数理解析研究所講究録
1039 巻 1998 年 170-179

サンプル数が少ない状況下におけるパターン認識

山口大・工 浜本義彦 (Yoshihiko Hamamoto)

1 まえがき

パターン認識とは, 任意に与えられたパターンを有限個のクラス (概念) のいずれかに対応づける機能である, と定義される. ここではコンピュータによるパターン認識に主眼を置いている. コンピュータによるパターン認識過程は, 観測系, 前処理系, 特徴抽出系, 識別系の四つの処理系からモデル化される (図 1 参照). 外界に存在する認識対象は, まず, 観測系に入力され, 多数のデータが観測され, 観測パターンとして表現される. 続いて, 観測パターンは, 前処理系に入力され, 正規化, ノイズ除去などの処理が施される. 処理の施された観測データの組はパターン空間を構成する. 認識対象は, パターン空間の 1 点として表され, パターン空間ではパターンとよばれる. 一般に, 観測データは互いに相関をもつため, パターン空間は冗長な空間である. 次の特徴抽出系では, 識別に有用な特徴を必要にしてかつ十分なだけパターンから抽出し, より簡潔にパターンが表現される. 特徴の組で構成される空間は特徴空間とよばれ, そこではパターンは特徴パターンとよばれる. 特徴空間において設計される識別器により, 特徴パターンは既知クラスのいずれかに識別される.

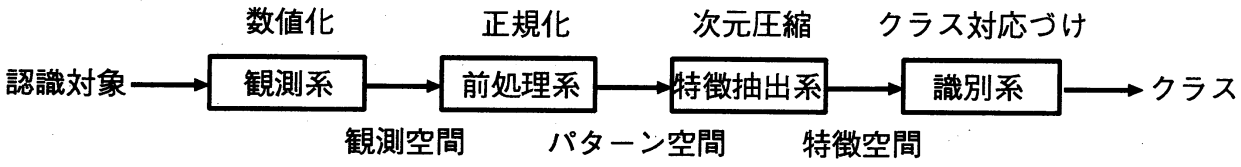

図 1. コンピュータによるパターン認識

ところで, パターン認識系の中で識別系以外の, 観測系, 前処理系, 特徴抽出系は, いずれも認識対象の性質に強く依存している. このため, パターン認識問題は認識対象に関する研究に基づいて解かれなければならず, 認識対象の性質を反映した個別的な手法が数多く提案されている. しかしながら, 認識対象の性質の解明が困難な場合がある. この問題に対して, パターンのなす統計的構造に着目し, それに基づいてパターン認識問題を解く理論がある. これは, 統計的パターン認識理論とよばれ, 数多くの研究がこれまでになされてきた [1]. 最近のパターン認識理論では, 実用に耐えうる強力な理論の確立に関心が高まっている. 本稿ではパターン認識理論における識別理論および特徴抽出理論の最近の発展を紹介する.

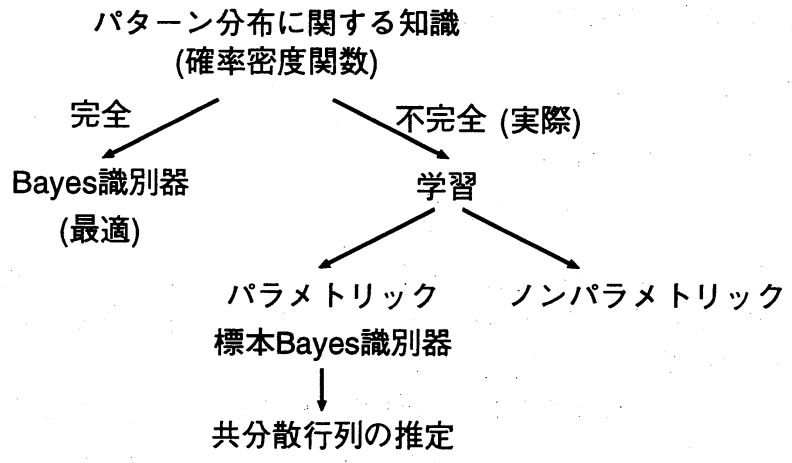

図 2. 識別理論

2 識別理論

統計的パターン認識理論によれば, 分布に関する知識, すなわち事前確率と確率密度関数が既知であるならば, 誤識別率を最小という意味で最適な識別器は Bayes 識別器である (図 2 参照). いま, 簡単のため 2 クラス問題を考えることにする. Bayes 識別器ではパターン $\boldsymbol{x}$ は

$$P(\omega_1)p(\boldsymbol{x}|\omega_1) \gtrless P(\omega_2)p(\boldsymbol{x}|\omega_2) \quad \begin{matrix} \rightarrow \boldsymbol{x} \in \omega_1 \\ \rightarrow \boldsymbol{x} \in \omega_2 \end{matrix} \tag{1}$$

と識別される. ここで, $P(\omega_i)$, $p(\boldsymbol{x}|\omega_i)$ はそれぞれクラス ω_i の事前確率, 確率密度関数である. 正規分布を仮定すると式 (1) は

$$\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\Sigma_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) - \frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_2)^T\Sigma_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2) + \frac{1}{2}ln\frac{|\Sigma_1|}{|\Sigma_2|} \lessgtr ln\frac{P(\omega_1)}{P(\omega_2)} \tag{2}$$

で表される. ここで, $\boldsymbol{\mu}_i$, Σ_i はそれぞれクラス ω_i の平均ベクトル, 共分散行列である. しかしながら, 実際は $P(\omega_i)$, $\boldsymbol{\mu}_i$, Σ_i は未知である. そこで分布から無作為に抽出されたサンプルを用いて $P(\omega_i)$, $\boldsymbol{\mu}_i$, Σ_i を推定し, 推定値を真値とみなして用いることにする. これは Plug-in アプローチと呼ばれる. 実際には利用できるサンプル数が限られているため推定誤差が生じ, これが識別器の性能劣化を招くことになる. 特に重大な影響を与えるのは共分散行列の推定誤差である. このため, 共分散行列の推定誤差を低減するさまざまな試みがなされてきた. ここでは, それらの中から代表的と思われる三つの手法を紹介するとともに, 計算機シミュレーションを通してそれらを比較する. なお, クラス毎に共分散行列は推定されるので, 簡単のためクラス名を省略することにする. まず, Hoffbeck らの手法 [2] を紹介する. この手法では共分散行列は次の形で表されるものとする.

$$\hat{\Sigma}_*(\alpha_1,\alpha_2,\alpha_3,\alpha_4) = \alpha_1 diag(\hat{\Sigma}) + \alpha_2\hat{\Sigma} + \alpha_3 S + \alpha_4 diag(S) \tag{3}$$

ここで $\hat{\Sigma}$ は標本共分散行列で, S は,

$$S = \frac{1}{2}(\hat{\Sigma}_1 + \hat{\Sigma}_2)$$

で与えられ, α_1, α_2, α_3, α_4 は

$$\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$$

を満たすパラメータで, 重みの役割を果たす. パラメータ α_1, α_2, α_3, α_4 の値は leave-one-out 法 を用いて選択される. 得られた $\hat{\Sigma}_*$ の逆行列と行列式を求め, それらを Bayes 識別器に用いるのである.

次に Friedman により提案された正則化法を紹介する [3]. 正則化法では共分散行列は

$$\hat{\Sigma}_*(\lambda, \gamma) = (1-\gamma)\hat{\Sigma}(\lambda) + \frac{\gamma}{n} tr\left[\hat{\Sigma}(\lambda)\right] I \tag{4}$$

で与えられる. ここで $\hat{\Sigma}(\lambda)$ は

$$\hat{\Sigma}(\lambda) = (1-\lambda)\hat{\Sigma} + S$$

であり, I は単位行列である. この $\hat{\Sigma}_*(\lambda, \gamma)$ から逆行列と行列式を求め, それらを Bayes 識別器に用いるのである. 問題はパラメータ λ, γ をいかに定めるかである. Friedman は leave-one-out 法により推定された誤識別率を最小にするパラメータ λ^*, γ^* を用いる方法を提案している.

最後にテプリッツ法について述べる [4]. テプリッツ法では共分散行列を

$$\hat{\Sigma}_* = \hat{\Gamma}\hat{R}\hat{\Gamma} \tag{5}$$

と表す. ここで

$$\hat{\Gamma} = \begin{bmatrix} \hat{\sigma}_1 & & & \mathbf{o} \\ & \hat{\sigma}_2 & & \\ & & \ddots & \\ \mathbf{o} & & & \hat{\sigma}_n \end{bmatrix}$$

$$\hat{R} = \begin{bmatrix} 1 & \hat{\rho} & \cdots & \hat{\rho}^{n-1} \\ \hat{\rho} & 1 & & \\ \vdots & & \ddots & \\ \hat{\rho}^{n-1} & & & 1 \end{bmatrix}$$

で, $\hat{\sigma}_i$ は標準偏差, $\hat{\rho}$ は相関係数である. このとき $\hat{\Sigma}_*^{-1}$ は

$$\hat{\Sigma}_*^{-1} = \hat{\Gamma}^{-1}\hat{R}^{-1}\hat{\Sigma}^{-1}$$

$$= \frac{1}{1-\hat{\rho}^2} \begin{bmatrix} \frac{1}{\hat{\sigma}_1} & & & \mathbf{o} \\ & \frac{1}{\hat{\sigma}_2} & & \\ & & \ddots & \\ \mathbf{o} & & & \frac{1}{\hat{\sigma}_n} \end{bmatrix} \begin{bmatrix} 1 & -\hat{\rho} & 0 & \cdots & 0 \\ -\hat{\rho} & 1+\hat{\rho}^2 & & & \vdots \\ 0 & & \ddots & & 0 \\ \vdots & & & 1+\hat{\rho}^2 & -\hat{\rho} \\ 0 & \cdots & 0 & -\hat{\rho} & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{\hat{\sigma}_1} & & & \mathbf{o} \\ & \frac{1}{\hat{\sigma}_2} & & \\ & & \ddots & \\ \mathbf{o} & & & \frac{1}{\hat{\sigma}_n} \end{bmatrix} \tag{6}$$

で与えられる. 従って, サンプルから推定すべきパラメータは n 個の $\hat{\sigma}_c$ と $\hat{\rho}$ のみである. この $\hat{\Sigma}_*^{-1}$ を Bayes 識別器に用いるのである. $|\hat{\Sigma}_*|$ については

$$|\hat{\Sigma}_*| = (1-\hat{\rho}^2)^{n-1} \tag{7}$$

で与えられる. このテプリッツ法は Hoffbeck らの手法, 正則化法に比べ計算コストの点で著しい優位性を有している.

以上紹介した Hoffbeck らの手法, 正則化法, テプリッツ法の比較を, 計算機シミュレーションにより行う.

実験 1: 本実験では, 4 次元 3 クラス からなる Iris データ [7] を用いた. まず, 各クラスに対して, ランダムに 50 個のサンプルを, 5 個のサンプルからなる訓練サンプル集合と 45 個のサンプルからなるテストサンプル集合に分割した. 次に各クラス $N(N=2,3,4,5)$ 個の訓練サンプルを用いて識別器を設計し, その識別器にテストサンプルを入力して誤識別率を推定した. 訓練サンプルとテストサンプルは独立であることに注意されたい. 以上の処理を独立に 10 回繰り返し, 誤識別率の平均値と標準偏差を求めた. 結果を表 1 に示す.

データ	:	Iris データ
クラス数	:	3
次元数	:	4
訓練サンプル数	:	各クラス 2, 3, 4, 5
テストサンプル数	:	各クラス 45
共分散行列の推定法	:	Hoffbeck らの手法, 正則化法, テプリッツ法
識別器	:	Bayes 識別器
$\alpha_i(i=1,2,3,4)$ の選択法	:	leave-one-out 法
λ, γ の選択法	:	leave-one-out 法
パラメータ $\alpha_i(i=1,2,3,4)$ の候補	:	0.0, 0.25, 0.5, 0.75, 1.0
パラメータ λ の候補	:	0.0, 0.125, 0.354, 0.650, 1.0
パラメータ γ の候補	:	0.0, 0.25, 0.5, 0.75, 1.0
試行回数	:	10

表 1. Iris データに対する各識別器の誤識別率 (%)

訓練サンプル数	2	3	4	5
Hoffbeck らの手法を用いた	N/A	6.00	7.85	6.81
Bayes 識別器		2.82	5.33	3.00
正則化法を用いた	6.89	5.11	5.11	4.15
Bayes 識別器	4.19	2.02	2.67	1.89
テプリッツ法を用いた	N/A	18.81	16.00	13.19
Bayes 識別器		10.78	4.93	4.26
通常の Bayes 識別器	N/A	N/A	N/A	N/A

(上段: 誤識別率の平均値, 下段: 誤識別率の標準偏差, N/A: 実行不能)

Iris データに対しては, 正則化法が最も良いことが分かる. Jain らによれば, 次元数に対する訓練サンプル数の比は 5 以上あることが望ましいと言われている [10]. ここでは, 上記の比を 5 よりも小さくし, これによって現実的な状況を想定している. 特筆すべきことは, 通常の Bayes 識別器が得られない現実的な状況下においても, 正則化法などにより Bayes 識別器を構成することができる, ということである.

実験 2: 本実験では, 8 次元 3 クラス からなる 8OX データ [9] を用いた. まず, 各クラスに対して, ランダムに 15 個のサンプルを 5 個のサンプルからなる訓練サンプル集合と 10 個のサンプルからなるテストサンプル集合に分割した. 次に各クラス $N(N = 2,3,4,5)$ 個の訓練サンプルを用いて識別器を設計し, テストサンプルを用いて誤識別率を推定した. 以上の処理を独立に 10 回繰り返し, 誤識別率の平均値と標準偏差を求めた. 結果を表 2 に示す.

データ	:	8OX データ
クラス数	:	3
次元数	:	8
訓練サンプル数	:	各クラス 2, 3, 4, 5
テストサンプル数	:	各クラス 10
共分散行列の推定法	:	Hoffbeck らの手法, 正則化法, テプリッツ法
識別器	:	Bayes 識別器
$\alpha_i(i = 1, 2, 3, 4)$ の選択法	:	leave-one-out 法
λ, γ の選択法	:	leave-one-out 法
パラメータ $\alpha_i(i = 1, 2, 3, 4)$ の候補	:	0.0, 0.25, 0.5, 0.75, 1.0
パラメータ λ の候補	:	0.0, 0.125, 0.354, 0.650, 1.0
パラメータ γ の候補	:	0.0, 0.25, 0.5, 0.75, 1.0
試行回数	:	10

表 2. 8OX データにおける各識別器の誤識別率 (%)

訓練サンプル数	2	3	4	5
Hoffbeck らの手法を用いた	N/A	19.33	17.33	13.67
Bayes 識別器		6.25	3.44	4.29
正則化法を用いた	19.67	14.67	14.33	11.00
Bayes 識別器	7.77	3.58	4.46	11.33
テプリッツ法を用いた	N/A	N/A	N/A	13.67
Bayes 識別器				9.22
通常の Bayes 識別器	N/A	N/A	N/A	N/A

(上段: 誤識別率の平均値, 下段: 誤識別率の標準偏差, N/A: 実行不能)

8OX データに対しても正則化法が優れていることが示された. 以上の実験結果より正則化法が Bayes 識別器の設計に役立つことが判明した.

3 特徴抽出理論

特徴抽出系は高次元であるパターン空間から識別に有用な情報で構成される低次元特徴空間への写像としてとらえることができる. 特徴抽出系の意義を最初に指摘したのは Rao[5] であった. Rao の指摘を識別理論の立場から説明することにしよう. そのためにはピーキング現象から話を始めることにする. ピーキング現象とは, 訓練サンプル数一定の下で, 最初は次元数を増加させると認識率は増加するが, 更に次元数を増加させると, 逆に認識率が低下してしまう, という現象である. 次元数を増加させるということには, 二つの側面がある. 一つは新しい識別情報が加えられることであり, このことは認識率の向上へとプラスに働く. もう一つは, マイナスとして働く面である. 一般に識別器の複雑さ, つまり学習により求めなければならないパラメータ数は次元数の増加とともに増える. 訓練サンプル数一定の下で未知パラメータ数が増加すれば, パラメータの推定誤差が増大し, 結果的にそれらのパラメータ値を用いる識別器の認識率が低下することになる. つまり, ピーキング現象とは, 最初次元数を増加していくとプラス面が支配的に働き, ある一定の次元数 (未知) を越えると今度は逆にマイナス面が支配的となる, ということである. 現実には訓練サンプル数は限られており, これを増加させることはサンプルに正解のクラス名をつけなければならず, この作業はコストの点で大変である. そこで, 与えられた訓練サンプル数に対して次元数の最適化が必要となる. これは, 特徴抽出系の役割である. このように, 特徴抽出系の存在意義は識別器の認識性能を向上させるという, より積極的な意味から与えられる.

特徴抽出系は, 用いる数学的手法によって線形手法と非線形手法に大別される. 非線形手法については, 理論上, 実際上の問題があり, あまり研究が進んでないのが現状である [1]. これに対し, 線形手法は見通しが良く, 文字認識など実際への適用も数多く報告されている. 代表的な線形手法に Karhunen-Loeve 展開 [6] と判別分析 [7] がある. Karhunen-Loeve 展開の特徴は, 平均 2 乗誤差基準の下で最適な直交展開である, ということである. この特徴ゆえ, Karhunen-Loeve 展開は, 類似志向の線形手法として位置づけられ, 部分空間法へと発展していった [1]. 一方, 判別分析は, 識別志向であり, 多くの研究者によって改良が加えられてきた [1]. 代表的な手法に, 正規直交判別ベクトル法がある [8]. これは, 判別分析の特徴軸数の制約問題を解決した手法である. 以下, 正規直交判別ベクトル法の概要を述べる.

正規直交判別ベクトル法とは特徴軸の直交性の下でフィッシャー評価関数を最大にする特徴軸 $\boldsymbol{\tau}$ を求める手法である. フィッシャー評価関数は次の形で表される.

$$J(\boldsymbol{\tau}) = \frac{\boldsymbol{\tau}^T S_w \boldsymbol{\tau}}{\boldsymbol{\tau}^T S_b \boldsymbol{\tau}} \tag{8}$$

ここで, S_w, S_b はそれぞれ,

$$S_w = \sum_{i=1}^{m} P(\omega_i)\Sigma_i$$

$$S_b = \sum_{i=1}^{m-1} \sum_{k=i+1}^{m} P(\omega_i)P(\omega_k)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_k)(\boldsymbol{\mu}_i - \boldsymbol{\mu}_k)^T$$

で, m はクラス数である. 式 (8) を最大にする特徴軸は, 固有値問題

$$S_w^{-1} S_b \boldsymbol{\tau} = \lambda \boldsymbol{\tau} \tag{9}$$

を解いて得られる最大固有値に対応する固有ベクトルである. 得られた特徴軸は第 1 特徴軸となる (図 3 参照). $\boldsymbol{\tau}_1$ 上ではクラス間の分離がフィッシャー評価関数の意味で最大となっている.

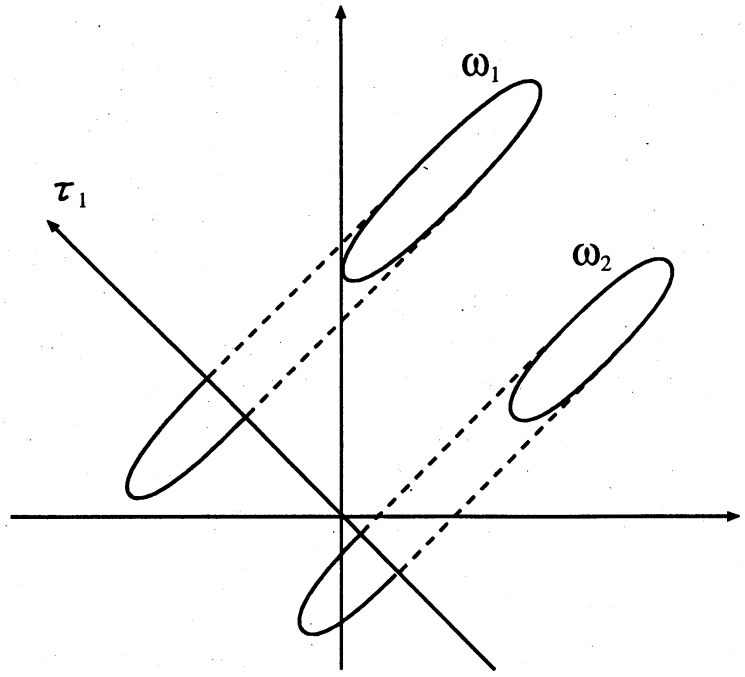

図 3. フィシャー評価関数の最大化の意味

一般に, 第 r 特徴軸 $\boldsymbol{\tau}_r$ は, $\boldsymbol{\tau}_k(1 \leq k \leq r-1)$ の直交補空間 S^{n-r+1} においてフィッシャー評価関数を最大にする特徴軸として, 次の手順で得られる.

手順 1: 直交補空間 S^{n-r+1} を生成する $n-r+1$ 個の n 次元正規直交基底ベクトル $\boldsymbol{\nu}_s(r \leq t \leq n)$ を次のグラム・シュミットの直交化法により求める.

$$\boldsymbol{\nu}_s = \alpha_s \left(I_n - \sum_{t=1}^{s-1} \boldsymbol{\nu}_t \boldsymbol{\nu}_t^T \right) \boldsymbol{\varphi}_s \tag{10}$$

ここで, $\boldsymbol{\nu}_t = \boldsymbol{\tau}_t(1 \leq t \leq r-1)$, $\boldsymbol{\varphi}_s$は $\boldsymbol{\nu}_t(1 \leq t \leq s-1)$ と 1 次独立な任意の n 次元ベクトル, α_s は $\|\boldsymbol{\nu}_s\| = 1$ とするための正規化定数, I_n は n 次の単位行列である.

手順 2: 正規直交基底ベクトル $\boldsymbol{\nu}_s$ を用いて

$$P_{r-1} = [\boldsymbol{\nu}_r \boldsymbol{\nu}_{r+1} \cdots \boldsymbol{\nu}_n] \tag{11}$$

なる $n \times (n-r+1)$ 行列 P_{r-1} を構成する.

手順 3: 直交補空間 S^{n-r+1} における S_w と S_b を求める.

$$S_{w_r} = P_{r-1}^T S_w P_{r-1} \tag{12}$$

$$S_{b_r} = P_{r-1}^T S_b P_{r-1} \tag{13}$$

手順 4: 次の固有値問題

$$S_{w_r}^{-1} S_{b_r} \boldsymbol{\tau}_k^{(r)} = \lambda_k^{(r)} \boldsymbol{\tau}_k^{(r)}, \quad \lambda_1^{(r)} \geq \cdots \geq \lambda_{n-r+1}^{(r)} \tag{14}$$

を解いて得られる最大固有値 $\lambda_1^{(r)}$ に対応する固有ベクトル $\boldsymbol{\tau}_1^{(r)}$ を式 (15) により n 次元ベクトルに変換し, それを第 r 特徴軸 $\boldsymbol{\tau}_r$ とする. すなわち

$$\boldsymbol{\tau}_r = P_{r-1}\boldsymbol{\tau}_1^{(r)}/\|\boldsymbol{\tau}_1^{(r)}\| \tag{15}$$

とする. 以上より

$$J(\boldsymbol{\tau}_1) \geq J(\boldsymbol{\tau}_2) \geq \cdots \geq J(\boldsymbol{\tau}_n) \tag{16}$$

に従う n 個の特徴軸が得られる. $\boldsymbol{\tau}_2$ を求める概要図を図 4 に示す.

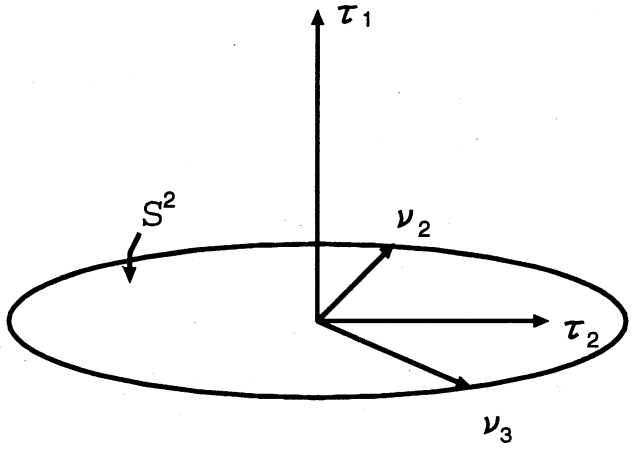

図 4. 直交補空間 S^2 上における $\boldsymbol{\tau}_2$ の探索

この特徴抽出においても, 次元数に対する訓練サンプル数の比が小さくなると共分散行列の正則化問題が生じる. クラス内共分散行列 S_w に対して, 正則化法, テプリッツ法を適用することを考える. 8OX データ (8 次元) に対して 2 次元への特徴抽出を行った. 結果を表 3 に示す.

データ	:	8OX データ
クラス数	:	3
次元数	:	8
訓練サンプル数	:	8
テストサンプル数	:	7
共分散行列の推定法	:	正則化法, テプリッツ法
識別器	:	フィシャーの線形識別器
λ, γ の選択法	:	leave-one-out 法
パラメータ λ の候補	:	0.0, 0.125, 0.354, 0.650, 1.0
パラメータ γ の候補	:	0.0, 0.25, 0.5, 0.75, 1.0
試行回数	:	10

表 3. 8OX データに対する 2 次元特徴抽出
2 次元特徴空間上におけるフィッシャー線形識別器の誤識別率 (%)

通常の正規直交	正則化法を用いた場合	テプリッツ法を用いた場合
22.50	13.75	15.00
13.21	9.43	7.14

表 3 より通常の正規直交判別ベクトル法に比べ, 正則化法などを用いることで, より識別情報に富んだ特徴空間が得られたことが分かる.

4 むすび

統計的パターン認識論での仮定とは異なり, 現実には次元数は大きく, その一方で訓練サンプル数は少ない. そのため, 従来の理論はそのままでは実際のパターン認識問題へ適用することはできない.

本稿ではパラメトリックな立場から, 共分散行列の推定問題が, 識別系だけでなく特徴抽出系においても重要であり, この問題を解決することでより広い範囲に理論が適用できることを示唆した.

謝辞 計算機シミュレーションに御協力頂いた本学内村俊二助手, 大学院生宮本貴宣, 水野圭の各氏に感謝致します.

文献

1. 浜本義彦, "パターン認識理論の最近の動向", 電子情報通信学会誌, Vol. 77, No.8, pp. 853-864 (1994).

2. J. P. Hoffbeck and D. A. Landgrebe, "Covariance matrix estimation and classification with limited training data", IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.18, No. 7, pp. 763-767 (1996).

3. J. H. Friedman, "Regularized discriminant analysis", J. of the American Statistical Association, Vol84, pp. 165-175 (1989).

4. K. Fukunaga, Introduction to statistical pattern recognition, Second Edition, Academic Press (1990).

5. C. R. Rao, "On some problems arising out of discrimnant with multiple characters", Sankhya, 9, pp. 343-364 (1949).

6. S. Watanabe, "Karhunen-Loeve expansion and factor analysis: theoretical remarks and applications", Trans. of 4th Prague Conf. Inf. Theroy, pp. 635-660 (1965).

7. R. A. Fisher, "The use of multiple measurements in taxonomic problems", Ann. Eugenics, 7, Part II, pp. 179-188 (1936).

8. 岡田敏彦, 富田真吾, "正規直交判別ベクトルによる特徴抽出論", 電子通信学会論文誌 (A), J65-A, 8, pp. 767-771 (1982).

9. A. K. Jain and R.C. Dubes, Algorithms for clustering data, Prentice Hall(1988).

10. A. K. Jain and B. Chandrasekan, "Dimensionality and sample size considerations in pattern recognition practice", in Handbook of Statistics 2, P. R. Krishnaiah and L.N. Kanal, Eds., North Holland, pp. 835-855(1982).