パターン認識における特徴選択*

3ZA－07

石川 慎也　福島 健一　矢口 博之　市野 学†

東京電機大学§

1　はじめに

一般に，観測によって得られたデータには，識別に重要な特徴と冗長な特徴とが含まれている．冗長な特徴が含まれていると，識別系の設計が複雑化し，信頼性も低下してしまう．

パタン認識系の設計者は，既知の知識に基づいて観測法を定めているが，未知のデータを扱う場合，特徴間の隠れた関係まで探り解析することは難しい．逆に，識別系の性能を向上させる目的で特徴数を必要以上に増やしてしまうことがあるが，処理時間の増加のみならず，構造が複雑化するため識別をより困難にする場合が多い．また，ベイズの決定境界や最近傍法等の識別手法においても，識別に用いる特徴はすべて有効に機能するという前提で設計されており，冗長な特徴が含まれると識別能力の低下は免れない．識別に必要な特徴を選択する一般的な方法も知られていない．

そこで本研究では，パタンクラス識別に有効な特徴群を選択する手法を提案する．提案手法は，データ構造や決定境界の構造にとらわれることなく適用可能であり，パタンクラス間構造的差異の明確化，識別率の向上が期待できる．

2　特徴選択アルゴリズム

有効な特徴選択とは，分離能力の高い特徴が選択され，その特徴によって一般性の高い記述が行えることをいう．分離能力とは，未知サンプルがどのパタンクラスに属するか決定する能力であり，記述の一般性とは，パタンクラス内での一般的なサンプルの様子を説明する能力である．

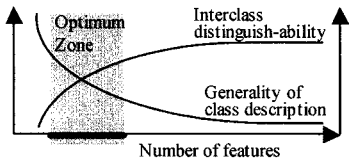

図1　記述の一般性と分離能力のトレードオフ

*A Study of Feature Selection in Pattern Recognition

† Shinya Ishikawa, Kenichi Fukusima, Hiroyuki Yaguchi, Manabu Ichino

§ Tokyo Denki University

2．1　特徴空間と矩形領域モデル

各サンプルがd次元のユークリッド空間Rd上に存在するとき，特徴集合を(1)式で記述する．

$$\mathbf{F}=\{\mathrm{F}_1,\mathrm{F}_2,\cdots,\mathrm{F}_d\} \tag{1}$$

識別したいパタンクラスをω_1，その他のパタンクラス群をω_2とすれば，多クラス問題を2クラス問題として考えることができる．

パタンクラスω_iに属すサンプル$\boldsymbol{x}_{ip}$は，d個の特徴成分からなる．

$$\boldsymbol{x}_{ip}=(x_{ip1},x_{ip2},\cdots,x_{ipd}) \tag{2}$$

N_i個のサンプルからなるパタンクラスω_iを(3)式で記述する．

$$\omega_i=\{\boldsymbol{x}_{i1},\boldsymbol{x}_{i2},\cdots,\boldsymbol{x}_{iN_i}\},\ i=1,2 \tag{3}$$

次に，矩形領域のモデルを定義する．パタンクラスω_iに属するサンプル$\boldsymbol{x}_{ip}$と$\boldsymbol{x}_{iq}$によって形成される閉区間を

$$I(\boldsymbol{x}_{ip},\boldsymbol{x}_{iq})_r=[\min(x_{ipr},x_{iqr}),\max(x_{ipr},x_{iqr})],\quad r=1,2,\cdots,d,\quad p\neq q \tag{4}$$

とすると，特徴集合$\boldsymbol{F}$におけるサンプル$\boldsymbol{x}_{ip}$と$\boldsymbol{x}_{iq}$によって張られる矩形領域は

$$RECT(\boldsymbol{x}_{ip},\boldsymbol{x}_{iq}|\mathbf{F})=I(\boldsymbol{x}_{ip},\boldsymbol{x}_{iq})_{F_1}\times I(\boldsymbol{x}_{ip},\boldsymbol{x}_{iq})_{F_2}\times\cdots\times I(\boldsymbol{x}_{ip},\boldsymbol{x}_{iq})_{F_d} \tag{5}$$

で記述される．

また，矩形領域$RECT(\boldsymbol{x}_{ip},\boldsymbol{x}_{iq}|\boldsymbol{F})$に含まれるサンプル数を示す指標として，$\alpha(\boldsymbol{x}_{ip},\boldsymbol{x}_{iq}|\boldsymbol{F})$と$\beta(\boldsymbol{x}_{ip},\boldsymbol{x}_{iq}|\boldsymbol{F})$を導入する．

$\alpha(\boldsymbol{x}_{ip},\boldsymbol{x}_{iq}|\boldsymbol{F})$は$RECT(\boldsymbol{x}_{ip},\boldsymbol{x}_{iq}|\boldsymbol{F})$に含まれるパタンクラス$\omega_i$に属するサンプル数を示す関数であり，

$$\alpha(\boldsymbol{x}_{ip},\boldsymbol{x}_{iq})=\left|\{\boldsymbol{x}_{iy}|\boldsymbol{x}_{iy}\in\omega_i,\boldsymbol{x}_{iy}\subset RECT(\boldsymbol{x}_{ip},\boldsymbol{x}_{iq}),\boldsymbol{x}_{iy}\neq\boldsymbol{x}_{ip},\boldsymbol{x}_{iy}\neq\boldsymbol{x}_{iq}\}\right| \tag{6}$$

で表される．ただし，｜・｜は集合の要素数を示す．

$\beta(\boldsymbol{x}_{ip},\boldsymbol{x}_{iq}|\boldsymbol{F})$はRECT$(\boldsymbol{x}_{ip},\boldsymbol{x}_{iq}|\boldsymbol{F})$に含まれるパタンクラス$\omega_j$に属するサンプル数を示す関数であり，

$$\beta(\boldsymbol{x}_{ip},\boldsymbol{x}_{iq})=\left|\{\boldsymbol{x}_{jy}|\boldsymbol{x}_{jy}\in\omega_j,\boldsymbol{x}_{iy}\subset RECT(\boldsymbol{x}_{ip},\boldsymbol{x}_{iq})\}\right|,i\neq j \tag{7}$$

で表される．

2．2　分離能力の算出

相互近隣およびSuccessive Intersection（以下SI）[1]を利用し，分離能力の算出アルゴリズムを提案する．

サンプル$\boldsymbol{x}_{ip}$と$\boldsymbol{x}_{iq}$が相互近隣であるとは，

$$\beta(\boldsymbol{x}_{ip},\boldsymbol{x}_{iq}|\boldsymbol{F})=0 \tag{8}$$

のときをいう．

SIは分離能力が局所的にしか有効でない特徴を削除し，

パタンクラスを識別可能な特徴を選択する手法であり，次アルゴリズムによって算出する．

$SFS_{pq}(t)$は，$RECT(x_{ip},x_{iq}|\boldsymbol{F})$に対して$\omega_j$のサンプル$x_{jt}$と分離可能な特徴組を示す関数であり，(9)式で示される．

$$SFS_{pq}(t)=\{r\,|\,x_{jt}\subset I(x_{ip},x_{iq})_r,\ r\in\boldsymbol{F}\},\quad x_{jt}\in\omega_j,\ t=1,2,\ldots,N_j \tag{9}$$

STEP1: SFS_{pq}はω_2サンプル数をN_2として，(10)式とする．

$$SFS_{pq}=\bigcap_{t}^{N_2}SFS_{pq}(t) \tag{10}$$

$SFS_{pq}(t)$が空になる場合，直前の結果を保存し，続行する．

STEP2: 全ての$SFS_{pq}(t)$に対して実施し，特徴集合の和を求める．

また，分離能力G_{mn}は$r\in\boldsymbol{F}$とし，(11)式で求める．

$$G_{mn}(r\,|\,\omega_1)=\sum_{p=1}^{N_1-1}\sum_{q=p+1}^{N_1}u\left(x_{1p},x_{1q}\,|\,SFS_{pq}\right)\times\left(1-u\left(\beta\left(x_{1p},x_{1q}\,|\,SFS_{pq}\right)\right)\right)\times|\{r\}\cap SFS_{pq}|$$
$$u(x)=\begin{cases}0, & x\leq 0\\ 1, & x>0\end{cases} \tag{11}$$

2．3　記述の一般性の算出

クラス間相互近隣である全サンプル対を対象に記述の一般性を算出する．

クラス間相互近隣(Interclass Mutual Neighbor:以下 IMN)は，関数$\lambda(x_{ip},x_{jq}|\boldsymbol{F})$を，$RECT(x_{ip},x_{jq}|\boldsymbol{F})$に含まれる$\omega_i$のサンプル数を示す関数として(12)式で定義する．

$$\lambda\left(x_{ip},x_{jq}\,|\,\mathrm{F}\right)=\left|\bigcup_{w=1}^{N_i}x_{iw}\ \middle|\ \{x_{iw}|x_{iw}\in\omega_i,x_{iw}\subset RECT(x_{ip},x_{jq}),x_{iw}\neq x_{ip}\},i\neq j\right. \tag{12}$$

パタンクラスω_i含まれるサンプルx_{ip}とパタンクラスω_jに含まれるサンプルx_{jq}がIMNであるとは，(13)式であるときをいう．

$$\lambda(x_{ip},x_{jq}|\boldsymbol{F})=0 \tag{13}$$

記述の一般性D_{imn}は，

$$d_{imn}=\begin{cases}\lambda(x_{ip},x_{jq}) & if\ \ \lambda(x_{jq},x_{ip})\neq 0\ \ \lambda(x_{ip},x_{jq})=0\\ \lambda(x_{jq},x_{ip}) & if\ \ \lambda(x_{ip},x_{jq})\neq 0\ \ \lambda(x_{jq},x_{ip})=0\\ 0, & otherwise\end{cases} \tag{14}$$

として，(15)式で求める．

$$D_{imn}(\omega_1,\omega_2\,|\,\mathrm{F})=\sum_{p=1}^{N_1}\sum_{q=1}^{N_2}d_{imn}(x_{1p},x_{2q}) \tag{15}$$

2．4　特徴選択アルゴリズム

分離能力と記述の一般性を同時に評価した特徴選択法を提案する．

STEP1: 特徴集合$\boldsymbol{F}$において，各特徴の分離能力$G_{mn}(k|\boldsymbol{F})$と記述の一般性D_{imn}を求める．

STEP2:分離能力の最も低い特徴を特徴集合$\boldsymbol{F}$から削除する．

STEP3:特徴集合の要素数がなくなるまで STEP1,STEP2 を 繰り返す．

STEP4: 特徴を削除する各段階で記述の一般性が最大であった時の特徴集合を解析結果とする．

3　検証実験

人工データを用いて，提案手法が識別に有効な特徴群が選択できるか検証する．実験に用いるデータセットは，(a)パリティデータ，(b)鍵盤型データ，(c)三角錐対峙データ，(d)タマゴ型データの 4 種類であり，それぞれに冗長な 50 特徴(F_{R1}～F_{R50})を混在させて特徴選択実験を行う．

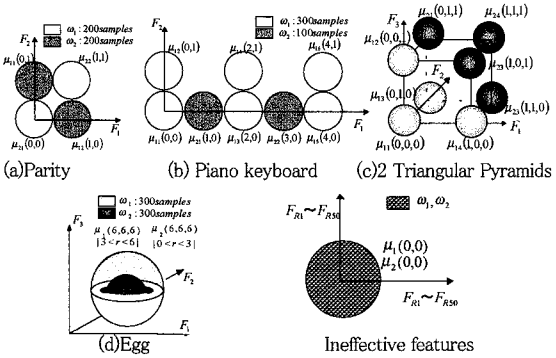

図2　人工データおよび識別に冗長な特徴群

実験結果を表1に示す．

表1　特徴選択実験結果

	(a)	(b)	(c)	(d)
All Features	$F_1,F_2,$ $F_{R1},\cdots,F_{R50}$ (52Features)	$F_1,F_2,$ $F_{R1},\cdots,F_{R50}$ (52Features)	$F_1,F_2,F_3,$ $F_{R1},\cdots,F_{R50}$ (53Features)	$F_1,F_2,F_3,$ $F_{R1},\cdots,F_{R50}$ (53Features)
Effective Features	F_1,F_2	F_1	F_1,F_2,F_3	F_1,F_2,F_3
Selected Features	F_1,F_2	F_1	F_1,F_2,F_3	F_1,F_2,F_3

提案した特徴選択手法は，いずれの人工データに対しても正しく特徴選択が行われている．また，複雑なデータ構造から識別に有効な少数個の特徴のみを選択することが可能であることを示しており，様々なデータ形式やパタンクラスの境界にも対応できる手法であると評価できる．

4　まとめ

領域モデル上にパタンクラスの記述の一般性，分離能力を定義し，記述の一般性と分離能力を同時に評価する特徴選択法を提案した．本特徴選択法は，データの分布構造をあらかじめ知ることなく適用可能な手法であり，パタン認識系の信頼性が向上すると期待できる．また、提案手法は、カルテシアンシステムモデル[2]上で記述することが容易であり、量質混在データにも対応できる。

参考文献

[1] M.Ichino: "Nonparametric feature selection method based on local interclass structure", IEEE Trans.on Syst.Man Cybern, vol. SMC-11, no.4, pp.289-296 (1981)

[2]市野:"量質混在の記述を許す一般的パターン認識法", 信学論(D)VolJ71-D No.1 pp92-101 (1988)