第16卷 第4期 地 震 学 报 Vol. 16, No. 4
1994年11月 (433－439) ACTA SEISMOLOGICA SINICA Nov., 1994

对模式识别ICHAM方法的进一步改进和应用*

赵卫明 金延龙

(中国银川 750001 宁夏回族自治区地震局)

摘 要

本文对模式识别ICHAM方法又做了进一步改进，提出了用迭代中心作为Hamming核，并给出了新的计算欧氏距离时的权系数，通过改变控制参数寻找最小误识率. 经对两个算例和南北地震带北段强震的分类识别，结果表明，本文改进后的方法识别效果优于ICHAM方法，更适合于对一般分布样本做分类识别.

关键词 模式识别；迭代中心；控制参数；最小误识率

引 言

自70年代初模式识别方法应用于地震预报领域后，其分类判别方法经不断地改进，已由最初的二值CORA-3方法，以及后来的CORA-3修改方法和连续Hamming方法(吕宏伯等，1986)，逐步优化为目前在我国常用的改进的连续Hamming方法(王碧泉、王春珍，1988)，简称ICHAM方法. 笔者在实际应用中发现，由于受客观条件的限制，所能选取到样本的分布往往是非均匀的，用ICHAM方法做分类识别的误识率还比较高. 本文对此原因分析后，对ICHAM方法中Hamming核的选取和计算欧氏距离时的权系数，又做了进一步的改进，使之更适合于对一般分布样本的分类识别.

1 对ICHAM方法的改进

用模式识别方法对样本进行分类识别，就是用与样本的属性有关的特征，通过对已知属性样本的分类学习，寻求一种对已知属性样本的最佳分类方法，再用此方法对未知属性样本做分类识别. 若仅从数学处理上讲，当有 m 个特征时，总可以构造出一个或多个 m 维曲面，对已知属性样本作出全部正确的分类. 由于构造高维曲面的困难，以及选取的样本数量往往不够多，就只好构造有限个比较简单的曲面. ICHAM方法就是用方差较小的那一类样本特征的均值作为球心(即Hamming核)，用两类样本特征的均值之差作为权系数，构造一个用于二类样本分类识别的 m 维椭球面.

* 1993年1月16日收到初稿，1994年1月2日决定采用.

http://www.dizhenxb.org.cn

若要用构造一个 m 维椭球面的方法对二类样本做分类识别，所选取的某一特征的 Hamming 核，就应从距 Hamming 核较小的半径而包含尽可能多的同属性样本数两方面来考虑. 由于往往受条件的限制，所选取的样本呈非均匀分布，用方差较小的那一类样本的特征均值做 Hamming 核还不是最佳的选择. 例如，当用小震活动作为特征对某一地区 $M\geqslant M_0$ 和 $M<M_0$ 地震做分类识别时，由于受该地区地震台网监测能力的限制，以及震级频次关系的客观存在，所选取的 D 类样本往往是高震级的少，低震级的多，样本呈非均匀分布. 若以 D 类样本特征的均值作为 Hamming 核，则 Hamming 核的位置就有可能偏于 D 类样本的低震级端. 当分类判别的阈值取得较小时，就会将较多的高震级 D 类样本误判为 N 类样本；当阈值取得较大时，又会将较多的 N 类样本误判为 D 类样本. 所以，有必要对 Hamming 核位置的选取做进一部的改进.

我们对模式识别 ICHAM 方法的进一步改进步骤如下：

(1) 迭代中心的计算. 设有一序列点 x_1，x_2，…，x_n，其均值 $\overline{X}$ 和几何中心 X 分别为

$$\overline{X}=\frac{\sum_{i=1}^{n}x_i}{n} \tag{1}$$

$$X=\frac{[\max(x_i)+\min(x_i)]}{2} \tag{2}$$

则满足用较小的半径包含尽可能多的序列点的中心点就只能在($\overline{X}$，X)或(X，$\overline{X}$)之间选取. 本文提出了一个用迭代寻找这个中心点的方法，即

$$X_p^{(k)}=\frac{(\sum_{i=1}^{n}W_i^{(k)}x_i+X_p^{(k-1)})}{2} \tag{3}$$

其中

$$W_i=\frac{|x_i-X_p^{(k-1)}|^p}{\sum_{i=1}^{n}|x_i-X_p^{(k-1)}|^p} \tag{4}$$

以上二式的意义为，当控制参数 $p>0$ 时，某序列点距上一次迭代中心 $X_p^{(k-1)}$ 的距离越远，则对本次迭代中心 $X_p^{(k)}$ 的贡献越大. $p<0$ 时则反之，$p=0$ 时贡献相同. 经对多个序列检验表明，当控制参数 $p<3.0$ 时，无论初值 $X_p^{(0)}$ 取何值，迭代均有较快的收敛性和唯一的极限值. 当 $p=0$ 时，$X_0=\overline{X}$. 随着 p 的增大，X_p 逐渐偏离均值 $\overline{X}$ 而靠近几何中心 X，故在 $0\leqslant p<3.0$ 时有

$$|\overline{X}|\leqslant|X_p|<|X| \text{ 或 } |X|<|X_p|\leqslant|\overline{X}|$$

满足我们在均值 $\overline{X}$ 和几何中心 X 之间寻找最佳中心点的要求. 图 1 给出了一个序列的迭代中心 X_p 和相对于迭代中心的方差 S_p 随 p 的变化曲线.

从其含义上来说，均值 $\overline{X}$ 为序列优势分布的最佳中心点，其残差平方和最小，但易受序列点分布均匀程度的影响；几何中心 X 为仅考虑序列点整体分布的中心点，仅与序列分布两端的两个点有关；式(3)的迭代中心 X_p 则兼顾了序列的优势分布和整体分布. 控制参数 p 小时主要反映了序列的优势分布，p 大时则主要反映了序列的整体分布.

ACTA SEISMOLOGICA SINICA

(2) Hamming 核的选取. 设有 n 个样本，m 个特征. 其中，D 类和 N 类样本数分别为 n_1 和 n_2. 经标准化处理后的特征矩阵为 x_{ij}，$i=1,\cdots,n, j=1,\cdots,m$. D 类和 N 类样本的各特征方差之和为 $S^2_{(D)}$ 和 $S^2_{(N)}$(王碧泉、王春珍，1988). 分别将 D 类和 N 类样本的每一特征值代入式(3)，求出某一 p 值时的迭代中心和相对于迭代中心的方差，分别记为 $X_{pj}(D)$、$X_{pj}(N)$ 和 $\sigma_{pj}(D)$、$\sigma_{pj}(N)$，$j=1,\cdots,m$. Hamming 核取

$$z_{pj}=\begin{cases}X^{(D)}_{pj} & S^2(D)\leqslant S^2(N)\\ & j=1,\cdots,m\\ X^{(N)}_{pj} & S^2(D)>S^2(N)\end{cases}\quad(5)$$

式中，Hamming 核的位置是随控制参数 p 变化的.

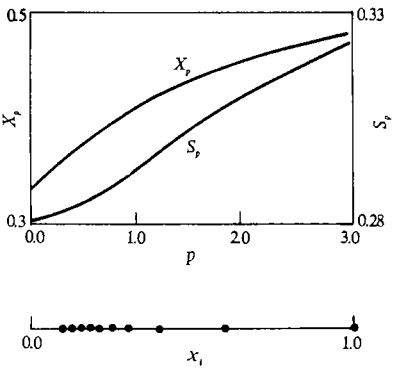

图 1　一个序列点的迭代中心 X_p 和迭代中心的方差 S_p 随 p 的变化

(3) 欧氏距离的计算. 本文仍采用加权欧氏距离计算各样本到 Hamming 核的距离 $d_i(X_i, Z, p, q)$

$$d_i(x_i, z, p, q)=\Big[\sum_{j=1}^{m}W_{qj}(x_{ij}-z_{pj})^2\Big]^{\frac{1}{2}}\quad(6)$$

权系数 W_{qj}取

$$W_{qj}=\frac{\left(\dfrac{|X^{(D)}_{pj}-X^{(N)}_{pj}|^q}{\sum_{j=1}^{m}|X^{(D)}_{pj}-X^{(N)}_{pj}|^q}+\dfrac{|\sigma^{(D)}_{pj}-\sigma^{(N)}_{pj}|^q}{\sum_{j=1}^{m}|\sigma^{(D)}_{pj}-\sigma^{(N)}_{pj}|^q}\right)}{|\sigma^{(D)}_{pj}+\sigma^{(N)}_{pj}|^g}\qquad j=1,\cdots,m\quad(7)$$

其中，q 为对权系数做控制的另一参数. 因为样本分布通常有两种典型的形式：一种是 D 类样本和 N 类样本各自成团分布，此时两类样本均值相差较大且方差之和较小的特征有利于分类，这相应于式(7)中的第一项；另一种是 D 类样本和 N 类样本成嵌套分布，两类样本的均值相差较小但方差相差较大. 此时均值对分类的作用减小，而方差对分类的作用增大. 两类样本方差相差较大的特征最有利于分类，这相应于式(7)中的第二项. 权系数 W_{qj}取式(7)就是为了对这两种形式的样本分布均能有所兼顾. 控制参数 q 的作用是为了进一步增大有利于分类的特征的贡献，而减小对分类作用较小的特征的贡献. q 值越大，其作用就越明显.

(4) 分类识别. 对用不同的控制参数 p 和 q 求出的 n 个样本的欧氏距离 $d_i(X_i, Z, p, q)$，其判别准则为

当 $S^2(D)\leqslant S^2(N)$时

$$\begin{cases}X_i\in D & d_i(X_i, Z, p, q)\leqslant T\\ X_i\in N & d_i(X_i, Z, p, q)>T\end{cases}\quad(8)$$

当 $S^2(D)>S^2(N)$时

$$\begin{cases}X_i\in N & d_i(X_i, Z, p, q)\leqslant T\\ X_i\in D & d_i(X_i, Z, p, q)>T\end{cases}\quad(9)$$

其中，T 为取定的阈值. 用下式计算出误识率 $\hat{\varepsilon}(p, q)$(王碧泉、王春珍，1988)

地震学报 ACTA SEISMOLOGICA SINICA

http://www.dizhenxb.org.cn

$$\hat{\varepsilon}(p, q) = \frac{\sum_{i=1}^{2} P(\omega_i)\tau_i}{n}$$

其中，$P(\omega_1)$、τ_1 分别为D类样本的先验概率和误判数，$P(\omega_2)$、τ_2 分别为N类样本的先验概率和误判数. 对于不同的控制参数 p 和 q，误识率 $\hat{\varepsilon}(p, q)$取最小值时的判别结果，即为最终分类判别结果.

控制参数 p 和 q 的作用是有所不同的. 误识率 $\hat{\varepsilon}(p, q)$取最小值时的 p 值越大，表明方差较小的那一类样本的分布均匀性越差，本文 Hamming 核距 ICHAM 方法 Hamming 核的距离就越远，q 值越大，则在一定程度上表现出了两类样本混杂的程度越高.

为了说明上述改进的特点和效果，我们取图2中的两个二维例子，分别用ICHAM

表1 两个二维算例的分类识别结果

样本序号	例1(n=27, n_1=14, n_2=13)					例2(n=28, n_1=14, n_2=14)				
	x_{i1}	x_{i2}	l_i	ICHAM	本 文	x_{i1}	x_{i2}	l_i	ICHAM	本 文
1	0.10	0.17	D	N	N	0.11	0.23	D	D	N
2	0.14	0.16	D	D	D	0.19	0.26	D	D	D
3	0.13	0.27	D	N	D	0.21	0.19	D	D	D
4	0.20	0.23	D	D	D	0.22	0.11	D	D	D
5	0.21	0.26	D	D	D	0.33	0.23	D	D	D
6	0.22	0.20	D	D	D	0.40	0.28	D	D	D
7	0.23	0.27	D	D	D	0.41	0.21	D	D	D
8	0.24	0.17	D	D	D	0.43	0.14	D	D	D
9	0.24	0.25	D	D	D	0.48	0.16	D	D	D
10	0.25	0.19	D	D	D	0.12	0.12	N	N	N
11	0.25	0.14	N	N	N	0.15	0.33	N	N	N
12	0.26	0.22	N	D	N	0.16	0.06	N	N	N
13	0.25	0.29	N	N	N	0.31	0.40	N	N	N
14	0.27	0.24	N	N	N	0.32	0.00	N	N	N
15	0.30	0.13	N	N	N	0.46	0.34	N	N	N
16	0.33	0.20	N	N	N	0.47	0.08	N	N	N
17	0.37	0.28	N	N	N	0.49	0.23	N	D	N
18	0.41	0.25	N	N	N	0.16	0.30	N	N	N
19	0.46	0.13	N	N	N	0.40	0.12	N	D	D
20	0.12	0.22	D	N	D	0.13	0.16	D	D	D
21	0.20	0.18	D	D	D	0.49	0.22	D	D	D
22	0.23	0.23	D	D	D	0.28	0.29	D	D	D
23	0.24	0.20	D	D	D	0.33	0.11	D	D	D
24	0.31	0.28	N	N	N	0.34	0.16	D	D	D
25	0.39	0.14	N	N	N	0.11	0.25	N	N	N
26	0.45	0.19	N	N	N	0.50	0.19	N	D	N
27	0.26	0.18	N	N	N	0.18	0.12	N	D	N
28						0.44	0.27	N	D	N
D类误判数 τ_1				3	1				0	1
N类误判数 τ_2				1	0				5	1
误识率 $\hat{\varepsilon}$				7.5%	1.9%				8.9%	3.6%

地震学报 ACTA SEISMOLOGICA SINICA

http://www.dizhenxb.org.cn

方法和本文改进的方法做了分类识别检验，表 1 中给出了重复检验时的分类结果. 例 1 中两类样本沿 x_1 轴各自成团分布. 用全部样本做重复检验时，ICHAM 方法的 Hamming 核为$(-0.70, 0.75)^T$，误识率 $\hat{\varepsilon}=7.5\%$. 本文的最终 Hamming 核为$(-0.77, 0.81)^T$，误识率 $\hat{\varepsilon}(p=0.7, q=4.8)=1.9\%$. 当用前 19 个样本学习、后 8 个样本检验时，ICHAM 方法的 Hamming 核为$(-0.66, 0.73)^T$，学习和检验的误识率分别为 $\hat{\varepsilon}=8.0\%$ 和 6.3%. 本文的最终 Hamming 核为$(-0.72, 0.80)^T$，误识率分别为 $\hat{\varepsilon}(p=0.5, q=4.6)=2.8\%$和 6.3%. 两种方法 Hamming 核位置的不同为识别结果不同的主要原因. 例 2 中两类样本成嵌套分布. 用全部样本做重复检验时，ICHAM 方法的 Hamming 核为$(0.021, -0.021)^T$，误识率 $\hat{\varepsilon}=8.9\%$. 本文的最终 Hamming 核为 $(0.023, -0.027)^T$，误识率 $\hat{\varepsilon}(p=0.4, q=0.7)=3.6\%$. 当用前 19 个样本学习、后 9 个样本检验时，ICHAM 方法的 Hamming 核为$(0.019, -0.017)^T$，误识率分别为 $\hat{\varepsilon}=5.3\%$和 22.2%. 本文的最终 Hmming 核为$(0.014, -0.021)^T$，误识率分别为 $\hat{\varepsilon}(p=0.1, q=9.8)=5.0\%$和6.2%. 两种方法 Hamming 核位置变化不大，权系数不同为识别结果不同的主要原因. 在这两个例子中，经本文改进后的方法的误识率均小于或等于 ICHAM 方法的误识率，且具有较好的稳定性，表明本文对 ICHAM 方法的改进是有一定效果的，更能适应于一般分布样本的分类识别.

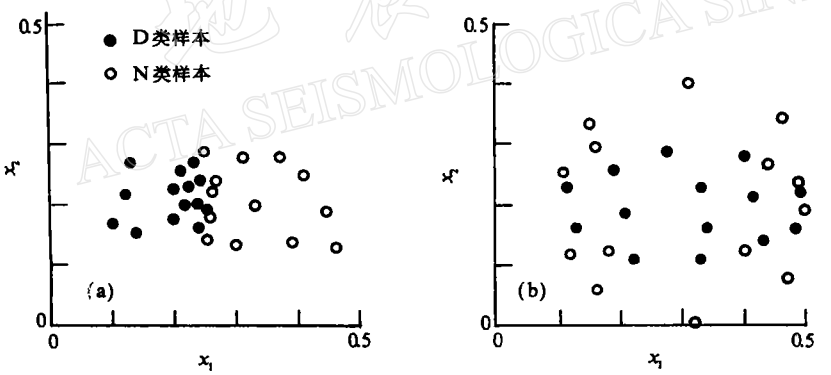

图 2　两个二维例子的样本分布

(a) 例 1，(b) 例 2

2 对南北地震带北段强震的分类识别

下面我们分别用 ICHAM 方法和本文进一步改进的方法，对南北地震带北段强震做了分类识别，以比较两种方法分类效果的差别：

(1) 资料选取. 考虑到地质构造特点、历史地震纪录情况和 $M_S \geqslant 6.0$ 地震的数量等，本文选取南北地震带北段(31°－40°N，100°－107°E)1881－1992 年 $M_S \geqslant 5.0$ 地震目录，并对余震做了统一删除. 删除余震的原则为：$M_S \geqslant 7.0$ 地震删除 2 年内的余震，6.0－6.9 级地震删除 1 年内的余震，5.5－5.9 级地震删除 6 个月内的余震，5.0－5.4 级地震删除 3 个月内的余震，震级相同的双震删除后一个地震.

http://www.dizhenxb.org.cn

样本的起止年份为 1901－1992 年，以每 2 年为一时段，共分为 46 段. 若在某时段内发生过 $M_S \geqslant 6.0$ 地震，则记该时段为 D 类样本(有震样本). 否则为 N 类样本(无震样本). 共有 D 类样本 20 个，N 类样本 26 个.

由于短时间内中强地震的数量有限，由此计算的统计参数稳定性差，本文仅选用中强地震频次作为分类特征. 在对各样本前 1－20 年的各种特征分析的基础上，并参照王碧泉等(1984)对地震活动特征的选取，共选取了以下 13 项单特征分类效果较好的特征：① 1 年内 $M_S \geqslant 5.0$ 地震次数；② 1 年内 $M_S \geqslant 5.5$ 地震次数；③ 1 年内 $M_S \geqslant 6.0$ 地震次数；④ 1 年内 $M_S \geqslant 6.5$ 地震次数；⑤ 8 年内 $M_S \geqslant 5.0$ 地震次数；⑥ 8 年内 $M_S \geqslant 5.5$ 地震次数；⑦ 12 年内 $M_S \geqslant 6.0$ 地震次数；⑧ 12 年内 $M_S \geqslant 6.5$ 地震次数；⑨ 12 年内 $M_S \geqslant 5.5$ 地震次数/$M_S \geqslant 5.0$ 地震次数；⑩ 12 年内 $M_S \geqslant 6.0$ 地震次数/$M_S \geqslant 5.0$ 地震次数；⑪ 12 年内 $M_S \geqslant 6.5$ 地震次数/$M_S \geqslant 5.0$ 地震次数；⑫ 8 年内 $M_S \geqslant 5.0$ 地震次数/16 年内 $M_S \geqslant 5.0$ 地震次数；⑬ 8 年内 $M_S \geqslant 5.5$ 地震次数/16 年内 $M_S \geqslant 5.5$ 地震次数.

(2) 分类检验. 对上面选取的 46 个样本，我们分别用 ICHAM 方法和本文改进后的方法做了重复检验和 U 方法检验，其分类和检验结果见表 2 和表 3. 在对全部样本的重复检验中，ICHAM 方法和本文改进后方法的误识率分别为 $\hat{\varepsilon}=10.6\%$ 和 $\hat{\varepsilon}(p=0.1,\ q=2.1)=8.4\%$；在 U 方法检验中，两种方法用 1960 年前的 30 个样本学习的误识率分别

表 2 南北地震带北段强震分类识别结果

序号	l_i	重复方法检验		U 方法检验		序号	l_i	重复方法检验		U 方法检验	
		ICHAM	本 文	ICHAM	本 文			ICHAM	本 文	ICHAM	本 文
1	N	N	N	N	N	24	N	N	N	N	N
2	D	N	N	N	N	25	D	N	D	N	D
3	N	N	N	N	N	26	D	D	D	D	D
4	N	N	N	N	N	27	D	N	N	N	D
5	N	N	N	N	N	28	N	N	N	N	N
6	N	N	N	N	N	29	D	D	D	D	D
7	N	N	N	N	N	30	D	D	D	N	D
8	N	N	N	N	N	31	N	N	N	N	N
9	N	N	N	N	N	32	N	N	N	N	N
10	D	N	N	N	N	33	N	D	D	N	N
11	N	N	N	N	N	34	D	N	N	N	N
12	D	D	D	D	D	35	N	D	N	N	N
13	N	D	D	D	D	36	N	N	N	N	N
14	D	D	D	D	D	37	D	D	D	N	N
15	D	D	D	D	D	38	D	D	D	D	D
16	N	N	N	D	D	39	N	N	N	N	N
17	D	D	D	D	D	40	N	D	D	D	D
18	D	D	D	D	D	41	D	D	D	D	D
19	D	N	N	N	N	42	N	N	N	D	N
20	N	N	N	N	N	43	D	D	D	D	D
21	N	N	N	D	N	44	D	D	D	D	D
22	N	N	N	N	N	45	D	D	D	D	D
23	N	N	N	N	N	46	N	N	N	N	N

ACTA SEISMOLOGICA SINICA

为 $\hat{\varepsilon}=14.3\%$ 和 $\hat{\varepsilon}(p=0.1, q=2.7)=8.1\%$；用 1961 年后的 15 个样本检验的误识率分别为 $\hat{\varepsilon}=12.5\%$ 和 $\hat{\varepsilon}(p=0.1, q=2.7)=9.1\%$. 可以看出，在此实例的重复检验和 U 方法检验中，本文改进后的方法识别效果均优于 ICHAM 方法，并且其识别结果具有较好的稳定性.

表 3　两种方法分类识别结果比较

检验方法	样　　本	识别方法	τ_1	τ_2	$\hat{\varepsilon}$
重复检验	全部样本用于学习和检验 ($n_1=20, n_2=26$)	ICHAM	6	4	10.6%
		本文	5	3	8.4%
U 方法检验	1960 年前样本用于学习 ($n_1=13, n_2=17$)	ICHAM	6	3	14.3%
		本文	3	2	8.1%
	1961 年后样本用于检验 ($n_1=7, n_2=9$)	ICHAM	2	2	12.5%
		本文	2	1	9.1%

3 结语和讨论

通过本文用两种方法对两个例子和一个应用实例做分类识别结果的差异，可以看出，当用构造一个椭球面的方法对二类样本做分类识别时，Hamming 核的位置和计算欧氏距离的权系数的选择都是很重要的，都会影响最终的分类识别结果. 本文用改变 Hamming 核的位置和权系数作用的方法，虽然能提高对一般分布样本的分类识别能力，但还不能说是对两类样本的最佳分类识别方法，并且计算量也比较大. 为了用较少的计算量而得到最优的计算结果，能否用样本的一些简单统计量，或是一些准则来给出最优 Hamming 核的位置和权系数. 比如，能否用构造一个有关 Hamming 的位置和距 Hamming 核任意距离内的同属性样本比例数的函数，对这个函数求极值的方法求取 Hamming 核.

另外，本文提出的迭代中心 X_p，也可以理解为一种不等权的均值，$|p|$ 越大，不等权的作用就越明显. $p>0$ 时为主要考虑全部序列点分布的不等权均值，如在本文中的应用. $p<0$ 时为主要考虑分布相对集中的部分序列点的不等权均值，可用于前兆数据处理时的均值计算.

参 考 文 献

吕宏伯、聂金宗、陈祖荫、马秀芳、王碧泉，1986. Hamming 分类方法的改进及其在地震危险区划中的应用. 地震学报，**8**，增刊，143－153.

王碧泉、陈祖荫、王春珍，1984. 用聚类分析法研究强震的孕震过程. 地震学报，**6**，121－128.

王碧泉、王春珍，1988. 对连续 Hamming 方法的改进及其在潜在震源判定中的应用. 地震学报，**10**，113－122.

http://www.dizhenxb.org.cn