Available online at www.sciencedirect.com

SCIENCE DIRECT®

Pattern Recognition Letters 27 (2006) 1151–1159

Pattern Recognition Letters

www.elsevier.com/locate/patrec

A local mean-based nonparametric classifier

Y. Mitani [a,*], Y. Hamamoto [b]

[a] *Ube National College of Technology, Department of Intelligent System Engineering, 2-14-1, Tokiwadai, Ube 755-8555, Japan*
[b] *Faculty of Engineering, Yamaguchi University, Ube 755-8611, Japan*

Received 21 February 2005; received in revised form 16 November 2005
Available online 23 February 2006

Communicated by Prof. F. Roli

Abstract

A considerable amount of effort has been devoted to design a classifier in practical situations. In this paper, a simple nonparametric classifier based on the local mean vectors is proposed. The proposed classifier is compared with the 1-NN, *k*-NN, Euclidean distance (ED), Parzen, and artificial neural network (ANN) classifiers in terms of the error rate on the unknown patterns, particularly in small training sample size situations. Experimental results show that the proposed classifier is promising even in practical situations.

Keywords: Classifier design; Nearest neighbor samples; Small training sample size; Dimensionality

1. Introduction

A classifier design in statistical pattern recognition is very fundamental. In small training sample size situations (Raudys and Jain, 1991), it is difficult to assume the normal distributions, in general. Thus, one may prefer nonparametric classifiers, e.g., a nearest neighbor classifier (called the 1-NN classifier Cover and Hart, 1967) and a Parzen classifier which are expected to be effective rather than parametric classifiers. Therefore, a large number of studies about 1-NN, *k*-NN, and Parzen classifiers have been published (Fukunaga, 1990; Hamamoto et al., 1997; Dasarathy, 1991; Jain and Ramaswami, 1988; Van Ness, 1980). However, we note that the number of training samples required by many standard nonparametric classifiers (e.g., the Parzen one) to meet desired values of probability of correct classification is known from the literatures (Raudys and Jain, 1991; Cover and Hart, 1967; Fukunaga, 1990) to grow exponentially with the number of features. The nonparametric classifiers need many samples because the parametric form of the density function cannot be assumed. This may be a severe drawback, when dealing with small-sample size situations.

It is well known that the nonparametric classifiers usually suffer from the existing outliers (Fukunaga, 1990). That is, the performance of nonparametric classifiers is severely in Influenced by the outliers, particularly in small training sample size situations. In order for a classifier to be practical, it should be robust to outliers. To design a reliable classifier, we explore the use of the local mean vectors when classifying unknown patterns (Mitani and Hamamoto, 2000). As compared with the previous conference proceedings (Mitani and Hamamoto, 2000), we add the following: one is the optimization of the parameter values of the proposed classifier by the cross validation (CV) method. We add the various experimental artificial and real data sets, and conduct the experiments of the influences of the dimensionality and the dependences of the parameter values of the proposed classifier. We also investigate the comparative study with the typical nonparametric classifiers, i.e., the Parzen and artificial neural network (ANN) classifiers. Furthermore, the robustness for outliers of the proposed classifier is discussed.

The classification performance of the proposed classifier is examined on both the artificial and real data sets. The

[*] Corresponding author. Fax: +81 836 354767.
E-mail address: mitani@ube-k.ac.jp (Y. Mitani).

0167-8655/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2005.12.016

proposed classifier is compared with the 1-NN, k-NN (Cover and Hart, 1967), Euclidean distance (ED), Parzen (Jain and Ramaswami, 1988), and ANN (Rumelhart et al., 1986) classifiers in terms of the error rate on unknown patterns. The experimental results show the effectiveness of the proposed classifier in such practical situations.

2. Proposed classifier

In order to design a simple and robust classifier, we will propose the following:

Proposed classifier: Let $X^i = \{\boldsymbol{x}_j^i | j = 1, \ldots, N_i\}$ be a training sample set from class ω_i, where N_i is the number of training samples from class ω_i. In the proposed classifier, a pattern $\boldsymbol{x}$ is classified into class ω_c by the following:

Step 1. Select r nearest neighbor training samples from $\boldsymbol{x}$ with Euclidean distance measure for each class ω_i. Here, a value of r must be selected by ranging from 1 to N_i.

Step 2. Compute the local mean vector, $\boldsymbol{y}^i$, using r nearest neighbor training samples, $\{\boldsymbol{x}_{k_1}^i, \boldsymbol{x}_{k_2}^i, \ldots, \boldsymbol{x}_{kr}^i\}$:

$$\boldsymbol{y}^i = \frac{1}{r}\sum_{j=1}^{r} \boldsymbol{x}_{k_j}^i.$$

Step 3. Classify $\boldsymbol{x}$ into class ω_c if

$$(\boldsymbol{x} - \boldsymbol{y}^c)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y}^c) = \min_i (\boldsymbol{x} - \boldsymbol{y}^i)^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{y}^i).$$

Note that the proposed classifier is equivalent to the 1-NN classifier when $r = 1$, and to the ED classifier when $r = N_i$. The computational cost of the proposed classifier may be relatively low. The parameter r must be optimized for each given data (Mitani and Hamamoto, 2000). The optimization approach with the cross validation method (Toussaint, 1974) is considered. The CV method is one of the tools in estimating the error rate. In the experiments, we will select the optimal value of r^* which minimizes the error rate estimated by the CV method.

The CV error rate is estimated as follows:

Step 1. For each class, divide N available training samples into the training and test samples at random. Here, the numbers of training and test samples are $N - N_T$ and N_T, respectively.

Step 2. Design the proposed classifier with the value of r using $N - N_T$ samples.

Step 3. Estimate the error rate $e_j(r)$ of the proposed classifier using N_T samples.

Step 4. Repeat steps 1–3 J times to get the following:

$$E_{\mathrm{CV}}(r) = \frac{1}{J}\sum_{j=1}^{J} e_j(r). \tag{1}$$

$E_{\mathrm{CV}}(r)$ can be influenced by N_T and J. In our experiments, we used $N_T = 1$ and $J = 100$. Thus, the optimal value of r^* obtained by the CV method is given by

$$r^* = \arg\min_r \{E_{\mathrm{CV}}(r)\}. \tag{2}$$

In the proposed classifier, the parameter r^* is obtained by the CV method by ranging from 1 to $N - N_T$, for each given data.

The four nonparametric classifiers, 1-NN, k-NN (Cover and Hart, 1967), Parzen (Jain and Ramaswami, 1988), and ANN (Rumelhart et al., 1986) classifiers, are compared, in terms of the average error rate. These are the typical nonparametric classifiers. Here, the average error rate is obtained by repeating the experiments. Fig. 1 shows the procedure of the error rate estimation. Note that the parameter optimization by the CV approach is made only by using the training samples statistically independent of the test samples. In the experiments, the k-NN with CV means that the k-NN classifier is optimized by the CV approach. The optimal value of k is selected in the same ranges for the parameter values of r. The Parzen and ANN classifiers are briefly described.

Parzen classifier: The Parzen classifier is known to be a Bayes type nonparametric classifier. First, we define the probability density function characterized by a kernel function $\kappa(\cdot)$ and its window width h,

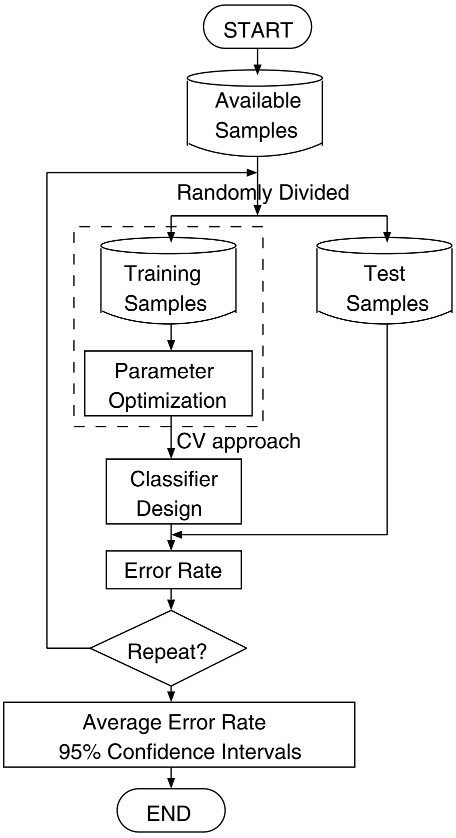

Fig. 1. Procedure of the error rate estimation.

$$\hat{p}(\boldsymbol{x}|\omega_i) = \frac{1}{N_i}\sum_{j=1}^{N_i}\frac{1}{h_i^n}\kappa\left(\frac{\boldsymbol{x}-\boldsymbol{x}_j^i}{h_i}\right),$$

where n indicates the dimensionality. The kernel function $\kappa(\cdot)$ is defined by the n-dimensional normal distribution:

$$\kappa(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\hat{\Sigma}_i|^{1/2}}\exp\left\{-\frac{1}{2}\boldsymbol{x}^{\mathrm{T}}\hat{\Sigma}_i^{-1}\boldsymbol{x}\right\}.$$

The sample covariance matrix $\hat{\Sigma}_i$ is usually estimated by

$$\hat{\Sigma}_i = \frac{1}{N_i-1}\sum_{j=1}^{N_i}(\boldsymbol{x}_j^i-\hat{\boldsymbol{\mu}}_i)(\boldsymbol{x}_j^i-\hat{\boldsymbol{\mu}}_i)^{\mathrm{T}},$$

where $\hat{\boldsymbol{\mu}}_i$ is the sample mean vector from class ω_i.

Next, the Bayes classifier is obtained by using the estimated probability density function, then a pattern $\boldsymbol{x}$ is classified into class ω_c if

$$P_c\hat{p}(\boldsymbol{x}|\omega_c) = \max_i P_i\hat{p}(\boldsymbol{x}|\omega_i),$$

where P_i is the a prior probability of class ω_i. In the experiments, we adopted that $h = h_1 = h_2 = \cdots = h_L$, for simplicity. Here, L is the number of classes. In the Parzen classifier, we must select the optimal value of h which influences the error rate. In the experiments, we chose the candidate value of h ranging from 0.2 to 2.2. In the Parzen classifier, the minimum error rate with the optimal value of h was found.

ANN classifier: The artificial neural network (ANN) classifiers with one hidden layer are used. The neurons in the input layer correspond to the dimensionality of the pattern to be classified. The hidden layer has m neurons. In the experiments, we examined the value of m ranging from 8 to 256, and the minimum error rate with the optimal value of m was used. The neurons in the output layer are usually associated with pattern class labels. In the case of a 2-class problem, the output layer has two neurons. The BP algorithm (Rumelhart et al., 1986) was used to train the ANN classifiers. The momentum term was not used for simplicity. The rate of convergence is affected by the learning rate c. We used $c = 0.1$. Initial weights were distributed uniformly in -0.5 to 0.5. Learning was terminated when the mean-squared error dropped below a specified threshold, or when there is little change in the mean-squared error. Here, the maximum number of iterations was set to 10,000. The features of real data sets were normalized to have zero mean and unit variance before being input to the ANN classifier. The ith feature was normalized by

$$\tilde{x}_i = \frac{x_i - \bar{x}_i}{\sigma_i}, \tag{3}$$

where $\bar{x}_i$ and σ_i are the sample mean and standard deviation of the ith feature, respectively. Each test sample was also normalized in a similar manner.

3. Experimental data

The error rate is the most effective measure of the performance of a classifier. In order for the estimated error rate to be reliable in predicting the future classification performance of the classifier, the training and test samples must be statistically independent (Devijver and Kittler, 1982). Therefore, the available samples must be divided into the training and test samples. In the experiments, the statistical independence between the training and test samples is maintained. In order to clarify the classification performance of the proposed classifier, we used both the artificial and real data sets.

3.1. Artificial data sets

With the artificial data sets, we can easily independently generate two sets of data, the training and test sets, from the same model by a computer. There are two advantages in using artificial data. (a) We can control the number of the available samples. Therefore, we chose the training sample size which directly influences difficulties of the pattern recognition problem, so as to set the small training sample size situations in the experiments. On the other hand, we used large test samples in order to reduce the influence of the test sample size on the error rate. In the experiments, we used 1000 test samples for each class. (b) For any two-class problem, the true Bayes error can be obtained (Fukunaga and Krile, 1969).

We briefly describe four Gaussian data sets. We used the I-Λ, I-4I, I-I (Fukunaga, 1990) and Ness data sets (Van Ness, 1980). In these artificial data sets, $\boldsymbol{\mu}_i$ is the mean vector and Σ_i is the covariance matrix from class ω_i.

[I-Λ data set]

The I-Λ data set consists of 8-dimensional Gaussian data. The true Bayes error is 1.9%.

$$\boldsymbol{\mu}_1 = \boldsymbol{0}, \quad \boldsymbol{\mu}_2 = [\mu_1, \mu_2, \ldots, \mu_8]^{\mathrm{T}}, \tag{4}$$

$$\Sigma_1 = I_8, \quad \Sigma_2 = \mathrm{diag}[\lambda_1, \lambda_2, \ldots, \lambda_8], \tag{5}$$

where I_k and diag[$\cdot$] denote the $k \times k$ identity and diagonal matrices, respectively. Values of μ_i and λ_i are shown in Table 1.

I-4I data set: The I-4I data set consists of 8-dimensional Gaussian data. The true Bayes error is about 9%.

$$\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{0}, \tag{6}$$

$$\Sigma_1 = I_8, \quad \Sigma_2 = 4I_8. \tag{7}$$

I-I data set: The I-I data set consists of p-dimensional Gaussian data. In this data, the dimensionality p can be controlled. The true Bayes error is 10%.

$$\boldsymbol{\mu}_1 = \boldsymbol{0}, \quad \boldsymbol{\mu}_2 = [2.56, 0, \ldots, 0]^{\mathrm{T}}, \tag{8}$$

$$\Sigma_1 = \Sigma_2 = I_p. \tag{9}$$

Table 1
Parameter values of I-Λ data

i	1	2	3	4	5	6	7	8
μ_i	3.86	3.10	0.84	0.84	1.64	1.08	0.26	0.01
λ_i	8.41	12.06	0.12	0.22	1.49	1.77	0.35	2.73

Ness data set: The Ness data set consists of p-dimensional Gaussian data.

$$\boldsymbol{\mu}_1 = \mathbf{0}, \qquad \boldsymbol{\mu}_2 = [\Delta/2, 0, \ldots, 0, \Delta/2]^{\mathrm{T}}, \tag{10}$$

$$\Sigma_1 = I_p, \quad \Sigma_2 = \begin{bmatrix} I_{p/2} & O \\ O & \frac{1}{2} I_{p/2} \end{bmatrix}. \tag{11}$$

In the Ness data set, the true Bayes error can be controlled by varying the values of Δ and p. In the experiments, we used $\Delta = 2$, 4, and 6, and also varied $p = 2$–50. The true Bayes errors are ranging from 21.7% to 7.6% ($p = 2$–50) when $\Delta = 2$, 6.2–2.5% ($p = 2$–50) when $\Delta = 4$, and 1.1–0.5% ($p = 2$–50) when $\Delta = 6$.

Our primary interests in studying the classification performance of the proposed classifier are in small training sample size situations. In such situations, the classification performance of the proposed classifier was examined in terms of the error rate. The number of trials was 100. Fresh samples were artificially generated by a computer on each trial. The 100 error rates were averaged and the 95% confidence interval was also calculated.

3.2. Real data sets

The real data sets were 8OX (Jain and Dubes, 1988), Iris (Fisher, 1936), IMOX (Jain and Ramaswami, 1988), and Gabor (Hamamoto et al., 1998) data sets. The 8OX, Iris, and IMOX data sets have been extensively used in the past. In (Schmidt et al., 1994), these have been used in the experimental comparison of the ANN classifier with the other classifiers. The 8OX data set consists of 45 8-dimensional data, which are extracted from 8, O, and X of the Munson's database. The Iris data set consists of 150 4-dimensional data: four features are sepal- and petal-length and breadth measured from Iris setosa, I. versicolor, and I. virginica. The IMOX data set consists of 192 8-dimensional data which is the IEEE data file of letters I, M, O, and X. The Gabor data set consists of 14,000 128-dimensional data. This was extracted from 10 handprinted numerals of the ETL-1 database (Yamamoto, 1968). The notation of the Gabor parameters follows (Hamamoto et al., 1998). The Gabor parameters are shown in Table 2.

In error estimation literature, the holdout method (Fukunaga, 1990) has been successfully used, because it maintains the statistical independence between the training and test samples. In order to evaluate the classification performance of the proposed classifier on various real data sets, the average of the error rate and the 95% confidence interval were obtained by the holdout method as follows. Note that the suffix to be represented as the class label was omitted.

Step 1. Divide randomly the available samples into N training and T test samples for each class.
Step 2. Design a classifier using N training samples, and estimate the error rate of the classifier using T test samples.
Step 3. Repeat t times from step 1 to step 2.
Step 4. Compute the average of the error rate and the 95% confidence interval.

Table 2
Values of the Gabor parameters

No. of the orientations	2
Wavelength λ	$4\sqrt{2}$
x-Direction standard deviation σ_x	0.5λ
y-Direction standard deviation σ_y	0.5λ
Extended frame size α	2

4. Experimental results

4.1. Experiment 1

The purpose of Experiment 1 is to investigate the dependence of the parameter r on the error rate on the unknown patterns. One must determine the optimal value of the parameter r which influences the error rate. Varying the values of the parameter r, we estimated the error rates on both the artificial and real data sets. In Experiment 1, assuming that each class has an equal prior probability, i.e., $P(\omega_1) = P(\omega_2)$, we set $p = 8$, $N_1 = N_2 = 16$ for the artificial data. For the real data, the experiments were conducted as shown in Table 3 (Type I).

Fig. 2 shows the plots of the error rate as a function of r on various data sets. The 95% confidence intervals with vertical line are also shown. Fig. 2 shows that the parameter r strongly influences the error rate. Experimental results suggest that the optimal value of r which minimizes the error rate may exist for the given data set. From any experimental results, the optimal value of r seems to have a global minimum of the error rate curve. This means that the optimization problem is not so difficult. In the following experiments, the value of r is optimized by the CV approach.

4.2. Experiment 2

The purpose of Experiment 2 is to investigate the classification performance of the proposed classifier as a function of the training sample size N. Our primary interest in studying the classification performance of the proposed classifier is in small training sample size situations. The effect of the training sample size needs to be addressed explicitly. We conducted the experiments on the I-Λ, I-4I,

Table 3
Experimental conditions on the real data sets

Real data set	8OX	Iris	IMOX	Gabor
No. of classes	3	3	4	10
Dimensionality p	8	4	8	128
(Type I) training sample size N/class	10	10	10	100
(Type II) training sample size N/class	5	5	5	50
Test sample size T/class	5	40	38	1200
No. of trials t	100	100	100	5

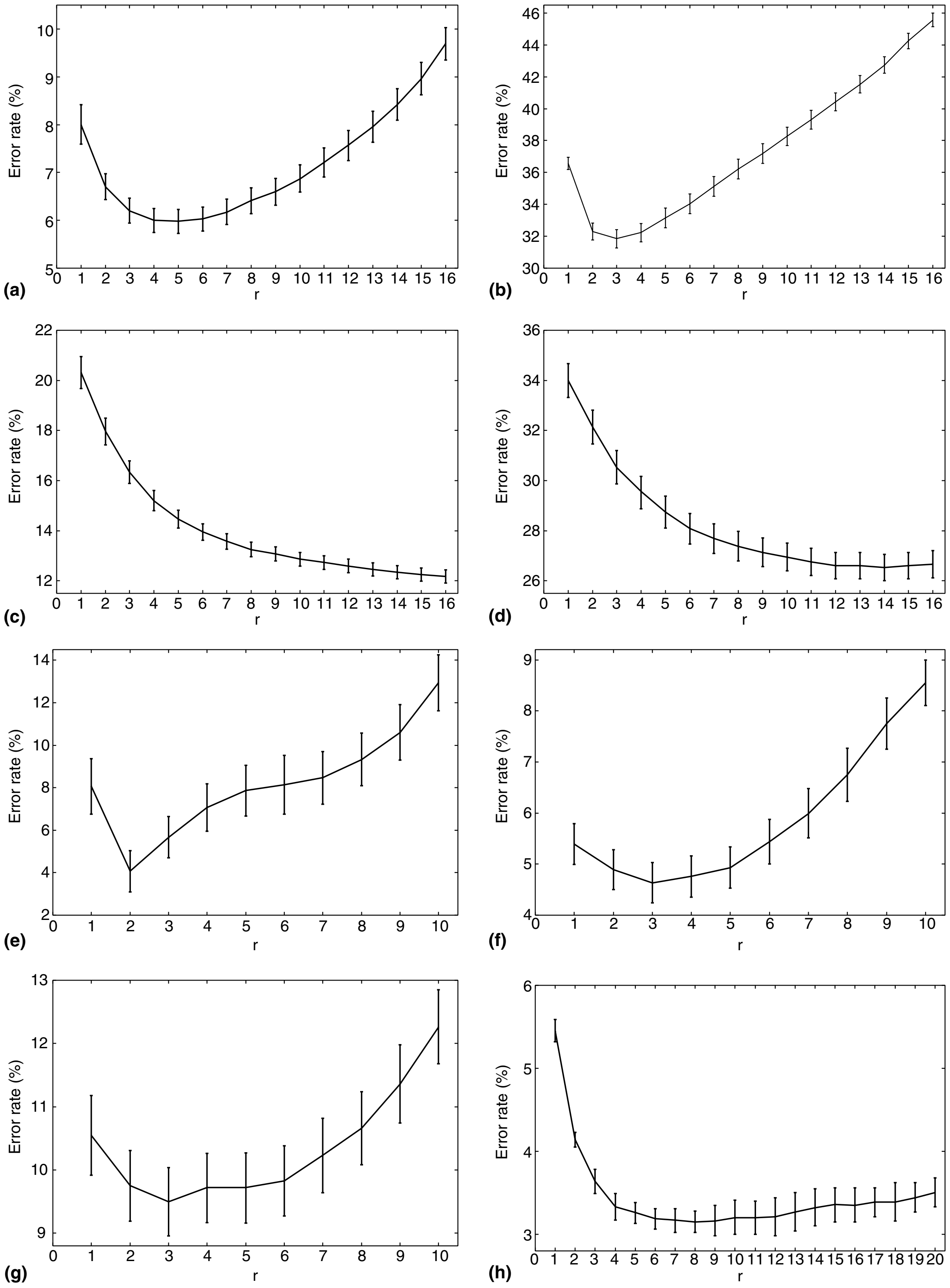

Fig. 2. Influences of the parameter r on the error rate (%). (a) I-Λ data set, (b) I-4I data set, (c) I-I data set, (d) Ness data set ($\Delta = 2$), (e) 8OX data set, (f) Iris data set, (g) IMOX data set and (h) Gabor data set.

I-I ($p = 8$ and 50), and Ness ($\Delta = 4$, $p = 8$ and 50) data sets. We varied $N = 4$–80 when $p = 8$, and $N = 10$–50 when $p = 50$. In Experiment 2, we compared the proposed classifier with the k-NN with CV, 1-NN, 3-NN, Parzen, and ANN classifiers, in terms of the error rate on the unknown patterns. Note that when the ratio of the training sample size to the dimensionality is less than 1, the error rates of the Parzen classifier are not available.

Fig. 3 shows the plots of the error rate as a function of N on various data sets. The 95% confidence intervals with

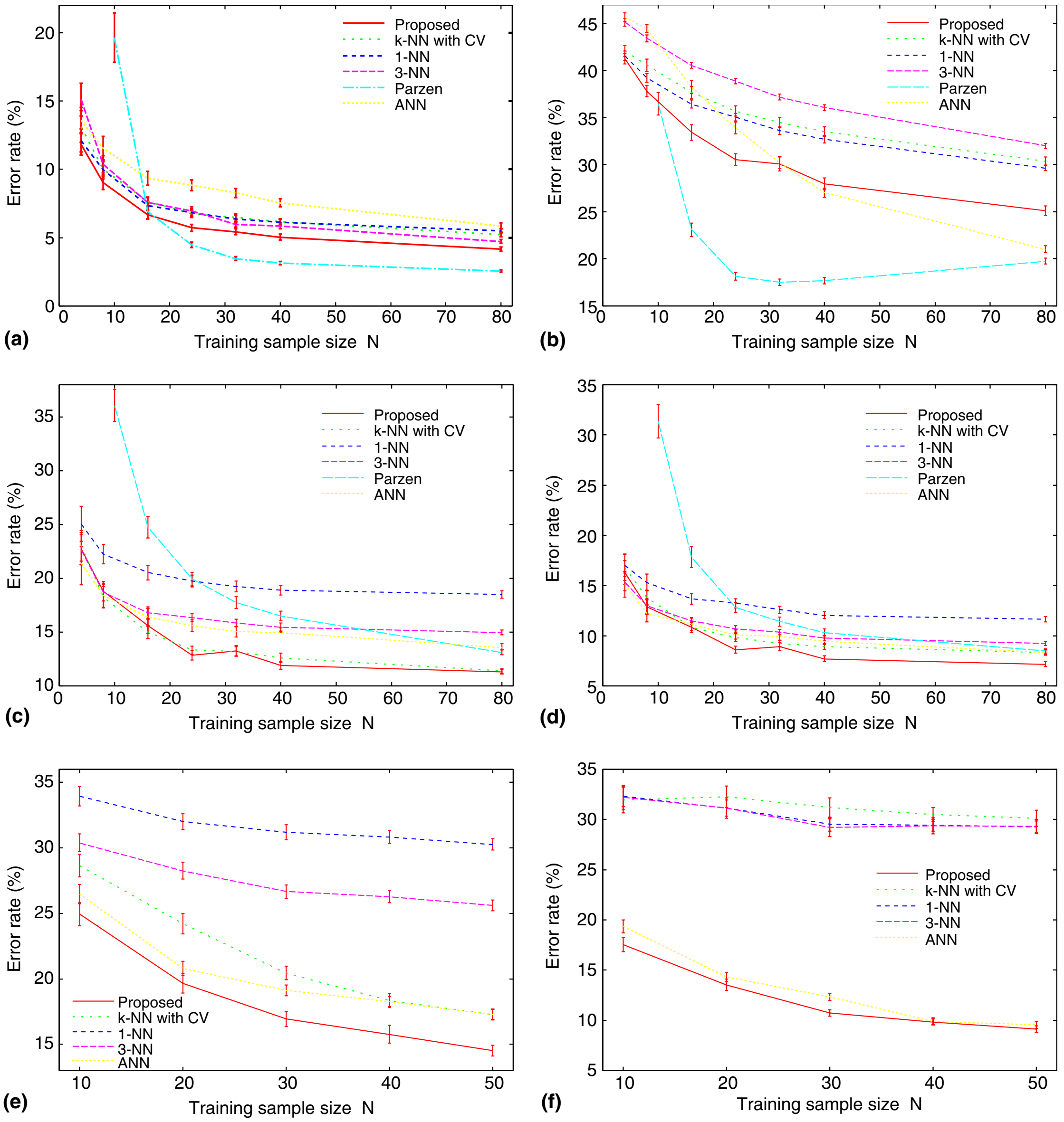

Fig. 3. Influences of the training sample size N on the error rate (%). (a) I-Λ data set, (b) I-4I data set, (c) I-I data set ($p = 8$), (d) Ness data set ($\Delta = 4$, $p = 8$), (e) I-I data set ($p = 50$) and (f) Ness data set ($\Delta = 4$, $p = 50$).

vertical line are also shown. From Fig. 3, the winner differs for the given data. For the I-Λ and I-4I data sets, the Parzen classifier works well. However, when the number of the training samples is small, the proposed classifier outperforms the other classifiers. In the experiments for the I-I and Ness ($\Delta = 4$) data sets, the proposed classifier is usually superior to the other classifiers. In high dimensional space, the proposed and ANN classifiers show a favorable behavior.

4.3. Experiment 3

The purpose of Experiment 3 is to study the classification performance of the proposed classifier in high dimensional spaces. The error rate of nonparametric classifiers suffers from the dimensionality p, which is called a 'curse of the dimensionality' (Jain and Chandrasekaran, 1982). Thus, we examined the behavior of the proposed classifier in high dimensions. The experiments were conducted by ranging from $p = 2$ to 50 on the I-I and Ness data ($\Delta = 2$, 4 and 6) sets, with the fixed number of training samples, $N_1 = N_2 = 10$. In Experiment 3, we compared the proposed classifier with the k-NN with CV, 1-NN, 3-NN, Parzen, and ANN classifiers, in terms of the error rate on the unknown patterns.

Fig. 4 shows the plots of the error rate as a function of the dimensionality p for each of six classifiers. The 95% confidence intervals with vertical line are also shown. From Fig. 4, we see that as the dimensionality increases, the error rates of the proposed and ANN classifiers increase more

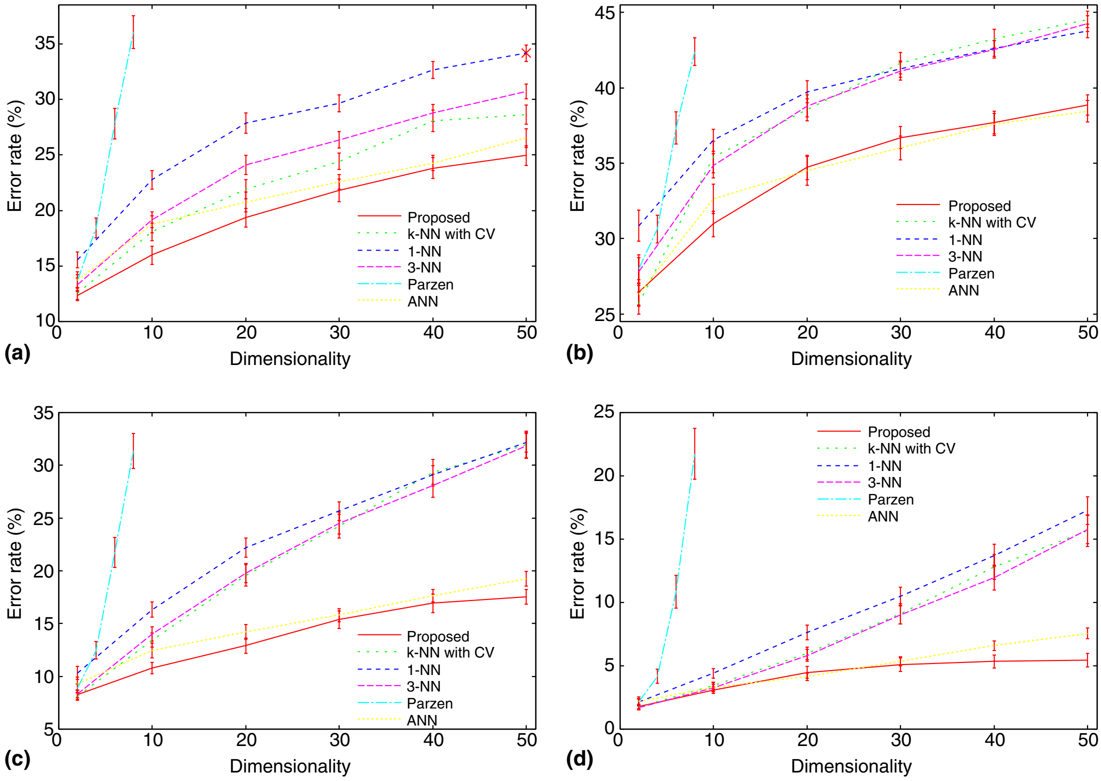

Fig. 4. Influences of the dimensionality p on the error rate (%). (a) I-I data set, (b) Ness data set ($\Delta = 2$), (c) Ness data set ($\Delta = 4$) and (d) Ness data set ($\Delta = 6$).

Table 4
Comparison of eight classifiers conducted by Table 3 (Type I) in terms of the error rate (%)

	8OX	Iris	IMOX	Gabor
Proposed	5.47	4.91	9.79	3.76
	4.26, 6.67	4.46, 5.37	9.23, 10.34	3.53, 4.00
k-NN with CV	9.67	5.75	12.24	6.19
	8.11, 11.21	5.19, 6.32	11.37, 13.11	5.76, 6.63
1-NN	8.07	5.39	10.55	6.07
	6.76, 9.37	4.99, 5.79	9.92, 11.18	6.56, 7.02
3-NN	12.73	5.17	13.84	6.79
	11.21, 14.26	4.75, 5.59	13.02, 14.65	6.56, 7.02
5-NN	18.07	5.73	17.01	7.22
	16.47, 19.66	5.26, 6.21	16.16, 17.86	7.06, 7.38
ED	12.93	8.55	12.26	11.84
	11.61, 14.26	8.10, 9.00	11.68, 12.85	10.98, 12.70
Parzen	26.13	5.98	24.61	NA
	23.96, 28.30	5.17, 6.78	23.21, 26.00	
ANN	8.67	4.83	13.23	4.43
	7.42, 9.91	4.40, 5.26	12.48, 13.99	4.08, 4.78

slowly than those of the *k*-NN with CV, 1-NN, 3-NN, and particularly Parzen classifiers. This suggests that the proposed and ANN classifiers may be more robust to the dimensionality than the *k*-NN with CV, 1-NN, 3-NN, and Parzen classifiers. The robustness to the dimensionality of the ANN classifier was shown in (Hamamoto et al., 1996). The proposed classifier, as well as the ANN classifier (Hamamoto et al., 1996), show a favorable behavior in high dimensions. The error rate of the Parzen classifier has been severely influenced in high dimensions. This means that the Parzen classifier is often an impractical approach. Our limited results show that the classification performance of the proposed classifier is superior to those of the other classifiers, or equivalent to the ANN classifier performance.

4.4. Experiment 4

The purpose of Experiment 4 is to investigate the classification performance of the proposed classifier on the real data sets well known in pattern recognition fields. Using the real data sets, the proposed classifier was compared with the *k*-NN with CV, 1-NN, 3-NN, 5-NN, ED, Parzen, and ANN classifiers, in terms of the error rate on the unknown patterns. In Experiment 4, we conducted the experiments as shown in Table 3 (Type I and Type II). The Type I and Type II are in the practical situations of the small training sample size. The Type II is half of the training sample size compared with the Type I.

Results are shown in Table 4 (Type I) and 5 (Type II). In Tables 4 and 5, the first line represents the average of the error rate and the second line is the 95% confidence interval. In most results, the proposed classifier is superior to the other classifiers, or equivalent to the ANN classifier, in terms of the error rate. The effectiveness of the proposed classifier was definitely verified on the real data sets.

Table 5
Comparison of eight classifiers conducted by Table 3 (Type II) in terms of the error rate (%)

	8OX	Iris	IMOX	Gabor
Proposed	10.47	6.65	13.49	5.48
	8.87, 12.06	6.05, 7.25	12.77, 14.21	4.74, 6.21
k-NN with CV	17.87	7.84	17.15	8.07
	15.94, 19.78	7.10, 8.58	16.07, 18.23	7.38, 8.76
1-NN	12.33	7.13	15.30	8.07
	10.90, 13.77	6.53, 7.73	14.56, 16.04	7.38, 8.76
3-NN	18.93	7.62	21.53	9.19
	17.13, 20.74	6.96, 8.27	20.58, 22.49	8.60, 9.77
5-NN	26.80	8.83	28.39	9.96
	24.71, 28.89	8.19, 9.48	27.26, 29.51	9.41, 10.52
ED	16.07	8.53	14.81	12.30
	14.14, 17.99	8.00, 9.05	14.18, 15.44	11.79, 12.81
Parzen	NA	NA	NA	NA
ANN	10.33	6.56	16.65	5.97
	8.79, 11.87	5.94, 7.18	15.62, 17.68	5.77, 6.17

5. Discussion

In statistical pattern recognition, it is well known that the outlier directly influences the classification performance, particularly in small training sample size situations (Fukunaga, 1990). For avoiding the influences of the outlier, the classifier which is taken into account of using the local mean vector of some nearest neighbor samples from the unknown pattern for each class has been proposed. It is difficult to explain theoretically why the proposed classifier is robust to outlier. Thus, we empirically explain a favorable situation when the proposed classifier works well. The Iris data set which is a well-known real data set in pattern recognition fields is used. Using the feature selection method, we reduce the dimensionality from 4 to 2 for visualization. In this paper, the Fisher criterion which measures the class separability is adopted. Assume that we have N_j training samples from class ω_j, and then the Fisher criterion $F(i)$ is defined by

$$F(i) = \frac{\sum_{j=1}^{L-1}\sum_{k=j+1}^{L} P(\omega_j)P(\omega_k)(\mu_{ij} - \mu_{ik})^2}{\sum_{j=1}^{L} P(\omega_j)\sigma_{ij}^2},$$
$$P(\omega_i) = \frac{N_i}{\sum_{k=1}^{L} N_k}, \qquad (12)$$

where L is the number of classes, and μ_{ij} and σ_{ij}^2 denote the average and variance on feature i for class ω_j, respectively. In this method, first, the value of the Fisher criterion is computed. Then, the features are ranked in the order of decreasing value of the Fisher criterion, and the first 2 features are used. Fig. 5 shows an example of the plots of the Iris data set. The plots include 10 training samples for each class and a test sample $\boldsymbol{x}$ from class 2. In this situation, the 1-NN classifier leads to misclassification, because an outlier from class 3 exists. On the other hand, the proposed classifier is expected to work well, in order to reduce the influence of the outlier. For each value of r, the proposed classifier classify $\boldsymbol{x}$ into the class label which has a nearest local mean vector $\boldsymbol{y}^i$. Table 6 denotes the distance $\sqrt{(\boldsymbol{x}-\boldsymbol{y}^i)^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{y}^i)}$ in terms of a value of r, and the result of classification. The underline shows the minimum distance for a value of r. From the results of Table 6, the proposed classifier with small values of r leads to misclassification. This means that a test sample $\boldsymbol{x}$ from class 2 is badly affected by the outlier from class 3. On the other hand, when $r \geqslant 6$, the proposed classifier successfully classifies a test sample $\boldsymbol{x}$. Then, it seems that the proposed classifier overcomes the outlier by using the local mean vector.

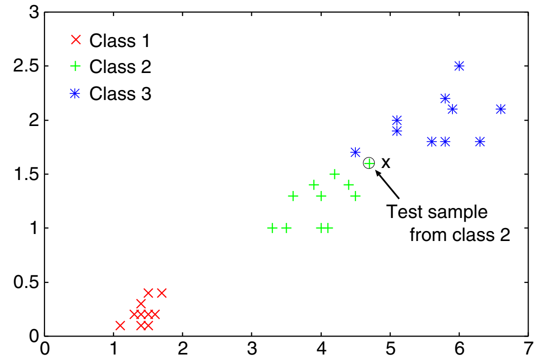

Fig. 5. An example of the plots of the Iris data set.

Table 6
Distance from a test sample to the local mean vector for a value of r, and the result of classification, in the case of Fig. 4

	Class 1	Class 2	Class 3	Result of classification
$r = 1$	3.23	0.36	$\underline{0.22}$	×
$r = 2$	3.77	0.69	$\underline{0.52}$	×
$r = 3$	3.87	0.92	$\underline{0.79}$	×
$r = 4$	3.91	1.07	$\underline{0.99}$	×
$r = 5$	3.94	1.17	$\underline{1.15}$	×
$r = 6$	3.96	$\underline{1.23}$	1.27	○
$r = 7$	3.99	$\underline{1.27}$	1.36	○
$r = 8$	4.01	$\underline{1.32}$	1.45	○
$r = 9$	4.03	$\underline{1.38}$	1.53	○
$r = 10$	4.06	$\underline{1.44}$	1.62	○

The underline denotes the minimum distance for a value of r.

6. Conclusions

In this paper, a local mean-based nonparametric classifier has been proposed. The proposed classifier was compared with the 1-NN, *k*-NN, ED, Parzen, and ANN classifiers in terms of the error rate on the unknown patterns, in small training sample size situations. Our limited experimental results show the following: (a) the proposed classifier usually outperforms the other classifiers in terms of the error rate, regardless of the training sample size and the dimensionality. (b) In most real data sets used, the proposed classifier is superior to the other classifiers, or equivalent to the ANN classifier, in terms of the error rate. Since the proposed classifier has been shown to be very effective even in practical situations, and the proposed classifier can be comparatively easily designed, we strongly recommend to use it in designing a practical pattern recognition system.

Acknowledgement

We would like to thank the Electro technical Laboratory in Japan for providing the ETL-1 database.

References

Cover, T.M., Hart, P.E., 1967. Nearest neighbor pattern classification. IEEE Trans. IT-13, 21–27.

Dasarathy, B.V. (Ed.), 1991. Nearest Neighbor Norms: NN pattern Classification Techniques. IEEE Computer Society Press.

Devijver, P.A., Kittler, J., 1982. Pattern Recognition: A Statistical Approach. Prentice Hall.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Ann. Eugen. 7, 179–188.

Fukunaga, K., 1990. Introduction to Statistical Pattern Recognition, second ed. Academic Press.

Fukunaga, K., Krile, T.F., 1969. Calculation of Bayes recognition error for two multivariate Gaussian distributions. IEEE Trans. C-18, 220–229.

Hamamoto, Y., Uchimura, S., Tomita, S., 1996. On the behavior of artificial neural network classifiers in high-dimensional spaces. IEEE Trans. PAMI-18 (5), 571–574.

Hamamoto, Y., Uchimura, S., Tomita, S., 1997. A bootstrap technique for nearest neighbor classifier design. IEEE Trans. PAMI-19 (1), 73–79.

Hamamoto, Y., Uchimura, S., Watanabe, M., Yasuda, T., Mitani, Y., Tomita, S., 1998. A Gabor filter-based method for recognizing handwritten numerals. Pattern Recognition 31 (4), 395–400.

Jain, A.K., Chandrasekaran, B., 1982. Dimensionality and sample size considerations in pattern recognition practice. In: Krishnaiah, P.R., Kanal, L.N. (Eds.), Handbook of Statistics, 2. North-Holland Publishing Company, pp. 835–855.

Jain, A.K., Dubes, R.C., 1988. Algorithms for Clustering Data. Prentice Hall.

Jain, A.K., Ramaswami, M.D., 1988. Classifier design with Parzen window. In: Gelsema, E.S., Kanal, L.N. (Eds.), Pattern Recognition and Artificial Intelligence. Elsevier Science Publishers B.V., North-Holland.

Mitani Y., Hamamoto Y., 2000. Classifier design based on the use of nearest neighbor samples, in: Proceedings of 15th International Conference on Pattern Recognition, Barcelona, vol. 2, pp. 773–776.

Raudys, S.J., Jain, A.K., 1991. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. IEEE Trans. PAMI-13 (3), 252–264.

Rumelhart, D.E., Hinton, G.E., Williams, R.J., 1986. Learning internal representations by error propagation. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition. MIT Press, Chapter 8.

Schmidt, W.F., Levelt, D.F., Duin, R.P.W., 1994. An experimental comparison of neural classifiers with 'traditional' classifiers. In: Gelsema, E.S., Kanal, L.N. (Eds.), Pattern Recognition in Practice IV. Elsevier Science B.V., pp. 391–402.

Toussaint, G.T., 1974. Bibliography on estimation of misclassification. IEEE Trans. IT-20, 472–479.

Van Ness, J., 1980. On the dominance of non-parametric Bayes rule discriminant algorithms in high dimensions. Pattern Recognition 12, 355–368.

Yamamoto, K., 1968. Present state of recognition method on consideration of neighbor points and its ability in common database. IEICE Trans. Inform. Systems E79-D5, 417–422.