第24卷 第5期
1998年9月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 24, No. 5
Sept., 1998

短 文

一种特征选择的动态规划方法[1)]

章新华

（海军大连舰艇学院 大连 116018）

（哈尔滨工程大学水声工程系 哈尔滨 150001）

摘 要 通过分析特征选择的机理，提出了一种特征选择性能指标和基于此指标的动态规划特征选择方法. 使复杂的多类特征信息选择的全局满意解寻求过程，转变成一个简单的阶段性最优化问题. 在一定条件下，由各阶段最优决策构成的整体策略等价于原问题的全局满意解. 本文法较好地应用于水声信号特征分析.

关键词 特征选择，动态规划，模式识别，信息融合.

1 引言

模型识别是利用特征向量中所包含的类别信息进行类别划分的过程. 为了实现鲁棒分类，分类器使用的特征向量必须含有足够的类别信息. 通常，针对某一特殊的模式识别问题，可从多个不同的角度（如时域和频域）得到多种类型的特征信息，这些不同类型特征信息的综合利用，对模式识别具有十分重要的作用. 多类特征信息的综合不是诸类特征集的简单合并，也不是常见的合并后的特征压缩，而应是有机理的融合. 研究发现，两类具有较好分类效果特征向量的简单联合，其分类效果并不一定优于单类特征. 究其原因主要有二：1）根据信息论的观点，诸类特征的类别信息既有互补性，又有矛盾性，特征集的简单合并所带来的互补信息可能不敌它们之间的矛盾信息；2）按分类原理，多类特征信息的简单合并使得特征空间维数增加，此时，若训练样本数目不变，则由此得到的分类器的类别可分离性会下降，且在训练集规模受限制时这种下降会更明显[1].

特征选择一直受到本领域的广泛关注：整数规划法[2]仅适用于逐段线性分类器；巴氏距离法对特征分布有一定要求[3]；基于互信息的贪婪算法[4]得到的不是全局满意解，被选择的特征维数是预先固定的；遗传算法[5]是一种随机优化搜索算法，可望得到全局满意解，但仍具有较大的计算复杂性. 本文从理论上分析了特征选择的机理，提出了一种近似最优的特征选择性能指标和基于此的动态规划选择方法.

1) 中国博士后科学基金资助项目.

收稿日期 1997-06-20

2 特征选择的机理

设 $\Omega_C=\{W,\sigma_w,P\}$是样本的类别概率空间,其中 $W=\{\omega_1,\cdots,\omega_C\}$表示模式类别集合,$C$ 是模式类别数,σ_w 是 W 的子集生成的一个 σ 代数,P 是定义在 σ_w 上的概率测度. 其中,各类别出现的先验概率为 $P(\omega_i),i=1,\cdots,C$. $\Omega_F=\{\boldsymbol{X},\beta_x,p(\boldsymbol{X})\}$是样本的特征概率空间,其中 $\boldsymbol{X}=(x_1,\cdots,x_F)^T$ 是样本的 F 维特征向量,β_x 是 $\boldsymbol{X}$ 的 Borel 域,$p(\boldsymbol{X})$是定义在 β_x 上的特征概率密度函数. 对于给定的模式分类问题,模式类别的初始不确定性

$$H(\Omega_C)=-\sum_{i=1}^{C}p(\omega_i)\log p(\omega_i) \tag{1}$$

是固定的. 但其后验熵

$$H(\Omega_C|\Omega_F)=-\sum_{i=1}^{C}\int_{\boldsymbol{X}}p(\boldsymbol{X},\omega_i)\log p(\omega_i|\boldsymbol{X})\mathrm{d}\boldsymbol{X}, \tag{2}$$

可根据已知的模式特征信息 Ω_F,通过特征选择加以改变. 其改变量

$$\begin{aligned}I(\Omega_C,\Omega_F)&=H(\Omega_C)-H(\Omega_C/\Omega_F)\\&=\sum_i\sum_j p(\omega_i,\boldsymbol{X}_j)\log\left(\frac{p(\omega_i,\boldsymbol{X}_j)}{p(\omega_i)p(\boldsymbol{X}_j)}\right),\end{aligned} \tag{3}$$

即为模式特征空间与类别空间之间的互信息,它反映了样本的特征向量与其各类别之间的整体相关性. 这里,$\boldsymbol{X}_j\in R^F$ 为 $\boldsymbol{X}$ 的第 j 种选择. 当特征分量之间不存在关于类别空间的互补信息时,它具有最简形式

$$I_s(\Omega_C,\Omega_F)=\sum_{i=1}^{C}\sum_{k=1}^{F}p(\omega_i,x_k)\log\left(\frac{p(\omega_i,x_k)}{p(\omega_i)p(x_k)}\right). \tag{4}$$

特征选择的目的是从不同信息含量的许多特征分量中选择若干分量,使(2)式表示的后验熵达到最小. 这等价于使(3)式表示的特征概率空间与类别概率空间的互信息达到最大. 即通过特征选择使模式类别的平均不确定性达到最小. 因为(3)式的计算量随 F 和变量的离散区间数目指数增长,用它直接作为特征选择的性能指标,计算十分复杂. 而(4)式又只是(3)式的特殊形式,并不满足一般特征分布条件. 实际上,(4)式表示的互信息没有考虑特征分量之间的信息冗余,它与(3)式有如下关系

$$I(\Omega_C,\Omega_F)\le I_s(\Omega_C,\Omega_F). \tag{5}$$

可见,将 $I_s(\Omega_C,\Omega_F)$减去一个正的修正项,可能得到互信息 $I(\Omega_C,\Omega_F)$的较好近似. 这一修正项(实际上是一惩罚项)应能较好地表征特征集内各分量之间的相关性,它可用各分量间的互信息表示. 这样就找到了一种既简单又能较好近似互信息 $I(\Omega_C,\Omega_F)$的特征选择性能指标.

综上所述,特征选择的基本思想是把各类特征向量组成一个大向量,然后从中选择使性能函数

$$V(m)=I_s(\Omega_C,\Omega_m)-\alpha\frac{1}{m}\sum_{\substack{x_i,x_j\in\Omega_m\\x_i\neq x_j}}I(x_i,x_j) \tag{6}$$

达到最大的 m 个分量. 其中,α 为特征相关性系数,一般取0.1—1.0. 如果 α 取值合适,(6)式能较好地近似(3)式表示的互信息,从而使特征空间与类别空间之间具有较好的整体相

关性.

3　基于动态规划的特征选择

设原始特征分量集合为 $O=\{x_1,\cdots,x_F\}$,F 为可供选择的特征总维数.取状态变量 $s_k\subset 0O$ 为第 k 阶段已选特征构成的集合,$F_k\subset O$ 为第 k 阶段可供选择的候选特征集合,$d_k(s_k)\in F_k$ 表示第 k 阶段的决策变量,即从候选特征集合中选择新的特征分量.则第 k 阶段的状态变量可表示为

$$s_k = T_{k-1}(s_{k-1},d_{k-1}(s_{k-1}))\,,$$

其中 T_{k-1}表示在状态 s_{k-1}下的状态转移变换.令 $P_{k,n-1}(s_k)$为 k 阶段到 $n-1$阶段所有允许策略的集合,用(6)式作为决策的性能函数.则从初始状态 s_0以策略 $p_{0,n}$到达状态 s_{n+1}时得到的性能指标函数 $V_{0,n}(s_0,p_{0,n})$可表示为

$$\begin{aligned} V_{0,n}(s_o,p_{0,n}) &= \sum_{x_j\in s_{n+1}}\sum_{i=1}^{C} p(\omega_i,x_j)\log\left(\frac{p(\omega_i,x_j)}{p(\omega_i)p(x_j)}\right) - \alpha\frac{1}{N_s}\sum_{\substack{x_i,x_j\in s_{n+1}\\ x_i\neq x_j}} I(x_i,x_j)\,, \\ &= V_{0,n-1}(s_0,p_{0,n-1}) + v_n(s_n,d_n)\,, \end{aligned} \tag{7}$$

其中

$$v_n(s_n,d_n) = \sum_{i=1}^{C} p(\omega_i,d_n)\log\left(\frac{p(\omega_i,d_n)}{p(\omega_i)p(d_n)}\right) - \alpha\frac{1}{N_s}\sum_{x_i\in s_n} I(x_j,d_n) \tag{8}$$

表示第 $n+1$阶段在状态 s_n 下采用决策 d_n 所得到的性能指标.这里 N_s 为已选特征集合的元素个数.

这样,特征选择就变成了一个动态规划问题.根据动态规划原理,允许策略 $p^*_{0,n-1}=(d^*_0,d^*_1,\cdots,d^*_{n-1})$是最优策略的充要条件是对任意$0<k<n-1$,$s_0$有

$$V_{0,n-1}(s_0,p^*_{0,n-1}) = \max_{p_{0,k-1}\in P_{0,k-1}(s_0)}\{V_{0,k-1}(s_0,p_{0,k-1}) + \max_{p_{n,k-1}\in P_{k,n-1}(s_k^p)} V_{k,n-1}(s_k^p,p_{k,n-1})\}\,, \tag{9}$$

其中 $s_k^p=T_{k-1}(s_{k-1},d_{k-1})$表示了由初始状态 s_0和子策略 $p_{0,k-1}$确定的第 k 阶段的状态.这就是说:如果 $p^*_{0,n-1}$是选择 n 个特征的最优策略,则它的前级子策略 $p^*_{0,n-2}$应该是选择 $n-1$个特征的最优策略.换言之,若 s_n^* 是经最优策略 $p^*_{0,n-1}=(d^*_0,d^*_1,\cdots,d^*_{n-1})$得到的最优状态(具有整体满意解的 n 个特征),则由最优策略$(p^*_{0,n-1},d^*_n)$得到的 $s^*_{n+1}=T_n(s^*_n,d^*_n)$也是最优状态.

特征选择算法

1)置 $k=0$,$s_k=s_0=\{\Phi\}$,$F_k=\{O\}$,$p_{0,k-1}=\{\Phi\}$,$V_{0,k-1}(s_0,p_{0,k-1})=0$;

2)若 $k>$pro-num,则转8;否则,转3;

3)对$\forall\ x\in F_k$,按(7)式计算性能函数 $v_k(s_k,x)$,并确定 $d_k=\arg\max\limits_{x\in F_k} v_k(s_k,x)$;

4)若 $v_k(s_k,d_k)<0$,则转8;否则,转5;

5)$V_{0,k}(s_0,p_{0,k})=V_{0,k-1}(s_0,p_{0,k-1})+v_k(s_k,d_k)$;

6)$s_{k+1}\leftarrow d_k$,,$F_k\leftarrow F_k\backslash d_k$,$p_{0,k}=\{p_{0,k-1},d_k\}$;

7)$k\leftarrow k+1$,转2;

8)结束.

其中 pro-num 为预定的特征数目.

算法中有两个结束条件:1)到达预定的特征数目;2)性能函数有减无增.条件2)表明,由于特征间存在相关性,甚至矛盾性,特征数目并非越多越好.算法可同时给出特征数目与被选特征分量的满意解.如无特殊要求,可设 pro-num$=F$,由算法自动确定应选择的特征数目.

特征选择属于 NP 难解的组合优化问题,被选特征空间维数的增加会引起组合爆炸.基于动态规划的特征选择算法,利用特征-模式样本集的内部信息,实现了自适应优化搜索,大大减少了特征选择的计算量.

4 应用实例

上述方法用于回音声纳目标的特征选择.通过对目标回波信号的时频分析,可提取多种类型的时频特征.在此,考虑两种类型的模式特征:一类是信号的倒谱特征(图1中的前30维),另一类是包含信号幅值分布、过零点分布等的波形结构特征(图1中的后13维).

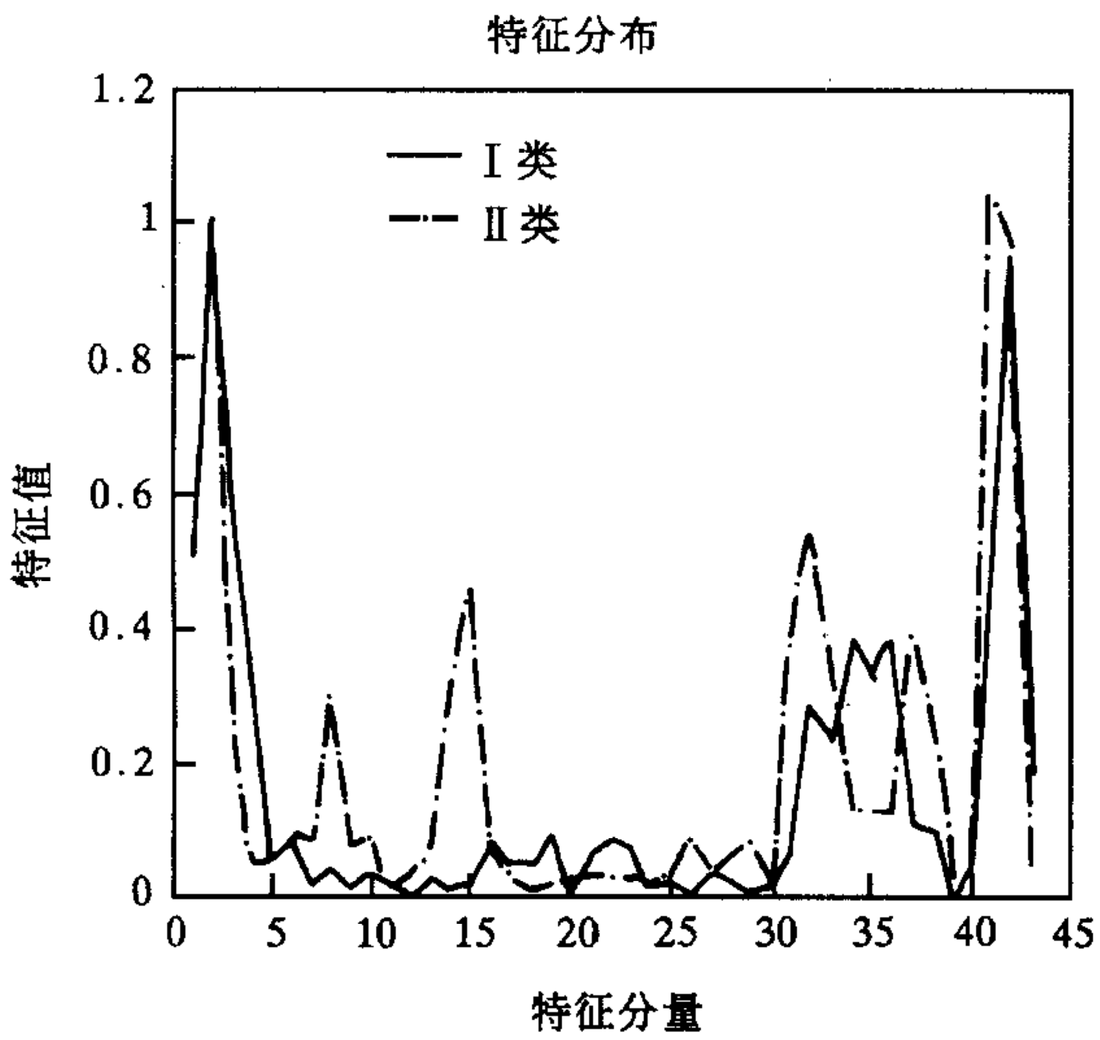

图1 两类目标的特征向量典例

本文考虑两类目标的分类问题,即 $C=2$.训练集取136个样本,测试集取396个样本.在同一训练集内,用30维倒谱特征、13维波形结构特征,以及它们简单合并后的43维特征分别训练结构自适应神经网络[6],得到三个品质最好的分类模型.它们在同一测试集的正确识别率如表1所示.

本文采用离散区间法估计互信息.考虑特征分量幅值的大与小所包含的信息量可能是等价的,文中取各特征分量的离散区间数目相同,均取10.这对融合从不同渠道得到的特征信息更有意义,因为不同类型特征之间难以实现一致性归一化.图2表示了43个特征分量各自包含的类别信息.

特征选择与相关性系数α有关.α的大小决定于模式的特征分布,文中采用试验法.

表1 各类特征集下测试结果比较

特征集	识别率(%)		
	Ⅰ类	Ⅱ类	Ⅲ类
30维倒谱	86.5	89.3	88.0
13维波形	87.1	83.9	85.4
43维合并	85.7	84.6	85.1

表2 特征融合后的测试结果

α	特征集	识别率(%)		
		Ⅰ类	Ⅱ类	平均
0.2	35	91.8	92.1	91.9
0.25	32	91.6	93.2	92.4
0.35	25	89.7	90.1	89.9

一般在不预定所选特征数目的情况下，α 越小选择的特征个数越多，反之越少. 用与计算表1相同的训练集、测试集和神经网络模型，选用不同的 α，得到的特征数目和融合性能如表2.

5 结　语

多类特征信息的综合选择是模式识别领域的重要环节，有效、快速的选择算法一直为人们所关注. 本文从理论上分析了特征选择的机理，提出了一种特征选择的性能指标. 基于这种性能指标，特征选择问题可用动态规划方法描述. 这样一个复杂的多类特征信息融合的全局满意解寻求过程，转变成一个简单的阶段性局部最优化问题，且由各阶段最优决策构成的整体策略等价于原问题的全局满意解. 本文法的特点是计算简单，所选特征数目可由算法自动确定，在本文的实例中取得了较满意的效果，其满意程度受相关性系数 α 的限制. 文中 α 需通过试验确定. 由于所采用的性能指标是一定前提下的近似最优，所以在一般意义下，本方法得到的不一定是最满意解.

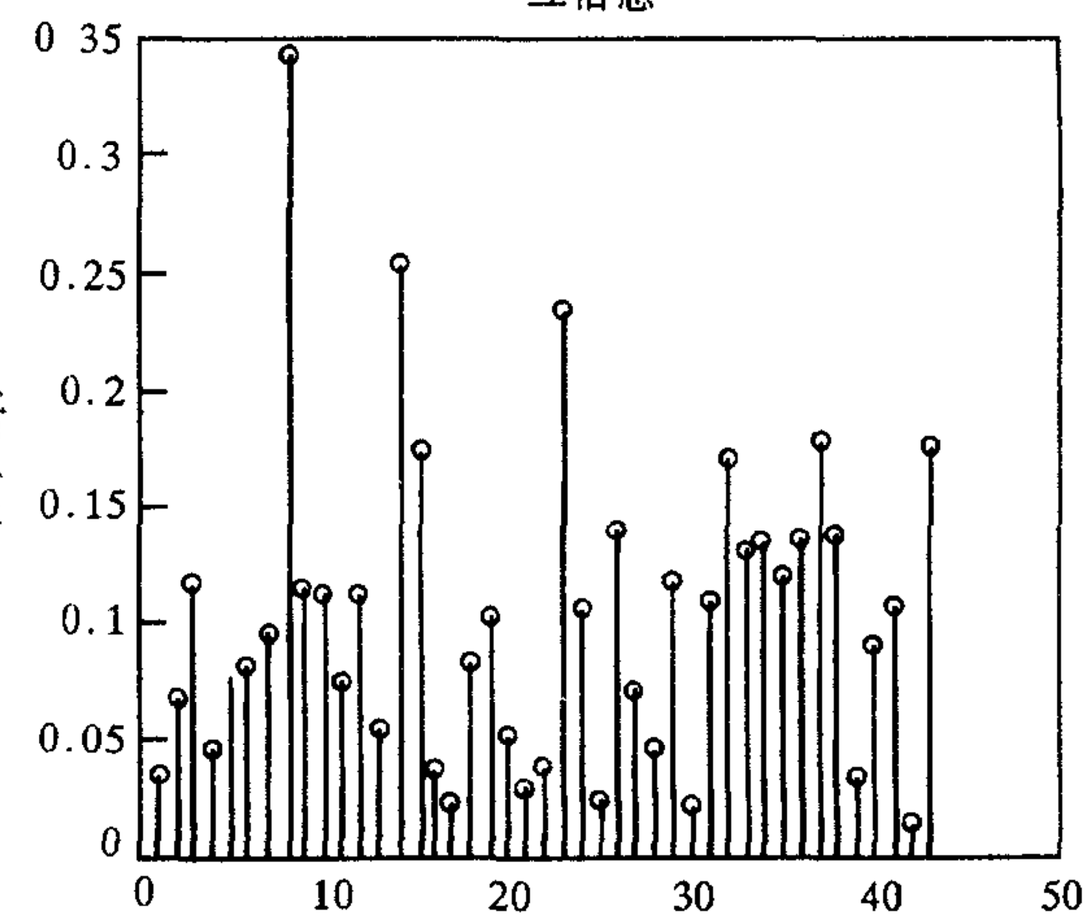

图2　各特征分量与类别空间的互信息

参 考 文 献

1 Elshaikl T S, Wacker A G. Effect of dimensionality and estimation on the performance of Gaussin classifiers. *IEEE trans. PAMI*, 1980, **2**(12):115—126

2 Foroutan I, Sklansky J. Feature selection for automatic classificatıon of non Gaussian data. *IEEE trans. SMC*, 1987, **17**(2):187—198

3 宣国荣，柴佩琪. 基于巴氏距离的特征选择，模式识别与人工智能，1996，**9**(4):324—329

4 Battiti R. Using mutual information for selecting features in supervised neural net learning. *IEEE trans. NN*, 1994, **5**(4):537—550

5 章新华. 遗传算法及其应用. 火力与指挥控制，1997，**22**(4):49—53

6 Zhang X H, Lin L, Wang J C. A neural network with self-organizing structure and its applications. In:Proc. of Inte. Conf. on Neural Information Processing, Beijing, 1995, **2**:906—909

DYNAMIC PROGRAMMING METHOD FOR FEATURE SELECTION

ZHANG XINHUA

(*Dalian Naval Academy*, *Dalian* 116018)

(*Dept. of Underwater Acoustical Engineering*, *Harbin Engineering University*, *Harbin* 150001)

Abstract The selection of multiple classes of features plays an important role in the field of pattern recognition. By analyzing the mechanism of feature selection, a performance measure of feature selection is proposed in this paper. Based on it, a dynamic programming method for feature selection is presented, by which a complex process of obtaining globally satisfactory solution to feature selection can be converted into a simple piecewise optimization. It is theoretically proved that the optimum strategy consisting of optimized decisions in each phase corresponds to the globally satisfactory solution of feature selection. The proposed approach is applied to feature selection of underwater acoustical signals.

Key words Feature selection, dynamic programming, pattern recognition, information fusion.