Small sample size generalization

Robert P.W. Duin
Pattern Recognition Group, Faculty of Applied Physics
Delft University of Technology, P.O. Box 5046
2600 GA Delft, The Netherlands
email: duin@ph.tn.tudelft.nl

Abstract

The generalization of linear classifiers is considered for training sample sizes smaller than the feature size. It is shown that there exists a good linear classifier, that is better than the Nearest Mean classifier for sample sizes for which Fisher's linear discriminant cannot be used. The use and performance of this small sample size classifier is illustrated by some examples.

Keywords: linear discriminants, classification error, small sample size

1. Introduction

It is a common tradition in pattern recognition that the direct computation of discriminant functions or classifiers is avoided for training sets that are not sufficiently large compared with the number of features. Generally one tries to reduce the number of features k such that the size of the training set n is a multiple of k: $n = \alpha\, k$. Typical values for α are 2, 5 or 10.

In this paper it is first shortly recapitulated why such large sample sizes are needed for most families of classifiers. Following it is argued that there is a need for classifiers operating in the range below $\alpha = 1$. In section 2 a linear classifier for this range is proposed. Its properties are illustrated by some examples in section 3, concluded with a discussion in section 4.

For a general discussion on the small sample size problem we refer to Jain and Chandrasekaran [1] and to Raudys and Jain [2]. Here it will be shortly discussed why the classic families of classifiers need sufficiently large training sets. See also Fukunaga [4].

Parametric classifiers based on the Bayes rule. Here one tries to estimate from the training set the class probability density functions assuming a parameterized model. The training set is used for estimating the parameters. For one of the most simple classifiers in this family, the *Fisher Linear Discriminant,* it is still necessary to estimate and invert the average class covariance matrix. This is only possible for $n > k$. In order to avoid numerical problems related to the matrix inversion, the sample size should be largely above this bound. For an accurate solution close to the Bayes error, which is not the topic of this paper, even much larger sample sizes are needed.

Nonparametric classifiers based on the Bayes rule. Now the class density functions are estimated in a nonparametric way, e.g. the *K-Nearest Neighbor Rule* or the *Parzen Window Classifier*. These methods generally need larger training sets than the parametric ones as the knowledge on the density model lacks. One can also understand from a geometrical interpretation that class sample size smaller than the dimensionality (the feature size) are insufficient. In these cases the set of samples of a single class are situated in a linear subspace of the feature space. One cannot expect to get any sensible density estimation from that. On the other side however, from a computational point it is possible to use these methods for any sample size. Therefore, we will use in our examples the 1-Nearest Neighbor rule as a reference.

Empirical error minimizing classifiers. This family is not based on the estimation of class density functions and the Bayes rule. Here directly a parameterized discriminant function is optimized by minimizing the empirical error on the training set. Perceptrons and multilayer perceptrons belong to this family, but also feed-forward neural networks using the MSE as an error measure instead of the frequency of misclassifications. A first analysis of the generalization possibilities for this family can be found by Cover [5], who showed that for $n <= k + 1$ always a zero error linear classifier can be found and that for a randomly labeled training set with $n <= 2k$ in at least 50% of the cases a zero error linear classifier exists. For this reason a linear classifier with $n <= 2k$ is said to have no generalization capacity. By the more general framework of Vapnik [7] an error bound can be given expressed in the classifier capacity and the training sample size. Here too, it has to be concluded that for $n <= k$ no generalization is possible.

Mean distance maximizing classifiers. According to this principle classifiers are found by optimizing their parameters such that the mean (squared) distance of the training samples is maximized. Distances to objects on the wrong side of the classifier have a negative sign. Sometimes distances are weighed by an arc tangent or sigmoid function in order to diminish the influence of remote objects. The principle of distance maximization has no direct connection with the classification error. However, relations with the Parzen window classifier and the aposteriori error probability can be formulated for some cases. Training rules for perceptrons and neural network classifiers belong to this family, see Pao [8]. The principle of distance maximization does not yield a lower bound on the number of training samples.

In many pattern recognition problems objects are initially given by images, time signals or spectra. The number of measurements representing a single objects can thereby easily be as large as 100 or even much more. Traditionally this amount should be condensed to a small number by a good definition of the features. If the knowledge to do this fails automatic feature extraction methods may be used. The neural network literature suggests that feature extraction and classification might be combined into a single approach. However, in that case it should be possible to train high dimensional classifiers with a relative small number of samples. It is the purpose of this paper to investigate some possibilities. See also a previous discussion [9].

That there should be something possible at very low sample sizes can be understood from the following example. Suppose we have for each of two classes 10 samples given in a 100 dimensional features space. According to the problem of density estimation and also accordinging to the principles formulated by Cover and Vapnik as discussed above, this is a case for which no generalization can be expected. However, suppose the class variances are small compared to the differences in class means. Now the Nearest Mean Classifier, generating the perpendicular bisector between the class means yields a perfect linear classifier. Even its generalization capacity can be checked by estimating the classification error using the leave-one-out estimator. So it is possible to find a good classifier and to know that it is good even if $n < k$. However, the Nearest Mean Classifier does not take into account differences in variances and covariances. That is done by then the Fisher's linear discriminant that needs sample sizes $n > k$. In the next section a linear classifier is discussed that can be used in between.

2. The small sample size classification problem

The Fisher Linear Discriminant is often defined as the optimal classifier for two Gaussian distributions with means μ_A and μ_B and with equal covariance matrix X:

$$\mathbf{x} \bullet \mathbf{w} = \mathbf{x}^T C_{-1}(\mu_A - \mu_B) = 0 \qquad (1)$$

It is assumed that the classes have equal apriori probabilities and that the origin is shifted such that E$\mathbf{x}$ = 0 or $\mu_A = -\mu_B$. Equation (1) is also the solution of

$$\min_{\mathbf{w}}\{E_A(\mathbf{x} \bullet \mathbf{w} - 1)^2 + E_B(\mathbf{x} \bullet \mathbf{w} + 1)^2\} \qquad (2)$$

which minimizes the expected difference between the discriminant function values and the target outcomes +1 for class A and -1 for class B. The use of equation (1) gives problems for small sample sizes, due to the fact that the use of the scatter for estimating the covariance matrix yields for $n < k$ a singular matrix that cannot be inverted. Equation (2) yields a direct, unique solution if we demand simultaneously that the distances of the given sample points $\mathbf{x}_i$ (i=1,n) have a maximum distance to the discriminant function. This is equivalent to stating that we want that solution $\mathbf{w}$ with has a minimum norm $||\mathbf{w}||$. This solution can be written in terms of the pseudo inverse:

$$\mathbf{x} \bullet \mathbf{w} = \mathbf{x}(C^T C)_{-1}\, C^T(\mu_A - \mu_B) = 0 \qquad (3)$$

in which we now for C substitute $\Sigma \mathbf{x}_i \mathbf{x}_i^T$ For values of $n > k$ this equation is identical to Fisher's linear discriminant (1). We will therefore call it the Pseudo Fisher discriminant.

Why should the linear discriminant that maximizes the distances to all given samples be a good discriminant? Why should it be any better than all other linear discriminants that give an exact solution for (2)? We will illustrate by some simulations that this is indeed questionable. However, there is one good argument: If $n < k$, all the given data samples are in a linear subspace. As (3) maximizes the distances between these samples and the discriminant plane, $\mathbf{w}$ should be in the same subspace as the data and the discriminant is perpendicular to that subspace. This makes sense as it corresponds with an indifference to directions for which no samples are given.

In the next section the behavior of the pseudo Fisher discriminant as a function of the sample size is illustrated by an artificial example (see Fig. 2.). For one sample per class this method is equivalent with the Nearest Mean and with the Nearest Neighbor method. If the total sample size is equal to or larger than the dimensionality ($n >= k$), the method is equal to Fisher's linear discriminant. For $n = k$, where Fisher's discriminant starts, the error shows a maximum. Surprisingly the error has a local minimum somewhere below $n = k$. This can be understood from the observation that the pseudo Fisher discriminant succeeds in finding hyperplanes with equal distances to all training samples until $n = k$. There is, however, no need to do that for samples that are already classified correctly. Therefore we modified this method to the following procedure in which we iteratively add misclassified objects until all objects in the training set are correctly classified:

1. Put all objects of the training set in set U. Create an empty set L.
2. Find in U those two objects, one from each class, that are most closest. Move them from U to L.
3. Compute de Pseudo Fisher discriminant D for L.
4. Compute the distance of all objects in U to D.
5. If all objects are correctly classified, stop.
6. Move of the objects in U that are misclassified the one at the largest distance to D (the "worst" classified object) from U to L.
7. If the number of objects in L is smaller than k (the dimensionality) go to 3.
8. Compute Fisher's linear discriminant for the entire set of objects, $L \cup U$.

This procedure, which we call the Small Sample Size Classifier (SSSC) uses only those objects for the Pseudo Fisher discriminant that are in the area between the classes and that are absolutely needed for constructing a linear discriminant that classifies all objects correctly. If this appears to be impossible, due to a too large training set relative to the class overlap, Fisher's discriminant is used.

The result of the SSSC for our example is shown in Fig. 2.It doesn't show any peaking, but it can still be improved, however. We noticed that from roughly $n = k/2$ to $n = 2k$, the SSSC showed just a minor improvement. We therefore decided to use subsets from the training set of the 'optimal' size, i.e. $n = k/2$: From each class at random $k/4$ objects were selected and were fed into the SSSC. This was repeated for a large number of randomly selected subsets. All resulting discriminant vectors were averaged. The SSSC-8 result in Fig. 2. is obtained by using subsets of 8 objects per class ($n = 16$).

3. Examples

In order to illustrate the possibilities of the two classifiers discussed in the previous section an artificial example has been constructed in which the covariance structure is such that the Nearest Mean classifier gives a bad result and the low sample size properties of the Nearest Neighbor rule are also bad. It is a two class Gaussian problem with equal covariance matrices in a 30 dimensional feature space. The first class has a zero mean in all directions. The mean of the second class is (3,3) for the first two features and 0 for all other features. The covariance matrix is a diagonal matrix with a variance of 40 for the second features and a unit variance for all other features. We studied this example in a rotated version in order to spread the separability over all features. A 2-dimensional projection is shown in Fig. 5. The intrinsic class overlap is 0.064 (Bayes error). Training sets with 1 to 200 samples per class were generated. The errors for the estimated Nearest Mean discriminants, and the Pseudo Fisher discriminants are computed analytically. The error for the 1-Nearest Neighbor rule is estimated by a test set. Experiments were repeated 50 times. The average results are presented in Fig. 2., showing the effects discussed above.

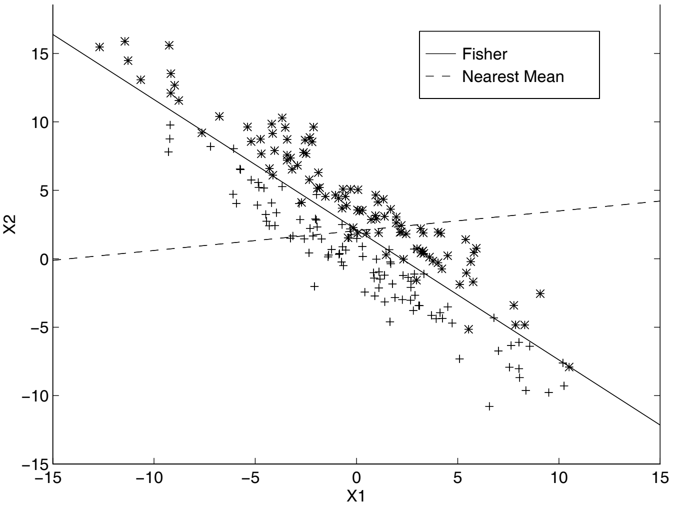

Fig. 1. Scatter plot of a 2-dimensional projection of the 30-dimensional example used for the experiments.

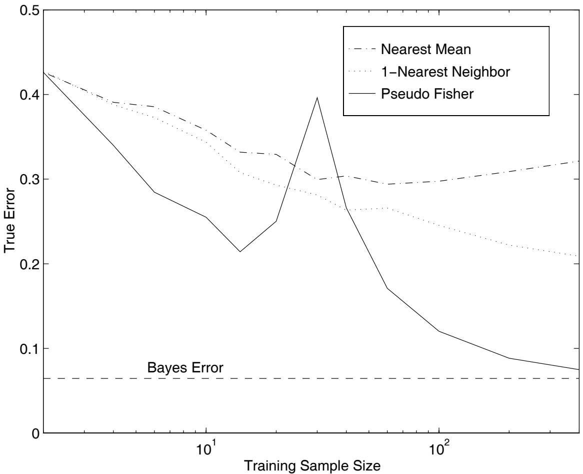

Fig. 2. True classification results for some classifiers as a function of the number of samples per class in an artificial 30-dimensional example. Results are averaged over 50 experiments.

The results for the Small Sample Size Classifier are presented in Fig. 3. This shows that this classifier can avoid the deterioration around $n = k$ as discussed above. It is noteworthy that it improves Fisher's discriminant up to $n = 10k$ in this example. That the SSSC is not a global improvement is shown by the results of a similar example in Fig. 4. Here the variances of the features 2 - 9 of the previous example set to 40, keeping all other variances at 1. Again the result was rotated in the 30-dimensional feature space.

Finally an example with real data is investigated. We used the sonar data set as originally published by Gorman [10]. The way we have adapted this data set is described in [11]. In total there are 208 objects in two classes, 111 from reflections of a metal cylinder, 97 from rock. Each object is 60 dimensional vector, representing the energy in 60 wave bands. In the present experiment training sets of 1 up to 80 objects per class are drawn randomly. The remaining objects are used for testing. This is repeated 50 times. The averaged results are presented in Fig. 5.The same effects as in the artificial example are found: Pseudo Fisher has a maximum for 30 samples per class ($n = 60$) at the point where starts to become equivalent to the Fisher Discriminant. The Small Sample Size Classifier with 8 objects per subset has around this point a much better performance and is the best linear classifier for the investigated domain. The result of the Nearest Neighbor classifier shows that the optimal classifier is probably nonlinear.

4. Discussion

The Pseudo Fisher linear discriminant presented in this paper can be used for all sample size. It has for very low sample sizes already a reasonable performance. The deterioration of the

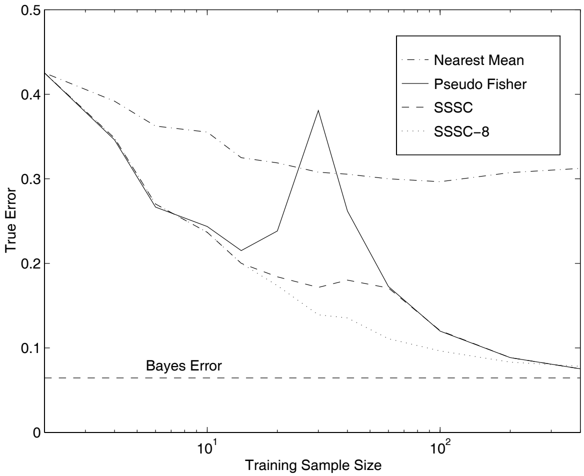

Fig. 3. True classification results of the Small Sample Size Classifier compared with the Pseudo Fisher method for the same artificial 30-dimensional example. Results are averaged over 50 experiments.

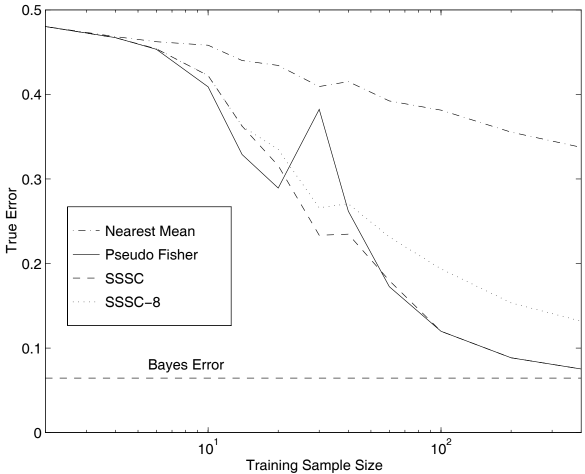

Fig. 4. True classification results of the Small Sample Size Classifier compared with the Pseudo Fisher method for the same artificial 30-dimensional example. Results are averaged over 50 experiments.

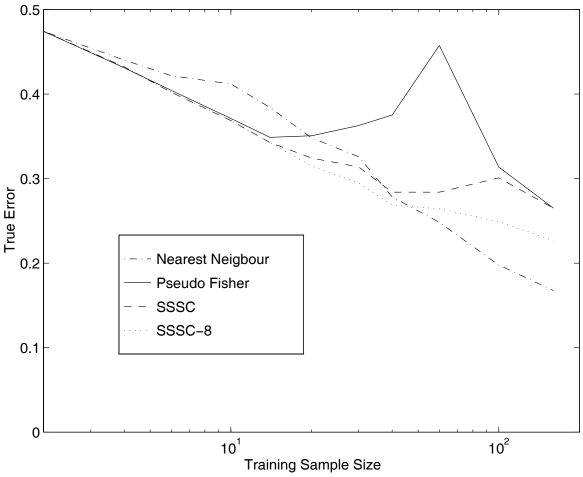

Fig. 5. Estimated true classification results for some classifiers as a function of the number of training samples per class for a real sonar data set with 60 features. Results are averaged over 50 experiments.

performance for samples sizes between half and three time the feature size can be overcome by the newly introduced Small Sample Size Classifier. Using averages over subsets better results can be obtained than Fisher's discriminant up to sample sizes of $n = 10k$. for some problems.

Let us reconsider why it is possible to find reasonably well performing linear discriminants for small sample sizes. First, density estimates are not necessary for constructing classifiers. In fact they offer a full description of the data and for classification purposes we are only interested in a generalizable separation. So density estimates are an overkill. Something might be gained by taking a less demanding approach.

On the other side, the arguments based on the work by Cover [5] and Vapnik [7] for the necessity of sample sizes larger than the dimensionality are entirely focussed on possible dichotomies of the data set neglecting any possible generalization based on some data structure. This approach treats the data as a structureless set of random points. Something might be gained by introducing some realistic types of structure.

In real applications there are at least two types of structure visible in almost any high dimensional data set:

1. Classes have different means and the means represent a significant mass of the class densities.
2. Classes are not just clouds of points with about the same variance in all directions, but have directions with low variance and directions with high variance. Sometimes these differences are that large that the data can be described as being structured in subspaces. This implies that the problem has an intrinsic dimensionality lower than the feature size.

Because of the first type of structure the Nearest Mean Classifier shows some generalization. Because of the second type of structure the Small Sample Size Classifier (SSSC) as described in this paper works. In a previous paper it has been suggested that in many real applications with large feature sets the intrinsic dimensionality of the data is much smaller than the feature size [9]. This might be one of the reasons that neural nets can be used as classifiers for these problems. The SSSC might help to understand why. See also a useful discussion by Raudys [3].

Due to the relation between the SSSC and subspaces, it is also related to other classifiers based on subspaces, e.g. the ones discussed by Oja [6]. An important difference is that Oja uses the subspaces as class descriptions, yielding quadratic classifiers, while in this paper the subspaces are used as data descriptions in which a linear classifier is found. However, the mathematical descriptions are close. As far as the author is aware of, the small sample size properties are not reported earlier.

5. 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. Yoh-Han Pao, Adaptive pattern recognition and neural networks, Addison-Wesley, Reading, Massachusetts, 1989.

[2] S.J. 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, 252-264.

[3] S. Raudys, "*Why do multilayer perceptrons have favorable small sample properties?*," in: Pattern Recognition in Practice IV: Multiple paradigms, comparative studies and hybrid systems, ed. L.S. Kanal, pp. 287 - 298, Elsevier, Amsterdam, 1994.

[4] K. Fukunaga, *Introduction to statistical pattern recognition*, second edition, Academic Press, New York, 1990.

[5] T.M. Cover, *Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition*, IEEE Trans.Elec.Comp, vol. EC-14, 1965, 326-334.

[6] E. Oja, *The Subspace Methods of Pattern Recognition*, Wiley, New York, 1984.

[7] V. Vapnik, *Estimation of Dependences based on Empirical Data*, Springer-Verlag, New York, 1982.

[8] Y.H. Pao, *Adaptive Pattern Recognition and Neural Networks*, Addison-Wesley, Reading, MA, 1989.

[9] R.P.W. Duin, *Superlearning capabilities of neural networks?*, in: Proc. of the 8th Scandinavian Conference on Image Analysis, pp. 547 - 554, NOBIM, Norwegian Society for Image Processing and Pattern Recognition, Tromso, Norway, 1993.

[10] P. Gorman and T.J. Sejnowski, *Learned Classification of Sonar Targets Using a Massively Parallel Network*, IEEE Transactions on ASSP, vol. 36, no. 7, July 1988.

[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, 391-402.