INFORMATICA, 1991, Vol.2, No.3, 434–454

ON THE EFFECTIVENESS OF PARZEN WINDOW CLASSIFIER

Šarūnas RAUDYS

Institute of Mathematics and Informatics,
Lithuanian Academy of Sciences,
2600 Vilnius, Akademijos St.4, Lithuania

Abstract. The smoothing constant λ is the most important characteristic of the nonparametric Parzen window classifier (PWC). The PWC tends to a one-nearest neighbour classifier as λ tends to zero and to a parametric linear Eucliden distance classifier as λ tends to infinity. An asymptotic probability of misclassification of the PWC decreases with the decrease in λ. A sensitivity of the PWC to a finiteness of the training data depends on a true-intrinsic dimensionality of the data, and it increases with the decrease in the value of λ. It is proposed to determine an optimal value of the smoothing constant from a smoothed empirical graph of the dependence of an expected probability of misclassification on the value of λ. The graph can be estimated by means of leaving-one-out or hold-out methods simultaneously for a number of values of λ chosen from the interval (0.001–1000) in a logarithmic scale.

Key words: Parzen window elascifier, the density estimate, probability density function, kernel function, smoothing parameter, classification error, probability of misclassification.

1. Introduction. Parzen window classifier (PWC) is one of the most popular classification rules in statistical pattern recognition. Its performance is rarely significantly less

than that of other classification rules (see, e.g. Table 1 of comparison of several classification rules by using 19 arrays of real data (Čepònis, 1980)). PWC is based on a Parzen or kernel estimate of the multivariate density function at X

$$\hat{f}(X|\Pi_i) = \frac{1}{N_i} \sum_{j=1}^{N_i} K\left(\frac{X - X_j^{(i)}}{\lambda} \right) \qquad (1)$$

where $X = (x_1, x_2, \ldots, x_p)'$ – is a p-variate vector to be classified; $X_j^{(i)}$ – is a j^{th} p-variate vector of the learning sample of the i^{th} class Π_i; $K(\cdot)$ is a window or kernel function with $\int_{-\infty}^{+\infty} K(t)dt = 1$, and λ is a window-width or smoothing parameter.

Most often the Gaussian kernel is used

$$K\left(\frac{X - X_j^{(i)}}{\lambda} \right) = C_1 \exp\left\{ -\frac{(X - X_j^{(i)})'(X - X_j^{(i)})}{\lambda^2} \right\}, \qquad (2)$$

where C_1 is a normalizing constant.

A more general form of the kernel function is

$$K\left(\frac{X - X_j^{(i)}}{\lambda} \right) = C_2 \exp\left\{ -\frac{(X - X_j^{(i)})' S^{-1} (X - X_j^{(i)})}{\lambda^2} \right\},$$

where S is a scaling matrix. One may use here a pooled sample covariance matrix or a diagonal variance matrix.

There are more modifications of PWC, where other types of kernel functions or only part of learning sample vectors are used (see, e.g. Fukunaga, 1972, section 6.1). Our experience with several arrays of real and artificial data and 13 types of kernel function have shown (Skurikhina, 1990) that the kernel (2) is the best after optimisation of λ, resulting the smallest

Table 1. The values of the learning quantity $\kappa = E_\chi P_\chi(\lambda)/P_{00}(\lambda)$ of the Parzen window classifier for Gaussian kernel and spherically Gaussian distributions of pattern vectors.

	$\delta = 2.56 \quad P_{00}^{(p)} = 0.1$			$\delta = 4.65 \quad P_{00}^{(p)} = 0.01$		
$k = N/p$	$p = 3$	$p = 5$	$p = 8$	$p = 3$	$p = 5$	$p = 8$
			$\lambda = 0.1$			
0.6	1.97	2.15	2.28	3.53	3.55	4.29
1	1.90	1.98	2.13	2.92	3.24	3.48
2	1.78	1.87	1.95	2.61	3.07	3.18
5	1.64	1.71	1.91	2.32	2.56	2.72
10	1.50	1.66	1.84	2.15	2.16	2.28
50	1.39	1.62	1.81	1.53	1.86	2.14
			$\lambda = 0.2$			
0.6	1.97	2.15	2.28	3.53	3.55	4.29
1	1.90	1.98	2.13	2.92	3.24	3.48
2	1.78	1.87	1.95	2.60	3.07	3.18
5	1.64	1.71	1.91	2.31	2.56	2.72
10	1.49	1.66	1.84	2.14	2.16	2.28
50	1.37	1.62	1.81	1.52	1.86	2.13
			$\lambda = 0.4$			
0.6	1.97	2.15	2.27	3.53	3.53	4.28
1	1.88	1.98	2.12	2.90	3.22	3.47
2	1.75	1.87	1.94	2.58	3.05	3.17
5	1.60	1.70	1.88	2.28	2.52	2.68
10	1.35	1.63	1.82	2.07	2.14	2.22
50	1.25	1.51	1.78	1.48	1.77	2.09
			$\lambda = 0.8$			
0.6	1.96	2.12	2.26	3.51	3.51	4.21
1	1.80	1.94	2.05	2.86	3.15	3.41
2	1.68	1.80	1.90	2.51	2.87	3.08
5	1.46	1.58	1.76	2.18	2.38	2.46
10	1.23	1.44	1.65	1.71	1.90	2.10
50	1.06	1.12	1.29	1.20	1.27	1.54

classification error. This paper analyses the simplest form of kernel function (2) with respect to its width.

Section 2 discusses the features of the PWC rule. The dependence of the classification error of PWC on the learning sample size, and the window width λ, is discussed in section 3. In section 4 we discuss the criteria and methods used to determine the optimal value of λ.

2. Features of the PWC rule. The Parzen window classification rule has some attractive features. If certain specific conditions for the kernel function $K(\cdot)$ and the smoothing parameter λ are satisfied (the most important one being $\lambda \to 0$ as $N \to \infty$), the estimate of the probability density function (1) is consistent and asymptotically unbiased. Then the classification error of the PWC rule tends to the Bayes error (Wolverton and Wagner, 1969).

Some other properties of the Parzen window classifier may be mentioned without detailed investigation. Suppose at first that $\lambda \to 0$, while the learning sample size N_i remains constant. When the parameter λ is very small, we have no smoothing and the PWC rule classifies an unknown vector X according to the class index of its nearest neighbour. Therefore, as $\lambda \to 0$, the PWC rule tends to the 1–NN classification rule.

Suppose now that $\lambda \to \infty$ and let us analyse the PWC rule with Gaussian kernel (2). As $\lambda \to \infty$,

$$\frac{(X - X_j^{(i)})'(X - X_j^{(i)})}{\lambda^2} \to 0.$$

Using only the first two terms of the Taylor expansion

$$\exp\left\{-\frac{(X - X_j^{(i)})'(X - X_j^{(i)})}{\lambda^2}\right\}$$

$$= 1 - \frac{(X - X_j^{(i)})'(X - X_j^{(i)})}{\lambda^2} + \ldots \qquad (3)$$

we have:

$$\begin{aligned}\widehat{f}(X|\Pi_i) =& 1 - \frac{1}{\lambda^2 N_i}\sum_{j=1}^{N_i}\left[(X-\bar{X}^{(i)})+(\bar{X}^{(i)}-X_j^{(i)})\right]' \\ & \times\left[(X-\bar{X}^{(i)})+(\bar{X}^{(i)}-X_j^{(i)}\right] \\ =& 1-\frac{1}{\lambda^2}\Big[\frac{1}{N_i}\sum_{j=1}^{N_i}(X-\bar{X}^{(i)})'(X-\bar{X}^{(i)}) \\ & + \frac{2}{N_i}\sum_{j=1}^{N_i}(\bar{X}^{(i)}-X_j^{(i)})'(X-\bar{X}^{(i)}) \\ & + \frac{1}{N_i}\sum_{j=1}^{N_i}(\bar{X}^{(i)}-X_j^{(i)})'(X^{(i)}-X_j^{(i)})\Big] \\ =& 1-(X-\bar{X}^{(i)})'(X-\bar{X}^{(i)})+\operatorname{tr} S_i,\end{aligned}$$

where $\bar{X}^{(i)} = \frac{1}{N_i}\sum_{j=1}^{N_i} X_j^{(i)}$ is the sample mean and

$$S_i = \frac{1}{N_i}\sum_{j=1}^{N_i}(X_j^{(i)}-\bar{X}^{(i)})(X_j^{(i)}-\bar{X}^{(i)})'$$

is the sample covariance matrix.

When sample covariance matrices are equal for all classes, the PWC rule classifies X according to the distance between X and means $\bar{X}^{(i)}$. When S_i are different, we obtain a bias of the hyperplane. Thus, with an increase in the smoothing parameter λ, the PWC tends to the well known Euclidean-distance classifier.

3. Dependence of the classification error on the learning sample size and the value of the smoothing parameter. One can distinguisn several sorts of the probability of misclassification (PMC) in classifier design.

If the underlying probability density functions $f(X|\Pi_i)$ are known an optimal Bayes classifier can be constructed. Its

performance (PWC), denoted by P_B, is referred to as *Bayes PMC.* In the case the classifier is determined on the particular learning sample χ, the PMC will depend on the characteristics of this sample. Then PMC may be regarded as a random variable $P_\chi(\lambda)$, the distribution of which depends on the learning sample size. This PMC will be called *conditional PMC.* Its expectation $E_\chi P_\chi(\lambda)$ over all learning samples will be called *expected PMC.* The *theoretical limit* $P_{00}(\lambda) = \lim_{N_i \to \infty} E_\chi P_\chi(\lambda)$ is called *asymptotic PMC.*

We shall analyse the dependence of the expected PMC on the learning sample size, the dimensionality p, the value of the smoothing parameter λ and the configuration of the pattern classes.

In the case of two equiprobable pattern classes Π_1 and Π_2 the expected PMC $E_\chi P_\chi(\lambda)$ can be written as

$$E_\chi P_\chi(\lambda) = \tag{4}$$

$$= \frac{1}{2}\int_\Omega f(X|\Pi_2)\mathrm{Prob}_\chi\left\{\hat{f}(X|\Pi_1) > \hat{f}(X|\Pi_2)\middle| X, X\epsilon\Pi_2\right\}dX$$

$$+ \frac{1}{2}\int_\Omega f(X|\Pi_1)\mathrm{Prob}_\chi\left\{\hat{f}(X|\Pi_1) \leqslant \hat{f}(X|\Pi_2)\middle| X, X\epsilon\Pi_1\right\}dX.$$

The difference $\hat{f}(X|\Pi_1) - \hat{f}(X|\Pi_2)$, asymptotically (when $N_1 \to \infty$ and $N_2 \to \infty$) has the Gaussian distribution. Therefore

$$P(X) = \mathrm{Prob}_\chi\left\{\hat{f}(X|\Pi_1) > \hat{f}(X|\Pi_2)\middle| X, X \in \Pi_2\right\}$$

$$= \Phi\left\{\frac{E_\chi\hat{f}(X|\Pi_2) - E_\chi\hat{f}(X|\Pi_1)}{\sqrt{V_\chi\hat{f}(X|\Pi_2) + V_\chi\hat{f}(X|\Pi_1)}}\right\}, \tag{5}$$

where the density estimate $\hat{f}(X|\Pi_1)$ at a fixed point of the multivariate space Ω is regarded as a random variable, whose

distribution depends on the learning sample size N_i, and $\Phi(\alpha) = \int_{-\infty}^{\alpha} n(\alpha, 0, 1)d\alpha$ is the standard univariate Gaussian distribution function. Here and below, $n(\alpha, \mu, \Sigma)$ denotes a multivariate Gaussian density with the mean vector μ and covariance matrix Σ. When we use the Gaussian kernel (2)

$$K\left(\frac{X - X_j^{(i)}}{\lambda}\right) = n(X, X_j^{(i)}, I \cdot \lambda^2)$$

and $f(X|\Pi_i) = n(X, \mu_i, \Sigma)$, the mean of the estimate

$$E_\chi \hat{f}(X|\Pi_i) = n(X, \mu_i, \Sigma + I \cdot \lambda^2) \tag{6}$$

and the variance

$$V_\chi \hat{f}(X|\Pi_i) = \frac{1}{N_i}\left[\frac{|2\Sigma + I \cdot \lambda^2|^{1/2}}{\lambda^2} n^2(X, \mu_i, 2\Sigma + I \cdot X^2) - E_\chi \hat{f}(X|\Pi_i)\right]. \tag{7}$$

Let T be a $p \times p$ orthogonal matrix such that $T\Sigma T' = D$ (D is a diagonal matrix with the elements $d_1, d_2, \ldots, d_p$). For *small values* of λ we can write

$$V_\chi \hat{f}(X|\Pi_i) \approx \frac{1}{N_i} \prod_{j=1}^{P} \sqrt{\frac{2d_j}{\lambda^2} + 1} \cdot C_i(X),$$

where $C_i(X)$ is a function of X. Then the probability (5) is asymptotically (when $N_i \to \infty$) proportional to

$$\Phi\left\{ - \frac{\sqrt{N}}{\prod_{j=1}^{P} \sqrt{\frac{2d_j}{\lambda^2} + 1}} \times \frac{n(X, \mu_2, \Sigma + I\lambda^2) - n(X, \mu_1, \Sigma + I\lambda^2)}{\sqrt{C_1(X) + C_2(X)}} \right\}. \tag{8}$$

Above we assumed $N_2 = N_1 = N$.

Equation (8) shows the relation between the learning sample size N and the dimensionality p. Let $d_1 = d_2 = \ldots = d_p = d$ and the number of variables, p, be increased. Then, in order to keep $P(X)$ constant, the learning sample size should be increased to the degree of p

$$N \equiv \left(\frac{2d}{\lambda^2} + 1\right)^P. \tag{9}$$

Equation (8) also shows that small eigenvalues d_j have little influence on an increase of the classification error. Therefore the sensitivity of the PWC rule to the learning sample size depends not on the formal dimensionality p but on the true – intrinsic dimensionality. The term "intrinsic dimensionality" is not defined exactly, and we shall not do this here. Note that when $d_1 = d_2 = \ldots = d_r = d$ and $d_{r+1} = d_{r+2} = \ldots d_p = d_0 << d$, the intrinsic dimensionality is r.

Equation (5) is an asymptotic one. For fixed N and small λ the distribution $f(\widehat{f}(X|\Pi_i))$ of a random variable $\widehat{f}(X)$ can have a very large positive asimmetry, e.g. when $p = 10$, the coefficient of asimmetry $\gamma' = \mu_3/(\mu_2)^{3/2}$ at the sphere $(X-\mu)'(X-\mu) = 4$ is

$\gamma' = 6651/\sqrt{N}$ when $\lambda = 0.01$,

$\gamma' = 13.1/\sqrt{N}$ when $\lambda = 0.8$.

Therefore approximate formulae or asymptotic expansions for the expected PMC of the PWC rule (see, e.g. Raudys 1976; Kharin, 1983) can show only the asymptotic qualitative conclusions. The only way to obtain quantitative results for finite learning sample sizes is a simulation experiment.

The most simple case: two spherically Gaussian distributions. $f(X|\Pi_i) = n(X, \mu_i, I), N_2 = N_1 = N$, equal the prior probabilities of the classes. It is obvious that, when $\lambda \to \infty$ and the PWC rule becomes similar to the Euclidean distance classifier (EDC), the sensitivity of the PWC rule to

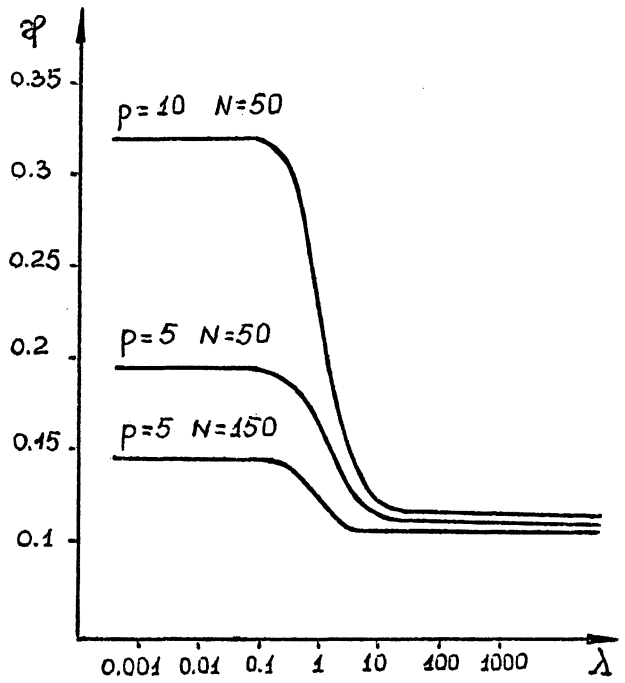

Fig. 1. Dependence of the learning quantity $\kappa = E_\chi P_\chi / P_\infty$ on the value of smoothing parameter λ two 5-variate spherically Gaussian populations.

the finiteness of the learning sample size coincides with that of EDC. The dependence of the expected PMC of EDC on N is obtained and tabulated in the paper of Raudys and Pikelis, 1980. When $\lambda \to 0$, the sensitivity increases to that of the NN classification rule. In Figure 1 we present three graphs of the dependence of a ratio $\kappa = E_\chi P_\chi(\lambda)/P_\infty(\lambda)$ on the value of the smoothing parameter λ.

We performed simulation studies for a number of different values of $P_B = P_{00}(\lambda)$, and dimensionality p. The results are presented in Table l. This data was used to find the relationship between the learning sample size N and dimensionality p. For getting this, we have drawn graphs $E_\chi P(\lambda) = f(N)$ and

found the learning sample size that was required to achieve a given expected PMC. E.g., for $P_{00}(\lambda = 0.8) = 0.1$ and $E_\chi P_\chi(\lambda = 0.8) = 0.15$ we got

$N = 12$ when $p = 3$;

$N = 38$ when $p = 5$;

$N = 210$ when $p = 8$.

Approximately for this data $N = 2.2 \cdot (1.77)^p$. When $P_{00}(\lambda = 0.8) = 0.01$ and $E_\chi P_\chi(\lambda = 0.8) = 0.015$

$N = 55$ when $p = 3$;

$N = 140$ when $p = 5$;

$N = 500$ when $p = 8$.

Approximately $N = 15 \cdot (1.55)^p$.

The same results ($N = a \cdot b^p$) were obtained for other combinations of the values of $P_{00}(\lambda)$, $E_\chi P_\chi(\lambda)$ and λ. Thus the above simulation studies confirm the theoretical conclusion: when λ is sufficiently small and fixed, the learning sample size required to achive the given learning accuracy should increase exponentially with the growth of dimensionality p.

A more complex case. Classes distributed near two arcs. In the simulation studies we generated bi-variate data according to the following formula

$$\begin{aligned} x_1 &= R_i \sin\alpha + \xi_1 \\ x_2 &= R_i \cos\alpha + \xi_2, \quad i = 1, 2, \end{aligned} \tag{10}$$

where ξ_1 and ξ_2 and α are independent random variables; $\xi_i \sim N(0, 1)$, and α is uniformly distributed at the interval $(-\Pi/3, +\Pi/3)$, $R_1 = 10$, $R_2 = 6.2$.

Here and below, before applying the Parzen window classifier, the data is normalized in such a manner that the variances of all variables are equal to 1. The Bayes error P_B, corresponding to this distribution, $P_B = \Phi(-\frac{1}{2}(R_1 - R_2)) \approx 0.028$. The dependence of the asymptotic PMC on the value of the smoothing parameter λ is estimated using very large learning and test sets and is depicted in Figure 2 (curve 1). We

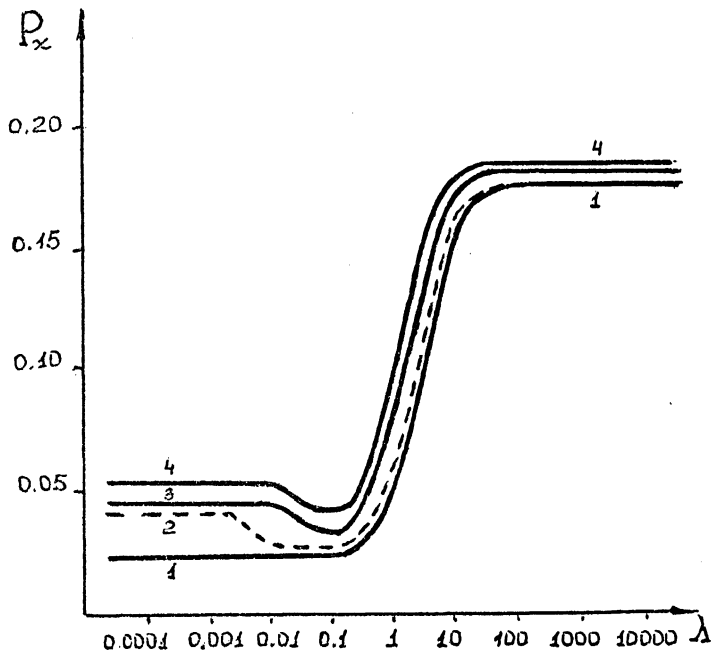

Fig. 2. Dependence of the expected PMC $E_\lambda P_\lambda$ on the value of smoothing parameter λ (articial bivariate data, learning sample size $N_L = N_1 = N_2 =$ 50, 200 and 1000).

can notice that when λ is small, the asymptotic PMC $P_{00}(\lambda)$ coincides with the Bayes error. The asymptotic error $P_{00}(\lambda)$ increases with an increases in λ until it reaches some limit. Such behaviour (the soft limited threshold function) of the asymptotic PMC is common for many real and artificial data sets studied in our experiments. The main difference between such curves is the interval of λ, where $P_{00}(\lambda)$ increases and the values $P_{00}(\lambda \to \infty)$ and $P_{00}(\lambda \to 0)$.

The dependence of the learning quantity $\kappa(\lambda) = E_\chi P_\lambda(\lambda)/P_{00}(\lambda)$ on λ is also a smoothed threshold function (see Fig. 1). Therefore the product of both functions – the expected PMC $E_\chi P_\lambda(\lambda) = P_{00}(\lambda) \cdot \kappa(\lambda)$, usually has a minimum (see curves 2, 3 and 4 in Fig. 2). Depending on the

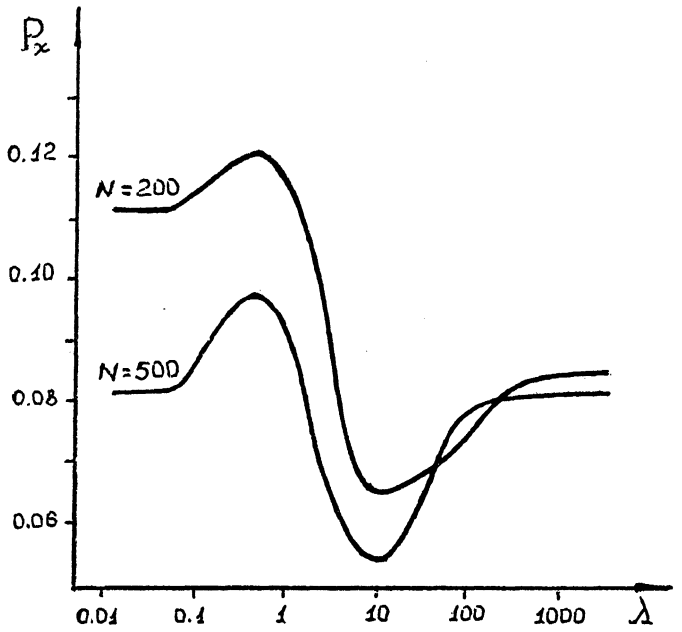

Fig. 3. Dependence of 5 conditional PMC P_χ on the value of smoothing parameter λ (real PRC–76 data).

interval, the curve $E_\chi P_\chi(\lambda) = f(\lambda)$ can have a maximum (see curves 1, 2 in Fig. 3, which were obtained for real data of Pattern Recognition Competition – 1976 (PRC–76 data)). The expected PMC EP_χ for artificial data (curves 2, 3, 4 in Fig. 2) were estimated as an arithmetic mean of 5 conditional probabilities of misclassification (CPM) $P_\chi{}^i (i = 1, 2, 3, 4, 5)$, obtained during 5 independent experiments. In each experiment the learning sampes ($N_1 = N_2 = N_L$) were different and the test sample ($N_1^t = N_2{}^t = N_T = 900$) was common. In Fig. 4 we present 5 graphs of CPM $P_\chi{}^i$ and their arithmetic mean EP_χ, obtained for $p = 2, N_L = 50$. We see that a deviation of values P_χ^i is rather great. However, the optimal value of the smoothing parameter is practically the same for all 5 curves. The graphs of EP_χ for the real data were obtained by the leave-one-out method.

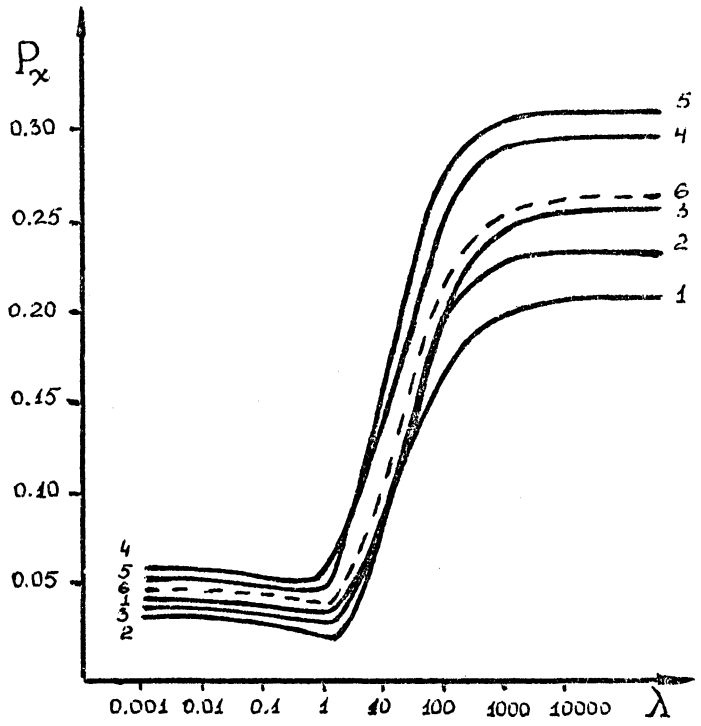

Fig. 4. Dependence of 5 conditional PMC P_λ (graphs 12345) and estimates of expected $E_\lambda P_\lambda$ (graph 6) on the value of smoothing parameter λ.

It is interesting to trace the influence of dimensionality on the character of the curve $P_\lambda(\lambda) = f(\lambda)$. In Fig. 5 we present graphs for artificial data obtained from the presented above bi-variate artificial data by adding 18 new variables

$$x_i = {x_2}^2 + \sigma \cdot \xi_i, \qquad i = 3, 4, \ldots, 20, \tag{11}$$

where ξ_i are independent standart Gaussian random variables and σ^2 is the variance of a noise, $\sigma = 1$.

In Fig. 6 we present graphs $P_\lambda(\lambda) = f(\lambda)$ for the real PRC–76 data after a linear transformation, perfomed with the help of the principal component method.

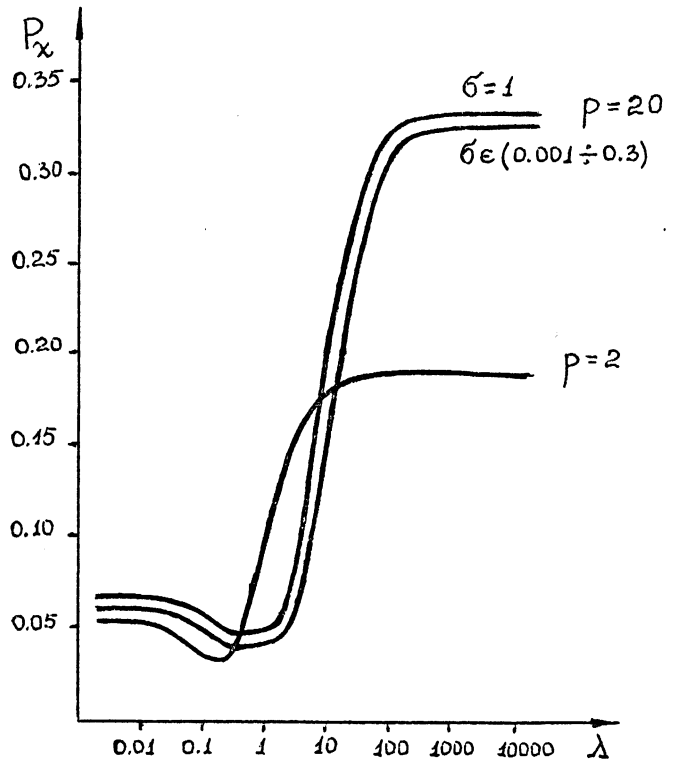

Fig. 5. Dependence of the conditional PMC P_χ on the value of smoothing parameter λ, dimensionality and noise level σ (artificial data, $p = 2$ and 20, learning sample size $N_L = 50$, test sample size $N_T = 900$).

At the beginning of this section it was shown that the sensitivity of the PWC rule to the learning sample size for small λ depends not on the true but on the intrinsic dimensionality of data. The graphs in Fig. 6 confirm this conclusion. This experiment was carried out with transformed data, where the variances of the last variables were many times less than those of the few first variables. Therefore the graphs $E_\chi P_\lambda(\lambda) = f(\lambda)$ for $p = 10$ and $p = 27$ are very close.

A special simulation experiment to verify this conclusion

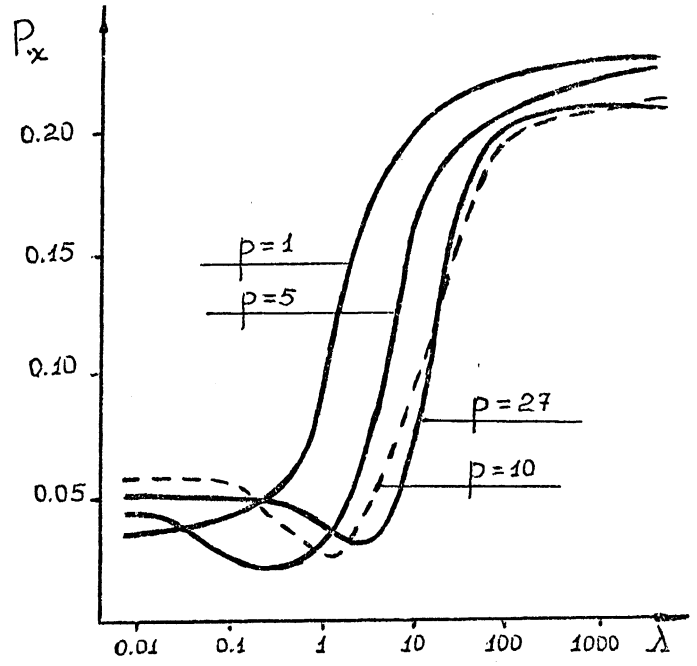

Fig. 6. Dependence of conditional PMC P_λ on the value of smoothing parameter λ (real PRC–76 data after orthogonal transformation, learning sample size $N_L = 500$, test sample size $N_T = 500$).

was performed on artificial data (formulae (11) and (10)) with 20 variables. In a series of experiments the variance σ^2 of the noise ξ_i in (11) was changed. The graps of dependence $P_\lambda(\lambda) = f(\lambda)$ for different values of σ (Fig. 5) in the case, when σ is very small ($\sigma = 0.001$), indicate that the values of the expected PMC for $p = 20$ and small values of λ are very close to the values of PMC, obtained for $p = 2$.

4. Determination .of the optimal value of a smoothing parameter. The above simulation results show that the value of a smoothing parameter appears to be the crucial factor in determining the performance of the PWC rule. In most experiments the optimal value of the smoothing pa-

rameter λ_{opt} was found in the interval (0.1,5.). However for transformed PRC–76 data, λ_{opt} has gradually changed from 0.003 (for p = 1) up to 2 (for p = 25, see Fig 6.).

Different criteria have been used to determine λ_{opt}. Popular criteria are the mean square error (MSE).

$$\begin{aligned} \text{MSE} &= E\big(f(X) - \widehat{f}(X)\big)^2 \\ &= \int_{\Omega} \big(f(X) - \widehat{f}(X)\big)^2 = f(X)dX \end{aligned} \tag{12}$$

and the integral square error (MISE)

$$\text{MISE} = \int \big(f(x) - \widehat{f}(x)\big)^2 dX. \tag{13}$$

When the true density $f(X)$ is multivariate Gaussian or a mixture of Gaussian densities and the Gaussian kernel is used, criteria (12) and (13) are easily computed. In Table 2 we present the optimal values of λ found by minimizing MISE criterion, when $f(X) = n(X, \mu, I)$.

Table 2. Optimal values of λ from MISE criteria (Raudys, 1977)

p *N*	30	80	200	500	2000	10000
1	.584	.468	.383	.315	.236	.170
5	.767	.663	.582	.515	.431	.353
20	1.08	1.01	.947	.894	.822	.748

The estimate of λ_{opt} were obtained for the MISE criterion too. However, both criteria resulted different estimates of λ_{opt}. In both criteria the values of λ_{opt} do not depend on the configuration of the data to be classified and the distance between the pattern classes. The values of λ_{opt} differ also

from the values of λ_{opt} found from the experimental graphs, presented in Fig. 1–7. E.g., for two spherical distributions the optimal value of the smoothing parameter should possibly be large (see Fig. 1), etc.

Therefore it follows a very important conclusion: *optimization of the PWC rule should be carried out from the criterion of minimum classification error.*

The simplest way to find λ_{opt} is to plot $P_\chi(\lambda)$ against λ and to choose that λ, which results the minimal value of $EP_\chi(\lambda)$.

The PWC rule is usually used in the case of complicated multimodal distributions of pattern classes. Until now theoretical results for a classification error of the PWC rule are not obtained even for the most simple distributions $f(X|\Pi_i)$. In the case when such a formula is known, it would be very difficult to use it since the true distributions $f(X|\Pi_i)$ are unknown. Therefore in order to find *in an optimal value of the smoothing parameter the nonparametric estimates of the classification error should be used.*

There are several nonparametric estimates of the expected PMC.

Hold out method H. Here the whole design sample is divided into two independent parts: a learning sample LS and a test sample TS.

Cross validation method CV. Here the whole design sample is divided into R equal parts. The learning sample consists of R–1 parts and the test sample consists of the remaining part. The procedure of designing and testing a classifier is carried out R times, changing the test sample each time.

Leaving-one-out method L is a version of the CV method when $R = N_1 + N_2$.

In the multidimensional case most of computer time in calculating the nonparametric Parzen window estimates

$\widehat{f}(X|\Pi_i)$ is vasted to calculate $N_1 + N_2$ distances

$$H(X, X_j^{(i)}) = (X - X_j^{(i)})'(X - X_j^{(i)}).$$

Thus, in order to reduce the computer time, it is convenient to calculate

$$H(X, X_j^{(i)})/\lambda_S \quad (S = 1, 2, \ldots, r)$$

right away after finding $H(X, X_j^{(i)})$ and estimate all 2r estimates of the multivariate density $\widehat{f}(X|\Pi_1, \lambda_1), \ldots, \widehat{f}(X|\Pi_1, \lambda_r), \widehat{f}(X|\Pi_2, \lambda_1), \ldots, \widehat{f}(X|\Pi_2, \lambda_r)$ simultaneously.

In order to determine the optimal value of the smoothing parameter of the PWC rule we propose the following procedure:

1) to select several values of the smoothing parameter λ in a logarithmic scale (from our experience with normalized data we propose to select 10 values at an interval (0.001, 100), e.g.: 0.001, 0.01, 0.03, 0.1, 0.3, 1.0, 3, 10, 30, 100).

2) to obtain unbiased estimates of the expected PMC for all values of λ simultaneously. If observations of the design sample are statistically independent, we propose to use the leaving-out-method which can very easily be implemented in the PWC rule. If observations are statistically dependent (e.g., the design sample consists of multivariate time series) CV or H methods are preferable. Note, when λ is very large, i.e., we have very large smoothing, all the observations of the design sample have approximately equal contribution into the sum (1) and sometimes the L method results an 100 % error.

3) to draw the graph $\widehat{P}_\chi = f(\ln \lambda)$ (when $N_1 + N_2$ exceeds 100, the graph $\widehat{P}_\chi = f(ln\lambda)$ is rather smooth (see Fig. 7).

4) to determine λ_{opt}, where $\widehat{P}_\chi = f(\ln \lambda)$ has its minimum.

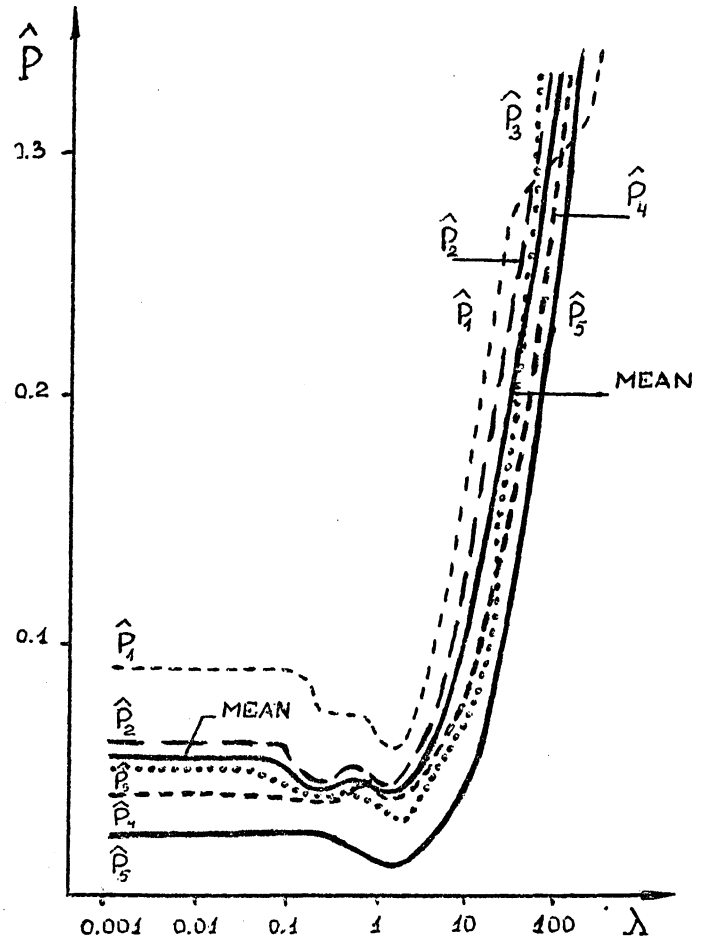

Fig. 7. Dependence of five leave-one-out estimates of the classification error of PWC on the value of smoothing parameter (artificial 5-variate data, 5 samples, $N_1 = N_2 = 50$).

5. Discussion. The Parzen window classification rule is one of the best among other statistical classification rules. When the learning sample size increases unlimitedly and the smoothing parameter λ tends to zero, the PWC rule tends to an optimal Bayes one. When $\lambda \to 0$, PWC tends to the

nearest neighbour classification rule, and when $\lambda \to \infty$ – to the Euclidean distance classifier. When the learning sample size is limited, the expected probability of misclassification increases. The sensitivity of the PWC rule to the learning sample size depends on the value of the smoothing parameter, on the true–intrinsic dimensionality and other characteristics of pattern classes. The sensitivity of the PWC rule to the finiteness of the learning sample size is lowest, when $\lambda \to \infty$, and the PWC rule becomes Euclidean distance classifier, and it increases with a decrease of λ. The asymptotic PMC is lowest for very small λ and it increases with λ. The dependence of the expected PMC on the value of smoothing parameter is a complex shaped function which has its minimum and maximum, and this shape depends on the data significantly.

Determining the optimal value of the smoothing parameter, where the expected PMC has its minimum, is the most important problem in practice. Therefore we have to obtain the graph of dependence of the expected PMC on parameter λ and to find λ_{opt} from its minimum. Since the nonparametric PWC rule is devoted to form complex nonlinear decision boundaries, we have to use nonparametric estimates of the expected probability of misclassification.

REFERENCES

Jain, A.K., and M.D. Ramaswami (1988). Classifier design with Parzen windows. In E.S. Gelsema and L.N. Kanal (Eds.), *Pattern Recognition and Artificial Intelligence,* Elsevier SSiPubl., pp. 211–228.

Čeponis, K. (1986). Experimental comparison of transformations approximating the distribution of initial features to the Gaussian distribution in discriminant analysis using real data. In *Statistical Problems of Control,* Vol.74. Inst. Math. Cyb. Press, Vilnius. pp. 67–75.

Fukunaga K. (1972). *Introduction to Statistical Pattern Recognition.* N.Y. Academic Press, New York.

Raudys, Š., and V.Pikelis (). On dimensionality, sample size, classification error and complexity of classification algorithm in recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence,* **PAMI-2**(2), 242–252.

Raudys, Š (1977). On the accuracy of some estimates of the multivariate density function. *Trans. of 7th Prague Conf. on Information Theory,* **A**, Publ. House of the Czekoslovak Acad. of Sciences, Prague. 429–439.

Skurikhina, M. (1990). Effect of the kernel function form on the quality of nonparametric Parzen window classifier. In *Statistical Problems of Control,* Vol.93. Inst. Math. Cyb. Press, Vilnius. pp. 167–181.

Received August 1991

Š. Raudys received the M.S. degree in electrical and computer engineering from Kaunas Polytechnical Institute in 1963, and the Candidate of Sciences and Doctor of Sciences degree from the Institute of Mathematics and Cybernetics, Academy of Sciences, Lithuania, in 1969 and 1978, respectively.

He is currently Head of the Department of Data Analysis in the Institute of Mathematics and Cybernetics and Professor in the Department of Control Systems at Kaunas Polytechnical Institute, Lithuania. His current research interests include statistical pattern recognition, artificial neural nets, expert systems, machine learning, and data analysis methods.

Dr.Raudys is a member of the Classification Societies of the U.S.S.R. and France. He is an Associate Editor of the Pattern Recognition Journal. He has been a member of the Program Committee of INTERFACE-90 and other Soviet and international conferences.