非线性判别函数的基于 MDL 标准的勒让德多项式构造法*)

A Constructing Method of Non-linear Discriminating Function Based on MDL Criterion Using Legendre Polynomials

翟正利　孙忠林　吴哲辉

（山东科技大学信息科学与工程学院　泰安271019）

Abstract In Model Recognition, although a non-linear discriminating function may be superior to linear or quadratic classifiers, it is difficult to construct. This paper proposes a constructing method using Legendre polynomials. The selection of an optimal set of Legendre is determined by the MDL criterion. Results using many data show the effectiveness of this method.

Keywords Non-linear discriminating function, Legendre polynomials, MDL

1. 引言

在模式识别中，分类决策问题是最基础也是最重要的内容，所谓分类决策就是根据被识别对象特征的观察值将其分到某个类别中去，其基本作法是在样本训练集基础上确定某个判决规则（即样本特征空间的一个函数，也称为判别函数或分类器），使按这种判决规则对被识别对象进行分类所造成的错误率最小或引起的损失最小。

当样本分布是正态分布时，从最小错误率来说二次分类器是最优分类器。此外，作为一种特殊情况，如果这些分布有相同的协方差矩阵，那么线性分类器就是最优的[5]。然而这些假定在很多实际问题中并不成立，因此线性或二次分类器对大量的问题是不能胜任的。另一方面，k-近邻或分段线性分类器不假设样本分布的统计模型并且决策规则是基于训练样本的，然而，在这两种方法中，决策规则通常太复杂以至于对未知样本无法工作。总之，线性和二次分类器显得太简单，k-近邻和分段线性分类器显得太复杂。

尽管非线性判别函数已被证明是理论上比较有效的，然而如何构造非线性判别函数却比较困难。本文基于最小描述长度（简记为 MDL）标准和勒让德多项式的正交性质给出了构造非线性判别函数的一种较为有效的方法。对许多数据的试验结果表明了该方法的有效性。

2. 非线性判别函数

首先考虑两类问题，即已知样本类别共两类（记为 ω_1 和 ω_2），要将被识别对象归为其中之一。判别函数分两个阶段来构造：①定义域的变换；②用多项式构造之。主要思想是：先把特征向量的定义域 x 变换为 y，此变换为非线性变换，即 y 是 x 的一个非线性函数；然后构造 y 的一个线性函数作为判别函数。例如，已经知道“排斥或”问题能通过定义域的变换来解决。在二维空间中，当两个样本(0,0)和(1,1)来自 ω_1，另外两个样本(1,0)和(0,1)来自 ω_2 时，这些样本不能被一个线性分类器判别。然而，我们可通过定义域的变换

$$y_1=x_1, y_2=x_2, y_3=x_1x_2$$

构造一个线性分类函数

$$f(y)=y_1+y_2-2y_3-1/3$$

来对其分类。

2.1 定义域的变换

设一个从 R^n 到 R^m 的定义域变换：

$$x=(x_1,x_2,\cdots,x_n)^T\rightarrow y(x)=(y_0(x),y_1(x),\cdots,y_m(x))^T$$

其中，$y_0(x)$ 是一个常量，$y_i(x)$ 是 x 的变换函数（$i=1,\cdots,m$）。

此时，我们用勒让德多项式来表示 $y_i(x)$，r-次勒让德多项式为：

$$P_r(x)=\frac{1}{2^r\cdot r!}\frac{d^r(x^2-1)^r}{dx^r},$$

在$[-1,1]^n$上的规范化勒让德多项式为：

$$Q_r(x)=\frac{P_r(x)}{\|P_r(x)\|},$$

其中，$\|P_r(x)\|^2=\int_{-1}^{1}P_r^2(x)dx$。

由勒让德多项式的正交性质知，

$$\{Q_1(x),Q_2(x),\cdots,Q_r(X),\cdots\}$$

构成了最简明的标准正交基。现定义 $y_i(x)$ $(i=0,1,2,\cdots,m)$ 如下：

$$y_0(x)=1/\sqrt{2}$$

$$y_i(x)=Q_{r_1^i}(x_1)Q_{r_2^i}(x_2)\cdots Q_{r_n^i}(x_n)$$

$y_i(x)$ 由序列 $r^i=(r_1^i,r_2^i,\cdots,r_n^i)$ 指定，$r_j^i\in\{0,1,2,\cdots\}$，$y_i(x)$ 的次数定义为 $\deg(y_i(x))=\sum_{j=1}^{n}r_j^i$，则 $\{y_i(x)\}$ 构成了[−

*)国家自然科学基金资助课题(60173053)。翟正利　讲师，硕士研究生，主要研究方向为 Petri 网理论与应用。孙忠林　副教授，主要研究方向为图像识别与数据库理论。吴哲辉　教授，博士生导师，主要研究方向为 Petri 网理论与应用、算法设计与分析等。

验结果。

参考文献

1 Bou-Ghazale S, Hansen J H L. A Comparative Study of Traditional and Newly Proposed Features for Recognition of Speech Under Stress. IEEE Trans. Speech and Audio Processing, 2000, 8(4): 429～442

2 Picone J W. Signal Modeling Techniques in Speech Recognition. Proceedings of IEEE, 1993, 81(9): 1215～1247

3 易克初，田斌，付强. 语音信号处理. 国防工业出版社，2000

4 吕成国，周健，诸光. 高噪声有变异语音库的建立. 见：第五届全国人机语音通讯学术会议论文集. 1998. 377～381

5 吕成国，王承发，张磊，韩纪庆. G-stress 和 Lombard 影响下变异语音谱图. 高技术通讯，2000(增刊)：377～381

6 马永林，韩纪庆，张磊，王承发. 基于修正 Mel 频率映射的应力影响下语音识别方法. 计算机工程与应用，2002(录用待发表)

$1,1]^n$ 上的标准正交基，即：

$$\langle y_i(x), y_j(x)\rangle_{[-1,1]^n} = \int_{-1}^{1}\cdots\int_{-1}^{1} y_i(x)y_j(x)dx = \delta_{ij} = \begin{cases} 1, & if\ i=j \\ -1, & otherwise \end{cases}$$

2.2 目标函数的勒让德多项式近似表示

定义在两类中的目标函数为：

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

我们用 $g(y(x))=y(x)^Ta$ 来近似表示 $f(x)$，其中 $a=(a_0,a_1,a_2,\cdots,a_m)^T$ 是系数向量，我们确定使 $\|f(x)-g(y(x))\|^2$ 取最小值的 a。用 $g(y(x))$ 作为判别函数，然后对 N 个训练样本 $\{x_i\}(i=1,2,\cdots,N)$，通过下面式子来确定 a

$$a=Y^{-1}F$$

$$Y=(y(x_1),y(x_2),\cdots,y(x_N))^T$$

$$F=(f(x_1),f(x_2),\cdots,f(x_N))^T$$

其中 Y^{-1} 是 Y 的广义逆矩阵。

2.3 用 MDL 标准选择特征

为了在每个训练样本中用 $g(y)$ 识别 $f(x)$，m 必须大于等于 N。然而，已经知道过度设计训练样本对识别未知样本是非常有害的，这就是通常所说的“过学习”问题[5]。因此，问题在于如何选择最优的 $y_i(x)(i=0,1,2,\cdots,m)$。这里用 MDL 标准来解决此问题。

首先，选择满足

$$m_0=|S|=\binom{n+l}{l}\geqslant N$$

的最小整数 l，并得到一个备选特征集

$$S=\{y_i\,|\,0\leqslant\deg(y_i(x))\leqslant l\}。$$

对于 S 的一个子集 T，它的 MDL 值通过下式计算：

$$MDL(T)=\frac{N}{2}\log_2\frac{\varepsilon^2(T)}{N}+\frac{m}{2}\log_2 N \qquad (1)$$

其中，$\varepsilon^2(T)=\sum_{i=1}^{n}(f(x_i)-g(y(x_i)))^2$，

$y=(y_0,y_1,\cdots,y_m)^T, y_i\in T$

在(1)式中，第一项表明了 T 对训练样本的解释程度，第二项表明 T 是几度复杂的。作为结果，MDL 标准选择扩充特征的一个较小的数目，对训练样本选择一个较小错误。

我们选择使 MDL 取最小值的 T。为了在限定时间内得到一个近似最优的 T，我们采用顺序搜索，算法如下：

步骤1 对 $\{y_i(x)\}$ 排序并重新命名使得

$\varepsilon^2(\{y_1(x)\})\leqslant\cdots\leqslant\varepsilon^2(\{y_{m_0}(x)\})$

步骤2 令 $T\leftarrow\{y_0(x)\}, i\leftarrow 1$

步骤3 $U\leftarrow T\cup\{y_i(x)\}$，$\triangle MDL\leftarrow -\frac{N}{2}\log_2\frac{\varepsilon^2(U)}{\varepsilon^2(T)}+\frac{1}{2}log_2N$

步骤4 如果△MDL＜0，则 T←U

步骤5 i←i＋1

步骤6 如果 i＞m_0则结束，否则转步骤3。

对于多类问题，我们可对每个类对都构造一个判别函数采用两两判别，并把一个未知样本判定其属于每个类对中的哪一类，然后根据多数原则进行分类。

3. 试验

通过用两组数据（注：所引用数据出自文[3,4]）把本方法同贝叶斯线性、贝叶斯二次和 k-近邻法（$k=1,5$）进行比较。

3.1 理想数据

首先，我们在三个理想数据集上进行试验。我们用 k-倍（$k=10$）交叉验证[2]（简记为10-倍 CV）来估计识别率。

·两类正态分布（记作 Norm1）：一类有均值(0,0)和协方差矩阵 I；另一类有均值(0,0)和协方差矩阵 $4I$

·两类正态分布（记作 Norm2）：一类有均值(0,0)和协方差矩阵 $\begin{pmatrix}4&0\\0&2\end{pmatrix}$；另一类有均值(3,2)和协方差矩阵 I

·两类均匀分布（记作 Square）：一类分布于内正方形，另一类分布在内外正方形之间。

每类都有100个样本，结果如表1所示。

表1 理想数据的结果（其中 m 为选择的特征数）

数据名	特征数	类数	10-倍交叉验证的识别率(%)				
			本方法(m)	线性	二次	1-近邻	5-近邻
Norm1	2	2	73.7(2.6)	58.9	78.1	64.1	73.6
Norm2	2	2	94.7(10.2)	91.0	96.3	92.0	93.2
Square	2	2	99.0(16.0)	47.5	85.0	94.3	90.9

3.2 实际数据

下面，处理许多实际数据集合，ETL3是一个26类的字母字符数据库[3]，其它所有的数据都来自机器学习数据库[4]。这里采用了数据集中的最初81个特征的 K-L 展开式的前10个系数。

结果如表2所示。识别率的排名如表3所示。正如在表中所见，本文提出的方法在很多情况下都优于其它方法。

表2 实际数据的结果（其中 m 为选择的特征数）

数据名	特征数	类数	每类的样本数	10-倍交叉验证的识别率(%)				
				本方法(m)	线性	二次	1-近邻	5-近邻
ETL3	10	26	每类都为100	98.5(46.5)	88.6	95.0	92.4	92.4
乳腺癌	9	2	458,241	96.3(46.8)	96.1	95.1	95.4	96.9
青光眼	9	6	70,17,76,13,9,29	65.1(25.7)	67.7	66.5	72.1	68.4
心脏病	14	4	303,29,123,200	89.4(24.4)	72.5	56.5	56.9	58.6
肝炎	19	2	32,123	76.2(4.6)	64.7	76.1	66.6	71.8
糖尿病	8	2	500,268	76.4(13.5)	74.9	72.9	66.8	70.7

（下转第145页）

$$L(H,i))\delta(h_2\in U(H,i))$$
$$=\sum_j\sum_k\frac{1}{NC(G_j,G_k)}\sum_{i\in C(G_j,G_k)}\sum_{h_1\in G_j}p(h_1,t)\delta(h_1\in L(H,i))\sum_{h_2\in G_k}p(h_2,t)\delta(h_1\in U(H,i))$$

由此，我们获得如下的定理2。

定理2(宏观精确模式定理) 在一点交叉无变异的条件下，定长 GP 模式 H 的总体传递概率为：

$$\alpha(H,t)=(1-p_{xo})p(H,t)+p_{xo}\sum_j\sum_k\frac{1}{NC(G_j,G_k)}\sum_{i\in C(G_j,G_k)}p(L(H,i)\bigcap G_i,t)p(U(H,i)\bigcap G_i,t) \quad (6)$$

集合 $L(H,i)\bigcap G_i$ 和 $U(H,i)\bigcap G_i$ 或者是定长模式，或者是空集。所以，该模式定理代表了具有低阶(或同阶)模式 H 的总体传递概率。宏观精确模式定理既是对式(2)的广义概括，也是对 Holland 的 GA 模式定理的精确表达。

可以看出，式(6)中 $p(L(H,i)\bigcap G_i,t)p(U(H,i)\bigcap G_i,t)$ 项就代表了模式 $L(H,i)\bigcap G_i$ 和 $U(H,i)\bigcap G_i$ 的积木块假设。随着 GP 的运行，如果变异率很小(此处为0)，基于一点交叉的 GP 群体开始收敛，规模和形状的多样性将会增加，程序包含的结点也越来越多。当和不同形状的程序交叉时，$p(L(H,i))\bigcap G_i,t)$和 $p(U(H,i)\bigcap G_i,t)$的概率将会增加，这正体现了积木块假设所揭示的内容。

结束语 定长精确模式定理清楚地阐明了 GP 模式创建机制，我们可以根据它推断算法收敛与否。同时，它是 John Holland 的 GA 模式定理在变异概率为0时更准确的表达，实现了由 GA 到 GP 模式定理的平滑过渡。通过对模型宏观和微观的分类，从不同角度分析了 GP 运行机理，超模式空间的引入也更好地支持了积木块假设。

参 考 文 献

1 Poli Riccardo. Microscopic and Macroscopic Schema Theories for Genetic Programming and Variable-length Genetic Algorithms with One-Point Crossover, their Use and their Relations with Earlier GP and GA Schema Theories: [Technical Report, CSRP-00-15]. The University of Birminham, UK 2000

2 Poli Riccardo. Exact Schema Theorems for GP with One-Point and Standard Crossover Operating on Linear Structures and their Application to the Study of the Evolution of Size: [Technical Report, CSRP-00-14]. The University of Birminham, UK 2000

3 Poli Riccardo. General Schema Theory for Genetic Programming with Subtree-Swapping Crossover: [Technical Report, CSRP-00-16]. The University of Birminham, UK 2000

4 McPhee Nicholas Freitag. A schema theory analysis of the evolution of size in genetic programming with linear representations: [Technical Report, CSRP-00-22]. The Unive-rsity of Minnesota, USA 2000

5 McPhee Nicholas Freitag. A schema theory analysis of mutation size biases in genetic programming with linear representations: [Technical Report, CSRP-00-24]. The Unive-rsity of Minnesota, USA 2000

6 Poli Riccardo. Why the Schema Theorem is Correct also in the Presence of Stochastic Effects. In: Proc. of the Congress on Evolutionary Computation(CEC2000), San Diego, USA. 2000. 487～492

7 Poli Riccardo, Langdon William B. Schema Theory for Genetic Programming with One-point Crossover and Point Mutation. EvolutionaryComputation, 1998, 6(3): 231～252

8 Poli Riccardo, Langdon William B. A Review of Theoretical and Experimental Results on Schemata in Genetic Programming. In: Proc. of the First European Workshop on Genetic Programming, Paris. 1998. 1～15

9 Poli Riccardo. Hyperschema Theory for GP with One-Point Crossover, Building Blocks, and Some New Results in GA Theory. In: Genetic Programming, Proc. of EuroGP'2000, Edinburgh. 2000. 163～180

10 Poli Riccardo, Langdon William B. A New Schema Theory for Genetic Programming with One-point Crossover and Point Mutation. In: Genetic Programming 1997, Proc. of the Seventh Intl. Conf. USA. 1997. 18～25

(上接第154页)

表3 识别率的排名

分类器	第一名	第二名	第三名	第四名	第五名
本方法	8	2	0	0	1
线性	0	3	4	0	4
二次	0	4	2	2	3
1-近邻	2	0	1	6	2
5-近邻	1	2	4	3	1

然而，对于数据集"青光眼"，该方法表现出了最低的性能，其原因在于当训练样本数很少时，MDL 标准倾向于选择简单判别规则而不是最优规则。

结束语 本文提出了基于勒让德多项式和 MDL 标准的构造非线性判别函数的方法，对很多问题该方法表现出了比其它方法好的性能。然而，当只有很少训练样本时，该方法不能构造一个好的规则。采用更有效的特征选择是可以进一步深入研究的课题。

参 考 文 献

1 Rissanen J. A Universal Prior for The Integers and Estimation by Minimum Discription Length. Ann. of Statistics, 1983, 11: 416～431

2 Stone M. Cross-Validation Choice and Assessment of Statistical Predictions. J. Roy. Statist. Soc., 1974, 36: 111～147

3 Japanese Technical Committee for Optical Character Recognition. ETL3 Character Database. Electrotechnical Laboratory, 1993

4 Murphy P M, Aha D W. UCI Repository of Machine Learning Databases: [Machine Readable Data Repository]. University of California, Department of Information and Computer Science, Irvine, California, 1991

5 边肇祺，张学工等编著. 模式识别(第二版). 清华大学出版社，2000