グラフィクスとCAD 60－14
(1992. 12. 18)

標本数が少ない状況下における識別器の評価

浜本義彦 † 内村俊二 ‡ 金岡泰保 † 富田真吾 †

† 山口大学工学部 ‡ 大島商船高等専門学校

755 山口県宇部市常盤台 2557 742-21 山口県大島郡大島町小松 1091-1

あらまし 統計的パターン認識における主要な問題は，識別器を設計することである。識別器は，有限個の訓練サンプルを用いて学習される。特に，識別性能の高い識別器を得るためには，大量の訓練サンプルが必要である。ところが，実際のパターン認識問題では，特徴数に対する訓練サンプル数の比が小さい場合が多い。そこで，そのような状況下で，誤識別率の意味で，いずれの識別器が最良かという疑問が生じる。本論文では，人工データを用いて，よく知られている識別器を識別性能の観点から比較する。

和文キーワード パターン認識，識別器，誤識別率，サンプルサイズ，特徴数

Evaluation of classifiers in a small sample size case

Yoshihiko Hamamoto † Shunji Uchimura ‡ Taiho Kanaoka † Shingo Tomita †

† Yamaguchi University ‡ Oshima National College of Maritime Technology

2557 Tokiwadai, Ube, 755, Japan 1091-1 Komatsu, Oshima-cho, 742-21, Japan

Abstract The main problem in statistical pattern recognition is to design a classifier. The classifier must be learned from the training samples. A large training sample is essential to design a classifier with a very low error probability. In many practical situations, the ratio of sample size to dimensionality is small. One may ask which of classfiers is best in terms of the error probability, when the ratio of sample size to dimensionality is small. We compare the classification performances of the well-known classifiers on three artificial data sets.

英文 key words pattern recognition, classifier, error probability, sample size, dimensionality

1 まえがき

パターン認識における識別器の目的は，任意に与えられたパターンを，しかるべきクラスに対応づけることである。その際，誤って識別されるパターンの数が少ない程，良い識別器と考えられる。統計的パターン認識論によると，誤識別率が最小となる識別器は，Bayes識別器であることが知られている。このBayes識別器は，クラスの条件付き確率密度関数が既知であることを前提にしている。しかし，現実のパターン認識問題では，確率密度関数が未知である。そこで，無作為に抽出された標本を用いて，識別器の学習というアプローチがとられた。学習には，例えば正規分布などの分布を仮定するパラメトリックな手法と，分布を仮定しないノンパラメトリックな手法がある。いずれにしても，有限個の標本を用いて，識別器の学習および評価を行わなければならない。特に，識別器の学習に用いられる標本は訓練サンプルと，また評価に用いられる標本はテストサンプルと呼ばれている。

さて，実際の認識においては，パターンから種々の特徴が抽出され，得られた特徴の組で構成される特徴空間上において識別器が学習される。このとき，特徴数に比べ大量の訓練サンプルを用いた識別器については，学習が十分なされていると考えられる。識別器の学習において，特徴数に対する訓練サンプル数の比が重要である，ということは多くの研究者により指摘されてきた[1][2][3][4]。Jainは，特徴数に対する訓練サンプル数の比が10以上あることが望ましいと述べている。上述のBayes識別器は，訓練サンプル数が無限にあるとき得られるものである。Bayes識別器は特殊な例としても，従来のパターン認識論では訓練サンプル数が十分多いということを暗に仮定してきた。しかしながら，現実のパターン認識問題では，特徴数に対する訓練サンプル数の比が小さい場合が多い。例えば，手書き漢字認識問題では，訓練サンプル数が高々160であるのに対し特徴数は64で，その比が2.5という状況で認識実験がなされている[5]。一般に，特徴数に対する訓練サンプル数の比が小さい状況下では，十分な学習が行われなく，結果として誤識別率の増加という問題が生じている。

このようなパターン認識論と現実のパターン認識問題との間のギャップを縮小しようとする研究が，近年盛んに行われるようになった[6][1]。本論文もこれらの研究と立場を同じくするもので，より現実の状況下で真に有効な識別器の解明をその目的としている。

2 準備

本論文では，2クラス問題のみを論じ，2クラスの事前確率は等しいとする。まず，諸記号の説明を行う。

n	:	特徴空間の次元数(特徴数)
$\boldsymbol{x}_{ij}$	:	クラスiのj番目の訓練サンプル
N_i	:	クラスiの訓練サンプル数
$N = N_1 + N_2$	:	全訓練サンプル数
$\boldsymbol{\mu}_i$	:	クラスiの母平均ベクトル
Σ_i	:	クラスiの母共分散行列
$\hat{\boldsymbol{\mu}}_i = \frac{1}{N_i}\sum_{j=1}^{N_i}\boldsymbol{x}_{ij}$	:	クラスiの標本平均ベクトル
$\hat{\Sigma}_i = \frac{1}{N_i-1}\sum_{j=1}^{N_i}(\boldsymbol{x}_{ij}-\hat{\boldsymbol{\mu}}_i)(\boldsymbol{x}_{ij}-\hat{\boldsymbol{\mu}}_i)^T$	:	クラスiの標本共分散行列
$\hat{W} = \frac{1}{2}(\hat{\Sigma}_1 + \hat{\Sigma}_2)$	:	標本クラス内共分散行列
λ_i^s	:	$\hat{\Sigma}_i$のs番目の固有値
$\boldsymbol{\phi}_i^s$	:	$\hat{\Sigma}_i$のs番目の固有ベクトル

また，$\boldsymbol{x}^T$ は $\boldsymbol{x}$ の転置を，$|A|$ は行列 A の行列式を，A^{-1} は A の逆行列を，$\|\boldsymbol{x}-\boldsymbol{y}\|$ は$\boldsymbol{x}$と$\boldsymbol{y}$との間のユークリッド距離を表わす。なお，固有値は $\lambda_i^1 \geq \cdots \geq \lambda_i^n$ と順序づけられている。

2.1 誤識別率の推定法

誤識別率の推定法としては，Resubstitution法，Leave-one-out法，Holdout法が提案されている[7]。

まず，Resubstitution法は，識別器を設計するために用いる訓練サンプルと誤識別率を推定するために用いるテストサンプルとを同一とするもので，低い方へ偏った誤識別率が得られる，という問題を含んでいる。これは，訓練サンプルとテストサンプルとの独立性が保たれていないことによる。この独立性が重要であることをFukunagaらは指摘している[8]。訓練サンプルとテストサンプルとの独立性を保つ手法が，Leave-one-out法とHoldout法である。

Leave-one-out法は，利用できる標本数が限られるとき，それを最大限に活用する手法として提案されたものである[9]。これは，いま N 個の標本が利用できるとすると，N 個の標本の中からテストサンプルとして1個の標本を取り出し，残りの $N-1$ 個の標本を訓練サンプルとして識別器の学習を行い，得られた識別器でテストサンプルの識別を行うものである。この処理を N 回繰り返して誤識別率を求める。しかし，Leave-one-out法により推定された誤識別率の分散は大きい，ということが指摘されている[10]。

一方，Holdout法は，利用できる標本を訓練サンプルとテストサンプルとに分割し，誤識別率を求めるものである。このHoldout法では，実在データを対象とした場合，利用できる標本数に限りがあるため，訓練サンプル数およびテストサンプル数の不足により偏りと分散の大きい誤識別率しか得られない。また，訓練サンプルとテストサンプルの分割法も問題となる。しかし，人工データを対象とした場合，利用できる標本数に制約がないため，上述の訓練サンプル数，テストサンプル数の不足および分割法の問題が解消される。また，実在データに対しては一般にBayes誤識別率が未知であるのに対し，人工データについてはBayes誤識別率が既知である，という長所もある。このため，真のBayes誤識別率が既知という条件のもとで，識別器の評価を行うことができる。

以上のことから，本論文では人工データ集合を用いたHoldout法により誤識別率の推定を行う。

2.2 人工データ集合

本論文では，母平均ベクトル$\boldsymbol{\mu}_i$，母共分散行列Σ_i $(i = 1,2)$ の正規分布に従う3種類の人工データ集合 [7] を用いる。いずれの人工データ集合も8次元特徴空間上で記述されている。いま，$\boldsymbol{\mu}_i$, Σ_i $(i = 1,2)$ をそれぞれ

$$\boldsymbol{\mu}_1 = [0 \cdots 0]^T, \quad \boldsymbol{\mu}_2 = [m_1 \cdots m_8]^T$$

$$\Sigma_1 = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}, \quad \Sigma_2 = \begin{pmatrix} r_1 & & 0 \\ & \ddots & \\ 0 & & r_8 \end{pmatrix}$$

とおくと，$m_i, r_i, (1 \leq i \leq 8)$ の値により，次の3種類のデータ集合が得られる。

1. データ集合1

   $$m_1 = 2.56,\ m_2 = \cdots = m_8 = 0,$$

   $$r_1 = \cdots = r_8 = 1$$

   このデータ集合は，2クラスの母平均ベクトルのみが異なり，母共分散行列は共に単位行列である。データ集合1については，真のBayes誤識別率が10%である。

2. データ集合2

   $$m_1 = \cdots = m_8 = 0,$$

   $$r_1 = \cdots = r_8 = 4$$

   このデータ集合は，2クラスの母平均ベクトルが等しく，母共分散行列のみが異なる。データ集合2については，真のBayes誤識別率が9%である。

3. データ集合3

   表1: データ集合3のパラメータ

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

   このデータ集合は，m_i, r_i $(1 \leq i \leq 8)$ が表1で示される値をとるもので，2クラスの母平均ベクトル，母共分散行列が共に異なる。データ集合3は，MarillとGreenによりアルファベット手書き文字A,Bのデータから得られたものである [11]。このデータ集合3については，真のBayes誤識別率が1.9%であることが報告されている [12]。

2.3 評価された識別器

(a) Fisherの線形識別関数 $G_j(\boldsymbol{x})$ [13]

これは，標本平均ベクトルの差異に基づいて識別がなされるもので，

$$G_j(\boldsymbol{x}) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \hat{W}^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) \tag{1}$$

で与えられる。Fisherの線形識別関数では，

$$G_i(\boldsymbol{x}) = \min\{G_1(\boldsymbol{x}), G_2(\boldsymbol{x})\} \tag{2}$$

ならば，入力パターン$\boldsymbol{x}$はクラスiに属するものと決定する。以下，この識別器をFLDFと呼ぶ。

(b) 二次識別関数 $H_j(\boldsymbol{x})$ [13]

これは，標本平均ベクトルおよび標本共分散行列の差異に基づいて識別がなされるもので，

$$H_j(\boldsymbol{x}) = (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)^T \hat{\Sigma}_j^{-1} (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j) + \log|\hat{\Sigma}_j| \tag{3}$$

で与えられる。二次識別関数では，

$$H_i(\boldsymbol{x}) = \min\{H_1(\boldsymbol{x}), H_2(\boldsymbol{x})\} \tag{4}$$

ならば，入力パターン$\boldsymbol{x}$はクラスiに属するものと決定する。以下，この識別器をQDFと呼ぶ。

(c) 修正二次識別関数 $M_j(\boldsymbol{x})$ [5]

これは，標本共分散行列の推定誤差を低減する工夫を取り入れたもので，

$$M_j(\boldsymbol{x}) = \frac{1}{\lambda_j^{\xi+1}} \Big[\|\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j\|^2 - \sum_{s=1}^{\xi} (1 - \frac{\lambda_j^{\xi+1}}{\lambda_j^s}) \{\boldsymbol{\phi}_j^{sT} \cdot (\boldsymbol{x} - \hat{\boldsymbol{\mu}}_j)\}^2 \Big] + \log(\prod_{s=1}^{\xi} \lambda_j^s \cdot \prod_{s=\xi+1}^{n} \lambda_j^{\xi+1}) \tag{5}$$

で与えられる。ここで，$\xi (0 \leq \xi \leq n-1)$ は実験的に定められる定数である。修正二次識別関数では，

$$M_i(\boldsymbol{x}) = \min\{M_1(\boldsymbol{x}), M_2(\boldsymbol{x})\} \tag{6}$$

ならば，入力パターン$\boldsymbol{x}$はクラスiに属するものと決定する。以下，この識別器をMQDFと呼ぶ。

(d) k最近傍識別器 [13]

これは，パターン分布の形を仮定しないノンパラメトリック識別器の一つである。

N個の全訓練サンプルの中で，入力パターン$\boldsymbol{x}$とのユークリッド距離がk番目に小さい訓練サンプルを$\boldsymbol{x}^k$とする。全訓練サンプルに対し，

$$\|\boldsymbol{x} - \boldsymbol{x}^*\| \leq \|\boldsymbol{x} - \boldsymbol{x}^k\| \tag{7}$$

なる訓練サンプル$\boldsymbol{x}^*$のうち，クラスjに属する訓練サンプルの個数を$v_k^j(\boldsymbol{x})$とする。k最近傍識別器では，

$$v_k^i(\boldsymbol{x}) = \max\{v_k^1(\boldsymbol{x}), v_k^2(\boldsymbol{x})\} \tag{8}$$

ならば，入力パターン$\boldsymbol{x}$はクラスiに属するものと決定する。以下，この識別器をk-NNと呼ぶ。

(e) ユークリッド距離識別器 ED

これは，最も簡単な識別器で，これでは，

$$\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}_i\| = \min\{\|\boldsymbol{x}-\hat{\boldsymbol{\mu}}_1\|, \|\boldsymbol{x}-\hat{\boldsymbol{\mu}}_2\|\} \tag{9}$$

ならば，入力パターン$\boldsymbol{x}$はクラスiに属するものと決定する。以下，この識別器をEDと呼ぶ。

(f) Parzen 識別器 [14][15]

Parzen識別器は，ノンパラメトリック識別器の一つである。

まず，訓練サンプル$\boldsymbol{x}_{ij}$を用いて，クラスiの条件付き確率密度関数$\hat{p}(\boldsymbol{x}|i)$を推定する:

$$\hat{p}(\boldsymbol{x}|i) = \frac{1}{N_i}\sum_{j=1}^{N_i} K(\boldsymbol{x}-\boldsymbol{x}_{ij}) \tag{10}$$

ここで，$K(\cdot)$は核関数と呼ばれるものである。本論文では，核関数Kに

$$K(\boldsymbol{x}-\boldsymbol{x}_{ij}) = \frac{1}{(2\pi)^{n/2}h^n|\hat{\Sigma}_i|^{1/2}}\exp\Big\{\frac{-1}{2h^2} (\boldsymbol{x}-\boldsymbol{x}_{ij})^T\hat{\Sigma}_i^{-1}(\boldsymbol{x}-\boldsymbol{x}_{ij})\Big\} \tag{11}$$

を用いる [15]。ここで，hは実験的に定められるウィンドウ幅である。Parzen識別器では，

$$\hat{p}(\boldsymbol{x}|i) = \max\{\hat{p}(\boldsymbol{x}|1), \hat{p}(\boldsymbol{x}|2)\} \tag{12}$$

ならば，入力パターン$\boldsymbol{x}$はクラスiに属するものと決定する。以下，この識別器をPARZENと呼ぶ。

3 計算機シミュレーション

3.1 手順

本論文で重要な役割を果たすパラメータηを次のように定義する。

$$\eta = N_i/n \tag{13}$$

すなわち，ηは特徴数に対する訓練サンプル数の比である。本実験では，$n=8, N_1=N_2$とし，ηを$1,2,4,6,8,10$と変え，それぞれについて識別器を設計し，その誤識別率を推定した。精度の高い誤識別率を推定するためには，大量のテストサンプルが必要である。そこで，訓練サンプルとは独立に発生させたテストサンプルを各クラスとも1000個用意した。以上の処理を独立に100回繰り返し，誤識別率の平均値，標準偏差および95%信頼区間を求めた。なお，ξの最適値は次のように求めた。ξの値を，$0,1,2,3,4,5,6,7$と変え，それぞれについて誤識別率を求めた。それらの中で誤識別率を最小とする値を，ξの最適値とした。hの最適値も，hの値を0.2から2.2まで0.2ずつ変えて求めた。

3.2 実験結果と検討

データ集合1に対する実験結果を表2に，データ集合2に対する実験結果を表3に，データ集合3に対する実験結果を表4にそれぞれ示す。ξの最適値とηとの関係，hの最適値とηとの関係をそれぞれ表5，表6に示す。

一般に，ηの値が小さくなるにつれて，パターン分布の正規分布からのずれが大きくなっていく。特に，ηの値が1,2の場合，正規性の仮定は適当でないと考えられる。

実験結果から以下のことが指摘される。

1. ηの値が小さくなるにつれて，いずれの識別器の誤識別率もBayes誤識別率からの偏りおよび標準偏差が大きくなっていく。標準偏差が大きいということは，識別器の信頼性が低い，ということを意味する。いずれのデータ集合についても，比較された中ではEDがηの影響を受けにくく，QDFが一番ηの影響を受ける。

2. 従来から指摘されているように，QDFよりFLDFの方がロバスト性が高い。QDFにおいては，共分散行列の推定誤差が問題となる。

3. ηの値が小さいときは，MQDFが有効である。$\eta=2$の場合における識別器の識別性能順位を表7に示す。

4. ηの値が大きくなるにつれて，QDFの誤識別率はMQDFのそれに接近してくる。

5. 表5より，データ集合2と3については，ηの値が小さくなるにつれて，ξの最適値も小さくなっていく。木村ら [16] によれば，MQDFは，$\xi=n-1$の場合，QDFに等しく，$\xi=0$の場合ユークリッド距離と等価になり，$0<\xi<n-1$の場合，低次のξ次元部分空間におけるQDFと，高次の$(n-\xi)$次元部分空間におけるユークリッド距離の荷重和となる。実験結果は，ηの値が小さいときは，QDFよりむしろユークリッド距離に比重をおいた識別の方が良い，ということを示している。修正二次識別関数では，共分散行列の推定誤差を低減しているため，QDFに比べ，良い識別結果が得られている。

6. 表6より，データ集合2と3については，ηの値が小さくなるにつれて，hの最適値は大きくなっていく。すなわち，訓練サンプル数が減少していくと，ウィンドウ幅を大きくとった方が，良い識別結果が得られる。

7. Van Ness [17] により，高次元のとき，すなわちηの値が小さいとき，たとえパラメトリックな仮定が正しくても，ノンパラメトリック識別器であるPARZENの方が，パラメトリック識別器であるQDFよりも，良い識別結果を与える，ということが指摘された。実験結果は，$\eta=2$の場合すべてのデータ集合について，PARZENの方がQDFより優れていることを示し，Van Nessの主張を支持している。

表 2: データ集合 1 に対する識別器の誤識別率 [%]

(上段:平均値, 中段:標準偏差, 下段:95%信頼区間, Bayes 誤識別率:10%)

識別器	η					
	1	2	4	6	8	10
1-NN	21.22 4.44 (20.35,22.09)	20.67 4.16 (19.86,21.49)	19.34 2.61 (18.82,19.85)	18.58 1.95 (18.20,18.96)	18.44 1.80 (18.09,18.79)	18.46 1.76 (18.11,18.80)
3-NN	18.31 3.57 (17.61,19.01)	17.24 3.29 (16.59,17.88)	15.83 2.00 (15.44,16.22)	15.12 1.60 (14.81,15.43)	15.08 1.42 (14.80,15.36)	14.95 1.27 (14.71,15.20)
5-NN	17.40 3.60 (16.70,18.11)	15.93 2.71 (15.40,16.46)	14.54 1.65 (14.22,14.86)	13.89 1.44 (13.61,14.17)	13.79 1.19 (13.55,14.02)	13.70 1.18 (13.47,13.93)
PARZEN	*****	24.84 4.84 (23.89,25.79)	17.73 3.00 (17.14,18.32)	14.95 2.40 (14.48,15.42)	13.84 1.49 (13.55,14.13)	13.31 1.54 (13.01,13.61)
ED	13.74 2.21 (13.30,14.17)	12.14 1.31 (11.88,12.40)	11.04 0.79 (10.88,11.19)	10.76 0.74 (10.61,10.91)	10.58 0.75 (10.44,10.73)	10.69 0.82 (10.53,10.85)
FLDF	21.98 5.68 (20.86,23.09)	15.67 3.08 (15.07,16.28)	12.38 1.29 (12.13,12.64)	11.63 1.00 (11.44,11.83)	11.12 0.91 (10.94,11.30)	11.14 0.88 (10.97,11.31)
QDF	*****	25.28 5.93 (24.12,26.45)	16.38 2.18 (15.95,16.81)	14.20 1.43 (13.92,14.49)	12.97 1.19 (12.74,13.20)	12.54 1.05 (12.33,12.74)
MQDF	15.92 3.30 (15.28,16.57)	12.98 1.84 (12.62,13.34)	11.50 0.94 (11.32,11.68)	10.96 0.82 (10.80,11.12)	10.75 0.75 (10.60,10.90)	10.79 0.83 (10.63,10.96)

表 3: データ集合 2 に対する識別器の誤識別率 [%]

(上段:平均値, 中段:標準偏差, 下段:95%信頼区間, Bayes 誤識別率:9%)

識別器	η					
	1	2	4	6	8	10
1-NN	38.74 1.90 (38.37,39.11)	36.44 2.24 (36.01,36.88)	33.34 1.51 (33.05,33.64)	31.67 1.55 (31.36,31.97)	30.66 1.49 (30.37,30.95)	29.70 1.54 (29.40,30.00)
3-NN	43.12 1.86 (42.75,43.48)	40.57 1.53 (40.27,40.87)	37.22 1.50 (36.93,37.51)	34.75 1.36 (34.48,35.01)	33.54 1.23 (33.30,33.78)	31.97 1.15 (31.74,32.19)
5-NN	45.99 2.02 (45.60,46.39)	43.95 1.53 (43.65,44.25)	40.90 1.44 (40.62,41.18)	38.37 1.25 (38.13,38.62)	36.82 1.26 (36.58,37.07)	34.99 1.15 (34.76,35.21)
PARZEN	*****	23.30 3.53 (22.61,24.00)	16.18 1.82 (15.83,16.54)	14.35 1.11 (14.13,14.57)	13.86 1.06 (13.65,14.07)	13.83 1.00 (13.63,14.02)
ED	44.61 2.62 (44.10,45.13)	45.49 2.38 (45.02,45.96)	47.40 2.09 (46.99,47.81)	47.54 1.81 (47.18,47.89)	47.88 1.50 (47.58,48.17)	48.09 1.48 (47.80,48.38)
FLDF	46.53 2.97 (45.94,47.11)	46.02 2.39 (45.55,46.49)	47.53 2.03 (47.13,47.93)	47.72 1.83 (47.36,48.08)	47.95 1.47 (47.66,48.24)	48.14 1.44 (47.86,48.42)
QDF	*****	29.93 4.10 (29.13,30.74)	17.50 2.09 (17.09,17.91)	13.97 1.51 (13.68,14.27)	12.32 1.15 (12.09,12.54)	11.56 1.04 (11.35,11.76)
MQDF	17.09 4.57 (16.19,17.98)	12.79 1.75 (12.45,13.13)	11.22 0.97 (11.03,11.41)	10.46 0.80 (10.30,10.61)	10.10 0.74 (9.95,10.24)	9.89 0.70 (9.76,10.03)

表 4: データ集合 3 に対する識別器の誤識別率 [%]

(上段:平均値, 中段:標準偏差, 下段:95%信頼区間, Bayes 誤識別率:1.9%)

識別器	η					
	1	2	4	6	8	10
1-NN	14.97 3.52 (14.28,15.66)	12.34 2.49 (11.85,12.83)	10.56 1.71 (10.23,10.90)	9.02 1.52 (8.73,9.32)	8.81 1.13 (8.55,9.06)	8.14 0.91 (7.96,8.32)
3-NN	15.48 4.34 (14.63,16.33)	11.40 2.23 (10.96,11.84)	9.34 1.38 (9.07,9.61)	8.36 1.24 (8.12,8.60)	7.76 1.09 (7.55,7.97)	7.37 0.92 (7.19,7.55)
5-NN	16.88 5.01 (15.90,17.86)	11.53 2.34 (11.07,11.99)	9.33 1.43 (9.05,9.61)	8.14 1.19 (7.90,8.37)	7.48 1.00 (7.28,7.68)	7.07 0.92 (6.88,7.25)
PARZEN	*****	6.58 2.58 (6.07,7.08)	3.69 0.66 (3.56,3.81)	3.00 0.53 (2.89,3.10)	2.83 0.53 (2.73,2.93)	2.64 0.42 (2.56,2.72)
ED	16.69 4.78 (15.75,17.63)	15.17 4.08 (14.37,15.97)	14.70 3.36 (14.04,15.36)	14.44 2.80 (13.89,14.98)	14.17 2.54 (13.67,14.66)	14.41 2.13 (13.99,14.82)
FLDF	14.28 5.41 (13.22,15.34)	9.33 2.14 (8.91,9.75)	7.43 1.27 (7.18,7.68)	6.67 0.88 (6.50,6.84)	6.28 0.65 (6.15,6.41)	6.13 0.63 (6.00,6.25)
QDF	*****	6.87 2.14 (6.45,7.29)	3.54 0.78 (3.39,3.69)	2.72 0.51 (2.62,2.82)	2.44 0.38 (2.36,2.51)	2.25 0.37 (2.18,2.32)
MQDF	11.29 3.13 (10.68,11.90)	6.10 1.94 (5.72,6.48)	3.39 0.74 (3.24,3.53)	2.68 0.53 (2.57,2.78)	2.44 0.38 (2.36,2.51)	2.25 0.37 (2.18,2.32)

表 5: MQDF のξの最適値

η	1	2	4	6	8	10
データ集合 1	2	1	2	1	1	1
データ集合 2	3	3	4	4	4	4
データ集合 3	3	6	6	6	7	7

表 6: PARZEN の h の最適値

η	1	2	4	6	8	10
データ集合 1	**	1.4	1.4	1.4	1.4	1.4
データ集合 2	**	1.2	0.8	0.8	0.8	0.8
データ集合 3	**	1.8	1.2	1.2	1.2	1.0

表 7: 識別性能による識別器の順位

	1 位	2 位	3 位	4 位	5 位	6 位	7 位	8 位
データ集合 1	ED	MQDF	FLDF	5-NN	3-NN	1-NN	PARZEN	QDF
データ集合 2	MQDF	PARZEN	QDF	1-NN	3-NN	5-NN	ED	FLDF
データ集合 3	MQDF	PARZEN	QDF	FLDF	3-NN	5-NN	1-NN	ED

8. 表中の*は，共分散行列の非正則性のため，識別器が得られなかったことを示す。このように，訓練サンプル数が少ないと，共分散行列が正則でなくなるため，QDF および PARZEN は実現できない，ということになる。

9. 最近傍識別器 (1-NN) については，訓練サンプル数が無限にあるとき，その誤識別率が Bayes 誤識別率の 2 倍を越えない，ということが理論的に示されている [18]。実験結果は，訓練サンプル数の有限性のため，偏りのある誤識別率が得られたことを示している。この偏りについては，Fukunaga らにより検討がなされている [19]。文献 [19] で，64 次元特徴空間上で，訓練サンプル数を 1000 個から 10000 個に増加させても，誤識別率の偏りは 6.9% しか減少しない，ということが示された。このことから，Fukunaga は，最近傍識別器の誤識別率を改善するためには，かなり大量の訓練サンプルが必要である，と述べている [7]。現実には，訓練サンプル数が限られているため，最近傍識別器の改良には，用いる距離を検討することが，重要と思われる。距離の改良については，文献 [20][21] がある。

4 むすび

本論文では，特徴数に対する訓練サンプル数の比が小さい状況下で，識別器の性能比較を誤識別率の観点から行った。比較された識別器は，Fisher の線形識別関数，二次識別関数，修正二次識別関数，k 最近傍識別器，ユークリッド距離識別器，Parzen 識別器である。これらの中では修正二次識別関数が有効であった。

実在データを用いて更に比較検討を行い，真に有効な識別器を設計することが，今後の課題である。

参考文献

[1] S. J. Raudys and A. K. Jain, "Small Sample Size Effects in Statistical Pattern Recognition", IEEE Trans. PAMI-13, 3, pp.252-264 (1991).

[2] D. H. Foley, "Consideration of Sample and Feature Size", IEEE Trans. IT-18, 5, pp.618-626 (1972).

[3] L. Kanal and B. Chandrasekaran, "On Dimensionality and Sample Size in Statistical Pattern Classification", in Proc. 1968 National Electronics Conf., pp.2-7 (1968).

[4] A. K. Jain and B. Chandrasekaran, "Dimensionality and Sample Size Considerations in Pattern Recognition Practice", in Handbook of Statistics, Vol.2, P. R. Krishnaiah and L. N. Kanal, Eds., North-Holland, pp.835-855 (1982).

[5] 鶴岡, 栗田, 原田, 木村, 三宅, "加重方向指数ヒストグラム法による手書き漢字・ひらがな認識", 信学論 (D),J70-D, 7, pp.1390-1397 (1987).

[6] K. Fukunaga and R. R. Hayes, "Effects of Sample Size in Classifier Design", IEEE Trans. PAMI-11, 8, pp.873-885 (1989).

[7] K. Fukunaga, "Introduction to Statistical Pattern Recognition", Second Edition, Academic Press (1990).

[8] K. Fukunaga and R. R. Hayes, "Estimation of Classifier Performance", IEEE Trans. PAMI-11, 10, pp.1087-1101 (1989).

[9] P. A. Lachenbruch and R. M. Mickey, "Estimation of Error Rates in Discriminant Analysis", Technometrics, 10, 1, pp.1-11 (1968).

[10] A. K. Jain, "Advances in Statistical Pattern Recognition", in Pattern Recognition Theory and Applications, eds., P. A. Devijver and J. Kittler, pp.1-19, Springer-Verlag Berlin Heidelberg (1987).

[11] T. Marill and D. M. Green, "On the Effectiveness of Receptors in Recognition Systems", IEEE Trans. IT-9, pp.11-27 (1963).

[12] K. Fukunaga and T. F. Krile, "Calculation of Bayes Recognition Error for Two Multivariate Gaussian Distributions", IEEE Trans. C-18, pp.220-229 (1969).

[13] A. K. Jain, R. C. Dubes and C. C. Chen, "Bootstrap Techniques for Error Estimation", IEEE Trans. PAMI-9, 5, pp.628-633 (1987).

[14] K. Fukunaga and D. M. Hummels, "Bayes Error Estimation Using Parzen and k-NN Procedures", IEEE Trans. PAMI-9, 5, pp.634-643 (1987).

[15] A. K. Jain and M. D. Ramaswami, "Classifier Design with Parzen Windows", Pattern Recognition and Artificial Intelligence, pp.211-228, Elsevier Science Publishers B. V. (1988).

[16] 木村, 高階, 鶴岡, 三宅, "2 次識別関数のピーキング現象とその防止に関する考察", 信学論 (D),J69-D, 9, pp.1328-1334 (1986).

[17] J. Van Ness, "On the Dominance of Non-parametric Bayes Rule Discriminant Algorithms in High Dimensions", Pattern Recognition, 12, pp.355-368 (1980).

[18] R. O. Duda and P. E. Hart, "Pattern Classification and Scene Analysis", John Wiley & Sons (1973).

[19] K. Fukunaga and D. M. Hummels, "Bias of Nearest Neighbor Estimates", IEEE Trans. PAMI-9, pp.103-112 (1987).

[20] R. D. Short and K. Fukunaga, "The Optimal Distance Measure for Nearest Neighbor Classification", IEEE Trans. IT-27, pp.622-627 (1981).

[21] K. Fukunaga and T. E. Flick, "An Optimal Global Nearest Neighbor Metric", IEEE Trans. PAMI-6, pp.314-318 (1984).