Stabilizing Classifiers for Very Small Sample Sizes

Marina Skurichina[1,2] and Robert P.W.Duin[1]

[1]Pattern Recognition Group, Faculty of Applied Physics, Delft University of Technology, P.O. Box 5046, 2600GA Delft, The Netherlands. E-mail: duin@ph.tn.tudelft.nl

[2]Dept. of Data Analysis, Inst. of Mathematics and Informatics, Akademijos 4, Vilnius 2600, Lithuania. E-mail: marina@ph.tn.tudelft.nl

Abstract

In this paper the possibilities for constructing linear classifiers are considered for very small sample sizes.

We propose a stability measure and present a study on the performance and stability of the following techniques: regularization by the Ridge-estimate of the covariance matrix [12], bootstrapping followed by aggregation ('bagging', [9]) and editing combined with pseudo-inversion [8]. It is shown that by these techniques a smooth transition can be made between the nearest mean classifier and the Fisher discriminant based on large samples sizes. Especially for highly correlated data very good results are obtained compared with the nearest mean method.

1. Introduction

In discriminant analysis one often runs into the small training sample size problem. Due to growing technical possibilities an increasing number of different measurements can be considered when studying real world object. As a result in many areas of application data sets may have a large numbers of features, e.g. 100-200 and more. The statistical dependency between single components of such data is often non-linear. It can thereby be essential to use additional polynomial features in order to evaluate the mentioned non-linear dependencies. In this way the dimensionality of the feature space becomes extremely large. If the number of objects that can be studied is still limited, large difficulties may arise in building a discriminant function [1], [2], [3].

Standard classical statistical methods require the inverse of the covariance matrix, e.g. in Fisher's Linear Discriminant function (FLD) [4], [5]:

$$g_F(\boldsymbol{x}) = \left[\boldsymbol{x} - \frac{1}{2}(\bar{\boldsymbol{X}}^{(1)} - \bar{\boldsymbol{X}}^{(2)})\right]' \boldsymbol{S}^{-1}(\bar{\boldsymbol{X}}^{(1)} - \bar{\boldsymbol{X}}^{(2)}) \qquad (1)$$

in which $\boldsymbol{S}$ is the standard maximum likelihood estimation of the pxp covariance matrix Σ, $\boldsymbol{x}$ is a p-variate vector to be classified, $\bar{\boldsymbol{X}}^{(i)}$ is the sample mean vector of the i-th class.

Recall that (1) is the mean square error solution for the linear coefficients ($\boldsymbol{w}$,$\boldsymbol{w}_0$) in

$$g_F(\boldsymbol{x}) = \boldsymbol{w} \bullet \boldsymbol{x} + \boldsymbol{w_0} = L \qquad (2)$$

with $\boldsymbol{x} \in \boldsymbol{X}$ and with L being the corresponding desired outcomes, 1 for class-1 and -1 for class-2. Direct computation is impossible when the number of features p exceeds the number of training vectors N [6]. For larger feature sizes decreasing to p the expected probability of misclassification rises dramatically [7]. There exist a number of different techniques to overcome this problem:

One is to reduce the number of features to $p < N$, using the knowledge of some experts or by the selection and extraction of the most informative features. A second approach is to modify the standard FLD in one way or another. It is the purpose of this paper to study a number of possibilities. The main ones are the Pseudo Fisher Linear Discriminant that studies the classifier in the subspace determined by the available objects [8] and the Nearest Mean Classifier (NMC) that ignores the covariances. We consider these and some modifications in the second section of current paper.

The third approach consists of a solution stabilization using different regularization techniques such as the ridge-estimate of the covariance matrix or bagging (bootstrapping and aggregating [9]). Often the optimal ridge-estimate of the covariance matrix or bagging can improve the generalization of a linear classifier. Sometimes these stabilization techniques don't help and, moreover, increase the generalization error. This is directly connected with the classifier stability on the particular data. The use and the performance of the ridge-estimate of the covariance matrix for FLD is considered in the third section. The use of bagging for the NMC is discussed in the fourth section. The stability of these two classifiers is experimentally investigated in the fifth section.

Conclusions can be found in the sixth section.

In our simulations we used following data:

The first data set consists of two correlated Gaussian classes with equal covariance matrices in a 30-dimensional feature space. For each class 200 vectors are given. The

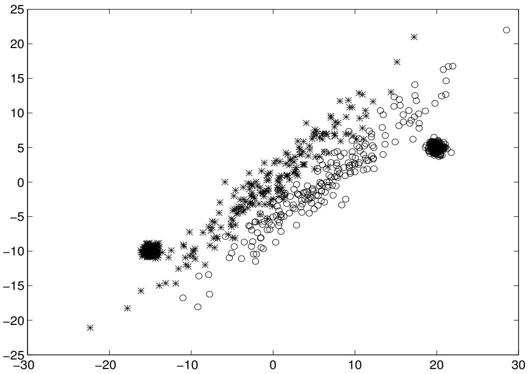

Fig.1. Scatter plot of a two-dimensional projecti of the 30-dimensional Non-Gaussian data.

first class has a zero mean for all features. The second class has a mean of 3 for the first two features and a mean of zero for all other ones. The common covariance matrix is a diagonal matrix with a variance of 40 for the second feature and a unit variance for all other features. The intrinsic class overlap is 0.064. This data set was rotated in order to spread the separability over all features. We will call this set further "Gaussian correlated data" (GC data).

The second data set is a mixture of the GC data described above and 200 additional 30-dimensional normal distributed vectors generated for each class with variances of 0.5 for all features and means of [-15,-10] and [20,5] for the first two features accordingly for the first and second classes and zero-means for all other features. The first two features are presented in Fig.1. Further it is called "Non-Gaussian data" (NG data).

Gaussian spherical data (GS data) consists of two classes of 30-dimensional normal distributed vectors (200 in each class) with common unit covariance matrix. The class means for all features are zero with except of the first feature that has the mean of 3.12 for the second class. The intrinsic class overlap is 0.064 the same as for the GC data.

The forth data set consists of real data. This is the sonar data set originally published by Gorman [10]. The way they were adapted is described in [11]. In total there are 208 objects in two classes, 111 from reflections of a metal cylinder, 97 from a rock. Each object is a 60-dimensional vector, representing the energy in 60 wave bands.

For the experiments training sets of 3 up to 100 (3 to 90 for the sonar data) objects per class were drawn randomly from the total set. All experiments were repeated 50 times. The averaged results over 50 repetitions are presented on all figures and we will not mention that further.

2. Linear Classifiers

One of the most simple modifications of the FLD is the Nearest Mean Classifier (NMC)

$$g_{NM}(\boldsymbol{x}) = \left[\boldsymbol{x} - \frac{1}{2}(\bar{\boldsymbol{X}}^{(1)} - \bar{\boldsymbol{X}}^{(2)})\right]'(\bar{\boldsymbol{X}}^{(1)} - \bar{\boldsymbol{X}}^{(2)}) \qquad (3)$$

Thus the NMC generates the perpendicular bisector between the class means and thereby yields the optimal linear classifier for classes with identical spherical normal distribution. The advantage of this classifier is that it is relatively insensitive to the number of training examples [2]. The NMC doesn't take into account, however, the differences in variances and covariances.

Another modification of the FLD (1), that allows to overcome the direct inverse of the ill-conditioned covariance matrix for the small training sample sizes ($N<p$), is the so-called Pseudo Fisher Linear Discriminant (PFLD), which we first studied in [8]. Here a direct solution of () is obtained by (using augmented vectors):

$$g_{PF}(\boldsymbol{x}) = (\boldsymbol{w}, \boldsymbol{w}_0) \bullet (\boldsymbol{x}, 1) = (\boldsymbol{x}, 1)(\boldsymbol{X}, I)^{(-1)}L \qquad (4)$$

in which $(\boldsymbol{x},1)$ is the augmented vector to be classified and $(\boldsymbol{X},\boldsymbol{I})$ is the augmented training set. The inverse $(\boldsymbol{X},\boldsymbol{I})^{(-1)}$ is the Moore-Penrose Pseudo Inverse which gives the minimum norm solution. Before the inversion the data is shifted such that it has zero mean. This method is closely related to singular value decomposition.

For values $N \geq p$ PFLD, maximizing the distance to all given samples, is identical to the FLD (1). However, for values $N<p$, the PFLD finds a linear subspace, where all the given data samples are located, and on this plane estimates the data means and the covariance matrix and builds a linear discriminant, that is perpendicular to this subspace in all other directions for which no samples are given. The behavior of the PFLD as a function of the sample size was illustrated in [8] (see also Fig. 2, 3 and 4).

The constant distance to all training samples found by the PFLD might be enlarged by editing the training set, allowing some samples to have larger distances. The resulting procedure, which we have called the Small Sample Size Classifier (SSSC) [8] is closely related to

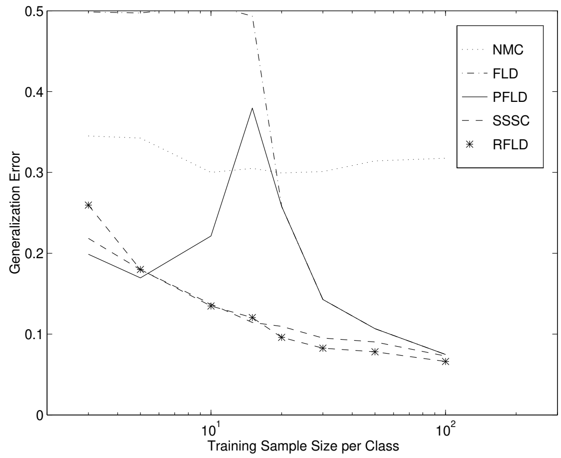

Fig.2. The generalization error of the FLD, the NMC, the PFLD, the SSSC and the RFLD, optimized on leave-one-out cross-validation, versus the training sample size for 30-dimensional GC data.

Vapnik's support vector classifier [ ]. In all figures we present the "bagged" version of the SSSC. The definition of bagging is given later in section 4.

In Fig. 2, 3 and 4 the behavior is presented for the 30-dimensional GC, NG data and the 60-dimensional non-Gaussian distributed sonar data of the standard FLD, the NMC, the PFLD and the SSSC as functions of the training sample size.

For the very small sample sizes the PFLD, regularization (RFLD) and the SSSC, including bagging, all outperform FLD. In this case the SSSC is slightly better than regularization, probably due to the fact that it is neither based on normal assumptions, nor it makes use of global statistics like the covariance matrix.

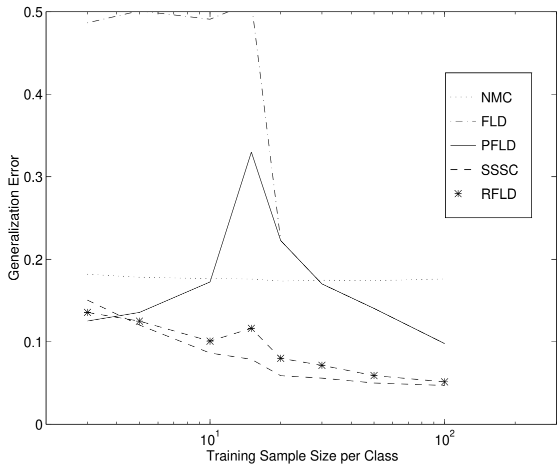

Fig.3. The generalization error of the FLD, the NMC, the PFLD, the SSSC and the RFLD, optimized on leave-one-out cross-validation, versus the training sample size for 30-dimensional NG data.

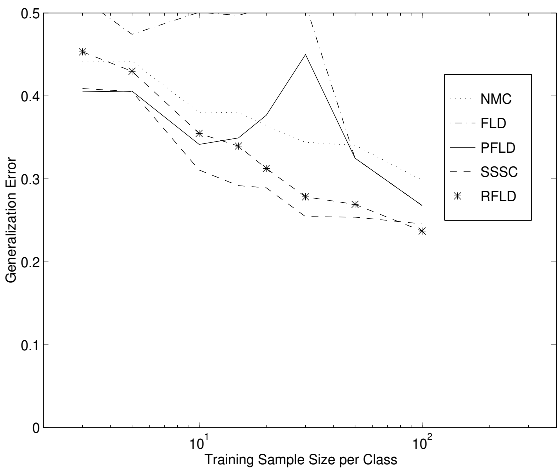

Fig.4. The generalization error of the FLD, the NMC, the PFLD, the SSSC and the RFLD, optimized on leave-one-out cross-validation, versus the training sample size for 60-dimensional sonar data.

3. Ridge-estimate of Covariance Matrix

The well-known technique to overcome the inverse of an ill-conditioned covariance matrix in building the standard FLD (1) is to add some constant values to diagonal elements of the estimate of the covariance matrix

$$\boldsymbol{S}_R = \boldsymbol{S} + \lambda \boldsymbol{I} \tag{5}$$

in which **I** is the *p*x*p* identity matrix, λ is a regularization parameter.

The new estimation $\boldsymbol{S}_R$ is called the ridge-estimate of the covariance matrix, which conception we have borrowed from regression analysis [12]. This approach is called regularized discriminant analysis. The modification (5) gives us the ridge or regularized discriminant function

$$g_R(\boldsymbol{x}) = \left[\boldsymbol{x} - \frac{1}{2}(\overline{\boldsymbol{X}}^{(1)} - \overline{\boldsymbol{X}}^{(2)})\right]'(\boldsymbol{S} + \lambda \boldsymbol{I})^{-1}(\overline{\boldsymbol{X}}^{(1)} - \overline{\boldsymbol{X}}^{(2)}) \tag{6}$$

It's evident that when $\lambda \to \infty$, we lose the values of covariances. In that case the classifier () approaches the

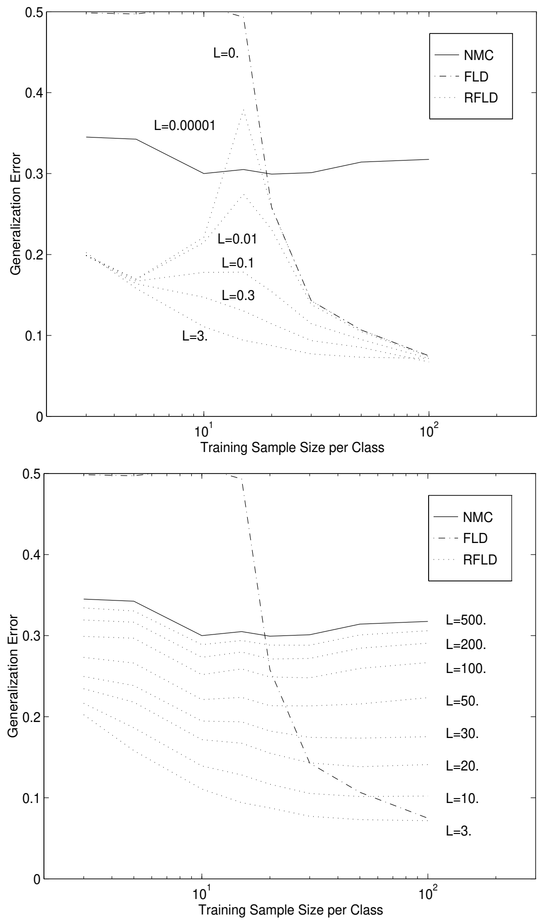

Fig.5a,b. The generalization error of the NMC and the RFLD with different values of the regularization parameter λ=L versus the training sample size for the 30-dimensional GC data.

NMC (6) (See Fig.5). Simultaneously the generalization error could essentially increase. Small values of the regularization parameter λ could be useful for stabilizing the solution. Very small values of λ may be not sufficiently effective. Usually two methods to choose the good value of the regularization parameter are used: cross-validation [13] and bootstrapping [14]. They both need much computation time and they are not enough precise for the small training sample size. See further Friedman [15], Raudys and Skurichina [16].

We investigated the use of the ridge-estimate of the covariance matrix in the FLD for the GC data. The dependence of the RFLD generalization error on the value of the regularization parameter λ is illustrated in Fig. 5a,b. The RFLD with λ=0 is identical to the standard FLD, the RFLD with λ=0.00001 appears to be similar to the PFLD.

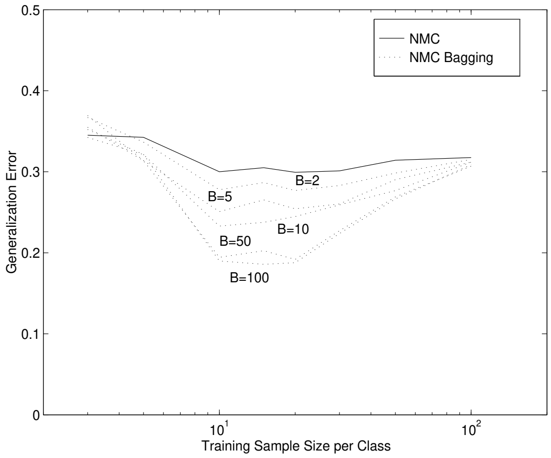

Fig.6. The generalization error of the NMC and the "bagged" NMC for different numbers of bootstraps B versus the training sample size for 30-dimensional GC data.

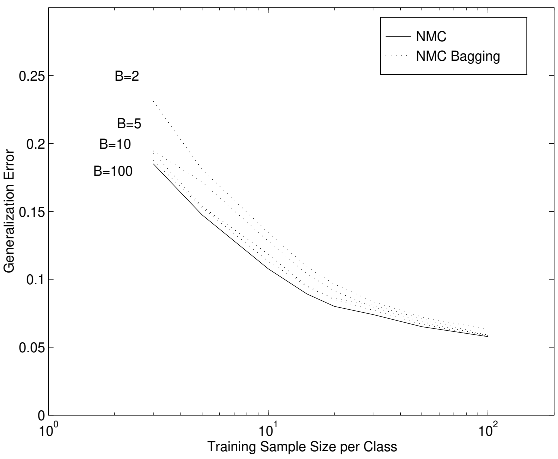

Fig.7. The generalization error of the NMC and the "bagged" NMC for different numbers of bootstraps B versus the training sample size for 30-dimensional GS data.

We can improve the generalization error increasing the value of λ till 3.0. At first, we have a peaking effect for p=N=30 that gradually disappears. The further increase of λ results in the generalization error increase, that approaches the generalization error of the NMC. In Fig.2, 3 and 4 the RFLD with the optimal regularization parameter, estimated by leave-one-out cross-validation, is presented. We see that for 30-dimensional GC and NG data the optimized RFLD reasonably improves the performance of the standard FLD. This also holds for the PFLD and the NMC in the case of the small training sample size. For 60-dimensional sonar data the optimized RFLD also helps to overcome the small sample size problem increasing the stability of the FLD.

4. Bagging

Another stabilizing technique is bagging, based on bootstrapping [17] and aggregating concepts as proposed by Breiman [9]. Bagging of the classifier was implemented by us by averaging the parameters of the same linear classifier built on several bootstrap replicates. The random selection with replacement of *N* samples from the set of *N* samples is called a bootstrap replicate. From each bootstrap replicate of the dataset a 'bootstrap' version of the classifier is built. The average of these 'bootstrap' versions gives us the 'bagged' classifier. Breiman has shown that bagging could reduce the error of linear regression and classification. It was noticed that bagging is useful for unstable procedures only. For stable ones it could even deteriorate the performance of the classifier.

We investigated the effect of bagging of the NMC for the GC, the GS and the sonar data (Fig. 6, 7 and 8). We see that bagging gives a reasonable improvement of the NMC generalization only for the GC data. It gives a small

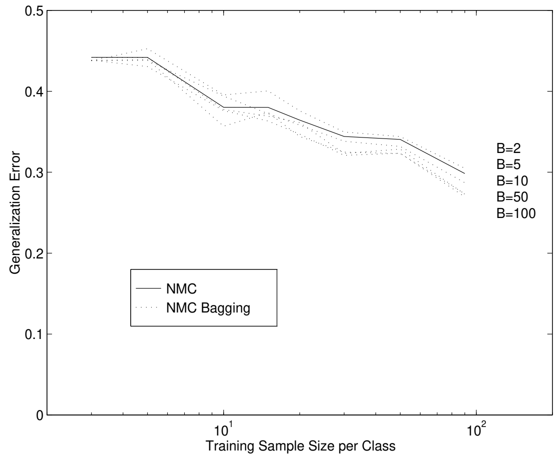

Fig.8. The generalization error of the NMC and the "bagged" NMC for different numbers of bootstraps B versus the training sample size for 60-dimensional sonar data.

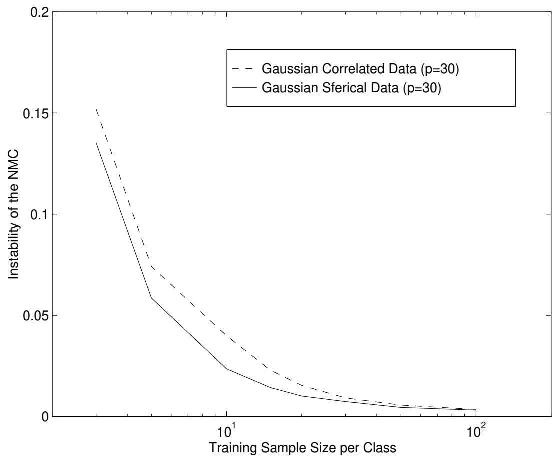

Fig.9. The NMC instability versus the training sample size for 30-dimensional GC and GS data with the same overlap of classes.

improvement for the sonar data and it changes to the worse for the GS data. This might be explained by considering the stability of the NMC. The NMC is optimal for the GS data and therefore it is stable for this data. We can hardly improve the performance of the NMC by bagging. It is even deteriorated when the number of bootstraps is not enough large. The NMC on GC data is not stable when the number of training vectors is comparable with the data dimensionality. That's why bagging can improve the generalization for this sample size, see Fig. 6. For the very small sample sizes the discriminants are very unstable and even bagging does not help. For large sample sizes the NMC is increasingly stable and bagging does not help anymore. The sonar data has a more complicated structure than the two data sets discussed above. However, the main conclusions are correct for this data as well. As the NMC is not optimal for this data we can improve its generalization error by bagging. We notice that different data sets require different numbers of bootstraps to reach the minimal generalization error of the classifier.

So the question of the stability or the instability of the classifier on particular data is very important. Measuring in some way the instability of the classifier on some data the use of bagging or the use of the ridge-estimate of the covariance matrix can be predicted for linear classifiers.

5. Stability of linear classifiers

Let us consider the instability of the NMC for examples presented in Fig. 6, 7, and 8. We measure the instability of the classifier, calculating the changes in the test data set classification after removing an arbitrary object from learning data set. Repeating this procedure several times and averaging the fraction of the objects for which the classification changes we get an estimation of the classifier instability. (This estimation could be shifted if we measure the instability differently.) Measured in this way and averaged over 50 repetitions the instability of the NMC is presented in Fig. 9 for the GC data and the GS data with the same overlap for the classes. This shows a nice illustration of the above conclusions. The NMC is not a stable procedure when the number of training vectors is much smaller than the data dimensionality. Its stability grows with the increase of the number of training vectors. As the NMC is the optimal classifier for the GS data, so we can consider the NMC instability on this data as some instability example that we can use for comparison with the NMC instability on other data. We see that the NMC on the GC data is more unstable (see Fig.9). Thus bagging of the NMC on this data is useful.

For sonar data we can't compare the NMC instability with some instability example as we did for Gaussian data, because we don't know the type of the sonar data set distribution.

Now let us compare for GC data the instability of the

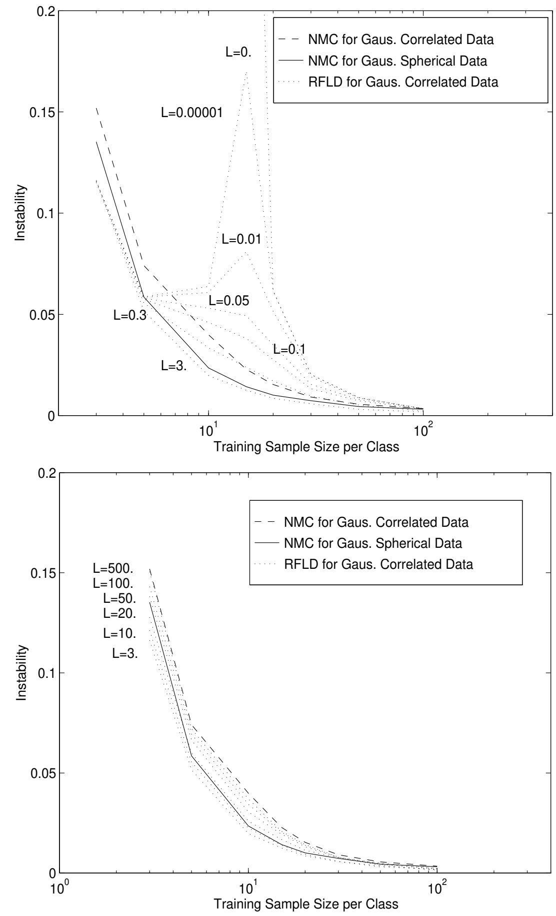

Fig.10a,b. The instability of the RFLD with different values of the regularization parameter λ=L for 30-dimensional GC data and the NMC instability for 30-dimensional GC and GS data versus the training sample size.

NMC and the RFLD with different values of the regularization parameter. Graphs presented on Fig.10a,b show that the RFLD instability depends very much on the value of the regularization parameter. We see that the RFLD instability behaves in a similar way as its generalization error (see Fig.5a,b). While the value of the regularization parameter increases till the optimal value $\lambda=3$, the RFLD instability decreases till its minimum that is even less than the NMC instability example obtained on GS data. For a further increase of λ the RFLD instability starts to grow and reaches the NMC instability obtained on the same GC data. For small values of the regularization parameter we have a peak for the number of training vectors equal to the number of features. Here the RFLD is very unstable due to the ill-conditioned covariance matrix. Only larger values of the regularization parameter could help to overcome this problem. The RFLD instability especially depends on the value of regularization parameter when the number of training samples N is comparable with the data dimensionality p.

6. Conclusions

In this paper some solutions are studied for the very small sample size problem: training set sizes that are below the feature size. Two methods that can always be used are the NMC and the Nearest Neighbor Classifier (NNC). The NMC does not use any information on the variability or distribution of the data. For the NNC it is known that it has bad small sample size properties. It is not considered in this study. In [8] we already showed that the SSSC might be much better for very small sample sizes than the NNC. Here we concentrated on various possibilities to stabilize FLD, which cannot be used in its vanilla version for the considered sample sizes due to the singularity of the covariance matrix.

It was shown that the PFLD based on a pseudo-inverse technique might give good results but shows a peak in the error for sample sizes that are about the feature size. The SSSC, which might be described as the application of the PFLD on an edited training set, removes this peak and improves the performance globally.

Next we studied a regularization of the FLD using the Ridge-estimate of the covariance matrix. It was shown that an automatic procedure that optimizes the regularization parameter using a leave-one-out technique might give very good results for closely normally distributed data.

As classifiers become instable for small sample sizes, a general stabilizing technique like bagging may yield improved classifiers.

Finally we introduced a stability measure and showed experimentally its relation with the behavior and performance of classifiers for small sample sizes. This measure may be used to judge whether stabilizing techniques like regularization and bagging may be useful.

References

1. A.K.Jain and B.Chandrasekaran, Dimensionality and Sample Size Considerations in Pattern Recognition Practice, in: Handbook of Statistics, vol. 2, ed.P.R.Krishnaiah and L.N.Kanal, pp. 835-855, North-Holland, Amsterdam, 1987.
2. S.Raudys. Statistical Pattern Recognition. Small Design Sample Problems, Institute of Mathematics and Cybernetics, Vilnius, 1984, pp. 1-460. A manuscript.
3. S.Raudys and A.K.Jain, Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.13, no.3, 1991, pp.252-264.
4. R.A.Fisher, The Use of Multiple Measurements in Tacomic Problems. Annals of Eugenics, 1936, vol. 7, no. 2.
5. R.A.Fisher, The Precision of Discriminant Functions. Annals of Eugenics, 1940, vol.10, no. 4.
6. C.R.Rao, On Some Problems Arising of Discrimination with Multiple Characters, Sankya, 1949, vol. 9, pp. 343-365.
7. S.Raudys and V.Pikelis, On Dimensionality, Sample Size, Clas-sification Error and Complexity of Classification Algorithm in Pattern Recognition, IEEE Transaction on Pattern Analysis and Machine Intelligence, 1980, vol. PAMI-2, no. 3, pp.242-252.
8. R.P.W.Duin, Small Sample Size Generalization, Proceedings of 9th Scandinavian Conference on Image Analysis, June 6-9, 1995, Uppsala, Sweden.
9. P.Gorman and T.J.Sejnowski, Learned Classification of Sonar Targets Using a Massively Parallel Network, IEEE Transactions on ASSP, 1988, vol. 36, no. 7.
10. L.Breiman, Bagging Predictors, Technical Report No. 421, 1994, University of California, Berkeley, California, USA.
11. W.F.Schmidt, D.F.Levelt, and R.P.W.Duin, An Experimental Comparison of Neural Classifiers with Traditional Classifiers, in: E.S.Gelsema, L.N.Kanal, Pattern Recognition in Practice IV, Multiple Paradigms, Comparative Studies and Hybrid Systems (Proc. Conf. Vlieland NL, June 1-3, 1994), Elsevier, Amsterdam, 1994, pp. 391-402.
12 S.A.Aivazian, V.M.Buchstaber, I.S.Yenyukov, L.D.Mechalkin. Applied Statistics. Classification and reduction of dimensionality. Reference Edition, Finansy i Statistika, Moscow, 1989.
13. P.A.Lachenbruch, Discriminant Analysis. Hafner Press, New York.
14. R.Peck and J.Van Ness, The Use of Shrinkage Estimators in Linear Discriminant Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, 1982, no. 4, pp. 530-537.
15. J.M.Friedman, Regularized Discriminant Analysis, JASA 84, 1989, pp. 165-175.
16. S.Raudys and M.Skurikhina, Small Sample Properties of Ridge-Estimate of Covariance Matrix in Statistical and Neural Net Classification, New Trends in Probability and Statistics, vol. 3, Tartu, Estonia, 1994, pp. 237-245.
17. B.Efron and R.Tibshirani, An Introduction to the Bootstrap, Chapman and Hall, 1993.