Effects of Many Feature Candidates in Feature Selection and Classification

Helene Schulerud[1,2] and Fritz Albregtsen[1]

[1] University of Oslo
PB 1080 Blindern, 0316 Oslo, Norway
[2] SINTEF
PB 124 Blindern, 0314 Oslo, Norway
hsc@sintef.no

Abstract. We address the problems of analyzing many feature candidates when performing feature selection and error estimation on a limited data set. A Monte Carlo study of multivariate normal distributed data has been performed to illustrate the problems. Two feature selection methods are tested: *Plus-1-Minus-1* and *Sequential Forward Floating Selection*. The simulations demonstrate that in order to find the correct features, the number of features initially analyzed is an important factor, besides the number of samples. Moreover, the sufficient 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. The two feature selection methods analyzed gave the same result. Furthermore, the simulations demonstrate how the leave-one-out error estimate can 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.

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 and that the number of samples available is limited. High-dimensional functions have the potential to be much more complicated then low-dimentional ones, and those complications are harder to discern. 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 initially analyzed is not much discussed in the pattern recognition literature.

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. In pattern recognition literature there is a large amount of papers addressing the problem of feature selection [4,5]. In this study two commonly used

T. Caelli et al. (Eds.): SSPR&SPR 2002, LNCS 2396, pp. 480–487, 2002.

suboptimal feature selection methods are analyzed, *Stepwise Forward Backward selection (SFB)* [11], also called *Plus-1-Minus-1*, and *Sequential Forward Floating Selection (SFFS)* [7]. The SFB method was chosen since it is commonly used for exploratory analyses and is available in statistical packages, such as SAS and BMDP. The SFFS method has been reported as the best sub-optimal feature selection method [5] and was therefore included.

An important part of designing a pattern recognition system is to evaluate how the classifier will perform on future samples. There are several methods of error estimation like leave-one-out and holdout. In the leave-one-out method [6], 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.

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, 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 goal of this study is to demonstrate how the number of correctly selected features and the performance estimate depends on the number of feature candidates initially analyzed.

2 Study Design

A Monte Carlo study was performed on data generated from two 200 dimensional normal distributions regarded as class one and two. The class means were $\mu_1 = (0, ..., 0)$ and $\mu_2 = (\mu_1^{'}, \mu_2^{'}, \mu_3^{'}, \mu_4^{'}, \mu_5^{'}, 0, ..., 0)$, $\mu_j^{'} = (\delta/\sqrt{r})$, $r = 5$ 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. We used the Stepwise Forward-Backward (SFB) feature selection method, also called Plus-1-Minus-1, with Wilk's λ as quality criterion ($\alpha - to - enter = \alpha - to - stay = 0.2$) [1], from the SAS statistical package. Sequential Forward Floating Selection (SFFS) was also analyzed, using the MATLAB based toolbox PRTools from Delft [3]. Sum of Mahalanobis distance was used as quality criterion.

Bayesian minimum error classifier [2] was applied, 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. The values of parameters tested are given in Table 1. For each set of parameters,

Table 1. Values of the different parameters tested

Symbol	Design variable	Values
n^{Tr}	No. of training samples	20, 50, 100, 200, 500, 1000
n^{Te}	No. of test samples	20, 100, 200, 1000
D	No. of feature candidates	10, 50, 200
δ^2	Mahalanobis distance	0, 1, 4

100 data sets were generated and the expected error rate,$\hat{P}_e^i$, and variance were estimated for i equal to the leave-one-out (L) and the holdout (H) method, using 50 % of the data as test samples. The expected number of correctly selected features was estimated by the mean number of correctly selected features of the k simulations, and is denoted $\hat{F}$.

3 Experimental Results

3.1 Feature Selection

The simulations show that the number of correctly selected features increases when the Mahalanobis distance between the classes increases, the number of samples increases and the number of feature candidates decreases, as shown in Figure 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.

Figure 1 shows the results of applying stepwise forward-backward (SFB) selection. Figure 1 left 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 Figure 1 right, the average number of correctly selected features for four different values of the ratio n^{Tr}/D is shown. Some additional simulations using 500 feature candidates were performed in order to complete this Figure.

Figure 2 shows the results of stepwise forward-backward (SFB) and sequential forward floating selection (SFFS) when the Mahalanobis distance equals 1 (left) and 4 (right).

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 increases, the number of correctly selected features increases.
- 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.
- The performance of the two feature selection methods analyzed is almost the same.

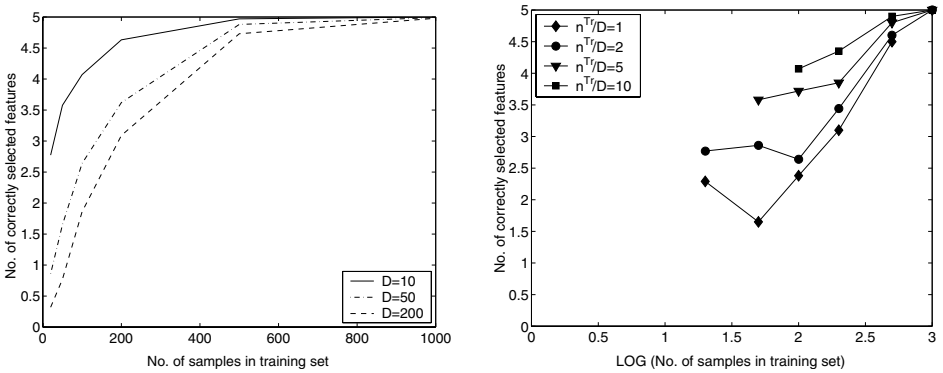

Fig. 1. The average number of correctly selected features, $\hat{F}$, when selecting 5 features and the Mahalanobis distance is 1.**Left:** $\hat{F}$ as a function of training samples for three different numbers of feature candidates. **Right:** $\hat{F}$ as a function of constant ratio

3.2 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.

Figure 3 left 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 using SFB. 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 Figure 3 left. 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

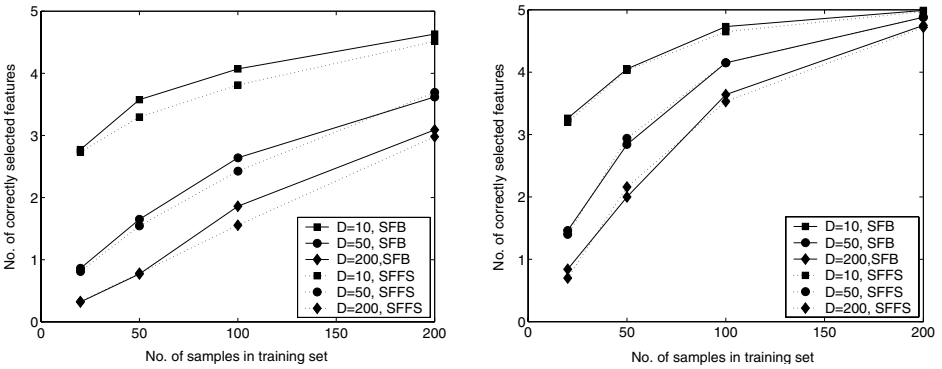

Fig. 2. The average number of correctly selected features as a function of training samples and feature candidates, when the Mahalanobis distance is 1 (left) and 4 (right) for Sequential Forward Backward (SFB) and Sequential Forward Floating Selection (SFFS). D = number of feature candidates

conditions as the leave-one-out estimates, see Figure 3 right. The holdout error estimate is also an unbiased estimate of the true error, but with some variance.

The bias of the three error estimates as a function of the number of feature candidates are shown in Figure 4 left. The Figure shows how the bias of the $\hat{P}_e^L$ error estimate increases with increasing number of feature candidates, while the two other estimates are not affected. Figure 4 right 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. 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.

4 Discussion

Our experiments are intendant to show the potential pitfalls of analyzing a large number of feature candidates on limited data sets. We have analyzed how the number of feature candidates and training samples influence the number of correctly selected features and how they influence different error estimates. Monte Carlo simulations have been performed in order to illustrate the problems.

The simulations show that when the number of training samples is less than 200, the number of feature candidates analyzed is an important factor and affects the number of correctly selected features. Moreover, 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,

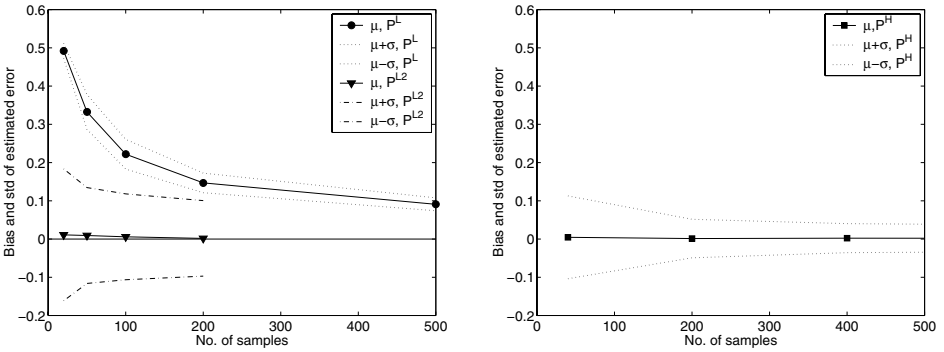

Fig. 3. Bias and variance of error estimates when the Mahalanobis distance between the classes is zero. **Left:** Leave-one-out error estimates. **Right:** Holdout error estimate with 50 % left out

Figure 2 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 separate the classes. This result corresponds only partially to previous work by Rencher and Larson [8]. 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. [9] 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. The two feature selection methods analyzed, *Stepwise Forward Backward selection (SFB)* and *Sequential Forward Floating Selection (SFFS)*, gave the same result.

Furthermore, the simulation results demonstrate the effect of performing feature selection before 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 is 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 $\hat{P}_e^{L2}$.

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

- The number of feature candidates analyzed statistically is critical when the number of training samples is small.
- Perform feature selection and error estimation on separate data, ($\hat{P}_e^{L2}, \hat{P}_e^H$), for small sample sizes.
- 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.

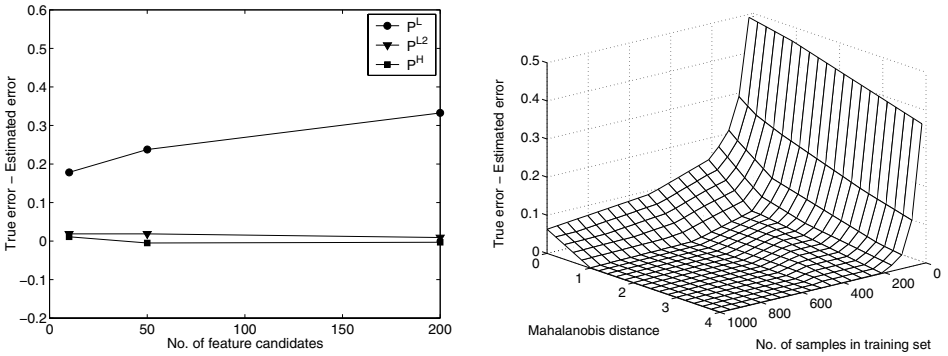

Fig. 4. Left: Bias of error estimates as a function of the number of feature candidates analyzed. **Right**: Bias of the $\hat{P}_e^L$ error estimate as a function of the Mahalanobis distance between the classes and the number of samples, when selecting 5 out of 200 feature candidates

– The traditional *Stepwise Forward Backwardselection (SFB)* gave the same results as the more advanced *Sequential Forward Floating Selection (SFFS).*

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 [10] the bias and variance of different error estimates have been analyzed in more detail and some of the main 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.

Acknowledgment

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

References

1. M. C. Constanza and A. A. Afifi. Comparison of stopping rules in forward stepwise discriminant analysis. *Journal of the American Statistical Association*, 74:777–785, 1979. 481
2. R. O Duda and P. E Hart. *Pattern classification and scene analysis.* A Wiley-interscience publication, first edition, 1973. 481
3. R. P. W. Duin. A matlab toolbox for pattern recognition. Technical Report Version 3.0, Delft University of Technology, 2000. 481
4. K. S. Fu, P. J. Min, and T. J. Li. Feature selection in pattern recognition. *IEEE Trans on Syst Science and Cybern - Part C*, 6(1):33–39, 1970. 480
5. A. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE Trans Pattern Anal Mach Intell*, 19(2):153–158, 1997. 480, 481
6. P. A. Lachenbruch and M. R. Mickey. Estimation of error rates in discriminant analysis. *Techometrics*, 10(1):1–11, 1968. 481
7. P. Pudil, J. Novovicova, and J. Kittler. Floating search methods in feature selection. *Pat Rec Let*, 15:1119–1125, 1994. 481
8. A. C. Rencher and S. F. Larson. Bias in Wilks' lambda in stepwise discriminant analysis. *Technometrics*, 22(3):349–356, 1980. 485
9. C. Rutter, V. Flack, and P. Lachenbruch. Bias in error rate estimates in discriminant analysis when setpwise variable selection is employed. *Commun. Stat., Simulation Comput*, 20(1):1–22, 1991. 485
10. H. Schulerud. The influence of feature selection on error estimates in linear discriminant analysis. *Submittet to Pattern Recognition.* 486
11. S. D. Stearns. On selecting features or pattern classifiers. *Proc. Third Intern. Conf. Pattern Recognition*, pages 71–75, 1976. 481