Regularization by Adding Redundant Features

Marina Skurichina and Robert P.W.Duin

Pattern Recognition Group, Department of Applied Physics, Faculty of Applied Sciences
Delft University of Technology, P.O. Box 5046, 2600GA Delft, The Netherlands.
Phone: +(31) 15 2786143, FAX: +(31) 15 2786740, E-mail: duin@ph.tn.tudelft.nl

Abstract

The Pseudo Fisher Linear Discriminant (PFLD) based on a pseudo-inverse technique shows a peaking behaviour of the generalization error for training sample sizes that are about the feature size: with an increase in the training sample size the generalization error at first decreases reaching the minimum, then increases reaching the maximum at the point where the training sample size is equal to the data dimensionality and afterwards begins again to decrease. A number of ways exist to solve this problem. In this paper it is shown that noise injection by adding redundant features to the data also helps to improve the generalization error of this classifier for critical training sample sizes.

Keywords: Pseudo Fisher linear discriminant, critical sample size, generalization error, peaking behaviour, noise injection.

1 Introduction

The main problem in building statistical parametric classifiers on small training sets is that they require the inverse of the covariance matrix, which is impossible to perform when the number of training objects *N* is less than the data dimensionality *p*. One of the ways to overcome the small sample size problem is to modify the standard classifiers in one way or another. However, even modified classifiers, such as the Pseudo-Fisher linear discriminant (PFLD) [1], may become very unstable and have a peaking effect of the generalization error when the training sample size is comparable with the data dimensionality [2, 3, 4].

The following ways are studied to solve this problem:

1. Removing features (decreasing *p*) by some feature selection method.
2. Adding objects (increasing *N*), either by using larger training sets, or, if it is not possible by generating additional objects (noise injection [5]).
3. Removing objects (decreasing *N*) brings the classifier out of the instable region. This method has been studied by us [2, 3] and is effectively being used in the Support Vector Classifier [13].

In this paper we will show that the fourth way is also effective:

4. Adding redundant features (increasing p). Like the third method this brings the classifier out of the instable region but now by enlarging the dimensionality by noise.

In this paper we concentrate on the injection of noise by adding redundant features to the data and its effect on the performance of the Pseudo Fisher linear discriminant. The data used in our simulation study are presented in section 2. The Pseudo Fisher linear discriminant is discussed in section 3. The use and the performance of noise injection in the data feature space is considered in section 4. Conclusions and discussion could be found in section 5.

2 Data

Two artificial data sets and one real data set are used for our experimental investigations. These data sets have a high dimension because we are interested in critical situations where the PFLD has a bad performance.

The first set is a 30-dimensional correlated Gaussian data set constituted by two classes with equal covariance matrices. Each class consists of 500 vectors. The mean of the first class is zero for all features. The mean of the second class is equal to 3 for the first two features and equal to 0 for all other features. 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 (Bayes error) is 0.064. In order to spread the separability over all features, this data set is rotated using a 30×30 rotation matrix which is $\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ for the first two features and the identity matrix for all other features. We call these data further “Gaussian correlated data”. Its first two features are presented in Fig. 1.

The second data set consists of two 30-dimensional Gaussian distributed data classes with unequal covariance matrices. Each data class contains 500 vectors. The first data class is distributed spherically with the unit covariance matrix and the zero mean. The mean of the second class is equal to 4.5 for the first feature and equal to 0 for all other features. The covariance matrix of the second class is a diagonal matrix with a variance of 3 for the first two features and a unit variance for all other features. We call these data further “Gaussian spherical data with unequal covariance matrices”. Its first two features are presented in Fig. 2.

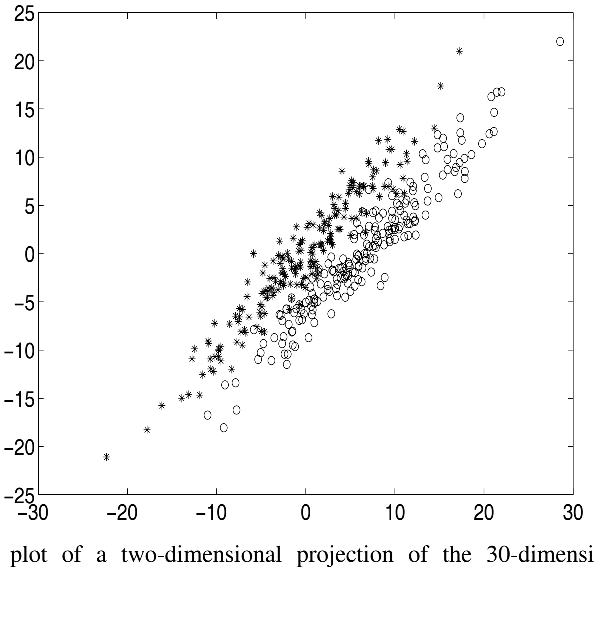

Fig. 1. Scatter plot of a two-dimensional projection of the 30-dimensional Gaussian correlated data.

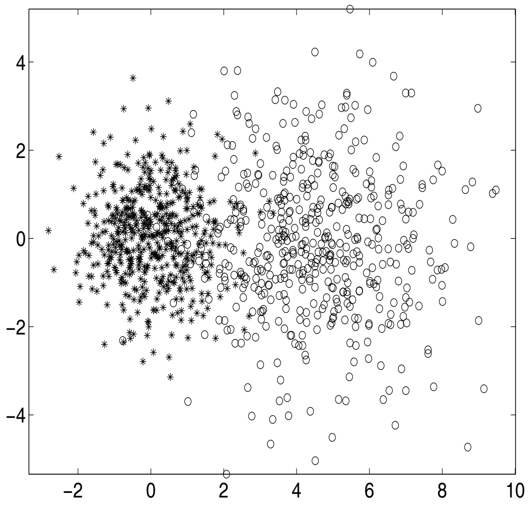

Fig. 2. Scatter plot of a two-dimensional projection of the 30-dimensional Gaussian spherical data with unequal covariance matrices.

The last data set consists of real data collected through spot counting in interphase cell nuclei (see, for instance, Netten *et al* [6] and Hoekstra *et al* [7]). Spot counting is a technique to detect numerical chromosome abnormalities. By counting the number of coloured chromosomes ('spots'), it is possible to detect whether the cell has an aberration that indicates a serious disease. A FISH (Fluorescence In Situ Hybridization) specimen of cell nuclei was scanned using a fluorescence microscope system, resulting in computer images of the single cell nuclei. From these single cell images 16×16 pixel regions of interest were selected. These regions contain either background spots (noise), single spots or touching spots. From these regions we constructed two classes of data: the noisy background and single spots, omitting the regions with touching spots. The samples of size 16×16 were considered as a feature vector of size 256. The first class of data (the noisy background) consists of 575 256-dimensional vectors and the second class (single spots) - of 571 256-dimensional vectors. We call these data "cell data" in the experiments.

Training data sets with 3 to 200 (with 3 to 300 for cell data) samples per class are chosen randomly from a total set. The remaining data are used for testing. These and all other experiments are repeated 10 times for independent training sample sets. In all figures the averaged results over 10 repetitions are presented and we do not mention that further.

3 The Pseudo Fisher Linear Discriminant

The most popular and commonly used linear classifier is the Fisher Linear Discriminant (FLD) [8, 9]:

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

where $\boldsymbol{S}$ is the standard maximum likelihood estimation of the $p \times p$ common covariance matrix Σ, $\boldsymbol{x}$ is a p-variate vector to be classified and $\bar{\boldsymbol{X}}^{(i)}$ is the sample mean vector of the i-th class, $i=1,2$.

Notice that (1) is the mean squared error solution for the linear coefficients $(\boldsymbol{w}, w_0)$ in

$$g_F(\boldsymbol{x}) = \boldsymbol{w} \bullet \boldsymbol{x} + w_{\boldsymbol{0}} = L \tag{2}$$

with $\boldsymbol{x} \in \boldsymbol{X}$ and with L being the corresponding desired outcomes, 1 for class-1 and -1 for class-2. When the number of data features p exceeds the total number of training vectors N, the estimate matrix $\boldsymbol{S}$ becomes singular and the direct inverse becomes impossible [10]. For increasing feature sizes the expected probability of misclassification rises dramatically [11].

The modification of the FLD, which allows to avoid the inverse of ill-conditioned covariance matrix, is the so-called Pseudo Fisher linear discriminant [1]. In the PFLD a direct solution of (2) is obtained by (using augmented vectors):

$$g_{PF}(\boldsymbol{x}) = (\boldsymbol{w}, \boldsymbol{w_0}) \bullet (\boldsymbol{x}, 1) = (\boldsymbol{x}, 1)(\boldsymbol{X}, \boldsymbol{I})^{-1} L, \tag{3}$$

where $(\boldsymbol{x},1)$ 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 are shifted such that they have zero mean. This method is closely related to singular value decomposition.

For values $N \geq p$ the PFLD, maximizing the distance to all given samples, is equivalent to the FLD (1). For values $N < p$, however, the Pseudo Fisher rule finds a linear subspace, which covers all the data samples. On this plane the PFLD estimates the data means and the covariance matrix, and builds a linear discriminant perpendicular to this subspace in all other directions for which no samples are given.

The behaviour of the PFLD as a function of the sample size is illustrated in [2, 4]. For one sample per class this method is equivalent to the Nearest Mean and to the Nearest Neighbour method. If the total sample size is equal to or larger than the dimensionality $N \geq p$, the method is equivalent to the FLD. It was noticed that the generalization error of the PFLD shows a peaking behaviour: with an increase in the training sample size the generalization error at first decreases reaching a local minimum somewhere below the point $N=p$, then increases reaching a maximum at the point $N=p$, where the training sample size is equal to the data dimensionality, and afterwards begins again to decrease (e.g., Fig. 3). This can be understood from the observation that the PFLD succeeds in finding hyperplanes with equal distances to all training samples until $N=p$. In [12] an asymptotic expression for the generalization error of the PFLD is derived which explains theoretically the such behaviour of the PFLD.

4 Noise Injection by Adding Redundant Features to the Data

In order to improve the generalization error of the PFLD for critical values of the training sample size (*N*=*p*), the number of techniques could be used.

One of the ways to solve this problem involves generating more training objects by noise injection to the training data. Usually, spherical Gaussian distributed noise is generated around each training object. However, this method requires to know the optimal variance of noise in order to get good results. The optimal value of the noise variance depends on many factors such as the training sample size, the data dimensionality and the data distribution [5]. It could vary dramatically for different data. As a rule, to find the optimal value of the noise variance is not an easy task and it goes on a long time.

To demonstrate the influence of the noise variance λ on the generalization error of the PFLD we considered the 30-dimensional Gaussian correlated data. The averaged results for some values of λ are presented in Fig. 3. We see that the performance of the PFLD strongly depends on the variance of the noise.

Considering small sample size properties (a learning curve) of the PFLD, one can reach another solution: decrease the number of training objects in order to avoid the critical training sample size problem. It could be also performed by noise injection in the data feature space instead of adding noise to the training objects. In this case the data dimensionality is enlarged by adding Gaussian distributed features with zero mean and variance of one. When increasing the data dimensionality *p* the training sample size *N* relatively decreases leaving a critical area *N*=*p*, where the PFLD has a high generalization error. For values $N < p$ the PFLD performs much better than for the critical sizes of the training set.

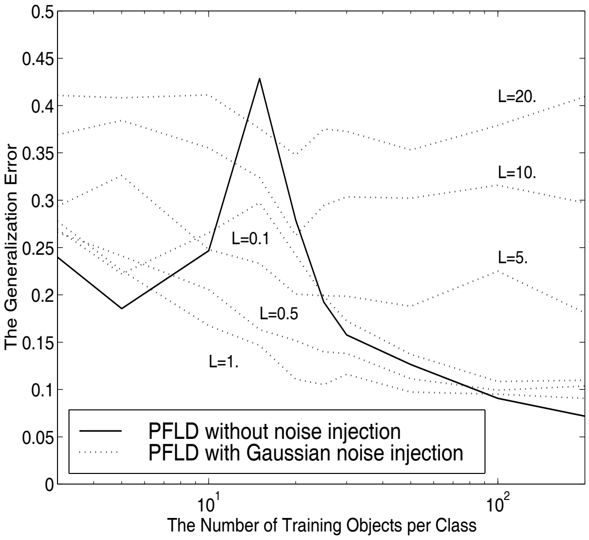

Fig. 3. The generalization error of the PFLD without and with noise injection to the training objects with different values of the noise variance λ=L versus the training sample size for 30-dimensional Gaussian correlated data.

Let us now investigate this approach for 3 examples of data described in section 2. In order to study the influence of injection of "noisy" features to the data, the additional redundant "noisy" features having Gaussian distribution with zero mean and variance of one were generated. The generalization error of the PFLD for 30-dimensional Gaussian correlated data and 30-dimensional Gaussian spherical data with unequal covariance matrices without noise injection in the feature space and with 20, 70 and 170 additional redundant "noisy" features is presented in Fig. 4 and Fig. 5, respectively. The generalization error of the PFLD obtained on cell data without noise injection in the feature space and on the cell data with 44, 100, 144 and 200 redundant "noisy" features is presented in Fig. 6.

For all data the PFLD shows a critical behaviour with a high maximum of the generalization error around critical training sample size *N*=*p*. Figures 4, 5 and 6 nicely demonstrate that noise injection in the data feature space helps to avoid the peaking effect of the generalization error of the PFLD. We see that redoubling of the data dimensionality by adding "noisy" features already twice improves the performance of the classifier at the point *N*=*p*. For cell data it was enough to add 44-100 "noisy" features for the same improvement. When the number of added "noisy" features was 4-5 times larger than the original dimensionality of the data, the peak of the generalization error was smoothed almost completely: the generalization error was reduced in a whole region around the critical training sample size. Adding redundant features is useless, however, for very small training sample sets. Adding noise to a highly dimensional feature space with only a few objects makes the training data set too "noisy" to represent the entire data set correctly. In this case it becomes difficult or even impossible to build a good discriminant function. All considered data nicely demonstrate that the more noise

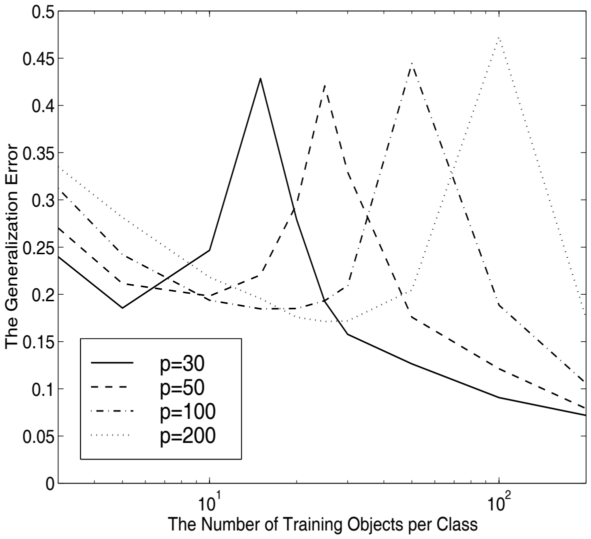

Fig. 4. The generalization error of the PFLD versus the training sample size for Gaussian correlated data without noise injection in the feature space (p=30) and with 20, 70, 170 additional redundant features (p=50, 100, 200).

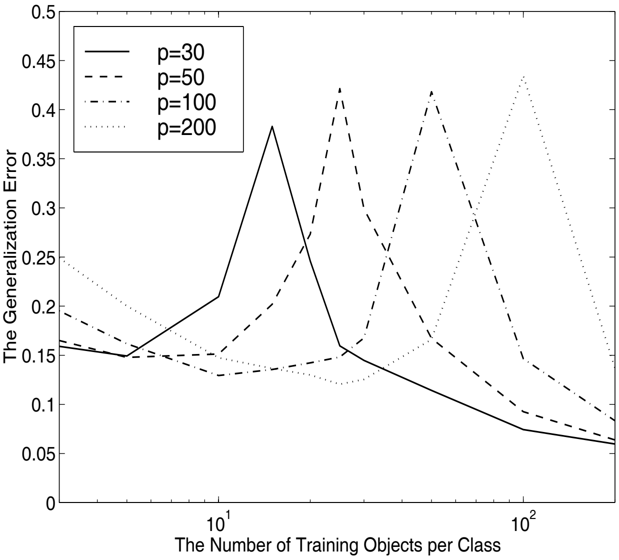

Fig. 5. The generalization error of the PFLD versus the training sample size for Gaussian spherical data with unequal covariance matrices without noise injection in the feature space (p=30) and with 20, 70, 170 additional redundant "noisy" features (p=50, 100, 200).

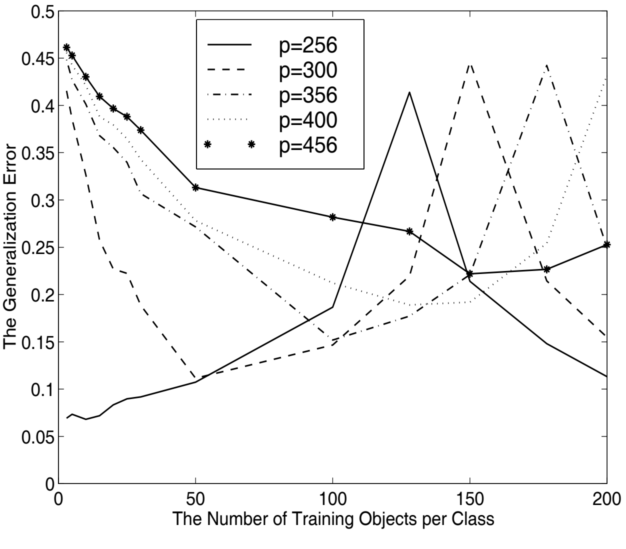

Fig. 6. The generalization error of the PFLD versus the training sample size for 256-dimensional cell data without noise injection (p=256) and with 44, 100, 144 and 200 additional redundant "noisy" features (p=300, 356, 400, 456).

is added to the data by adding redundant features the larger generalization error is obtained in the case of very small training sample sizes. For critical training data sizes adding redundant features helps to avoid the peaking effect of the generalization error of the PFLD.

However, one can notice that the improvement obtained in the generalization error also depends on the number of additional “noisy” features used for each data set. Obviously, this question requires to be investigated in future. Nevertheless, our simulation study completely proved the possible usefulness of noise injection in the data feature space in order to reduce the generalization error of the PFLD for critical training sample sizes.

5 Conclusions and Discussion

The PFLD might have a peaking behaviour of the generalization error for training sample sizes that are about the feature size. Based on the small sample size properties of the PFLD in this paper it was suggested to inject noise to the data feature space in order to improve the generalization error of the PFLD for critical training sample sizes. This approach was studied for two artificial data sets and one example of real data. Simulation results have shown that adding redundant “noisy” features to the data allows to reduce dramatically the generalization error of the PFLD in the region of critical training sample sizes.

Finally we make the following suggestion for future research. It has been observed previously [5] that the use of artificially generated normally distributed data is equivalent to regularizing the covariance matrix $(\Sigma + \lambda \boldsymbol{I})$ in case of the FLD. A similar type of regularization, but now on the inner product matrix $(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})$ might be equivalent to the stabilizing of the PFLD by the generation of redundant features discussed in this paper. This demands for a more thorough mathematical analysis than possible in this paper.

Acknowledgement

This work was supported by the Foundation for Applied Sciences (STW) and the Dutch Organization for Scientific Research.

Reference

1. K. Fukunaga, Introduction to Statistical Pattern Recognition. Academic Press, 400-407 (1990).
2. R.P.W. Duin, Small sample size generalization, *Proceedings of 9th Scandinavian Conference on Image Analysis*, Uppsala, Sweden, 957-964 (1995).
3. M. Skurichina and R.P.W. Duin, Stabilizing classifiers for very small sample sizes, *Proceedings of ICPR*, Vienna, Austria, 891-896 (1996).
4. M. Skurichina and R.P.W. Duin, Bagging for Linear Classifiers, *Pattern Recognition*, vol. 31, no. 7 (1998), in press.
5. Š. Raudys, M. Skurichina, T. Cibas and P. Gallinari, Optimal Regularization of Neural Networks and Ridge Estimates of the Covariance Matrix in Statistical

Classification, In: *Pattern Recognition and Image Analysis: Advances in Mathematical Theory and Applications* (an Int. Journal of Russian Academy of Sciences), Vol. 5, No. 4, 1995, pp. 633-650.

6. H. Netten, I.T. Young, M. Prins, L.J. van Vliet, H.J. Tanke, J. Vrolijk, W. Sloos, Automation of Fluorescent dot counting in cell nuclei, *Proceedings of the 12th Int. Conference on Pattern Recognition,* Vol.1, Jerusalem, 84-87 (1994).
7. A. Hoekstra, H. Netten and D. de Ridder, A neural network applied to spot counting, *Proceedings of ACSI'96, the Second Annual Conference of the Advanced School for Computing and Imaging*, Lommel, Belgium, 224-229 (1996).
8. R.A. Fisher, The Use of multiple measurements in taxonomic problems, *Annals of Eugenics* **7**, no. 2 (1936).
9. R.A. Fisher, The precision of discriminant functions, *Annals of Eugenics* **10**, no. 4 (1940).
10. R. Rao, On some problems arising of discrimination with multiple characters, *Sankya* **9**, 343-365 (1949).
11. Š. Raudys and V. Pikelis, On dimensionality, sample size, classification error and complexity of classification algorithm in pattern recognition, *IEEE Transaction on Pattern Analysis and Machine Intelligence* **PAMI-2**, no. 3, 242-252 (1980).
12. Š. Raudys and R.P.W. Duin, On expected classification error of the Fisher linear classifier with pseudo-inverse covariance matrix, *Pattern Recognition Letters* (1998), in press.
13. C. Cortes and V. Vapnik, Support-vector networks, *Machine Learning*, Vol. 20, No. 3, pp. 273-297 (1995).