discriminant analysis on limited data sets will be discussed. It is well known that the number of training samples affects the feature selection and the error estimation, but the effect of the number of feature candidates is not discussed in the pattern recognition literature.

We present simulation results demonstrating how the number of feature candidates, sample size and the Mahalanobis distance influence the number of correct selected features and the different error rate estimates.

*Corresponding author. Tel.: +47-22-06-7722;
fax: +47-22-06-7350.
E-mail address: helene.schulerud@sintef.no (H. Schulerud).

0169-2607/$ — see front matter © 2003 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/S0169-2607(03)00018-X

Section 2 describes common feature selection methods and in Section 3 is the optimal error rate and some error estimates described. In Section 4 the simulation study is presented and in Section 5 the results are given. Section 6 contains a discussion of the results.

2. Feature selection

The goal of the feature selection is to find the subset of features which best characterizes the differences between groups and which is similar within the groups by maximizing the ratio of between-class variance to that of within-class variance. In pattern recognition literature there is a large amount of papers addressing the problem of feature selection [1,2]. Some of the methods are mentioned below.

For a given quality criterion, exhaustive search (testing all possible combinations of d out of D candidates) finds the optimal feature set. However, the number of possible sets grows exponentially with the number of feature candidates, particularly if we do not know a priori the maximum number (m) of feature to select: $N = \sum_{d=1}^{m} \frac{D!}{(D-d)!d!}$. Thus, the method is impractical even for a moderate number of features. A suboptimal approach is the selection of the best single features according to some quality criterion. However, the individually best features are often correlated and thus may give a clearly suboptimal discrimination [3—5]. Sequential forward or backward and stepwise forward—backward selection are other suboptimal approaches [5]. Sequential forward selection [6] adds one new feature to the current set of selected features in each step. Sequential backward selection [7] starts with all the possible features and discards one at the time. The main drawback with forward and backward selection is that when a feature is selected or removed this decision can not be changed. This is called the nesting problem. Stepwise forward—backward selection [8] combines a forward and backward selection strategy, and thus overcomes the nesting problem. Stepwise forward—backward selection is a special case of 'plus l-take away r', where a given number of features (l) are included and another given number of features (r) are excluded in each step. A more recently developed suboptimal method is the Floating Search method where the number of features added and removed is allowed to change in each step [2,9,10].

Moreover, an optimal algorithm does not necessary finds the 'best' subset. It does select the optimal feature subset for the particular data set, which has been used in the training process, but how well this feature set will work on new data is uncertain. Like optimal search, advanced suboptimal feature selection methods with a high degree of freedom, have a higher risk of overfitting than more simple methods, when the data set is limited. By overfitting we mean that we select a subset of features which separates well the classes included in the training data, but it does not differentiate new cases.

In this study, we have used stepwise forward selection (plus 1 take away 1) which is a commonly use method and which is easily available in statistical packages such as SAS, SPSS and BMDP.

3. Error estimation

An important part of designing a pattern recognition system is to evaluate how the classifier will perform on future samples. The optimal error rate (${P_e}^O$), also called the Bayes optimal error rate, is the probability of misclassifying a new observation, when the optimal classification rule is used [11]. Since we do not normally know this optimal classification rule, we only get estimates of the optimal error rate. The true error rate (${P_e}^T$) is in this paper defined as the Bayes error using the selected feature set.

There are several methods of error estimation like resubstitution, leave-one-out and holdout. For the resubstitution method [12], the classifier is designed and the error rate of the classifier is estimated using the same data. In the leave-one-out method [13], one sample is omitted from the dataset of n samples, and the n—1 samples are used to design a classifier, which again is used to classify the omitted sample. This procedure is repeated until all the samples have been classified once. For the holdout method, the samples are divided into two mutually exclusive groups (training data and test data). A classification rule is designed using the training data, and the samples in the test data are used to estimate the error rate of the classifier.

Several authors have shown that the resubstitution error estimate is optimistically biased [13—17]. It is also shown that the leave-one-out error estimate is less biased than the resubstitution method, but more variable [18,19]. The holdout method gives slightly pessimistically biased error rates, and it has desirable stability properties [20].

The leave-one-out error estimate can be applied in two different ways. The first approach is to first perform feature selection using all data and afterwards perform leave-one-out to estimate the error, here denoted P^L, using the same data. The second approach is to perform feature selection and leave-one-out error estimation in one step. Then one sample is omitted from the data set, feature selection is performed and a classifier is designed and the omitted sample is classified. This procedure is repeated until all samples are classified. The estimate is denoted P^{L2}.

4. Study design

A Monte Carlo study was undertaken to evaluate the feature selection and error estimation under different conditions. We have analyzed how the number of features in the classification rule (d), the number of feature candidates (D) and the number of samples (n) influence the stepwise feature selection and different error estimates.

Data were generated from two 200 dimensional normal distributions regarded as class one and two. The class means were $\mu_1 = (0, \ldots, 0)$ and $\mu_2 = (\mu_1', \ldots, \mu_r', 0, \ldots, 0)$, $\mu_j' = (\delta/\sqrt{r})$, $r = 5$ being the number of features separating the classes and δ^2 being the Mahalanobis distance between the classes. The data sets consisted of an equal number of observations from each class. The number of samples in training and test sets are denoted n^{Tr} and n^{Te}, respectively, and the total number of samples available is denoted n.

A forward—backward stepwise feature selection method [21] (plus 1 take away 1) was used with α-to-enter equal to α-to-stay equal to 0.2 [22]. The first f features selected by the method were included in the classifier.

We used a Bayesian minimum error classifier [23], assuming Gaussian distributed probability density functions with common covariance matrix and equal apriori class probabilities. The covariance matrix is equal to the identity matrix. The Bayesian classification rule then becomes a linear discriminant function,

$$D(x^i) = [x^i - (\bar{x}_1 - \bar{x}_2)/2]'S^{-1}(\bar{x}_1 - \bar{x}_2)$$

where $\bar{x}_1$, $\bar{x}_2$ and S are unbiased estimates of the distribution parameters using training samples.

$P_e{}^i$ is the number of classification errors divided by the number of samples classified for a given data set. For each set of parameters, 100 data sets were generated and the expected mean error rate and variance were estimated for i equal to the leave-one-out method (L) and the holdout (H). More details of the design of the simulations are given in the Appendix A.

5. Experimental results

The Mahalanobis distance between the classes is an important factor when performing feature selection and error estimation. When the classes are well separated, less samples are needed in order to both find the features which separate the classes and to obtain a good estimate of the error rate. Consequently, the most difficult situation to analyze is when there is only a small or no difference between the classes. Normally, the analyst does not know the Mahalanobis distance between the classes, and therefore, the worst case is important to consider.

5.1. Feature selection

Stepwise forward—backward feature selection was used on the simulated data in order to analyze how the number of feature candidate, training sample size and Mahalanobis distance influence on the number of correct selected features.

The simulations show that the number of correctly selected features increase when the Mahalanobis distance between the classes increase, the number of samples increase and the number of feature candidate decrease, as shown in Figs. 1 and 2. Normally we do not know the Mahalanobis distance between the classes, so we need to analyze the number of training samples (n^{Tr}) and feature candidates (D) and their relation.

Fig. 3 shows the average number of correctly selected features as a function of the number of training samples for three different values of the number of feature candidates. In Fig. 4, the average number of correctly selected features for four different values of the ratio n^{Tr}/D is shown. We observe that:

- if the number of samples is low (less than 200), the number of feature candidates is of great importance, in order to select the correct features.
- When the number of training samples increase, the number of correctly selected features increase.
- For a given number of training samples, it is beneficial to have a low number of feature candidates.
- The optimal ratio n^{Tr}/D depends on the Mahalanobis distance, the number of training samples and feature candidates. Hence, recommending an optimal ratio is not advisable.

In order to see what happens if the number of feature candidates is greater than 200, an additional study was performed. Using the same method and variables as described in Section 4, 100 simulations with 500 feature candidates were performed.

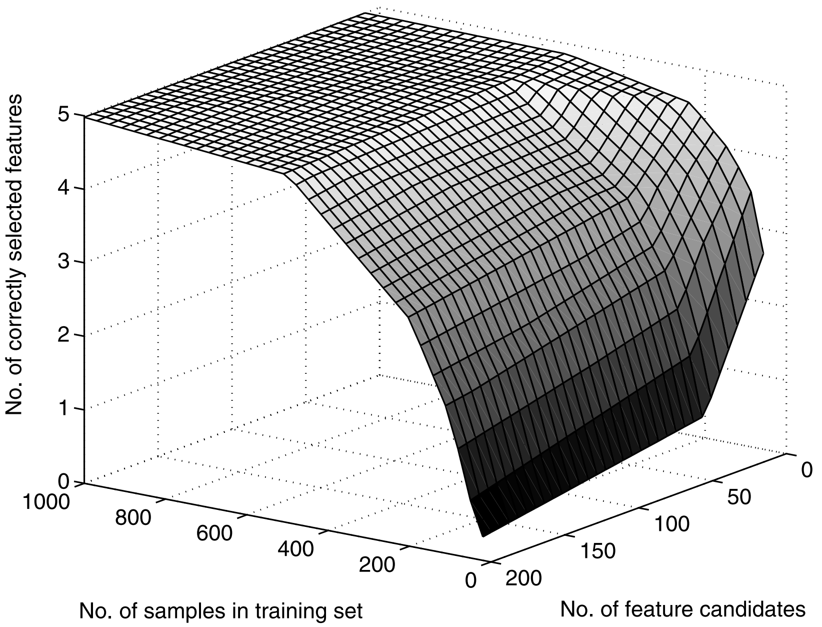

Fig. 1 Number of correctly selected features using different number of training samples and feature candidates when selecting five features out of 200 and the Mahalanobis distance is 1.

Fig. 5 shows that interpolation of the results as shown in Figs. 1 and 2, gives a good indication of $\hat{F}$. When analyzing 500 feature candidates on a training set of 500 samples, the number of correctly selected features was still five. Fig. 6 shows the relation between feature candidates and training samples in order to select a minimum of four correct features. On the diagonal of the figure the number of samples equals the number of feature candidates. Below the diagonal the number of samples is greater than the number of feature candidates. The figure shows that when the classes are overlapping and the number of samples is less than 400, the ratio n^{Tr}/D differs between 1 and 10 and the ratio decrease when the sample size increase.

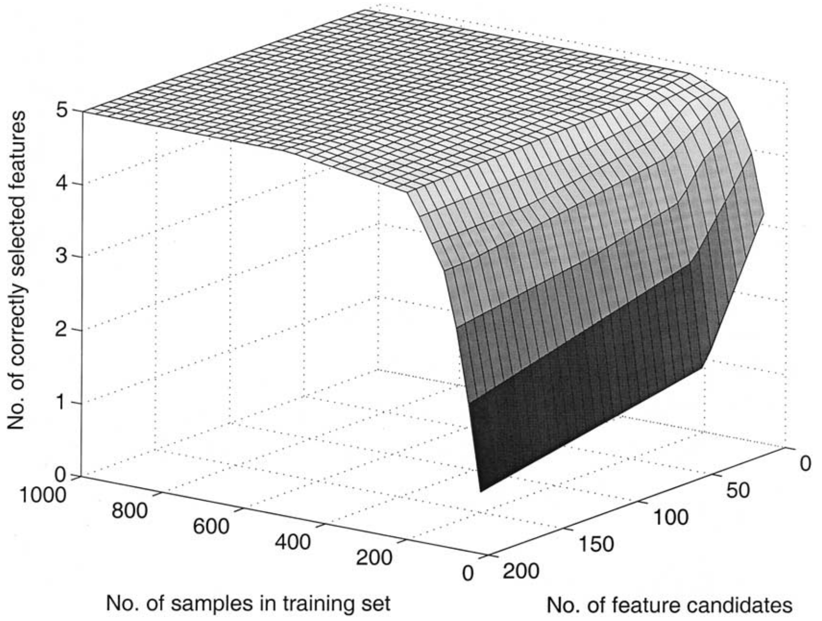

Fig. 2 Number of correctly selected features using different number of training samples and feature candidates when selecting five features out of 200 and the Mahalanobis distance is 4.

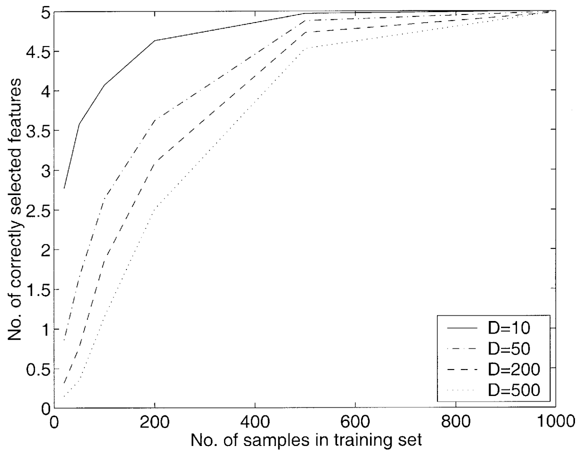

Fig. 3 The average number of correctly selected features, $\hat{F}$, when selecting five features and the Mahalanobis distance is 1. $\hat{F}$ as a function of training samples for three different numbers of feature candidates.

5.2. Performance estimation

The bias of the resubstitution error estimate introduced by estimating the parameters of the classifier and the error rate on the same data set, is avoided in the leave-one-out, since the sample to be tested is not included in the training process. However, if all data are first used in the feature selection process and then the same data are used in error estimation using, e.g. the leave-one-out method ($\hat{P}_e^L$), a bias is introduced. To avoid this bias feature selection and leave-one-out error estimation can be performed in one process ($\hat{P}_e^{L2}$). We have analyzed the bias and variance of these two variants of the leave-one-out error estimate and of the holdout error estimate.

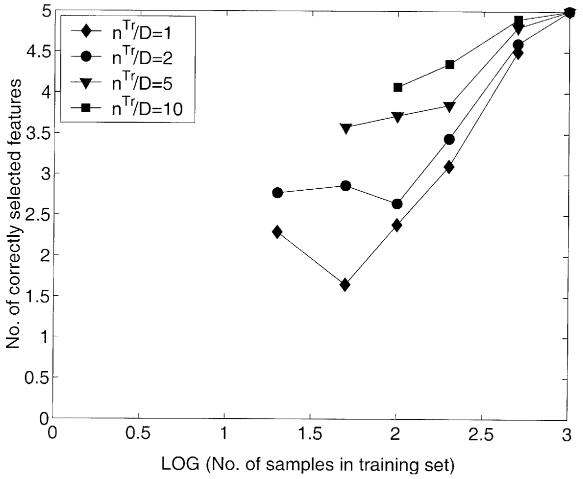

Fig. 4 The average number of correctly selected features, $\hat{F}$, when selecting five features and the Mahalanobis distance is 1. $\hat{F}$ as a function of constant ratio.

Fig. 7 shows the bias and variance of the two leave-one-out error estimates when there is no difference between the classes and we select 5 out of 200 feature candidates. The simulations show that when the number of samples is low (less than 200), the $\hat{P}_e^L$ estimate tends to give a highly optimistic error estimate. Moreover, when analyzing many features on a small data set, the $\hat{P}_e^L$ estimate can indicate complete separation of the classes, while no real difference between the classes exists. As the number of samples increases, the $\hat{P}_e^L$ approaches the true error. The number of samples necessary to get a good estimate of the true error depends on the Mahalanobis distance between the classes and the number of feature candidates. However, the simulation results show that if the number of training samples is greater than 200, the bias of the leave-one-out estimate is greatly reduced.

Performing feature selection and leave-one-out error estimation in one process results in an almost unbiased estimate of the true error, but the $\hat{P}_e^{L2}$ estimate has a high variance, see Fig. 7. When the number of samples is less than 200, the $\hat{P}_e^{L2}$ gives a clearly better estimate of the true error than $\hat{P}_e^L$. The bias and variance of the holdout error estimate ($\hat{P}_e^H$) were analyzed under the same conditions as the leave-one-out estimates, see Fig. 8. The holdout error estimate is also an unbiased estimate of the true error, but with some variance.

The bias of the three error estimate as a function of the number of feature candidates are shown in Fig. 9. The figure shows how the bias of the $\hat{P}_e^L$ error estimate increase with increasing number of feature candidates, while the two other estimates are not affected. Fig. 10 shows the bias of the $\hat{P}_e^L$ estimate as a function of Mahalanobis distance and number of training samples. The figure shows how the bias of the $\hat{P}_e^L$ estimate increases when the Mahalanobis distance decreases, towards the case shown in Fig. 7 ($\hat{P}_e^L$ for $\delta^2 = 0$). We note that for a small number of training samples (less than 200), this leave-one-out error estimate has a significant bias, even for high class distances.

6. Discussion

Our experiments are intendent to show the potential pitfalls of using feature selection on relative small data sets in a high dimensional space. We have analyzed how the number of feature candidates and training samples influence the number of correctly selected features and different error estimates. Monte Carlo simulations have been performed in order to illustrate the problems.

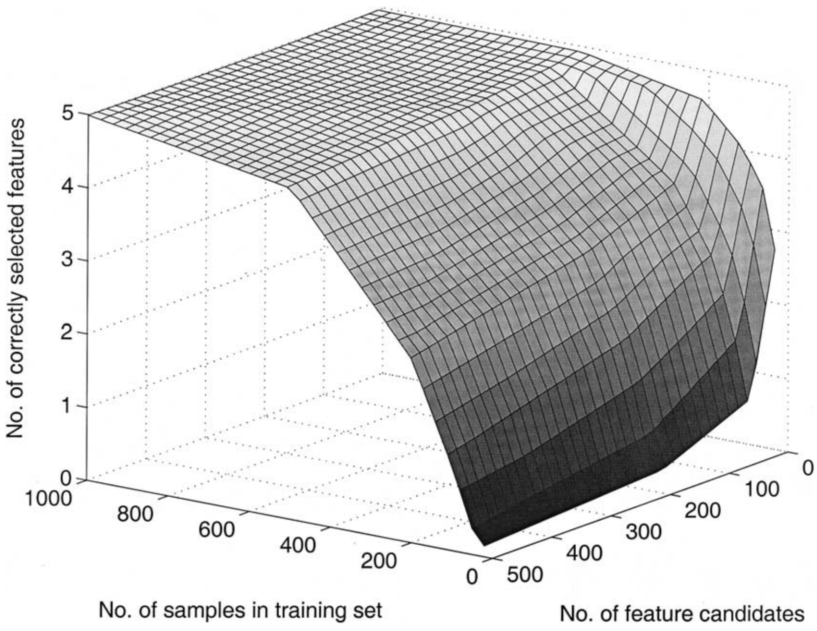

Fig. 5 The average number of correctly selected features as a function of training samples and feature candidates, when the Mahalanobis distance is 1.

The simulations show that few of the correct features are found when the number of samples is low (less than 100). To find most of the correct features the ratio n^{Tr}/D (number of training samples/number of feature candidates) differs between 1 and 10, depending on the Mahalanobis distance, the number of feature candidates and the number of training samples. Hence, to give a recommended general ratio n^{Tr}/D is not possible. However, Figs. 5 and 6 could be used to indicate if the given number of samples and feature candidates used in a stepwise feature selection is likely to find the features which separates the classes.

This result corresponds only partially to previous work by Rencher and Larson [24]. They state that when the number of feature candidates exceeds the degrees of freedom for error [$D > (n^{Tr}-1)$] in stepwise discriminant analysis, spurious subsets and inclusion of too many features can occur. Rutter et al. [20] found that when the ratio of sample size to number of feature candidates was less than 2.5, few correct features were selected, while if the ratio was 5 or more, most of the discriminative features were found.

Moreover, the simulation results demonstrate the effect of performing feature selection before

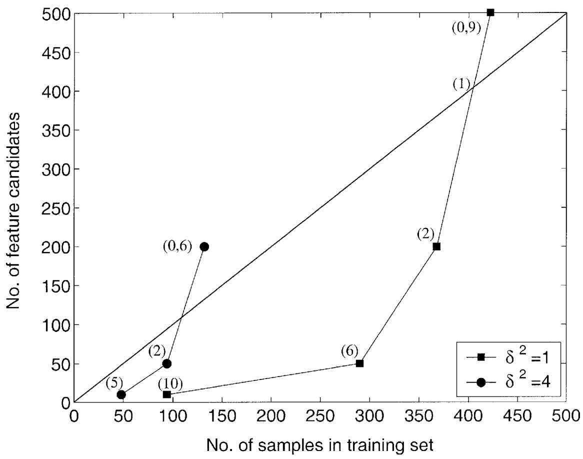

Fig. 6 The relation between feature candidates and training samples necessary to select a minimum of four correct features. The n^{Tr}/D ratio is shown in brackets.

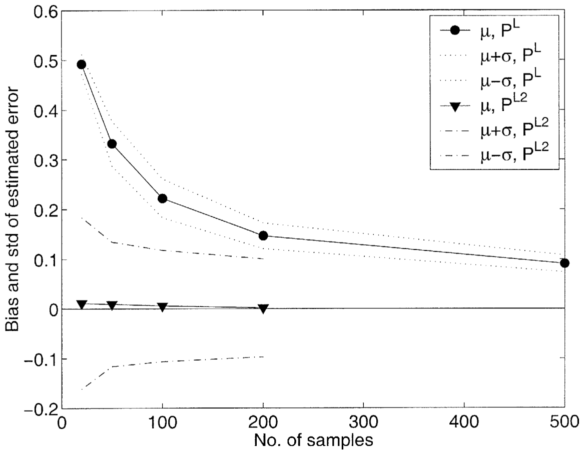

Fig. 7 Bias and variance of the leave-one-out error estimates when the Mahalanobis distance between the classes is zero.

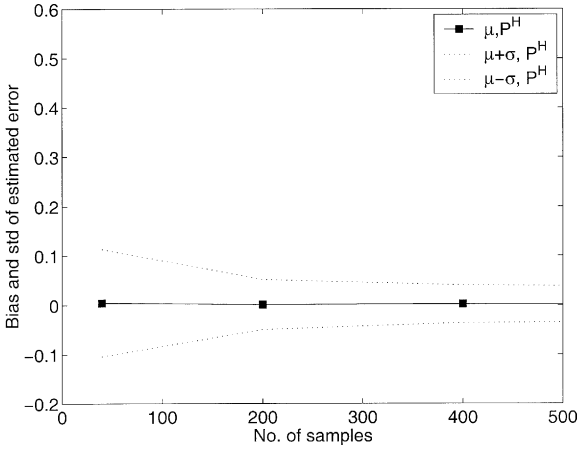

Fig. 8 Bias and variance of the holdout error estimates when the Mahalanobis distance between the classes is zero.

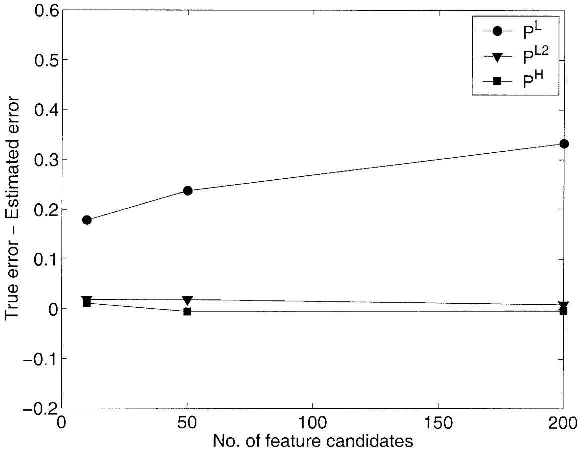

Fig. 9 Bias of error estimates as a function of the number of feature candidates analyzed.

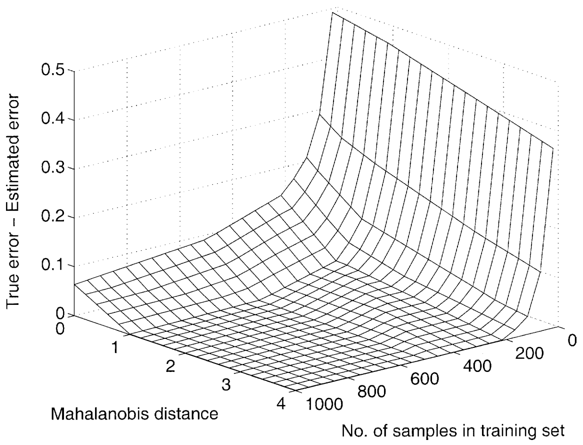

Fig. 10 Bias of the $\hat{P}_e^L$ error estimate as a function of the Mahalanobis distance between the classes and the number of samples, when selecting 5 out of 200 feature candidates.

leave-one-out error estimation on the same data. If the classes are overlapping, the number of training samples is small (less than 200) and the number of feature candidates are high, the common approach of performing feature selection before leave-one-out error estimation on the same data ($\hat{P}_e^L$) results in a highly biased error estimate of the true error. Performing feature selection and leave-one-out error estimation in one process ($\hat{P}_e^{L2}$) gives an unbiased error estimate, but with high variance. The holdout error estimate is also an unbiased estimate, but with less variance than $\hat{P}_e^{L2}$.

When performing a proper leave-one-out, feature selection is performed within each cycle of the leave-one-out procedure. Thus, if we are selecting d out of D features in this manner, up to n^{Tr} different feature sets of dimension d may be selected. The correct classification rate given is a result of n^{Tr} different classifiers. Hence, two questions arise: which features should be used in the classification system for future data, and which performance estimate should be used to give a realistic estimate of the training data? One solution to the first question is to select the most frequently selected features, as suggested by [25]. Another possibility is to weight the most frequently selected features by their rank. A solution to the second question is to rerun the leave-one-out procedure using only the d selected features. Moreover, the simulations showed that if we use the proper leave-one-out estimate, it suffers of high variance. To reduce the variance one could use an n-fold crossvalidation, meaning that we leave n samples out instead of one in the feature selection and error estimation process.

The following conclusions can be drawn based on the simulation results:

- Perform feature selection and error estimation on separate data ($\hat{P}_e^{L2}$, $\hat{P}_e^H$), for small sample sizes ($n^{Tr} < 200$).
- The number of feature candidates is critical when the number of training samples are small.
- In order to find the correct features the n^{Tr}/D ratio differs depending on the number of training samples, feature candidates and the Mahalanobis distance.

A method often used to eliminate feature candidates is to discard one of a pair of highly correlated features. However, this is a multiple comparison test, comparable to the tests performed in the feature selection process. So, the number of feature candidates analyzed will actually not be reduced. If the n^{Tr}/D ratio is low for a given sample size, one should either increase the sample size or reduce the number of feature candidates using non-statistical methods.

In a previous work [26] the bias and variance of different error estimates have been analyzed when applying different feature selection methods. The study showed that using Floating Search feature selection gave the same result as presented here. Some of the results from this study are included here.

Some of the results presented here may be well known in statistical circles, but it is still quite common to see application papers where a small number of training samples and/or a large number of feature candidates render the conclusion of the investigation doubtful at best. Statements about the unbiased nature of the leave-one-out error estimate are quite frequent, although it is seldom clarified whether the feature selection and the error estimation are performed on the same data ($\hat{P}_e^L$) or not ($\hat{P}_e^{L2}$). Finally, comparison between competing classifiers, feature selection methods and so on are often done without regarding the heightened variance that accompanies the proper unbiased error estimate, particularly for small sample sizes. The key results of this study are the importance of the number of feature candidates and that the proper n^{Tr}/D ratio in order to select the correct features is not a constant, but depends on the number of training samples, feature candidates and the Mahalanobis distance, when stepwise feature selection is performed.

Acknowledgements

This work was supported by the Norwegian Research Council (NFR).

Appendix A. Monte Carlo simulations

Samples were generated from two 200 dimensional normal distributions regarded as class one and two. The class means were $\mu_1 = (0, \ldots, 0)$ and $\mu_2 = (\mu_1', \ldots, \mu_r', 0, \ldots, 0)$, $\mu_r' = (\delta/\sqrt{r})$, r being the number of features separating the classes and δ^2 being the Mahalanobis distance between the classes. We used $r = 5$ and a common covariance matrix equal to the identity matrix. The data sets consisted of an equal number of observations from each class, $n_1 = n_2$. The number of samples in training and test sets are denoted n^{Tr} and n^{Te}, respectively, and the total number of samples available is denoted N. Five parameters were varied: n^{Tr}, n^{Te}, D, δ^2. The values of each parameter tested are given in Table A.1.

Table A.1 Values of the different parameters tested

Symbol	Design variable	Values
n^{Tr}	Number of training samples	20, 50, 100, 200, 500, 1000
n^{Te}	Number of test samples	20, 100, 200, 1000
D	Number of feature candidates	10, 50, 200
δ^2	Mahalanobis distance	0, 1, 4

$P_e{}^i$ is the number of classification errors divided by the number of samples classified for a given data set. For each set of parameters, 100 data sets were generated and the expected error rate and variance were estimated for i equal to the leave-one-out method (L) and the holdout (H). $\hat{P}_e^i = 1/k \sum_{x=1}^{k} P_{e,x}^i$ and $\widehat{VAR}\{P_e^i\} = 1/(k-1) \sum_{x=1}^{k} (P_{e,x}^i - \hat{P}_e^i)^2$. When the feature selection method failed to select any features, no classification was performed. Consequently, for some situations the number of simulations (k) were less than 100. This occurred, e.g. when trying to select five out of ten features when the Mahalanobis distance is 0. The expected number of correctly selected features was estimated by the mean number of correctly selected features of the k simulations, and is denoted $\hat{F}$.

References

[1] K.S. Fu, P.J. Min, T.J. Li, Feature selection in pattern recognition, IEEE Trans. Syst. Sci. Cybern. Part C 6 (1) (1970) 33—39.

[2] A. Jain, D. Zongker, Feature selection: evaluation, application, and small sample performance, IEEE Trans. Pattern Anal. Mach. Intell. 19 (2) (1997) 153—158.

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

[4] L. Kanal, Patterns in pattern recognition: 1968—1974, IEEE Trans. Inf. Theory 20 (1974) 697—722.

[5] P. Devijver, J. Kittler, Pattern Recognition. A Statistical Approach, first ed, Prentice-Hall International, London, 1982.

[6] A. Whitney, A direct method of nonparametric measurement selection, IEEE Trans. Comput. 20 (1971) 1100—1103.

[7] T. Marill, D.M. Green, On the effectiveness of receptors in recognition systems, IEEE Trans. Inf. Theory 9 (1963) 11—17.

[8] S.D. Stearns, On selecting features or pattern classifiers. Proceedings of Third International Conference on Pattern Recognition, 1976, pp. 71—75.

[9] P. Pudil, J. Novovicova, J. Kittler, Floating search methods in feature selection, Pattern Recogn. Lett. 15 (1994) 1119—1125.

[10] P. Pudil, J. Novovicova, J. Kittler, Adaptive floating search methods in feature selection, Pattern Recogn. Lett. 20 (1999) 1157—1163.

[11] P.A. Lachenbruch, Discriminant Analysis, first ed, Hafner Press, New York, 1975.

[12] C.A.B. Smith, Some examples of discrimination, Ann. Eugen. 18 (1947) 272—273.

[13] P.A. Lachenbruch, M.R. Mickey, Estimation of error rates in discriminant analysis, Techometrics 10 (1) (1968) 1—11.

[14] L.E. Larsen, D.O. Walter, J.J. McNew, W.R. Adey, On the problem of bias in error rate estimation for discriminant analysis, Pattern Recogn. 3 (1971) 217—223.

[15] D.H. Foley, Considerations of sample and feature size, IEEE Trans. Inf. Theory 18 (1972) 618—626.

[16] G.T. Toussaint, Bibliography on estimation of misclassification, IEEE Trans. Inf. Theory 20 (4) (1974) 472—479.

[17] S. Raudys, A. Jain, Small sample size effects in statistical pattern recognition: recommendations for practitioners, IEEE Trans. Pattern Anal. Mach. Intell. 13 (1991) 252—264.

[18] B. Efron, Estimating the error rate of a prediction rule: improvement on cross-validation, J. Am. Stat. Assoc. 78 (382) (1983) 316—331.

[19] S.M. Snapinn, J.D. Knoke, Estimation of error rates in discriminant analysis with selection of variables, Biometrics 45 (1) (1989) 289—299.

[20] C. Rutter, V. Flack, P. Lachenbruch, Bias in error rate estimates in discriminant analysis when setpwise variable selection is employed, Commun. Stat., Simul. Comput. 20 (1) (1991) 1—22.

[21] W.R. Klecka, in: Discriminant Analysis, Sage Publications, Beverly Hills, 1980, pp. 07—019.

[22] M.C. Constanza, A.A. Afifi, Comparison of stopping rules in forward stepwise discriminant analysis, J. Am. Stat. Assoc. 74 (1979) 777—785.

[23] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, first ed, Wiley-Interscience, 1973.

[24] A.C. Rencher, S.F. Larson, Bias in Wilks' lambda in stepwise discriminant analysis, Technometrics 22 (3) (1980) 349—356.

[25] G.D. Murray, A cautionary note on selection of variables in discriminant analysis, Appl. Stat. 26 (3) (1977) 246—250.

[26] H. Schulerud, F. Albregtsen, Effects of many feature candidates in feature selection and classification, Lect. Notes Comput. Sci. 2396 (2002) 480—487.