KYBERNETIKA — VOLUME *34* (1998), NUMBER 4, PAGES 461–466

INTRINSIC DIMENSIONALITY AND SMALL SAMPLE PROPERTIES OF CLASSIFIERS

ŠARŪNAS RAUDYS

Small learning-set properties of the Euclidean distance, the Parzen window, the minimum empirical error and the nonlinear single layer perceptron classifiers depend on an "intrinsic dimensionality" of the data, however the Fisher linear discriminant function is sensitive to all dimensions. There is no unique definition of the "intrinsic dimensionality". The dimensionality of the subspace where the data points are situated is not a sufficient definition of the "intrinsic dimensionality". An exact definition depends both, on a true distribution of the pattern classes, and on the type of the classifier used.

1. INTRODUCTION

In statistical literature, it is well known that small sample properties of statistical classifiers heavily depend on dimensionality of the data. Estimates exist that show that in high-dimensional cases, the learning-set size should be very large. Practice, however, indicates that often some of the statistical classifiers have been perfectly trained in cases when learning-set sizes were small in comparison with a number of dimensions [1]. Most often such comments are related with a usage of artificial neural nets. This paper develops an idea originally presented by Duin [1] concerning effect of the ıntrinsic dimensionality on the small sample properties of statistical classifiers. We analyze known theoretical results concerning dimensionality-sample size relationships and show that for several parametric and non-parametric classifiers, as well as a non-linear single-layer perceptron not the real, but an intrinsic dimensionality of the data should be taken into account while determining the small sample properties.

2. SMALL SAMPLE PROPERTIES OF PARAMETRIC CLASSIFIERS

The simplest statistical classifier is *the Euclidean distance (the nearest means) classifier*. It is a linear discriminant function (DF) designed to classify two spherical multivariate Gaussian populations differing in mean vectors μ_1, μ_2, but sharing the same *identity* covariance matrix $\Sigma = \mathbf{I}\sigma^2$.

In [2] the generalization error was first considered in asymptotic, when the dimensionality p and learning set sizes N_1, N_2 are large and are increasing simultaneously. Moreover, true distributions of the pattern classes was considered to be Gaussian with common covariance matrix Σ(GCCM): $N(\mu_i, \Sigma)$.

Note that while designing the EDC classifier one assumes, the covariance matrix $\Sigma = \mathbf{I}\sigma^2$, and in the analysis of the generalization error, we consider the case when *the probabilistic model of the pattern classes is different*, i.e., $\Sigma \neq \mathbf{I}\sigma^2$. This approach implies that asymptotically conditional distribution of the random variable $g(\mathbf{X}, \overline{\mathbf{X}}^{(1)}, \overline{\mathbf{X}}^{(2)})$ tends to Gaussian distribution and allows us to obtain very simple, however very accurate estimates.

Let $N_2 = N_1 = N$, $q_2 = q_1$. For large p and N, following expression for the expected PMC was obtained

$$EP_N^{(E)} \approx \Phi\left\{ -\frac{\delta^*}{2}\frac{1}{\sqrt{T_\mu^*}}\right\}, \tag{1}$$

where $\Phi\{a\} = \int_{-\infty}^{a}(2\pi)^{-1/2}\sigma^{-1}\exp\{-t^2/(2\sigma^2)\}\,\mathrm{d}t$, $\delta^* = \frac{\mu'\mu}{\sqrt{\mu'\Sigma\mu}}$, $\mu = \mu_1 - \mu_2$, $T_\mu^* = 1 + \frac{2p^*}{\delta^{*2}N}$, $p^* = \frac{(\mu'\mu)^2(tr\Sigma^2)}{(\mu\Sigma\mu)^2}$ is an effective dimensionality.

Asymptotically, as $N \to \infty$ we obtain the asymptotic PMC of EDC: $P_\infty^{(E)} = \Phi\{-\delta^*/2\}$. Equation (1) shows that small learning-set properties of EDC heavily depend on true distributions of the pattern classes (parameters μ and Σ). For the spherical Gaussian case we have $\Sigma = \mathbf{I}\sigma^2$. Then $p^* = p$, $\delta^* = \delta$, where $\delta^2 = \mu'\Sigma^{-1}\mu$ is a squared Mahalanobis distance.

In a more general case (when $\sigma \neq \mathbf{I}\sigma^2$), $\delta^* \leq \delta$, and $p^* \neq p$. In principle, p^* can be arbitrary large. An example is two 100-variate ($p = 100$) Gaussian classes with common covariance matrix; unit variances; $\mu_1 = -\mu_2 = 0.0018805 \times (1, 1, \ldots, 1)$, correlations between all the variables $\rho = -0.0101$, and $P_\infty^{(E)} = 0.03$. Then $p^* \approx 10^8$. From (1) for $N = 200$ we calculate $EP_N^{(E)} = 0.497$. We have obtained the same result by simulation experiments too. It is a very high generalization error. Another example is two pattern classes that are distributed on two very close parallel straight lines in the multivariate feature space.

Theoretically, situations exist where p^* is close to 1. It means that distributions of the pattern classes lie in a one-variate linear subspace, i.e. the intrinsic dimensionality of the data is equal to 1. An example is two 100-variate ($p = 100$) Gaussian classes sharing common covariance matrix: unit variances; correlations between all the variables $\rho = 0.3$, $\mu_1 = -\mu_2 = 1.042 \times (1, 1, \ldots, 1)$. For this data $p^* \approx 1.05$, $\delta^* = \delta = 3.76$ and $P_\infty^{(E)} = 0.03$. Due to the small effective dimensionality p^*, for this specific choice of parameters, we can train the EDC with very small learning-sets: from (1) for $N = 5$ we calculate $EP_N^{(E)} = 0.0318$. Simulation experiments confirm this theoretical estimate. We see that for this very favorable case, in spite of the high formal number of variables ($p = 100$), only five vectors per class are sufficient to train the classifier.

Another popular parametric classification rule is *the standard Fisher linear DF*. It is an asymptotically optimal classifier designed to classify two multivariate GCCM populations. Its generalization error can be expressed by (1) with $T_\mu T_\Sigma$ instead of T_μ^* [5]. The term $T_\mu = 1+\frac{2p}{\delta^2 N}$ arises from the inexact sample estimation of the mean vectors of the classes and the term $T_\Sigma = 1 + \frac{p}{2N-p}$ arises from the inexact sample estimation of the covariance matrix. For GCCM model, however, the generalization error depends on the rue dimensionality p, and not on p^*: for both above examples with $p = 100$ from asymptotical formula we obtain $EP_N^{(E)} = 0.0577$.

3. SMALL SAMPLE PROPERTIES OF NON-PARAMETRIC CLASSIFIERS

The most popular version of a non-parametric *Parzen window (PW) classifier* is based on following estimate of the multivariate density function

$$\hat{f}_{PW}(\mathbf{x}|\pi_i) = \frac{1}{N_i}\sum_{j=1}^{N_i} N(\mathbf{x}, \mathbf{X}_j^{(i)}, \mathbf{I}\lambda^2), \tag{2}$$

where $N(\mathbf{x}, \mu, \Sigma)$ stands for multivariate density function and λ is a smoothing constant.

At the fixed point $\mathbf{x}$ of the multivariate feature space Ω, a value of the Parzen window distribution density estimate depends on N_i random vectors of the learning-set $\mathbf{X}_1^{(i)}, \dots, \mathbf{X}_{N_i}^{(i)}$. Therefore it can be analyzed as a random variable. According to the central limit theorem when $N_i \longrightarrow \infty$ the sum (2) of N_i random contribution terms $N(\mathbf{x}, \mathbf{X}_j^{(i)}, \mathbf{I}\lambda^2)$ tends to the Gaussian distribution. Thus, at one particular point $\mathbf{x}$, a conditional probability of misclassification approximately is determined by means E and variances V of estimates $\hat{f}_{PW}(\mathbf{x} \mid \pi_1)$ and $\hat{f}_{PW}(\mathbf{x} \mid \pi_2)$

$$P(\mathit{misclassification}|\mathbf{x}, \mathbf{x} \in \pi_i) \approx \Phi\left\{\frac{E\hat{f}(\mathbf{x}|\pi_1) - E\hat{f}(\mathbf{x}|\pi_2)}{\sqrt{(V\hat{f}(\mathbf{x}|\pi_1) + V\hat{f}(\mathbf{x}|\pi_2))/2}}(-1)^i\right\}. \tag{3}$$

Consider the GCCM model with parameters μ_i and Σ. The conditional mean of the nonparametric density estimate (conditioned at fixed point $\mathbf{x}$) *with respect to all possible learning sets*, consisting of N_i observations, is

$$\begin{aligned} E\hat{f}(\mathbf{x}|\pi_i) &= \frac{1}{N_i}\sum_{j=1}^{N_i}\int N(\mathbf{X}_j^{(i)}, \mu_i, \mathbf{I})N(\mathbf{x}, \mathbf{X}_j^{(i)}, \mathbf{I}\lambda^2)\, d\mathbf{X}_j^{(i)} = \\ &= N(\mathbf{x}, \mu_j, \Sigma + \mathbf{I}\lambda^2). \end{aligned} \tag{4}$$

For above model of the true densities the variance of the PW density estimate is

$$V\hat{f}(\mathbf{x}|\pi_i) = \frac{1}{N_i}\left[\frac{|2\Sigma + \mathbf{I}\lambda^2|^{1/2}}{\lambda^p}\left(N(\mathbf{x}, \mu_i, 2\Sigma + \mathbf{I}\lambda^2)\right)^2 - \left(E\hat{f}(\mathbf{x}|\pi_i)\right)^2\right]. \tag{5}$$

Let $\mathbf{T}$ be a $p * p$ orthonormal matrix such that $\mathbf{G}\Sigma\mathbf{G}' = \mathbf{D}$ ($\mathbf{D}$ is a diagonal matrix of eigenvalues with elements $d_1, d_2, \dots, d_p$). Then

$$V\hat{f}(\mathbf{x}|\pi_i) = \frac{1}{N_i}\left[\prod_{j=1}^{p}\sqrt{1 + \frac{2d_j}{\lambda^2}}\left(N(\mathbf{x}, \mu_i, 2\Sigma + \mathbf{I}\lambda^2)\right)^2 - \left(E\hat{f}(\mathbf{x}|\pi_i)\right)^2\right]. \tag{6}$$

For very small λ^2, the variance of the PW estimate is determined primarily by the term

$$\frac{1}{N_i}\prod\sqrt{1\frac{2d_j}{\lambda^2}}. \tag{7}$$

This term decreases with an increase in a value of the smoothing parameter λ^2 and decreases with an increase in N_i, the number of learning examples. Let the eigenvalues of covariance matrix Σ are equal: $d_1 = d_2 = \ldots = d_p = d$ and let the number of features p be increased. Then for small λ^2 we can conclude that in order to keep variance (6) constant, the number of learning vectors N_i should increase as a degree of the dimensionality p:

$$N_i \equiv \left(1 + \frac{2d_j}{\lambda^2}\right)^{p/2} \tag{8}$$

Let now several eigenvalues of the covariance matrix Σ be very small: $d_1 = d_2 = \ldots = d_r = d$, $\quad d_{r+1} = d_{r+2} = \ldots = d_p = \varepsilon 0$. We call number r, *the intrinsic dimensionality* of the data for the GCCM model. For this data model instead of (8) we have

$$N_i \equiv \left(1 + \frac{2d}{\lambda^2}\right)^{r/2} \tag{9}$$

It means that *small learning-set properties of the nonparametric Parzen window density estimate* (2) *are determined not by the formal dimensionality of the data, but by the true-intrinsic dimensionality r.* Therefore the number of learning vectors required to design this classifier should increase as a degree of the intrinsic dimensionality r. Note definition of r differs from that of p^*.

For the GCCM model, similar conclusions can be obtained also for a k-NN classification rule that uses the Euclidean distance to determine distances between the pattern vectors in the multivariate feature space.

4. SMALL SAMPLE PROPERTIES OF A NON-PARAMETRIC LINEAR ZERO EMPIRICAL ERROR CLASSIFIER

The non-parametric linear *zero empirical error classifier* is obtained when while training the minimum empirical error classifier, we succeed to discriminate the learning-set vectors without errors. Different criteria and optimization techniques are used to design the classifier that classifies the learning-set with a minimal number of misclassifications. In small learning-set analysis, a useful training model is a random search optimization procedure.

The *random search optimization procedure* generates many (say, t times) random discriminant hyperplanes $\mathbf{w}'\mathbf{x} + w_0 = 0$ according to a certain prior distribution of the weights, determined by a *priori* density $f_{prior}(\mathbf{w}, w_0)$, and selects those that classify learning sets LS^1 and LS^2, each of size N, without error.

In [4], an equation for a *mean expected probability of misclassification* for pattern vectors which did not participate in the training was derived. The pattern classes were considered to be spherical Gaussian, and the prior density $f_{prior}(\mathbf{w}, w_0)$ of the

$(p+1)$-variate weight vector was considered to be $N(\mathbf{0},\ \mathbf{I})$. The derivation is based on following representation

$$Prob\{\mathbf{w}'\mathbf{x} + w_0 < 0 \mid \mathbf{x} \in \pi_i\} = \Phi\left\{(-1)^i \frac{\mathbf{w}'\mu_i + w_0}{\sqrt{\mathbf{w}'\mathbf{w}}}\right\} = P. \tag{10}$$

Consider the GCCM model $N(\mu_i, \Sigma)$ with $\mu_1 = -\mu_2 = \mu$, $\Sigma = \mathbf{G}'\mathbf{D}\mathbf{G}$, where G is $p \times p$ orthonormal matrix of eigenvalues of $\Sigma, \mathbf{D}$ is a $p \times p$ diagonal matrix of eigenvectors, such that $\mathbf{D} = \begin{bmatrix} \mathbf{I}_r & 0 \\ 0 & \varepsilon \mathbf{I}_{p-r} \end{bmatrix}$, ε is small such that $(p-r)\varepsilon \ll 1$, components m_{2j} of $(p-r)$-variate vector $\mathbf{m}_2$, $(\mu'\mathbf{G}' = (\mathbf{m}', \mathbf{m}_2'))$, $m_{2j} \ll \varepsilon$, and can be ignored. In this model of the data, we have *the intrinsic dimensionality* equal to $r < p$. Then

$$P = \Phi\left\{-\frac{\mathbf{w}'\mu_1 + w_0}{\sqrt{\mathbf{w}'\Sigma\mathbf{w}}}\right\} = \Phi\left\{-\frac{\mathbf{w}'\mathbf{G}'\mathbf{G}\mu_1 + w_0}{\mathbf{w}\mathbf{G}'\mathbf{G}\Sigma\mathbf{G}'\mathbf{G}\mathbf{w}}\right\} = \Phi\left\{-\frac{\mathbf{w}_1'\mathbf{m}/2 + w_0}{\sqrt{\mathbf{w}_1'\mathbf{w}_1}}\right\},$$

where $\mathbf{w}_1 = \mathbf{g}\mathbf{w}$, an r-variate subvector of vector $\mathbf{G}\mathbf{w}$, and $\mathbf{G} = \begin{bmatrix} \mathbf{g} \\ \mathbf{g}_2 \end{bmatrix}$.

Therefore for this model with the intrinsic dimensionality equal to r, the small learning-set properties of the zero empirical error classifier can be analyzed in the r-variate space. In this space, the r-variate vector $\mathbf{Y} = \mathbf{g}\mathbf{X}$ is Gaussian $N(\mathbf{m}/2, \mathbf{I}_r)$, or $N(-\mathbf{m}/2, \mathbf{I}_r)$. It means that the small sample properties of the zero empirical error classifier are determined not by the real but by the intrinsic dimensionality of the data r.

5. THE NON-LINEAR SINGLE-LAYER PERCEPTRON CLASSIFIER

Recently it was shown that while training the non-linear SLP the weights are increasing. Therefore during the iterative training process, a cost function used to obtain the weights changes its statistical properties. In principle, under certain conditions, the SLP pereptron can realize decision boundaries of seven known statistical classifiers, beginning with the simplest EDC, following the regularized discriminant analysis, the standard linear Fisher DF, a generalized Fisher linear DF, the minimum empirical error and the most complex – the maximum margin classifiers [3]. Small sample properties of some of these classifiers are determined not by the real but by the intrinsic dimensionalities of the data, p^* or r. We performed numerous simulation experiments with a singular multivariate Gaussian data that lies in the linear r-variate subspace, and the nonlinear SLP classifier with a sigmoid activation function, and targets $t = 0$ and 1. The experiments have confirmed that the small sample properties of the nonlinear SLP are determined by the intrinsic dimensionality p^* at the beginning, the formal dimensionality p later (if one uses non-limiting target values, e.g. 0.1 and 0.9), and the intrinsic dimensionality r at last.

6. CONCLUDING REMARKS

It is a common belief that in real world problems, there exist comparatively small number of "main factors" that determine a variability of patterns in the multivariate feature space. Thus it is postulated that pattern vectors lie in the non-linear subspace of low dimensionality. Abundant experimental investigations confirm this belief. Unfortunately, in addition to the "main factors" mentioned, a number of extra "noisy factors" influence the data. Therefore the data lies in a "non-linear blanket of a certain thickness" [1]. Extra, non-zero width directions worsen the small sample properties of the classification algorithms.

We have demonstrated that small learning-set properties of several classification rules depend on the "intrinsic dimensionality" of the data. There is no unique definition of the "intrinsic dimensionality". *The dimensionality r of the subspace where the data points are situated is not a sufficient definition of the intrinsic dimensionality. An exact definition depends both, on a true distribution of the pattern classes, and on the type of the classifier used.* Therefore the definition of the "intrinsic dimensionality" p^* of EDC for the GCCM model is different from the definition of the "intrinsic dimensionality" of the Parzen window classifier for the same GCCM model. One such example has been presented above: two Gaussian pattern classes that are distributed on two close parallel lines in the multivariate feature space. In this model, the data is distributed in the one-variate subspace. Only one eigenvalue of the covariance matrix is different from zero. Thus, $r \approx 1$. Nevertheless, the effective dimensionality p^* for EDC can be arbitrarily high.

(Received December 18, 1997.)

REFERENCES

[1] R. P. W. Duin: Superlearning capabilities of neural networks. In: Proc. of the 8th Scandinavian Conference on Image Analysis NOVIM, Norwegian Society for Image Processing and Pattern Recognition, Tromso 1993, pp. 547–554.

[2] Š. Raudys: On determining the training sample size of a linear classifier. In: Computing Systems (N. Zagoruiko, ed.), Vol. 28, Nauka, Institute of Mathematics, Academy of Sciences USSR, Novosibirsk 1967 (in Russian), pp. 79–87,

[3] Š. Raudys: Linear classifiers in perceptron design. In: Proceedings 13th ICPR, Vol. 4, Track D, Vienna 1996, IEEE Computer Society Press, Los Alamitos, pp. 763–767.

[4] Š. Raudys: On dimensionality, sample size and classification error of nonparametric linear classification algorithms. IEEE Trans. Pattern Analysis Machine Intelligence *PAMI-19* (1989), 6, 669–671.

[5] Š. Raudys and A. K. Jain: Small sample size effects in statistical pattern recognition: Recommendations for practitioners. IEEE Trans. Pattern Analysis Machine Intelligence *PAMI-13* (1991), 252–264.

*Prof. Dr. Šarūnas Raudys, Institute of Mathematics and Informatics, Akademijos 4,
2600 Vilnius. Lithuania.
e-mail:raudys@ktl.mii.lt*