# 模糊特征选择新算法：Ⅱ *

尚修刚 蒋慰孙
（华东理工大学自动化研究所·上海，200237）

**摘要**：用模糊似然函数计算类内及类间距离，得到任意特征子集的模糊特征选择系数，用于特征子集的选择，从而得出最能区分和表征模式类的特征子集. 举例说明了该方法的具体用法，表明具有好的实用性.

**关键词**：模糊特征选择；模糊似然函数；特征选择系数

## 1 引 言

我们在文献[1]中利用模糊似然函数和模糊交互熵作度量工具，计算模式类内及类间距离，给出基于类内及类间距离的模糊特征选择系数，用以模糊特征选择. 用此法可以对描述模式的诸多特征的重要性进行有效的排序，从而得到最能表征和区分模式类的特征.

本文将用与文献[1]类似的方法，求出任意特征子集的模糊特征选择系数，用以计算各个特征子集的重要性. 当特征子集只有一个特征时，就回到文献[1]的情况.

## 2 特征选择系数

设有基本集合 $X=\{x_1,x_2,\cdots,x_n\}$，$A,B$ 是 $X$ 的任意两个模糊子集，$x_i$ 对 $A,B$ 的隶属度用 $\mu_A(x_i),\mu_B(x_i)(i=1,2,\cdots,n)$ 表示. 则模糊子集 $A,B$ 间的模糊似然函数定义为[3]

$$S(A,B)=\frac{1}{n}\sum_{i=1}^{n}\frac{\mu_{A\cap B}(x_i)}{\mu_{A\cup B}(x_i)}. \tag{1}$$

其值域为 $[0,1]$. 当 $A=B$ 时，其值为 1；当 $A,B$ 差异变大时，其值变小.

模糊似然函数能表示出集合 $A,B$ 间的差异程度，因此可以用于模糊特征选择.

设 $C_1,C_2,\cdots,C_j,\cdots,C_m$ 是 $N$ 维特征空间 $(X_1,X_2,\cdots,X_q,\cdots,X_N)$ 中的 $m$ 个模式类且类 $C_j$ 有 $n_j$ 个元素. 类 $C_j$ 的第 $i(i=1,2,\cdots,n_j)$ 个元素沿第 $q$ 个特征分量的隶属度函数记为 $\mu_{iq}^j$. 若有 $n$ 个特征的特征子集记为 $T_n=(t_1,t_2,\cdots,t_n)$，则以模糊似然函数 $S$ 度量的类 $C_j$ 中任意两个（如第 $s$ 和第 $t$ 个）元素以特征子集 $T_n$ 表征的差异程度为

$$S^j_{T_n st}=\frac{1}{n}\sum_{q=t_1}^{t_n}\frac{\mu_{sq}^j\wedge\mu_{tq}^j}{\mu_{sq}^j\vee\mu_{tq}^j}. \quad (s\neq t) \tag{2}$$

对于类 $C_j$ 的 $n_j$ 个元素的总平均值为

$$S^j_{T_n}=\frac{2}{n_j(n_j-1)}\sum_{s=1}^{n_j}\sum_{t=1}^{n_j}\left(\frac{1}{n}\sum_{q=t_1}^{t_n}\frac{\mu_{sq}^j\wedge\mu_{tq}^j}{\mu_{sq}^j\vee\mu_{tq}^j}\right). \quad (s\neq t) \tag{3}$$

类似地，可以求得以模糊似然函数 $S$ 度量的类 $C_k$ 的各元素沿这 $n$ 个特征分量的差异程度之平均值 $S^k_{T_n}$. 而类 $C_j$ 和类 $C_k$ 间所有的元素沿这 $n$ 个特征分量的差异程度的平均值定义为

$$S^{jk}_{T_n}=\frac{1}{n_jn_k}\sum_{s=1}^{n_j}\sum_{t=1}^{n_k}\left(\frac{1}{n}\sum_{q=t_1}^{t_n}\frac{\mu_{sq}^j\wedge\mu_{tq}^k}{\mu_{sq}^j\vee\mu_{tq}^k}\right). \tag{4}$$

* 国家自然科学基金重点资助项目(69334012).
本文于 1996 年 12 月 9 日收到. 1997 年 7 月 7 日收到修改稿.

注意，$S^{jk}_{T_n}$ 中的第 $s$ 和第 $t$ 个元素分别取自类 $C_j$ 和类 $C_k$.

令 $S'^{j}_{T_n}=1-S^{j}_{T_n}$，$S'^{k}_{T_n}=1-S^{k}_{T_n}$ 和 $S'^{jk}_{T_n}=1-S^{jk}_{T_n}$. $S'^{j}_{T_n}$，$S'^{k}_{T_n}$ 表示类内各元素间的平均“距离”，而 $S'^{jk}_{T_n}$ 表示类 $C_j$ 和类 $C_k$ 间各元素的平均“距离”，我们定义特征选择系数为[2]

$$(\mathrm{FEI})^{jk}_{T_n}=\frac{S^{j}_{T_n}+S^{k}_{T_n}}{S^{jk}_{T_n}}. \tag{5}$$

则特征子集 $T_n$ 用于表征和区分第 $j$ 和第 $k$ 个模式类的可靠性（或优势）增加时，$S'^{j}_{T_n}$，$S'^{k}_{T_n}$ 将递减，而 $S'^{jk}_{T_n}$ 将递增（即 $(\mathrm{FEI})^{jk}_{T_n}$ 递减），因而 $(\mathrm{FEI})^{jk}_{T_n}$ 值越小，表示特征子集 $T_n$ 对于模式分类越重要.

$(\mathrm{FEI})^{jk}_{T_n}$ 只是表示出特征子集 $T_n$ 用于表征和区分两个模式类的可靠性，考虑所有 $m$ 个类时，特征子集 $T_n$ 的平均可靠性定义为

$$(\mathrm{FEI})^{av}_{T_n}=\sum_{j}\sum_{k}(\mathrm{FEI})^{jk}_{T_n}W_jW_k. \tag{6}$$

其中，$W_j=n_j/n$，$W_k=n_k/n$，$n=\sum_{l}n_l$（$j,k=1,2,\cdots,m$ 且 $j\neq k$）. $W_j$ 和 $W_k$ 表示第 $j$ 和第 $k$ 个模式类的权重系数，它表明两对具有相同（FEI）值的类对 $(\mathrm{FEI})^{av}$ 贡献是不一样的，具有较多元素的那一对的贡献大. 这是符合一般逻辑的.

## 3　举　　例

为计算简单方便，这里仅考虑有两个模式类，每个模式类有三个特征的情况. 设模式 $C_1$ 有 40 个样本点，它们是以（0.35 0.4 0.4）为中心的一些点；模式 $C_2$ 也有 40 个样本，它们是以（0.65 0.5 0.65）为中心的一些点（为突出重点，将这些样本点放在附录中列出）.

下面用上述方法计算特征选择系数. 计算数据如下表所示.

表 1　特征选择系数

| | 一二特征 $S'$ 值 | 二三特征 $S'$ 值 | 一三特征 $S'$ 值 |
|---|---|---|---|
| $C_1$ 内部距离 | 0.1396 | 0.1334 | 0.1377 |
| $C_2$ 内部距离 | 0.09762 | 0.09806 | 0.08624 |
| $C_1+C_2$ 间距离 | 0.3319 | 0.2940 | 0.4191 |
| （FEI）值 | 0.715 | 0.787 | 0.534 |

从表可以看出，$(\mathrm{FEI})_{13}<(\mathrm{FEI})_{12}<(\mathrm{FEI})_{23}$，第一三特征组成的特征子集的（FEI）值最小，因而其区分和表征 $C_1$，$C_2$ 两模式类的重要性就最大. 就是说，用第一三特征进行分类时，更易将两个类分离开来. 第一二特征次之，二三特征重要性最小.

## 4　结　　论

本文用模糊似然函数计算各模式类内及类间的差异程度，以此定义模糊特征选择系数，可以用于计算任意特征子集对于区分模式类的重要程度. 该算法具有较强的适应性，可以用于任何情况的模糊特征选择. 而且，该算法象文献[1]一样，不局限于用 $S$ 函数作为基本特征选择函数，其它能用于表示两模糊集差异程度的函数亦可以用于本方法，使得本方法具有很强的广泛性.

## 参 考 文 献


1 尚修刚，蒋慰孙. 模糊特征选择新算法. CDC'97 控制与决策年会，庐山，1997，491－494

2 Pal，S. K. and Chakraborty，B.. Fuzzy set theoretic measure for automatic feature evaluation. IEEE Trans. Syst. Man and Cybern.，1986，16(5)：754－760

3 Shang，X. G. and Jiang，W. S.. A note on fuzzy information measures. Pattern Recognition Letters，1997，18：425－432

4 尚修刚，蒋慰孙. 一种新的模糊似然函数. 模式识别与人工智能，1997，10(1)：9－14


**附录** 模式类 $C_1$ 和 $C_2$ 的样本点

类 $C_1$ 的样本点：

(0.35 0.4 0.4)，(0.3 0.4 0.4)，(0.4 0.4 0.4)，(0.35 0.35 0.4)，(0.35 0.45 0.4)，(0.31 0.36 0.4)，(0.31 0.44 0.4)，(0.39 0.36 0.4)，(0.31 0.4 0.36)，(0.39 0.4 0.35)，(0.39 0.4 0.44)，(0.35 0.35 0.45)，(0.39 0.45 0.4)，(0.35 0.4 0.35)，(0.35 0.4 0.45)，(0.31 0.4 0.44)，(0.35 0.35 0.35)，(0.35 0.45 0.44)，(0.35 0.45 0.36)，(0.35 0.4 0.5)，(0.35 0.4 0.3)，(0.35 0.5 0.4)，(0.35 0.3 0.4)，(0.25 0.4 0.4)，(0.45 0.4 0.4)，(0.35 0.49 0.49)，(0.35 0.49 0.3)，(0.35 0.31 0.49)，(0.35 0.3 0.31)，(0.44 0.4 0.49)，(0.44 0.4 0.31)(0.26 0.4 0.5)，(0.26 0.4 0.3)，(0.26 0.49 0.4)，(0.26 0.3 0.4)，(0.45 0.5 0.4)，(0.45 0.3 0.4)，(0.35 0.42 0.42)，(0.33 0.38 0.4)，(0.37 0.37 0.38).

类 $C_2$ 的样本点：

(0.65 0.5 0.65)，(0.65 0.5 0.7)，(0.65 0.55 0.65)，(0.7 0.5 0.65)，(0.65 0.55 0.7)，(0.65 0.55 0.6)，(0.65 0.45 0.7)，(0.65 0.45 0.6)，(0.7 0.5 0.7)，(0.7 0.5 0.6)，(0.6 0.5 0.7)，(0.6 0.5 0.6)，(0.7 0.55 0.65)，(0.7 0.45 0.65)，(0.6 0.55 0.65)，(0.6 0.45 0.6)，(0.75 0.5 0.65)，(0.55 0.5 0.65)，(0.65 0.6 0.65)，(0.65 0.4 0.65)，(0.65 0.5 0.75)，(0.65 0.5 0.55)，(0.65 0.59 0.74)，(0.65 0.59 0.56)，(0.65 0.41 0.75)，(0.65 0.41 0.56)，(0.74 0.5 0.74)，(0.74 0.5 0.56)，(0.56 0.5 0.74)，(0.56 0.5 0.56)，(0.74 0.6 0.65)，(0.74 0.41 0.65)，(0.55 0.58 0.65)，(0.55 0.41 0.65)，(0.63 0.52 0.62)，(0.67 0.48 0.67)，(0.62 0.52 0.67)，(0.67 0.53 0.66)，(0.63 0.47 0.64)，(0.65 0.47 0.63).

# New Fuzzy Feature Selection Algorithm: Ⅱ


SHANG Xiugang and JIANG Weisun

(Research Institute of Automatic Control，East China University of Science and Technology · Shanghai，200237，PRC)



**Abstract**: The fuzzy feature selection index are defined using the distances of intraset and interset of pattern classes，which are calculated by fuzzy likelihood function. The algorithm can be used to compare the importance of any subset of features，so that the most important feature subset can be gotten. An example indicates how to use them in concrete situations，which shows that these methods are applicable.



**Key words**: fuzzy feature selection; fuzzy likelihood function; feature selection index


## 本文作者简介


**尚修刚** 1966年生. 1992年于华东化工学院获硕士学位. 1997年于华东理工大学自动化研究所获工业自动化专业博士学位. 现在上海邮电管理局工作. 主要研究方向：模式识别，信息论方法在控制中的应用，模糊控制等.

**蒋慰孙** 1926年生. 1947年毕业于上海交通大学化学系. 现任华东理工大学教授，博士生导师，自动化研究所名誉所长. 目前主要兴趣是：复杂工业过程建模，控制与优化，柔性过程系统，故障诊断及智能控制等.