3

Measures of Data and Classifier Complexity and the Training Sample Size

Šarūnas Raudys

Summary. The size of the training set is important in characterizing data complexity. If a standard Fisher linear discriminant function or an Euclidean distance classifier is used to classify two multivariate Gaussian populations sharing a common covariance matrix, several measures of data complexity play an important role. The types of potential classification rules cannot be ignored while determining the data complexity. The three factors — sample size, data complexity, and classifier complexity — are mutually dependent. In situations where many classifiers are potentially useful, exact characterization of the data complexity requires a greater number of characteristics.

3.1 Introduction

Today it is generally recognized that the complexity of a pattern recognition algorithm should be chosen in accordance with the training sample size. The more complex the classifier is, the more data are required for estimating its parameters reliably. Conversely, if the sample size is small, one is obliged to use the simplest classification algorithms (e.g., [9]). In addition, the complexity of the most suitable classifier depends also on the complexity of the data.

Theory shows that the difference between generalization and asymptotic errors of sample-based classifiers depends on both the sample size and the data configuration. Consequently, data complexity affects both the sensitivity of the classifier to training-set size and the complexity of the resultant decision boundary. For that reason, there is no wonder that numerous attempts to introduce general measures of data and classifier complexity that satisfy a majority of researchers did not lead to definite success (see, e.g., comments in [6, 8]).

We believe that *the concept "complexity of the data" does not exist without reference to a concrete pattern recognition problem and a concrete decision-making method.* The measure of the data complexity depends on the purpose for which this measure will be used. Three factors — the sample size, data complexity, and classifier complexity — are mutually related. An objective of this chapter is to examine the complexity of the classification rule and that of the data from the point of view of the sample size necessary to train the classifier.

The study of data complexity is a complicated issue, and we do not expect to obtain immediate success. For that reason, we restrict our main analysis to very simple data models and two standard statistical classification algorithms. We will consider linear decision boundaries and multivariate Gaussian distribution with a common covariance matrix for two pattern classes.

3.2 Generalization Errors of Two Statistical Classifiers

In this section we present definitions of distinct types of classification errors: the Bayes error, which is the asymptotic, conditional, and expected probabilities of misclassifications (PMC). We also present expressions of expected PMC (generalization error) for two typical statistical classifiers frequently used in applications: the standard Fisher linear discriminant function and the Euclidean distance classifier.

Suppose one knows probability density functions (PDF) of the input vectors and the prior probabilities of the pattern classes C_1 and C_2. One can then design the optimal Bayes classifier B, which, in classifying all possible vectors from C_1 and C_2, results in a minimal probability of misclassification. This PMC is called the *Bayes error* and is denoted by P_B. The probability of misclassification of a classifier designed from one particular training set using the classifier training algorithm A is conditioned on this specific algorithm and on this particular training set. The error rate for classifying the pattern vectors from the general population is called the *conditional probability of misclassification* and is denoted by P_N^{A}, where the index A indicates that the classifier training algorithm A is utilized and the index N symbolizes that the training set size is fixed. In the equation below, I assume $N = N_1 = N_2$, *where N_1 and N_2 are the sample sizes of the two classes*. In theoretical analysis, vectors of the training set may be considered as random ones; however, the sample size (N_1 and N_2) is fixed. Then the conditional PMC, P_N^{A}, may be considered as a random variable too.

Let $f(P_N^{\mathrm{A}})$ be the probability density function of the random variable P_N^{A} and let $\bar{P}_N^{\mathrm{A}}$ be its expectation over all possible randomly formed training sets of size N_1 and N_2 for each of the two classes, respectively. This expectation is called an *expected probability of misclassification.* The limit $P_\infty^{\mathrm{A}} = \lim\limits_{N_1 \to \infty, N_2 \to \infty} \bar{P}_N^{\mathrm{A}}$ is called *an asymptotic probability of misclassification*. In the neural network literature, both the conditional and expected PMC frequently are called *generalization error*, often without mentioning a proper distinction between the two notions.

The mathematical model of the data to be considered below are two multivariate Gaussian distributions with different means, $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$, and a common covariance matrix for both classes, $\boldsymbol{\Sigma}$ (the GCCM data model). The linear discriminant function (DF)

$$g(\mathbf{X}) = \boldsymbol{W}^T\mathbf{X} + w_0 \tag{3.1}$$

is an asymptotically optimal (when $N_1 \to \infty, N_2 \to \infty$) decision rule for this data model. In equation (3.1) $\boldsymbol{W} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2), w_0 = \boldsymbol{W}^T\boldsymbol{\mu}, \boldsymbol{\mu} = -\frac{1}{2}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)$, and ^{T} denotes a transpose operation. The Bayes error rate can be expressed as

$$P_B = \Phi\{-\frac{\delta}{2}\} = P_\infty^{\mathrm{A}}, \tag{3.2}$$

where $\Phi(c)$ is the cumulative distribution function of a standard N(0,1) Gaussian random variable and δ is the Mahalanobis distance, $\delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

In sample-based classifiers, the true values of the parameters $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2,$ and $\boldsymbol{\Sigma}$ are unknown and are substituted by sample-based estimates. If one makes use of maximum likelihood estimates, $\hat{\mu}_1, \hat{\mu}_2, \hat{\Sigma}$, one obtains the standard Fisher discriminant function, F, with $\boldsymbol{W}^{\mathrm{F}} = \hat{\boldsymbol{\Sigma}}^{-1}(\hat{\mu}_1 - \hat{\mu}_2), w_0 = \boldsymbol{W}^T\hat{\mu}$, and $\hat{\mu} = -1/2(\hat{\mu}_1 + \hat{\mu}_2)$. The Fisher DF was proposed 70 years ago; however, up to this day it remains one of the most often used classification rules [4, 5, 12]. Approximately two dozen alternative ways have been suggested to estimate the unknown coefficients of the linear discriminant function in situations where either the data is non-Gaussian or the number of training vectors is too small to estimate the covariance matrix reliably (see references in [2, 3, 10, 11, 12]).

One easy way to develop a simple linear discriminant function is to ignore the covariance matrix. Then one obtains the Euclidean distance (nearest mean) classifier E. Here only the mean vectors are used to calculate the weights: $\boldsymbol{W}^{\mathrm{E}} = \hat{\mu}_1 - \hat{\mu}_2, w_0 = \boldsymbol{W}^T\hat{\mu}$. Therefore, it is less sensitive to sample size.

The expected (generalization) error of Fisher DF can be approximated by a rather simple expression (see, e.g., [12, 14] and references therein)

$$\bar{\mathrm{P}}_N^{\mathrm{F}} \approx \Phi\Big\{ -\frac{\delta}{2}\frac{1}{\sqrt{T_M T_\Sigma}}\Big\} \tag{3.3}$$

where the term $T_{\boldsymbol{M}} = 1 + \frac{4p}{\delta^2 n}$ arises due to the inexact sample estimation of the mean vectors of the classes and the term $T_{\bar{\Sigma}} = 1 + \frac{p}{n-p}$ arises due to the inexact sample estimation of the covariance matrix, p denotes the number of input variables (the dimensionality of the feature vector), and n is the sample size: $n = N_1 + N_2$. In equation (3.3) we assume $N_1 = N_2 = N$. An asymptotic error of Fisher linear classifier, $\bar{P}_\infty^{\mathrm{F}}$, can be computed when $N_1 \to \infty$, $N_2 \to \infty$, and thus $\bar{P}_N^{\mathrm{F}} \to \bar{P}_\infty^{\mathrm{F}}$. Note, for GCCM data model, the Bays error is $\bar{P}_\infty^{\mathrm{F}} = P_B$.

In an analogous expression for the Euclidean distance classifier (EDC) we have to skip the term $T_{\bar{\Sigma}}$. This analytical expression for the generalization error of EDC is valid if the data distribution is spherically symmetric Gaussian, i.e.,

$$\boldsymbol{\Sigma} = \sigma^2 \begin{bmatrix} 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ & & \cdots & \\ 0 & 0 & & 1 \end{bmatrix} = \sigma^2 \mathbf{I},$$

where $\mathbf{I}$ is the identity matrix and σ^2 is a positive scalar constant.

In real-world applications, the input variables are often correlated. In those cases the asymptotic errors differ:

$$P_\infty^{\mathrm{E}} = \Phi\{-\frac{\delta^*}{2}\} \geq \bar{P}_\infty^{\mathrm{F}} = P_B, \tag{3.4}$$

where P_{∞}^{E} is the asymptotic error of EDC, and δ^* is an effective distance between pattern classes,

$$\delta^* = \frac{(\mu_1-\mu_2)^T(\mu_1-\mu_2)}{\sqrt{(\mu_1-\mu_2)^T\bar{\Sigma}(\mu_1-\mu_2)}}. \tag{3.5}$$

The expected PMC of EDC is also affected by an effective number of features p^*,

$$\bar{P}_N^{\mathrm{E}} \approx \Phi\Big\{-\frac{\delta^*}{2}\frac{1}{\sqrt{1+\frac{4p^*}{n(\delta^*)^2}}}\Big\}, \tag{3.6}$$

where

$$1 \le p^* = \frac{((\mu_1-\mu_2)^T(\mu_1-\mu_2))^2\mathrm{tr}(\Sigma^2)}{((\mu_1-\mu_2)^T\mathbf{\Sigma}(\mu_1-\mu_2)^T)^2} \le \infty. \tag{3.7}$$

3.3 Complexities of the Classifiers and the Data

Perhaps the least complicated theoretical data model in pattern recognition is the spherically symmetric Gaussian distribution. Here all variables are uncorrelated and have identical variances. In such a situation, $P_{\infty}^{\mathrm{E}} = P_{\infty}^{\mathrm{F}} = P_B,\ p^* = p$, and $\delta^* = \delta$. Only two parameters are needed to describe the data distribution: the dimensionality p and the Mahalanobis distance δ. This data model is illustrated with the pair of classes C1, C2 in Figure 3.1.

In a majority of known generalization error studies, the complexity of the data is characterized by the dimensionality only (see, e.g., [1, 17]). The equations presented in the previous section advocate that, from the point of view of statistical pattern recognition, a difference between the asymptotic and expected probabilities of misclassification also is inducing the complexity of the data. The situation becomes much more complicated in the case where input variables are mutually correlated. Then it may happen that $\delta^* < \delta$ or even $\delta^* << \delta$ (for illustration see the pair of classes C6, C7 in Fig. 3.1). It may also happen that the effective dimensionality p^* is close to 1 (the pair of classes C4, C5 in Fig. 3.1). In this case, for EDC the actual dimensionality of the data is one. Thus, looking from a perspective of a difference between expected and asymptotic error rates, the pair of classes C4, C5 is very simple for EDC; however, it becomes more complex for the Fisher classifier. In another extreme case, p^* tends to infinity (the pair of classes C3, C4 in Fig. 3.1). In the latter case, the distribution of pattern classes is much more "complicated" for the EDC. For the Fisher classifier, however, the complexity of the data remains the same.

The following measures of complexity could be useful in characterizing the complexity of the data from the point of view of classification error:

$$\begin{aligned} &\bar{P}_N^{\mathrm{F}}/\bar{P}_N^{\mathrm{E}},\ \bar{P}_N^{\mathrm{F}}/P_{\infty}^{\mathrm{F}},\ \bar{P}_N^{\mathrm{E}}/\bar{P}_{\infty}^{\mathrm{E}}, \\ &\min(\bar{P}_N^{\mathrm{F}},\bar{P}_N^{\mathrm{E}})/P_B,\ P_{\infty}^{\mathrm{E}}/P_B, P_{\infty}^{\mathrm{F}}/P_B,\ P_{\infty}^{\mathrm{E}}/P_{\infty}^{\mathrm{F}}, p/p^*. \end{aligned} \tag{3.8}$$

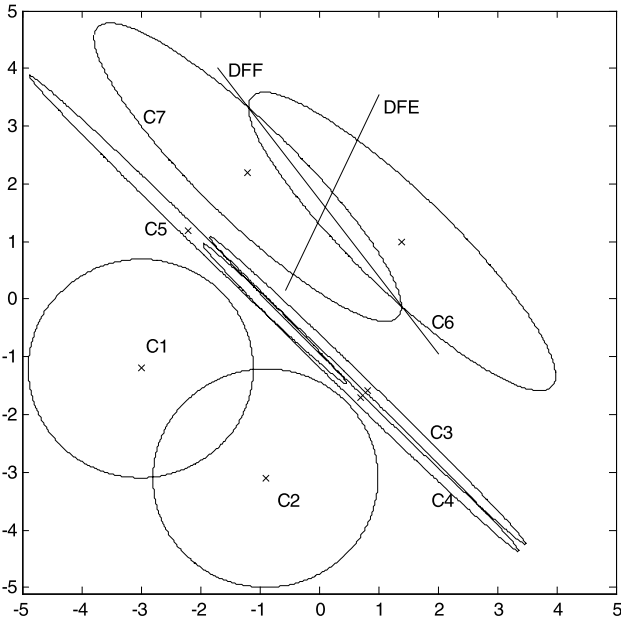

Fig. 3.1. Effect of the covariance matrix and the difference between the mean vectors, $\hat{\mu}_1 - \hat{\mu}_2$, on the effective dimensionality p^* and the effective distance δ^*: for classes C1 and C2, $p^* = p = 2, \delta^* = \delta$; for C3 and C4, $p^* >> p,\ \delta^* = \delta$; for C4 and C5, $p^* << p, \delta^* = \delta$. For classes C6 and C7, $\delta^* << \delta,\ p^* = 1.8$ (DFE and DFF are the decision boundaries of the EDC and the Fisher classifiers, respectively).

Instead of the ratios, absolute differences between each pair of quantities could be utilized too. At first sight, in terms of these measures, the simplest data model is the spherically symmetric Gaussian distribution where $P_\infty^{\mathrm{E}}/P_\infty^{\mathrm{F}} = 1$ and $p^* = p$. From this viewpoint, it seems that the measure $\gamma^{\mathrm{EF}\infty} = P_\infty^{\mathrm{E}}/P_\infty^{\mathrm{F}}$ is quite reasonable. Condition $\gamma^{\mathrm{EF}\infty} = 1$ indicates that the data set is simple enough and the simple classifier EDC can be used instead of the more complex Fisher classifier. It is true only if one does not take into account the fact that the training sample size is finite. A deeper examination reveals, however, that the situation could exist where $p^* >> p$. In such cases, instead of EDC, the Fisher classifier could become more useful.

If $P_\infty^{\mathrm{E}}/P_\infty^{\mathrm{F}} = 1$ and p^* is close to 1, one prefers to use EDC. In those cases, the intrinsic dimensionality for the data is equal to 1, and such data models should be considered as very simple. For that reason, the parameter $\gamma^{\mathrm{EF}\infty}$ alone is insufficient to characterize the data complexity. In fact *all four parameters*, p, δ, and $p^*,\ \delta^*$, *jointly determine the complexity of a pattern recognition task* if the data distribution is Gaussian with a common covariance matrix for the two classes, such that either EDC or Fisher potentially could be used for classification.

The data complexity problem becomes even more complicated if more types of classification rules are considered as potential candidates for decision making. Consider the GCCM data model. Let $p = 200, N_1 = N_2 = 100$, and $\delta = 3.76\ (P_\infty^{\mathrm{F}} =$

0.03). Then the Fisher classifier will result in approximately 11% error. Such a high generalization error rate means that the sample size is too small. If the features are correlated, it could happen that $P_{\infty}^{\mathrm{E}}/P_{\infty}^{\mathrm{F}} >> 1$. Therefore, EDC will be an inappropriate classifier too.

A number of ways could be attempted in order to design the classifier given small sample size and high feature dimensionality. Examples are dimensionality reduction by feature selection or feature extraction, different regularization methods, and structuralization of the covariance matrix with a small number of parameters [12, 16].

A rather universal structuralization method is to approximate the dependence structure among the variables by the first-order tree dependence. Here the probability density function is approximated by the product of $p-1$ conditional and one marginal density:

$$p(x_1, x_2, ..., x_p) = \prod_{j=1}^{n} f(x_j|x_{mj})(1 \leq m_j \leq p). \tag{3.9}$$

In representation equation (3.9) the sequence m_1, $m_2, ..., m_p$ constitutes an unknown permutation of the integers, and $f(x_1|m_1)$, by definition, is equal to $p(x_1)$. In a general case, the covariance matrix has $p \times p$ nonzero elements. An inverse of this matrix $\boldsymbol{\Sigma}^{-1}$, however, has only $2p-1$ distinct nonzero elements. Thus, to design the Fisher classifier with the first-order tree type structuralized covariance matrix (denoted by FT_1), we estimate $2p$ parameters that are different in the opposite pattern classes and $2p-1$ parameters that are common for both classes. In addition, we need to know the permutation m_1, $m_2, ..., m_p$.

In practice, unknown permutations have to be found from the sample covariance matrix. The theory shows that the expected PMC of this classifier is expressed by equation (3.3) with $T_{\bar{\Sigma}} = 1$ [12, 18]. Experiments with a dozen real-world data sets indicated that in a majority of cases, such a decision-making rule outperforms both the Euclidean distance and the Fisher classifiers [13].

For a layman, it seems that such a data model is very complex; we have to understand the permutation structure and know how to estimate it and the coefficients of the structuralized covariance matrix. For an expert in pattern recognition equipped with well-organized software, such a model (the case where the dependencies between p input variables are determined by the first-order tree dependence model, so that the asymptotic errors $P_{\infty}^{\mathrm{FT}_1} \approx P_{\infty}^{\mathrm{F}}$) implies that the classifier FT_1 has very favorable small sample properties: $\bar{P}_N^{\mathrm{FT}_1}/P_{\infty}^{\mathrm{FT}_1} \approx \bar{P}_N^{\mathrm{E}}/\bar{P}_{\infty}^{\mathrm{E}}$. For him or her the data set for classifier FT_1 is rather simple. In contrast, for the Euclidean distance and the Fisher classifiers this data set is complex. The above considerations about the first-order tree dependence model advocate that the complexity of data depends also on the researcher's knowledge about this model and on the presence of the software for estimating the structuralized covariance matrix.

3.4 Other Classifiers and Concluding Remarks

Our main concern in this chapter is to show that the complexity of data should be evaluated from the standpoint of the classifier utilized for decision making. In sections 3.2 and 3.3 we considered the unimodal GCCM model where a linear decision rule is asymptotically optimal. Even in such a simple data model, we have found that a number of issues are important while evaluating the data complexity.

If the covariance matrices of the classes are different, i.e., $\boldsymbol{\Sigma}_2 \neq \boldsymbol{\Sigma}_1$, the asymptotically optimal classifier is a quadratic disriminant function (DF). A good alternative in the two-class case is a linear classifier suggested by Anderson-Bahadur (known as the AB procedure) where the weight vector is expressed as (see, e.g., [12])

$$\boldsymbol{w}_{AB} = (\boldsymbol{\Sigma}_1\alpha + \boldsymbol{\Sigma}_2)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{3.10}$$

The unknown coefficients α_1 and threshold weight w_0 have to be found to minimize certain selected classification performance criteria. If $\boldsymbol{\Sigma}_2 \neq \boldsymbol{\Sigma}_1$, the differences among the asymptotic errors of the quadratic DF, the AB procedure, and the Fisher linear DF will affect the evaluation of the data complexity. If the sample size is taken into account, one needs to remember that the quadratic DF is very sensitive to sample size if the dimensionality of the input feature space is high. In a relatively small region of a multidimensional feature space where the pattern classes overlap, the quadratic DF can be approximated by a hyperplane. In the remaining space, we will have relatively few overlapping vectors. Therefore, in a major part of the multivariate feature space, an exact position of the decision boundary is not important. Simulations show that small sample properties of the AB procedure are much more favorable than that of the quadratic DF. Therefore, in many real-world two-class problems, the AB procedure works as well as or even better than the quadratic classifier [12, 14].

A very important concern in considering the quadratic DF is the fact that the expected classification error depends on the sample sizes of both classes, i.e., N_1 and N_2. Such a situation is characteristic of nonoptimal statistical classifiers trained with the plug-in design principle [5]. Another example where both sample sizes N_1 and N_2 are affecting the generalization error is a multinomial classifier, such as the one used in the behavior knowledge space method (see, e.g.. [7, 12]). In certain situations, an increase in the number of training vectors of one pattern class increases the generalization error instead of decreasing it! [12, 15]. Therefore, while characterizing data complexity; both sample sizes N_1 and N_2 are important.

In addition to statistical pattern recognition and heuristically based methods, linear classifiers can be obtained by other procedures, such as by training a single layer perceptron (SLP). In this way, one may obtain the EDC, regularized and robust discriminant analysis, the standard Fisher rule and that with covariance matrix pseudoinversion, the minimum empirical error classifier, as well as the support vector machine [11, 12]. Which classifier will be obtained depends on the training parameters and, most importantly, on stopping moment. Thus, the SLP is not a single classification rule. It is a set of different rules of diverse complexity. Here the classifier's complexity is measured in terms of a number and a type of parameter of distribution

density function if statistical methods would be utilized to estimate the weights of the linear classifier. Some data sets could be very difficult for SLP training at the very beginning; however, it becomes "easier" later. For example, the GCCM data with low p and very high p^* could become very difficult for the classical SLP training procedure [12].

In difficult pattern classification problems, we deal with multimodal distribution densities of input pattern vectors. In such situations nonparametric (local) classification rules (k-NN, Parzen window, decision trees, etc.) have to be applied [5]. Complexities of the local classification rules could be characterized by values of smoothing parameters like the number of nearest neighbors, k, in the k-NN rule or the kernel width in the Parzen window classifier. Optimal values of these parameters have to be chosen according to the sizes of the training set (N_1 and N_2). Consequently, we have a vicious circle: the complexity of the data depends on the complexity of the optimal local classifier. Optimal parameters of these classifiers depend on training sample size and data complexity.

Generally speaking, each new pattern classifier that potentially could be utilized for classification introduces one or several measures of complexity. Yet *a large number of characteristics is impractical* to determine data complexity in concrete work. As a compromise, a question arises: could some of the measures be clustered into a smaller number of groups? The question remains unanswered. Factors similar to that presented in equation (3.8) should be taken into account while trying to taxonomize the data sets obtained in real experiments — comparative measures of asymptotic errors of distinct classifiers, and the small sample properties of them.

In this chapter we discussed the possibilities and difficulties in estimating data complexity assuming some simple data models, such as the spherically symmetric Gaussian, where certain simple classifiers are known to be suitable. Based on the behavior of several popular classifiers, a number of possible measures for the difficulty of a classification task were given, each representing the perspective of some specific classifiers. Examples were shown where the same problem may appear easy or difficult depending on the classifier being used, through the influences of the density parameters and the sample sizes on the relevant measures such as the effective dimensionality and the effective Mahalanobis distance.

Returning to the GCCM data model, it is worth mentioning also that the estimation of complexity parameters from experimental data may become very problematic. For example, in estimating the effective number of features p^* [equation (3.7)], we have to estimate the means and the covariance matrix. The confidence interval in estimating parameter p^* is too wide to be practically useful. Thus, the estimation of parameter p^* is more difficult and less reliable than training the linear Fisher classifier. This fact suggests once more that the estimation of data complexity measures is not necessarily easier than training a classifier for the task.

References

[1] S. Amari, N. Fujita, S. Shinomoto. Four types of learning curves. *Neural Computation*, 4, 605–618, 1992.

[2] M. Basu, T.K. Ho. The learning behavior of single neuron classifiers on linearly separable or nonseparable input. *Proc. of IEEE Intl. Joint Conf. on Neural Networks*, July 10-16, 1999, Washington, DC.

[3] J. Cid-Sueiro, J.L. Sancho-Gomez. Saturated perceptrons for maximum margin and minimum misclassification error. *Neural Processing Letters*, 14, 217–226, 2001.

[4] R.O. Duda, P.E. Hart, D.G. Stork. *Pattern Classification and Scene Analysis*. 2nd ed. New York: John Wiley, 2000.

[5] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. 2nd ed. New York: Academic Press, 1990.

[6] T.K. Ho, M. Basu. Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24, 289–300, 2002.

[7] Y.S. Huang, C.Y. Suen. A method of combining multiple experts for the recognition of unconstrained handwritten numerals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1), 90–94, 1995.

[8] M. Li, P. Vitanyi. *An Introduction to Kolmogorov Complexity and its Applications*. New York: Springer, 1993.

[9] S. Raudys. On the problems of sample size in pattern recognition. In V. S. Pugatchiov, ed. *Detection, Pattern Recognition and Experiment Design*, volume 2, pages 64–76. Proc. of the 2nd All-Union Conference Statistical Methods in Control Theory. Moscow: Nauka, 1970 (in Russian).

[10] S. Raudys. On dimensionality, sample size and classification error of nonparametric linear classification algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 669–671, 1997.

[11] S. Raudys. Evolution and generalization of a single neuron. I. SLP as seven statistical classifiers. *Neural Networks*, 11, 283–296, 1998.

[12] S. Raudys. *Statistical and Neural Classifiers: An Integrated Approach to Design*. New York: Springer-Verlag, 2001.

[13] S. Raudys, A. Saudargiene. Tree type dependency model and sample size - dimensionality properties. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 23, 233–239, 2001.

[14] S. Raudys. Integration of statistical and neural methods to design classifiers in case of unequal covariance matrices. *Lecture Notes in Computer Science*, New York: Springer, 3238, 270–280, 2004.

[15] S. Raudys, D. Young. Results in statistical discriminant analysis: A review of the former Soviet Union literature. *Journal of Multivariate Analysis*, 89, 1–35, 2004.

[16] A. Saudargiene. Structurization of the covariance matrix by process type and block diagonal models in the classifier design. *Informatica* 10(2), 245–269, 1999.

[17] V. N. Vapnik. *The Nature of Statistical Learning Theory*. New York: Springer, 1995.

[18] V.I. Zarudskij. The use of models of simple dependence problems of classification. In S. Raudys, ed. *Statistical Problems of Control*, volume 38, pages 33–75, Vilnius: Institute of Mathematics and Informatics, 1979 (in Russian).