文章编号: 1001-0920 (2013) 04-0495-06

基于证据理论的纠错输出编码解决多类分类问题

周进登[1,2], 王晓丹[1], 崔永花[2], 任宏洋[2]

(1. 空军工程大学 导弹学院, 陕西 三原 713800; 2. 复杂系统第四实验室, 北京 100076)

摘 要: 针对多类分类问题, 利用纠错输出编码作为分解框架, 把多类问题转化为多个二类问题加以解决; 同时提出一种基于证据理论的解码策略, 把每一个二分器的输出作为证据之一进行融合, 并讨论在两种编码类型 (二元和三元编码矩阵) 下证据融合的不同策略. 通过实验分别对 UCI 数据集和 3 种一维距离像数据集进行测试, 并与几种经典的解码方法进行比较, 验证了所提出的方法能有效提高纠错输出编码特别是三元编码矩阵的分类正确率.

关键词: 模式识别; 多类分类; 纠错输出编码; 证据理论

中图分类号: TP181 **文献标志码:** A

Error-correcting output codes based on evidence theory for multi-class classification

ZHOU Jin-deng[1,2], *WANG Xiao-dan*[1], *CUI Yong-hua*[2], *REN Hong-yang*[2]

(1. Missile Institute, Air Force Engineering University, Sanyuan 713800, China; 2. The Fourth Lab of Complex System, Beijing 100076, China. Correspondent: ZHOU Jin-deng, E-mail: zhoujin198417@yahoo.com.cn)

Abstract: To model multi-class classification problems, error correcting output codes(ECOC) are used as decomposing frame to reduce multi-class to binary. A decoding strategy based on DS evidence theory is proposed, which takes every binary learner's output as evidence to fusion and discusses different DS evidence fusion strategies based on two kinds of coding type(binary ECOC and ternary ECOC). Experimental results on UCI and three kinds of HRRP show that the proposed scheme provides better performance of error correcting output codes especially ternary ECOC than other state-of-the-art encoding strategies.

Key words: pattern recognition; multi-class classification; error-correcting output codes; DS evidence theory

0 引 言

利用纠错输出编码 (ECOC) 解决多类分类问题是当前模式识别领域中的研究热点. 它利用二元纠错输出编码 (binary ECOC) 或三元纠错输出编码矩阵 (ternary ECOC) 作为分解框架, 把多类分类问题转化为多个二类问题, 从而能有效地利用经典的两类分类方法, 目前已成功地应用于人脸识别[1]、文本识别[2]、手写数字分类[3]以及交通指示牌识别[4]等领域, 并取得了很好的识别效果.

在利用纠错输出编码解决多类分类问题领域中, 编码和解码是应用此类方法最重要的两个步骤. 在编码方面的研究有: BCH 编码[5-6]、无遗编码[5,7]、随机编码 (按编码阵中选取元素的不同又分为密集随机编码和稀疏随机编码)[8]以及搜索编码[9]. 此外, 由文献 [7] 可知, 一些经典的多类分类方法, 如 one-versus-all[10]和 one-versus-one[11]也可以认为是 ECOC 框架下编码方法的一种.

解码作为基于纠错输出编码决策融合阶段, 其融合策略的选取直接影响着分类性能的好坏. 该阶段可视为通信处理过程中对信号的解码. 一个学习任务可建模为一种信号的传输, 信源端信号的编码对应于利用纠错输出编码对多类进行分解, 信宿端的信号解码对应于二分器分类结果的融合过程. 这方面的研究有经典的汉明距离解码[5]; 文献 [12] 在此基础上利用二分器软输出向量提出了一种改进的汉明距离解码 —— 逆汉明距离解码; 欧氏距离解码是另一种基于距离的解码策略[11], 它引入空间欧氏距离作为类别差异性大小的度量, 该解码方法能有效地克服三元纠错

收稿日期: 2011-09-18; **修回日期:** 2012-05-31.

基金项目: 国家自然科学基金项目(60975026, 61273275).

作者简介: 周进登(1984−), 男, 博士生, 从事智能信息处理和机器学习的研究; 王晓丹(1966−), 女, 教授, 博士生导师, 从事智能信息处理和机器学习等研究.

输出编码中由“0”元素带来的解码误差[8]. 此外, 还有中心距解码[11], 基于损失函数解码[7, 13]以及后验概率解码[14]等.

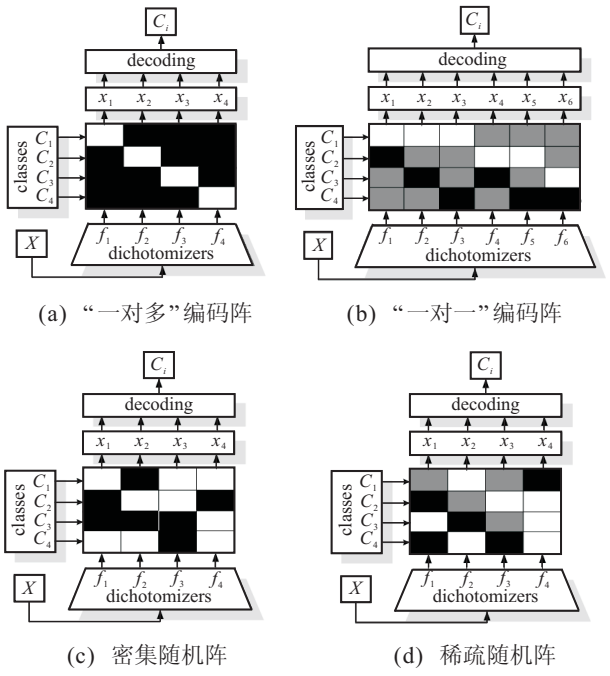
注意到利用纠错输出编码对多类进行分解后, 原来的多类问题转换成了二类分类问题. 在解码阶段(即决策融合阶段), 得到的是每一个二分器的输出组成的一组结果向量, 而每个二分器的结果可以看成是测试样本属于某一类或几类的概率大小. 因此本文引入证据理论作为融合这一结果向量的解码策略, 把每一个二分器的输出作为一条证据, 利用证据理论具有深厚的数学理论基础, 能够依靠证据(二分器输出向量)的积累不断缩小假设集的特点, 使解码的正确率逐步提高, 并利用证据理论具有区分“不知道”和“不确定”的能力, 对决策分类提供包含拒绝域的功能, 此特点对于一些具有分类风险大的样本尤为重要(如医学领域中对癌细胞的辨别分类).

本文首先简要介绍ECOC框架解决多类分类和证据理论的相关知识; 然后给出两种不同类型编码矩阵的DS解码策略; 最后给出了实验结果和分析.

1 相关概念介绍

1.1 纠错输出编码(ECOC)

ECOC框架利用一种二元或三元编码矩阵实现多类类别分解和基分类器集成. 在其编码矩阵中, 二元码用$\{-1, +1\}$表示, 三元码用$\{-1, 0, +1\}$表示. 其中: “−1”代表一类, “+1”代表另一类, “0”表示该码字位所对应的类在其列所形成的二类划分中被忽略(即不参与由该列所产生的基分类器的训练). 图1给出了4种常见的ECOC分类系统示意, 以编码矩阵来区分, 分别是: “一对多”编码阵、“一对一”编码阵、密集随机阵以及稀疏随机阵.



(a) “一对多”编码阵 (b) “一对一”编码阵

(c) 密集随机阵 (d) 稀疏随机阵

图 1 4种常见的ECOC

图1中所有编码阵的每一行代表某一类$C_i\,(i = 1, 2, 3, 4)$的码字, 每一列代表样本的一种二类划分, 码元“1”、“−1”和“0”分别用白色、黑色和灰色表示. 在训练阶段, 每一个基分类器$f_i\,(i = 1, 2, \cdots, 6)$的训练样本先根据其在编码阵中对应的列重新划分, 把多类转化成二类; 然后分别训练并得到与该列对应的二类分类器. 例如, 在图1(d)中对基分类器f_3进行训练时, 白色对应的C_2为一类, 黑色对应的C_4为另一类, 而灰色对应的C_1和C_3不参与该基分类器的训练. 由此依次训练, 得到4个二类分类器$\{f_1, f_2, f_3, f_4\}$. 在测试阶段, 给定一个测试样本X, 同时利用这4个二类分类器对其进行分类, 结果为一码字向量(x_1, x_2, x_3, x_4)(其中$x_i \in \{-1, +1\}$), 最后根据某种解码规则(即融合策略)对其进行解码即可得到最终分类结果.

1.2 DS证据理论

在证据理论中, 一个样本空间称为一个辨识框架, 常用Θ表示, 它是关于命题的彼此独立的可能答案或假设的一个有限集合, Θ是完备的且其中的元素互不相容. Θ的幂集记为2^{Θ}. 证据理论的基本问题是: 在已知辨识框架Θ的条件下, 判明Θ中的一个先验的未定元素属于Θ中某一个子集的程度. 如果集函数$m: 2^{\Theta} \to [0, 1]$满足

$$m(\varnothing) = 0,\ \sum_{A \subseteq \Theta} m(A) = 1, \tag{1}$$

则称m为Θ上的基本概率赋值(BPA)函数或称mass函数. $\forall A \subseteq \Theta, m(A)$称为$A$的基本可信度, A为焦元[15]. 在证据理论中, 对事件A的不确定度采用区间$[\mathrm{Bel}(A), \mathrm{Pl}(A)]$描述, Bel和Pl分别称为信任函数和似真函数, 与BPA存在如下关系:

$$\begin{aligned} \mathrm{Bel}(A) &= \sum_{B \subseteq A} m(B), \\ \mathrm{Pl}(A) &= \sum_{A \cap B \neq \varnothing} m(B). \end{aligned} \tag{2}$$

设Bel_1和Bel_2为同一辨识框架下的两个信任函数, m_1和m_2为相应的BPA函数, 其焦元分别为$A_1, A_2, \cdots, A_p$和$B_1, B_2, \cdots, B_q$. 则对于给定命题$A \subseteq \Theta$, 两个证据可通过如下Dempster规则进行组合:

$$\begin{aligned} m(A) &= m_1 \oplus m_2(A) = \\ &\frac{1}{1-q} \sum_{A_i \cap B_j = A} m_1(A_i) m_2(B_j), \\ q &= \sum_{A_i \cap B_j = \varnothing} m_1(A_i) m_2(B_j), \end{aligned} \tag{3}$$

其中q反映了m_1和m_2之间冲突的大小. $q = 0$意味着m_1和m_2之间完全没有冲突; 而$q = 1$则说明m_1

和 m_2 之间是完全冲突的. 系数 $1/(1-q)$ 称为归一化因子, 其作用是避免在合成时将非 0 的信任赋给空集.

2　基于 DS 证据理论的 ECOC 解码策略

由 1.1 节可知, 纠错输出编码的每一列对应于多类(假设有 $K \geqslant 3$ 类)的一个二类划分, 而每一个二类划分都对应于此 K 类组成的幂集(2^K)中的一个元素. 因此, 可以将 K 类视为一个辨识框架 $\Theta = \{\text{class}_1, \text{class}_2, \cdots, \text{class}_K\}$, 把基于 ECOC 的多类分类过程视为一个证据推理过程, 而 ECOC 每一列的输出对应于一个证据. 在解码阶段, 解码策略将对应于证据理论的融合策略. 下面将讨论基于 ECOC 的 BPA 获取; 然后就两种不同类型的纠错输出编码——二元码(binary ECOC)和三元码(ternary ECOC)分别讨论其解码策略, 即不同的证据组合策略.

2.1　基于 ECOC 的 BPA 获取

假设 ECOC 为 $M \in \{-1, 0, +1\}^{K\times L}$, 由于 BPA 需满足式(1)的要求, 在基于 ECOC 获取每个证据的 BPA 时, 首先利用 Platt 概率估计模型来获得第 l 个基分类器的概率输出[16], 即

$$p_l(y=1|f_l) = \frac{1}{1+\exp(Cf_l + D)}. \tag{4}$$

其中: f_l 为第 l 个基分类器的输出量, C 和 D 通过求负 log 似然函数的极小值得到. 假定第 l 个基分类器的分类精度为 r_l, 则可定义该基分类器 BPA 函数为

$$\begin{aligned} &m_l(A_l) = r_l p_l,\\ &m_l(\bar{A}_l) = r_l(1-p_l),\\ &m_l(\Theta) = 1 - r_l, \end{aligned} \tag{5}$$

其中 A_l 为编码矩阵中第 l 列对应的标签为"+1"的类, 也称 BPA 函数 m_l 的一个焦元. 因此, 每一个 BPA 函数对应的焦元数都是相同的, 且均为 4 个, 即 $\{\varnothing, A_l, \bar{A}_l, \Theta\}$. 对于两种不同类型编码矩阵, ternary ECOC 由于"0"元素的引入, 导致某个证据对某种类别不提供任何信息, 在证据融合过程中 ternary ECOC 较 binary ECOC 更容易产生证据冲突, 其在融合过程中将会遇到不同的问题, 下面将分别加以讨论.

2.2　Binary ECOC 下的 DS 证据理论解码策略

在编码矩阵为 binary ECOC 的情况下, 由于每一个基分类器对应的 BPA 函数中的焦元 A_i 与其他基分类器对应的 BPA 函数中的焦元 $\{A_j, \bar{A}_j\}$ 中的某一个总存在交集, 各 BPA 之间不会存在严重的冲突, 可采用 Dempster 规则进行组合(如式(3)所示). 对第 k 类的组合 BPA 值计算步骤如下.

Step 1: 首先计算 f_1 和 f_2 的组合值

$$\begin{aligned} &m_{12}(\text{class}_k) = \\ &\frac{1}{1-q}[m_1(A_1)m_2(A_2)(\text{or } m_1(A_1)m_2(\bar{A}_2)) + \cdots + \\ &m_1(A_1)m_2(\Theta) + m_1(\Theta)m_2(\Theta)]. \end{aligned} \tag{6}$$

其中

$$\begin{aligned} &\text{class}_k \in A_1 \bigcap A_2 \ (\text{or } \text{class}_k \in A_1 \bigcap \bar{A}_2),\\ &q = m_1(A_1)m_2(A_2)(\text{or } m_1(A_1)m_2(\bar{A}_2)),\\ &A_1 \bigcap A_2 = \varnothing (\text{or } A_1 \bigcap \bar{A}_2 = \varnothing). \end{aligned}$$

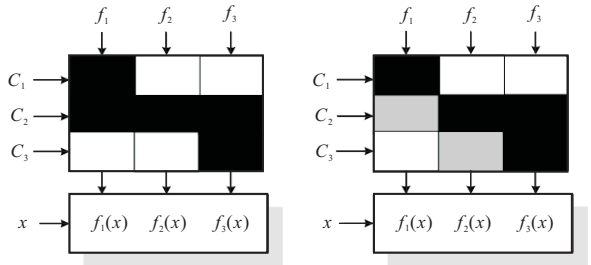
Step 2: 归一化处理

$$\begin{aligned} &\bar{m}_{12}(\text{class}_k) = \frac{m_{12}(\text{class}_k)}{\sum_{k=1}^{K} m_{12}(\text{class}_k) + m(\Theta)},\\ &\bar{m}(\Theta) = 1 - \sum_{k=1}^{K} \bar{m}_{12}(\text{class}_k). \end{aligned} \tag{7}$$

Step 3: 因为该规则满足交换律和结合律, 所以多个证据可通过逐次应用该规则进行组合, 把计算得到的 $m_{12}(\text{class}_k)$ 再与 f_3 的 BPA 值进行组合; 如此循环, 直到所有基分类器对应的 BPA 值都组合到上一项组合值中, 从而得到最终的 BPA 组合值为

$$m(\text{class}_k) = ((m_1 \oplus m_2) \oplus m_3) \oplus \cdots \oplus m_L. \tag{8}$$

例如, 假设有 3 类问题, 其 ECOC 编码矩阵如图 2(a)所示, 且假定各基分类器的分类正确率和对测试样本 x 正类的输出概率如表 1 所示, 则可计算各基分类器对应各焦元的 BPA 值如表 2 所示.



(a) Binary ECOC　　(b) Ternary ECOC

图 2　3 类 ECOC 编码矩阵

表 1　各基分类器分类正确率及概率输出

基分类器	分类正确率(r_i)	对x正类概率输出(p_i)
f_1	0.83	0.34
f_2	0.76	0.66
f_3	0.58	0.85

表 2　各焦元对应的 BPA 赋值

焦元	m_1	焦元	m_2	焦元	m_3
$A_1(C_1, C_2)$	0.55	$A_2(C_2)$	0.50	$A_3(C_2, C_3)$	0.49
$\bar{A}_1(C_3)$	0.28	$\bar{A}_2(C_1, C_3)$	0.26	$\bar{A}_3(C_1)$	0.09
Θ	0.17	Θ	0.24	Θ	0.42

根据式(6)～(8), 组合m_1和m_2可得$m'(C_1)=0.32$, $m'(C_2)=0.48$, $m'(C_3)=0.16$, $m'(\Theta)=0.04$; 然后, 组合m'和m_3可得$m(C_1)=0.22$, $m(C_2)=0.56$, $m(C_1)=0.20$, $m(\Theta)=0.02$. 因此, 根据最大信念决策规则, 样本x属于第2类. 注意到由基分类器输出组成的向量与第2类的汉明距离最小, 若使用汉明距离解码, 则最终分类结果也为C_2类.

2.3 Ternary ECOC下的DS证据理论解码策略

当编码矩阵为ternary ECOC时, 由于在基分类器中存在被忽略的一种或几种类别数据, 导致存在两列不同基分类器对应的焦元集的交集可能为空集, 在组合此两列基分类器对应的BPA时可能存在严重的冲突, 适用于binary ECOC的组合公式(6)～(8)将不再具有适用性. 例如, 仍然假设有3类分类问题, 编码矩阵如图2(b)所示, 且各基分类器的分类正确率和对测试样本x正类的输出概率仍如表1所示, 则计算各基分类器所对应的各焦元BPA值如表3所示.

表3 各焦元对应的BPA赋值

焦元	m_1	焦元	m_2	焦元	m_3
$A_1(C_1)$	0.55	$A_2(C_2)$	0.50	$A_3(C_2,C_3)$	0.49
$\bar{A}_1(C_3)$	0.28	$\bar{A}_2(C_1)$	0.26	$\bar{A}_3(C_1)$	0.09
Θ	0.17	Θ	0.24	Θ	0.42

根据式(8)的组合规则, 首先组合m_1和m_2得到$m'(C_1)=0.56$, $m'(C_2)=0.20$, $m'(C_3)=0.17$, $m'(\Theta)=0.07$; 然后, 组合m'和m_3得到$m(C_1)=0.44$, $m(C_2)=0.28$, $m(C_3)=0.25$, $m(\Theta)=0.03$. 此时, 根据最大信念决策规则, 样本x属于第1类, 很明显, 这个结果与事实不符. 可以观察到在第1次组合时, 因为基分类器f_1和f_2中只存在交集C_1, 不存在交集C_2和C_3, 所以测试样本属于C_1的概率被加强, 属于C_2和C_3的概率被减弱, 即使第3条证据强烈支持C_2和C_3. 因此, 必须考虑新的组合策略作为ternary ECOC的解码策略.

为解决上述问题, 本文引入加权组合策略, 每一次组合对最终BPA形成的贡献的重要度由一个权值系数来衡量, 该权值系数按下式获取:

$$\gamma_{i,j}^k=\begin{cases}\dfrac{K-\mathrm{count}\{\mathrm{class}_{\bar{k}}|M(\bar{k},i)=0 \text{ or } M(\bar{k},j)=0\}}{K},\\ \mathrm{class}_k\in A_i\bigcap A_j \text{ or } \mathrm{class}_k\in A_i\bigcap\bar{A}_j;\\ \dfrac{\mathrm{count}\{\mathrm{class}_{\bar{k}}|M(\bar{k},i)=0 \text{ or } M(\bar{k},j)=0\}}{K},\\ \mathrm{class}_k\notin A_i\bigcap A_j \text{ and } \mathrm{class}_k\notin A_i\bigcap\bar{A}_j.\end{cases}\tag{9}$$

其含义是: 当计算第k类的BPA时, 需要组合的第i条和第j条证据所对应的权值系数与该两条证据对应的焦元集与是否都包含该类有关, 当$\mathrm{class}_k\in A_i\bigcap A_j$ or $\mathrm{class}_k\in A_i\bigcap\bar{A}_j$时, $\gamma_{i,j}^k$等于与之对应的编码矩阵中不被"忽略"(即类别码元值为"+1"或"−1")的类别比率; 反之, 则为$1-\gamma_{i,j}^k$. 即当两条证据中不支持的类别数目越多时, 组合后的证据对各类别最终BPA的形成贡献越小; 当需要组合的两条证据对所有类别都提供支持度时, 其权值为1. 于是, 本文得到编码矩阵为ternary ECOC下的DS组合策略为

$$m_{12}(\mathrm{class}_k)=\frac{\gamma_{ij}^k}{1-q}[m_i(A_i)m_j(A_j \text{ or } \bar{A}_j)+\cdots+m_i(A_i)m_j(\Theta)+m_i(\Theta)m_j(\Theta)].\tag{10}$$

其中

$$\mathrm{class}_k\in A_i\bigcap A_j \ (\text{or } \mathrm{class}_k\in A_i\bigcap\bar{A}_j),$$
$$q=m_i(A_i)m_j(A_j)(\text{or } m_i(A_i)m_j(\bar{A}_j)),$$
$$A_i\bigcap A_j=\varnothing(\text{or } A_i\bigcap\bar{A}_j=\varnothing).$$

利用式(10)、(7)和(8)重新计算本节开头所述案例, 得到最终结果为: $m(C_1)=0.27$, $m(C_2)=0.38$, $m(C_3)=0.32$, $m(\Theta)=0.03$. 根据最大信念决策规则, 样本x属于第2类, 结果与2.2节一样, 因此, 本节引入的加权组合策略在编码矩阵为ternary ECOC时能较好地避免证据之间的冲突, 从而使结果更加准确.

3 实验研究

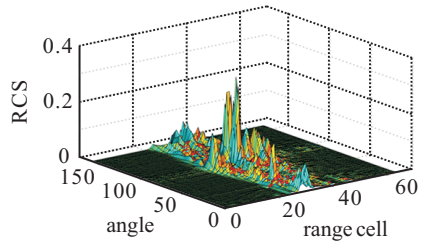
为验证本文提出的两种不同类型编码矩阵下基于DS证据理论的解码策略的应用效果, 本节将采用两种不同数据进行验证.

3.1 实验数据

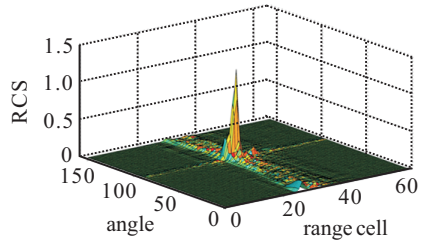
在实验中将用到两类数据集: 第1类为UCI数据集, 第2类为3种不同目标的一维距离像数据集. 表4为UCI数据集及各类数据描述, 其中部分UCI数据集进行了归一化处理, 并删除了一些样本数很小的类; 同时为了提高分类速度, 实验中对高维数据使用主成分分析法(PCA)对其进行降维处理. 图3为3种不同飞行器(B-52、J-6和J-7)的一维距离像示意图.

表4 UCI数据集各数据描述

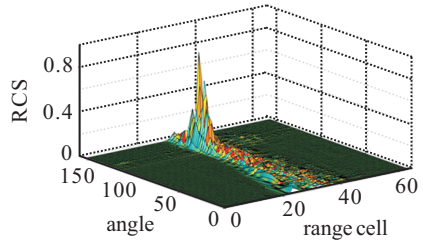
problem	train	attributes	classes
Yeast	1 484	8	10
Segmentation	2 310	19	7
Sat	6 435	36	6
Glass	214	9	7
Page-blocks	5 743	10	5
Vehicle	846	18	4
Zoo	101	18	7
Shuttle	14 500	9	7



(a) B-52



(b) J-6



(c) J-7

图 3　3 种不同飞行器的一维雷达距离像

3.2　实验设计

首先, 利用 UCI 数据集检验基于 DS 证据理论的解码策略在两种不同类型编码矩阵下的分类效果, 并分别与几种经典的解码方法进行比较, 它们分别是: 汉明距离解码 (HD), 逆汉明距离 (IHD), 欧氏距离解码 (ED), 泛欧氏距离解码 (AED), 损失函数解码 (包括线性损失函数解码 (LLB) 和指数函数解码 (ELB)) 以及最小二乘解码 (LS). 然后, 将该一整套方法用于多类目标一维距离像的目标识别中, 以检验此类方法的实际应用效果.

在 UCI 数据中选择两种具有代表性的二元编码和三元编码矩阵, 即密集随机编码和稀疏随机编码, 两种随机编码的码长分别取 $\lceil 10\log_2(Nc)\rceil$ 和 $\lceil 15\log_2(Nc)\rceil$. 在对 3 种一维距离像的识别过程, 分别提取 3 个不同角度范围 (0°～100°, 80°～155° 和 0°～155°) 的距离像加以识别. 实验中选取决策树分类器作为基分类器, 在估计分类错误率时为保证估计的准确性, 样本数据个数大于 500 时采用 10 重交叉验证, 小于 500 时采用 5 重交叉验证来进行. 利用双边估计 t 检验法来计算置信水平为 0.95 的分类错误率并作为最终结果, 计算公式如下:

$$\frac{|\bar{x}-u|}{\sigma/\sqrt{n}} \geqslant t_{0.025}(n-1). \tag{11}$$

其中: u 和 σ 分别表示 n 重交叉验证的均值和标准差, $t_{0.025}(4)=2.776\,4$, $t_{0.025}(9)=2.262\,2$. 实验中所用分类器均来自 PRTool(http://www.prtools.org) 工具箱, 实验机器配置为 1 G 内存, 2.30 G CPU, 算法基于 Matlab7.7 (R2008b) 实现.

3.3　实验结果与分析

表 5 和表 6 分别为在两种不同类型编码矩阵下各解码方法对不同 UCI 数据集的分类错误率, 其中加粗显示的数据为当前数据下最小分类错误率. 从两个表显示的结果看, 基于 DS 证据理论的解码策略在大部分数据中的分类效果好于其他解码策略; 只有当数据为 Zoo 和 Sat 时, DS 解码方法的错误率大于其他解码策略, 其原因可能是该两类数据集分布不均匀. 表 7

表 5　编码矩阵为密集随机编码时各解码方法的分类错误率及置信水平为 0.95 的置信区间

	HD	IHD	ED	AED	LLB	ELB	LS	DS
Yeast	60.35±3.80	49.88±1.16	51.21±2.70	50.70±3.31	50.32±1.93	49.02±2.07	49.27±2.59	**44.06±1.38**
Segmentation	14.52±0.74	13.99±0.68	14.72±1.42	14.91±0.98	12.42±1.15	11.65±0.82	9.78±0.89	**8.41±0.60**
Sat	20.71±1.16	24.91±1.12	16.53±1.18	**16.02±0.42**	20.45±1.42	19.64±0.69	28.0±1.31	20.00±0.57
Glass	9.46±1.44	10.14±5.03	9.57±2.06	8.89±3.79	8.29±3.42	8.98±4.51	10.8±3.55	**8.32±5.68**
Page-blocks	4.93±0.52	5.41±0.69	5.2±0.86	5.34±1.12	5.56±0.42	**3.81±0.42**	5.28±0.76	4.57±0.48
Vehicle	30.31±1.36	29.96±2.66	31.18±2.24	29.02±2.34	29.96±1.73	29.49±2.81	25.59±2.34	**25.44±2.02**
Zoo	30.52±8.76	24.51±17.99	29.5±11.61	31.48±5.67	25.75±8.18	27.53±3.78	**22.53±5.32**	23.14±9.12
Shuttle	15.23±2.34	15.43±3.65	16.53±3.33	16.24±5.19	18.52±6.90	17.26±4.44	**14.63±4.64**	15.63±4.52

表 6　编码矩阵为稀疏随机编码时各解码方法的分类错误率及置信水平为 0.95 的置信区间

	HD	IHD	ED	AED	LLB	ELB	LS	DS
Yeast	60.44±3.11	49.98±0.98	50.49±2.16	50.43±3.45	50.47±1.27	49.18±2.21	48.94±2.58	**43.39±0.99**
Segmentation	14.62±0.87	14.13±0.79	13.94±1.45	15.03±0.84	12.36±1.08	11.23±0.66	9.14±0.78	**7.63±0.49**
Sat	19.94±0.44	24.92±1.00	16.03±1.32	**15.24±0.83**	20.02±1.37	19.62±0.31	27.68±0.94	19.83±0.95
Glass	9.40±0.90	10.00±4.87	8.92±1.36	8.57±3.94	7.82±3.19	8.39±4.45	**7.24±3.24**	8.20±5.76
Page-blocks	5.07±0.84	4.73±0.27	4.64±0.89	4.78±1.12	4.99±1.32	4.35±0.59	4.72±0.67	**4.23±0.69**
Vehicle	30.33±1.13	29.69±2.76	30.65±2.18	28.96±1.91	29.71±0.99	28.73±2.53	**24.55±2.46**	24.76±1.77
Zoo	30.17±7.96	24.03±17.34	29.48±11.11	31.19±5.02	**21.54±7.63**	27.37±3.65	25.46±4.96	22.41±8.53
Shuttle	15.33±1.68	15.44±3.37	16.71±2.59	15.87±4.48	18.67±6.09	17.22±4.44	16.68±3.91	**15.22±3.97**

表 7 HRRP 数据分类错误率

angle	airplane	HD	IHD	ED	AED	LLB	ELB	LS	DS
$0^\circ \sim 100^\circ$	B-52	55.43±3.54	56.78±1.16	49.21±2.31	27.32±4.32	90.68±1.82	33.84±3.06	75.96±2.93	**28.36±1.39**
	J-6	22.22±1.11	26.46±0.97	37.94±1.87	20.00±1.74	56.84±1.57	60.16±1.00	**15.15±1.15**	18.75±1.27
	J-7	38.91±0.68	25.31±1.10	29.23±2.26	30.85±1.41	26.00±1.60	24.94±1.13	29.22±0.98	**16.73±1.60**
$80^\circ \sim 155^\circ$	B-52	42.07±4.35	95.80±1.51	21.04±3.18	30.82±4.94	73.06±2.03	44.80±3.53	69.04±3.78	**22.51±1.62**
	J-6	37.10±1.34	43.39±1.29	30.23±2.30	54.45±2.65	37.45±2.01	46.82±1.26	**26.88±1.74**	27.88±1.87
	J-7	**16.58±0.90**	21.79±1.39	19.66±2.69	35.62±1.49	18.12±2.41	24.06±2.06	17.31±1.47	19.37±1.84
$0^\circ \sim 155^\circ$	B-52	65.83±5.32	52.20±2.03	47.99±3.67	69.54±5.62	44.93±2.40	34.44±3.57	35.60±4.69	**26.40±1.72**
	J-6	44.95±1.67	53.78±1.42	41.87±2.40	54.06±3.14	60.82±2.72	53.93±2.15	56.91±2.44	**36.82±1.90**
	J-7	49.90±1.40	56.19±2.30	**37.96±3.30**	41.40±2.30	35.43±2.59	61.26±2.95	38.17±1.96	44.41±2.82

为 3 种不同角度范围下各解码策略的错误率. 从表中结果可以看出, 当角度范围较小时其分类错误率也较小. 但从总体看, 基于 DS 证据理论的解码策略相比其他解码方法在不同角度范围下的错误率绝大部分情况会更小, 这进一步说明在面对实际分类问题时, 基于 DS 证据理论的解码方法具有更好的实用效果.

4 结 论

利用纠错输出编码将多类问题转化为二类问题, 进而利用经典的二类分类方法对多类分类问题进行求解, 已成为模式识别领域中解决多类分类的一种重要手段. 在融合阶段如何利用各基分类器的输出结果进行最终的分类决策也成为能否利用好此类方法的一个关键点. 对此, 本文利用 DS 证据理论, 把基分类器的输出视为一条证据, 通过证据融合规则来实现对各基分类器输出的综合处理, 并形成最终决策. 实验结果表明, 基于 DS 解码策略相比其他经典方法, 能很好地辅助两种不同类型的编码矩阵以获得较好的分类效果.

参考文献(References)

[1] Windeatt T, Smith R S, Dias K. Weighted decoding ECOC for facial action unit classification[C]. The 18th European Conf on Artificial Intelligence. Patras, 2008: 26-30.

[2] Ghani R. Combining labeled and unlabeled data for text classification with a large number of categories[C]. Proc of Int Conf on Data Mining. California, 2001: 597-598.

[3] Zhou J, Suen C. Unconstrained numeral pair recognition using enhanced error correcting output coding: A holistic approach[C]. Proc of Int Conf on Document Analysis and Recognition. Seoul, 2005, 1: 484-488.

[4] Pujol O, Radeva P, Vitria J. Discriminate ECOC: A heuristic method for application dependent design of error correcting output codes[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2006, 28(6): 1001-1007.

[5] Dietterich T G, Bakiri G. Solving multiclass learning problems via error-correcting output codes[J]. J of Artificial Intelligence Research, 1995, 2(3): 263-286.

[6] Peterson W W, Weldon J R. Error-correcting codes[M]. Cambridge: MIT press, 1972: 233-235.

[7] Allwein E L, Shapire R E, Singer Y. Reducing multiclass to binary: A unifying approach for margin classifiers[J]. J of Machine Learning Research, 2000, 1(5): 113-141.

[8] Escalera S, Pujol O, Radeva P. On the decoding process in ternary error-correcting output codes[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2010, 32(1): 120-134.

[9] Jiang Yan-huang, Zhao Qiang-li, Yang Xue-jun. A search coding method and its application in supervised classification[J]. J of Software, 2005, 6(11): 1081-1088.

[10] Nilsson N J. Learning machines[M]. New York: McGraw-Hill, 1965: 123-124.

[11] Hastie T, Tibshirani R. Classification by pairwise coupling[J]. The Annals of Statistics, 1998, 26(4): 451-471.

[12] Windeatt T, Ghaderi R. Coding and decoding for multi-class learning problems[J]. Information Fusion, 2003, 4(8): 11-21.

[13] 周进登, 王晓丹. 加权解码在解决纠错输出编码 consistent-diverse 平衡问题的应用[J]. 电子学报, 2011, 39(7): 1514-1522.
(Zhou J D, Wang X D. Application of weighted decoding for the consistent-diverse balance problem of error correcting output codes[J]. Acta Electronica Sinica, 2011, 39(7): 1514-1522.)

[14] Passerini A, Pontil M, Frasconi P. New results on error correcting output codes of kernel machines[J]. IEEE Trans on Neural Networks, 2004, 15(1): 45-54.

[15] Shafer G A. Mathematical theory of evidence[M]. Princeton: Princeton University Press, 1976: 456-458.

[16] Platt J. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods[C]. Advances in Large Margin Classiers. Cambridge: MIT Press, 1999: 32-35.

[17] 尹安容, 谢湘, 匡镜明. Hadamard 纠错码结合支持向量机在多分类问题中的应用[J]. 电子学报, 2008, 36(1): 122-126.
(Yin A R, Xie X, Kuang J M. Application of Hadamard ECOC in multi-class problems based on SVM[J]. Acta Electronica Sinica, 2008, 36(1): 122-126.)