非線形識別関数のための特徴選択

4Q－2

佐藤 学　　工藤峰一　　外山 淳　　新保 勝

北海道大学大学院 工学研究科 システム情報工学専攻

1. はじめに

非線形識別関数の構成法の一つとして、原特徴の非線形関数を新しい特徴として追加することで原特徴空間を非線形的に拡張し、拡張特徴空間で識別関数を構成する手法がある。この手法では特徴空間の拡張方法が重要となる。筆者らはこれまでに、訓練集合を完全に説明できる程度に特徴を追加した後にMDL基準を用いて有効な特徴を選択し、この選択した特徴により線形識別関数を構成する手法を提案している[1]。しかし、その特徴選択方法は個々の特徴の独立評価をもとにしており、十分な特徴選択とは言えない。したがって、本研究は従来提案されている特徴選択手法の幾つかを用いて、より有効な識別関数を構成することを目的とする。

2. 非線形識別関数の構成 [1]

いま、n 個の原特徴からなる特徴ベクトルを $\boldsymbol{x}=(x_1, x_2,\ldots,x_n)^T$、拡張した特徴ベクトルを $\boldsymbol{y}=(y_0(\boldsymbol{x}),y_2(\boldsymbol{x}),\ldots,y_m(\boldsymbol{x}))^T$ とし (通常 $m \geq n$)、識別関数 $g(\boldsymbol{y})$ を

$$
\begin{aligned}
g(\boldsymbol{y}) &= \boldsymbol{y}^T\boldsymbol{a} \\
\boldsymbol{a} &= (a_0,a_1,\ldots,a_m)^T
\end{aligned}
$$

と線形関数で構成することを考える。$\boldsymbol{a}$は係数ベクトル、^{T}は転置を示す。

ここでは、2クラス問題を考え、$\boldsymbol{x}$をクラスω_1とω_2に識別するための目的関数 $f(\boldsymbol{x})$ を

$$
f(\boldsymbol{x}) = \begin{cases} 1 & (\boldsymbol{x}\in\omega_1) \\ -1 & (\boldsymbol{x}\in\omega_2) \end{cases}
$$

とする。この関数を訓練サンプル集合 $\{\boldsymbol{x}_i\}(i=1,2,\ldots,N)$ に対して二乗誤差$\sum_{i=1}^{N}|f(\boldsymbol{x}_i)-g(\boldsymbol{y}_i)|^2$が最小になるように$\boldsymbol{a}$を決定する。ここで、$\boldsymbol{y}_i=(y_0(\boldsymbol{x}_i),y_2(\boldsymbol{x}_i),\ldots,y_m(\boldsymbol{x}_i))^T$である。また、識別は $g(\boldsymbol{y})$ の正負により判定する。

本研究では拡張特徴集合 $\{y_i(\boldsymbol{x})\}$ として、Legendre多項式による正規直交関数系 (O.N.S.) を用いる。一変数 x_i に関する O.N.S. は

$$
\left\{\frac{P_0(x_i)}{|P_0(x_i)|},\frac{P_1(x_i)}{|P_1(x_i)|},\ldots,\frac{P_r(x_i)}{|P_r(x_i)|},\ldots\right\}
$$

で与えられる。ここで、$P_r(x_i)$ は r次の Legendre 多項式である。次にこれらを多変数に拡張して、

$$
y_i(\boldsymbol{x}) = \frac{P_{r_1^i}(x_i)}{|P_{r_1^i}(x_i)|}\frac{P_{r_2^i}(x_i)}{|P_{r_2^i}(x_i)|}\cdots\frac{P_{r_n^i}(x_i)}{|P_{r_n^i}(x_i)|}
$$

を得る。また、$y_i(\boldsymbol{x})$ の次数は $\deg(y_i(\boldsymbol{x}))=\sum_{j=1}^{n}r_j^i$ と定義される。この $\{y_i(\boldsymbol{x})\}$ は区間 $[-1,1]^n$で O.N.S. をなす。

なお、多クラスの識別方法は2クラス間の識別をすべての組合せについて行い、多数決により識別を行う。

2.1. 特徴選択法

続いて拡張した特徴 $\{y_i(\boldsymbol{x})\}$ から識別に有効な特徴を選択し、識別関数を構成する方法を述べる。まず、次数が0からℓ次までのすべての特徴を要素とする拡張特徴集合 Sを用意する。ここで、ℓは $N\leq|S|$ を満たすようにとる。その Sから MDL 基準 [2] を用いていくつかの特徴を選び、その集合を選択特徴集合 $T(\subseteq S)$ とする。MDL 基準の評価式には次式を用いた。

$$
\begin{aligned}
\mathrm{MDL}(T) &= \frac{N}{2}\log_2\frac{\varepsilon^2}{N}+\frac{|T|}{2}\log_2 N \\
\varepsilon^2(T) &= \sum_{i=1}^{N}|f(\boldsymbol{x}_i)-g(\boldsymbol{y}_i)|^2 \\
\boldsymbol{y}_i &= (y_0(\boldsymbol{x}_i),y_2(\boldsymbol{x}_i),\ldots,y_m(\boldsymbol{x}_i))^T \quad y_i\in T
\end{aligned}
$$

ここで、訓練サンプルに関する MDL(T) を最小にする Tを用いて構成した $g(\boldsymbol{y})$ が最適モデルとなる。従来法では単独での二乗誤差の小さい特徴から順に $\mathrm{MDL}(T+\{y_i(\boldsymbol{x})\})<\mathrm{MDL}(T)$ を満たす $y_i(\boldsymbol{x})$ を順に加えて Tを構成した。その計算量は$\Theta(N)$ である。

3. 特徴選択方法の変更

一般に、サンプルの分布の偏り等により単独の特徴による二乗誤差から、特徴を組み合わせた場合の二乗誤差を推定することは難しい。そのため、従来法は十分な特徴選択を行っているとはいえない。そこで、今回は幾つかの準最適な特徴選択法を用いて、従来の方法との比較を行う。今回、新たに用いる選択方法は以下に示す。

SFS Sequential Forward Search Method. 選択していない特徴の中から最も良い特徴を選び選択特徴集合に追加する方法 (図1)(計算量 :$O(N^2)$)[3]。

Feature Selection for Nonlinear Discrimination Functions.

Manabu Sato, Mineichi Kudo, Jun Toyama and Masaru Shimbo
gaku@huie.hokudai.ac.jp
Division of Systems and Information Engineering,
Graduate School of Engineering,
Hokkaido University, Sapporo 060, Japan

SFFS Sequential Forward Floating Search Method.
SFS により特徴を一つ選択したあと、選択した特徴を一つづつ削除し、サイズの一つ小さい部分集合のうち従来よりも良いものが見つかれば一段戻り、同じ処理を繰り返しつつ、特徴を追加する方法 (図 2)(計算量：$O(2^N)$)[4]。

なお、両手法ともに MDL の値が 10 回連続で上昇した場合、処理を打ち切る。

4. 実験

前述の特徴選択法を用いて人工データに対し、実験を行った。各選択方法で選択した特徴集合に対する MDL の値と評価回数、選択特徴数、MDL の評価回数および識別率を求めた。実験に用いた人工データは平均 (0,0)、共分散行列 I,4I(I は単位行列) の正規分布データ “Norm” (図 3) と、各クラスのデータが弧状に入り組んだデータ “Arc”(図 4) を用いた。データ “Norm” については訓練サンプルを変えて実験を行った。なお、すべてのデータは未知サンプルの広がりを考慮して訓練サンプルが区間 $[-0.6, 0.6]^n$ に収まるように前処理で正規化した。

4.1. 結果と考察

表 1 に実験結果を示す。特徴選択法としては、やはり SFFS、SFS、従来法の順でより小さな値を求めており、計算量においては逆順であった。この意味では時間がかかることを除いて、SFFS を用いる意義は大きい。しかし、MDL 値と識別率との関係を見ると、必ずしも小さな MDL 値で良い識別をしているとは限らない。これは、サンプル数の少ない時に MDL の識別子の複雑さを示す項がきき過ぎるせいと思われる。また、サンプル数が増えるに従い、MDL 値はほぼ、識別子の性能を示している。結果として、サンプル数がそれ程多くない場合を考慮に入れると、SFS 程度の評価回数のものを用いることにより、従来法より良い識別子が得られると考えることができる。

5. おわりに

非線形識別関数の構成において特徴選択法の有効性を調べた。人工データに対する実験においては、SFS が特徴選択法として適していることがわかった。しかし、従来法と比べ、識別率が向上した反面、計算コストは増えている。今後は実データに対して実験を行うとともに、計算コストの削減について検討する。

文献

[1] M.Sato, M.Kudo, J.Toyama and M.Shimbo, Construction of a Nonlinear Discrimination Function Based on the MDL Criterion. *Proceedings of the 1st IAPR Workshop on Statistical Techniques in Pattern Recognition*, Prague, 1997, 141–146.

[2] J.Rissanen, A Universal Prior for Integers and Estimation by Minimum Description Length. *Annals of Statistics*, **11**(1983), 416–431.

[3] J.Kittler, Feature Set Search Algorithms. *Pattern Recognition and Signal Processing* (ed. by C.H.Chen), Sijthoff and Noordhoff, Netherlands, 1978, 41–60.

[4] P.Pudil, J.Novovičová and J. Kittler, Floating Search Methods in Feature Selection. *Pattern Recognition Letters*, **15**(1994), 1119–1125.

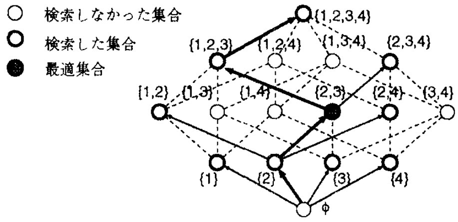

図 1: SFS

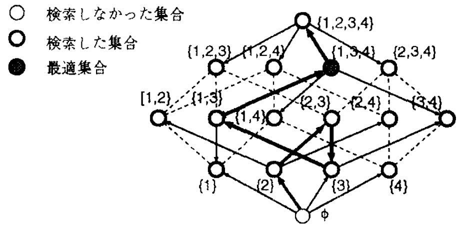

図 2: SFFS

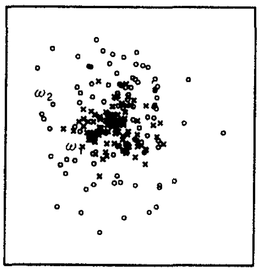

図 3: Norm

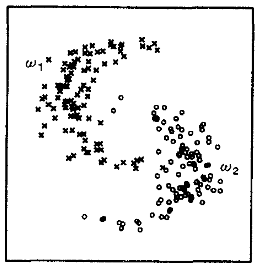

図 4: Arc

実験結果

データ	n^{*1}	c^{*2}	N_1^{*3}	N_2^{*4}	従来法				SFS				SFFS			
					MDL	m	k	R*5	MDL	m	k	R	MDL	m	k	R
Norm	2	2	50	500	-11.8	3	104	72.2	-12.3	3	505	**73.2**	**-14.2**	3	837	71.3
			100	500	-37.9	4	209	71.7	**-42.2**	4	1030	**73.2**	**-42.2**	4	1471	**73.2**
			200	500	-70.1	28	405	73.8	**-83.8**	5	6344	**74.4**	**-83.8**	5	7719	**74.4**
Arc	2	2	100	500	-483.6	19	209	98.3	-489.8	13	4345	**98.4**	**-498.3**	11	6376	98.2

*1：特徴数、*2：クラス数、*3,*4：訓練サンプル数、検査サンプル数 (個/クラス)
*5 MDL：MDL の最小値、m：選択特徴数、k：MDL の評価回数、R：識別率 [%]