基于 ROC 的三元再编码研究

雷 蕾[①] 王晓丹*[①] 罗 玺[②]
[①](空军工程大学防空反导学院 西安 710051)
[②](空军工程大学信息与导航学院 西安 710077)

摘 要：针对三元编码矩阵中基分类器不包含被忽略样本类别先验知识的问题，该文提出一种基于接收机工作特性(ROC)曲线的矩阵再编码方法。首先基于 ROC 曲线寻找构造拒绝域的阈值对，从而获得最优分类器；然后利用最优分类器对训练样本中被忽略的类别进行分类，将经典的二值输出变为三值输出，从而对初始编码矩阵的码元“0”进行重新编码。在解码阶段，采用经典的汉明距离解码方法对未知样本进行决策。该方法能够避免基分类器的二次训练，适用于任意的三元纠错输出编码，具有良好的普适性和实用性。基于人工和 UCI 公共数据集的实验结果表明该方法简单高效，在不增加训练时间的基础上，能够提高解码的速度和精度，促进分类效果的提升。

关键词：三元纠错输出编码；二次编码；最优分类器；拒绝域；接收机工作特性
中图分类号：TP391 **文献标识码**：A **文章编号**：1009-5896(2016)10-2515-08
DOI: 10.11999/JEIT151343

Recoding Error-correcting Output Codes Based on Receiver Operating Characteristics

LEI Lei[①] WAGN Xiaodan[①] LUO Xi[②]
[①](*Institute of Air and Missile Defense*, *Aire Force Engineering University*, *Xi'an* 710051, *China*)
[②](*Institute of Information and Navigation*, *Aire Force Engineering University*, *Xi'an* 710077, *China*)

Abstract: As to the problem that the base classifiers in ternary Error Correcting Output Codes (ECOC) matrix do not contain the prior information of classes which are ignored in binary splits, a new recoding ECOC based on Receiver Operating Characteristic (ROC) curve is presented. To recode the ternary matrix, the two thresholds of reject region are obtained based on ROC to build the optimal classifiers. Then, the optimal classifiers are used to classify the ignored classes based on bipartition in training phase. In so doing, the classical two-symbol output expands to three-symbol to recode the zeros. Finally, the Hamming decoding strategy is adopted for decision in decoding. This method can avoid a second training and is applied to any kind of ternary matrix. The experiments based on Synthetic and UCI datasets validate the better efficiency and remarkable promotion without increasing training complexity of the proposed approach.

Key words: Ternary error-correcting output codes; Recoding; Optimal classifier; Reject option; Receiver Operating Characteristics (ROC)

1 引言

纠错输出编码(Error-Correcting Output Codes, ECOC)[1]，作为一种分而治之的多类与二类分类问题的连接桥梁，已成功应用到生物学数据识别[2]、疾病诊断[3]和计算机视觉识别[4]等诸多领域。而三元纠错输出编码(ternary ECOC)[5]作为一种更为通用的 ECOC 类型，几乎能统一现有的所有多类分类框架。矩阵中“0”符号的引入增加了二类划分的多样性和基分类器之间的差异性，使 ECOC 在多类分类问题领域的性能得以提升，并已受到众多学者的关注。文献[6]讨论了一种新的编码设计准则 —— 最大化三符号编码距离准则(ternary distance)，通过该准则作者提出了一种新的稀疏编码矩阵，实验结果表明基于该准则获得的 ECOC 分类效果有显著提高，且与解码方法的选取相关性较小。2008 年，文献[7]针对样本集线性不可分问题，提出对基类子集再分割的 Subclass ECOC 编码方法(SECOC)。文献[8]利用数据分布的先验知识，基于混淆矩阵和 Fisher 准则得到最优的子类划分，对子类采用“一对多”划

收稿日期：2015-12-01；改回日期：2016-06-06；网络出版：2016-08-26
*通信作者：王晓丹 afeu-wang@163.com
基金项目：国家自然科学基金(61273275, 61503407)
Foundation Items: The National Natural Science Foundation of China (61273275, 61503407)

分方法；在单个子类内部每次只选择两种类别作为正负类，而子类的其他类别将被忽略，即在矩阵中编码为“0”，从而弥补了单个基分类器识别的误差。文献[9]认为利用编码矩阵中二类划分的先验原始类结构信息可以提高 ECOC 分类性能，并给出了在流形假设和聚类假设的情况下将先验结构信息融入基分类器决策函数的方法。文献[10]针对经典的“一对一”三符号编码矩阵中码元“0”会引入分类偏差的问题，提出了避免二次训练的码元“0”再编码方法，并将该分类结果作为权值融入基于损失函数的解码过程中。文献[11]利用遗传算法来优化编码矩阵的构造，将初始 ECOC 编码矩阵看作遗传个体，经过交叉和变异形成新的编码矩阵。相关的研究还有文献[12-14]等，这些成果都有力地促进了三元 ECOC 多类分类的发展。

在三元编码矩阵中，码元“0”所对应的类别不参与训练，因此构造得到的基分类器将不包含此类样本的先验信息。故当该基分类器作用于测试样本时，就有可能因为信息缺失而输出片面的结果，导致分类错误。本文针对此分类偏差问题，提出了一种基于 ROC 的再编码方法(re-coding error-correcting output codes based on ROC)。该方法能够在不进行二次训练的基础上，通过基于 ROC 构造的最优基分类器对码元“0”进行重新编码，使得矩阵尽可能包含所有类别样本的结构信息；同时拒绝域的构造实现了三元 ECOC 对样本的选择性分类，矩阵中的“0”也在解码过程中变得更为具体，从而减小解码误差。决策时，将输出向量与新的类别编码进行 Hamming 距离解码。该方法采用的初始编码矩阵可以为任意的三元纠错输出编码矩阵，具有很好的普适性。

论文结构安排如下：第 2 节首先简要介绍基于 ROC 的三元再编码思想；第 3 节提出基于 ROC 再编码思想的步骤和方法，并对该方法存在的问题进行一一阐述；第 4 节给出实验结果和分析；最后给出总结说明。

2 再编码思想

ECOC框架中基分类器不能对样本进行选择性分类，而对于未参与基分类器训练的样本而言，直接对其分类将引入误差。

问题1 假设有符合高斯混合分布的5类数据(ecoli数据集[1])，如图1所示。当采用“一对一”方法对其进行编码时，由 C_1 和 C_5 作为训练样本产生的基分类器的决策面也能在一定程度上对类别 C_2 ,C_3 和 C_4 进行正确的划分。这样，基分类器 h_{15} 就能在测试时对其做出正确的硬判决输出。针对此问题，文献[10]提出了一种不需要重新训练的再编码方法，通过已训练的基分类器对码元“0”所对应的类别进行分类，通过设定分类正确率的阈值，将其重新编码为$\{1,-1\}$。但对于编码矩阵中遗留的码元“0”仍采用传统的损失函数进行解码。

问题2 假设一个三元编码矩阵 $\boldsymbol{M}_{4\times 7}$ 和输出向量 $\boldsymbol{x}$ 。

从图2中可以看出，根据Hamming距离解码和欧式距离解码，未知样本 $\boldsymbol{x}$ 被判定为 C_1 。但事实上 $\boldsymbol{x}$ 属于第2类。因为只有当 C_2 类在基分类器训练中没有被忽略时，其类别属性与 C_2 类一致；当 C_2 类被忽略时基分类器的决策面不能对样本 $\boldsymbol{x}$ 进行正确识别。出现这样的情况是由于编码矩阵中码元“0”的干扰，使训练得到的分类器不包含对应类别的分布信息，在解码过程中引入了偏差，从而不能做出正确的判决。本文结合问题1和问题2，将“拒绝域”引入每一个基分类器，对样本输出值落入拒绝域中的样本予以拒绝，不对其进行分类识别，使基分类器的输出由二元值扩展到三元值，从而减少基分类器对“0”标识的类别样本直接分类带来的误差。具体来讲：就是用训练好的基分类器对码元‘0’所对应的样本进行分类，根据一定的拒绝域准则，将“0”重新编码为“1”或“-1”；如果码元“0”所对应样本的输出落入拒绝域，此时选择性拒判就产生了，此类别在新的编码矩阵中码元依旧为“0”，这样就从本质上减小了利用Hamming距离解码时带来的误差，从而克服了问题2。

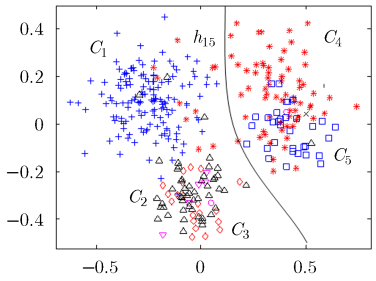

图 1 5 类数据集分布

$\boldsymbol{M}_{4\times 7}$

	h_1	h_2	h_3	h_4	h_5	h_6	h_7	HD	ED
C_1	1	1	−1	−1	1	−1	−1	1	4
C_2	1	−1	0	0	0	0	0	2.5	5
C_3	−1	1	1	1	−1	−1	1	6	12
C_4	0	1	−1	0	−1	1	−1	4	8
$\boldsymbol{x}$	1	−1	−1	−1	1	−1	−1		

图 2 $\boldsymbol{M}_{4\times 7}$ 编码矩阵

[1]对 ecoli 数据集进行了归一化处理，并删除了一些样本数很小的类

要实现选择性拒判，关键是构造拒绝域。作为分类器性能评估的有效手段，ROC曲线判决直观、概念清楚，对样本数据的先验分布知识和错分代价矩阵都不敏感，为解决前面提到的拒绝判决问题提供了强大的工具。为此本文引入ROC来设计拒绝域，构造最优分类器进行再编码设计。根据以上分析，图3给出了基于ROC的三元再编码的多类分类结构框图。其中 S_T 和 S_V 分别代表训练样本的训练子集和验证子集。值得注意的是，再编码过程仍然在样本训练阶段进行，这样就避免了二次训练，减少了训练时间。

3　基于 ROC 的三元再编码研究方法

本节利用 ROC 构造阈值对，进而得到带拒绝域的最优分类器，通过其对训练阶段所忽略的类别样本进行分类再编码，最终得到新的编码矩阵。

3.1 基于 ROC 的最优分类器

为了说明 ROC 的生成及相关特性，首先简要介绍两类样本混淆矩阵的相关概念。混淆矩阵描绘了样本数据的真实类别属性与识别结果之间的关系，是评价分类器性能的一种常用的有效方法。假定一个二分器的分类结果为一个 2×2 的混淆矩阵，如表 1 所示，其中行元素代表样本的真实属性，列元素代表分类器的分类结果。矩阵中：TP 为正确分类的正类样本数；FP 为被错误分类的负类样本数；FN 为被错误分类的正类样本数；TN 为正确分类的负类样本数。定义代价矩阵 $\boldsymbol{C}_{\mathrm{M}}$，其中 TP 和 TN 的分类代价为 0。$\mathrm{cr}=c_{21}/c_{12}$ 为代价矩阵的费效比(cost ratio)。由此定义该分类器的分类损失代价为

$$\mathrm{cost}=\frac{\mathrm{FN}c_{12}+\mathrm{FP}c_{21}}{\mathrm{TP+FN+FP+TN}} \tag{1}$$

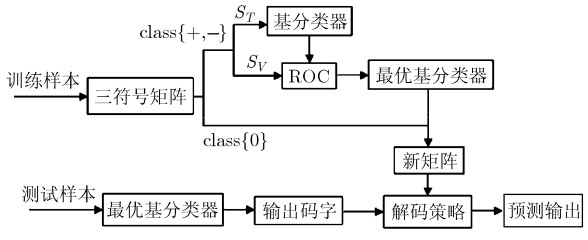

图 3 基于 ROC 的三元再编码多类分类结构框图

表 1 混淆矩阵和代价矩阵

混淆矩阵 $\boldsymbol{C}$				(b) 代价矩阵 $\boldsymbol{C}_{\mathrm{M}}$			
	+	−			+	−	0
+	TP	FN	P	+	0	c_{12}	c_{13}
−	FP	TN	N	−	c_{21}	0	c_{23}

其中，FP+TN=N，TP+FN=P。ROC 为一个 2 维图形，如图 4 所示，y 轴表示为 tp=TP/(TP+FN)，x 轴表示为 fp = FP/(FP+TN)。利用式(1)对 FP 求导，并令其等于 0 可得：$f'_{\mathrm{ROC}}(\mathrm{fp}^*)=\mathrm{cr}\dfrac{N}{P}$。

对于带拒绝域的基分类器而言，要找到最优的阈值对 $(\mathrm{fp}^*,\mathrm{tp}^*)$，文献[15]探讨了一系列算法，本文就对其采用的方法进行简单介绍.

假设 ROC 曲线上两点 $(\mathrm{fp}_\alpha,\mathrm{tp}_\alpha)$ 和 $(\mathrm{fp}_\beta,\mathrm{tp}_\beta)$ 对应的基分类器为 C_α 和 C_β，设 $\mathrm{fp}_\alpha\le\mathrm{fa}_\beta$，如图 4。$C_\alpha$ 和 C_β 是分类器根据不同阈值设定而得到的，属于一个家族，其数据分布具有一致性，即使存在分类不一致的情况，对分类结果也不会产生太大的影响。故通过此 ROC 曲线得到的最优分类器 C_{opt} 能满足下列条件：

$$C_{\mathrm{opt}}(x)=\begin{cases}+, & C_\alpha=+\wedge C_\beta=+\\ 0, & C_\alpha=-\wedge C_\beta=+\\ -, & C_\alpha=-\wedge C_\beta=-\end{cases} \tag{2}$$

同时假设 C_α 和 C_β 的混淆矩阵为 $(\mathrm{TP}_\alpha,\mathrm{FN}_\alpha,\mathrm{FP}_\alpha,\mathrm{TN}_\alpha)$ 和 $\left(\mathrm{TP}_\beta,\mathrm{FN}_\beta,\mathrm{FP}_\beta,\mathrm{TN}_\beta\right)$，分类代价矩阵为 $\boldsymbol{C}_{\mathrm{M}}$，式(1)的分类损失代价函数可改写为

$$\begin{aligned}\mathrm{rc}(N+P)=&\underbrace{\left(\mathrm{FP}_\beta-\mathrm{FP}_\alpha\right)c_{23}}_{C_\alpha,C_\beta\text{disagree,}-}+\underbrace{\left(\mathrm{FN}_\alpha-\mathrm{FN}_\beta\right)c_{13}}_{C_\alpha,C_\beta\text{disagree,}+}\\&+\underbrace{\mathrm{FP}_\alpha c_{21}}_{\text{FP for both}}+\underbrace{\mathrm{FN}_\beta c_{12}}_{\text{FN for both}}\\=&\left(\mathrm{FN}_\alpha c_{13}+\mathrm{FP}_\alpha\left(c_{21}-c_{23}\right)\right.\\&\left.+\mathrm{FN}_\beta\left(c_{12}-c_{13}\right)+\mathrm{FN}_\beta c_{23}\right)\\=&P\left(1-f_{\mathrm{ROC}}\left(\frac{\mathrm{FP}_\alpha}{N}\right)\right)c_{13}+\mathrm{FP}_\alpha\left(c_{21}-c_{23}\right)\\&+P\left(1-f_{\mathrm{ROC}}\left(\frac{\mathrm{FP}_\beta}{N}\right)\right)\left(c_{12}-c_{13}\right)+\mathrm{FP}_\beta c_{23}\end{aligned} \tag{3}$$

可以看出式(3)是关于 FP_α 和 FP_β 两个变量的函数，为了得到局部最小值利用式(3)分别对该两个变量求偏导可得：

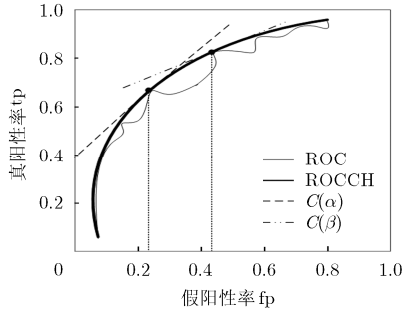

图 4　ROC 曲线特性

$$\left.\begin{aligned}\frac{\partial \mathrm{rc}}{\partial \mathrm{FP}_\alpha}(N+P)&=-\frac{P}{N}f'_{\mathrm{ROC}}\left(\frac{\mathrm{FP}_\alpha}{N}\right)c_{13}+c_{21}-c_{23}\\ \frac{\partial \mathrm{rc}}{\partial \mathrm{FP}_\beta}(N+P)&=-\frac{P}{N}f'_{\mathrm{ROC}}\left(\frac{\mathrm{FP}_\beta}{N}\right)(c_{12}-c_{13})+c_{23}\end{aligned}\right\} \tag{4}$$

令式(4)等于 0 得到最终结果为

$$\left.\begin{aligned}f'_{\mathrm{ROC}}\left(\mathrm{fp}^*_\alpha\right)&=\frac{c_{21}-c_{23}}{c_{13}}\frac{N}{P}\\ f'_{\mathrm{ROC}}\left(\mathrm{fp}^*_\beta\right)&=\frac{c_{23}}{c_{12}-c_{13}}\frac{N}{P}\end{aligned}\right\} \tag{5}$$

由式(5)就可获得 ROC 曲线上满足要求的两点，以此作为该基分类器产生拒绝域的阈值对$(\mathrm{fp}^*_\alpha,\mathrm{fp}^*_\beta)$。由前面的假设可知$\mathrm{fp}^*_\alpha\le\mathrm{fp}^*_\beta$，$f_{\mathrm{ROC}}(\mathrm{fp}^*_\alpha)\le f_{\mathrm{ROC}}(\mathrm{fp}^*_\beta)$。由于 ROC 的凸包是递增和凸的，所以它的一阶导数是非负和非增的，也就是$f'_{\mathrm{ROC}}(\mathrm{fp}^*_\alpha)\ge f'_{\mathrm{ROC}}(\mathrm{fp}^*_\beta)\ge 0$。所以代价矩阵$\boldsymbol{C}_{\mathrm{M}}$需要满足以下条件：

$$(c_{21}\ge c_{23})\wedge(c_{12}>c_{13})\wedge(c_{21}c_{12}\ge c_{21}c_{13}+c_{23}c_{12}) \tag{6}$$

3.2 新编码矩阵的生成

上一节通过 ROC 曲线构造出带拒绝域的基分类器并获得阈值对$(\mathrm{fp}^*_\alpha,\mathrm{fp}^*_\beta)$。其再编码过程为

$$M(i,j)=\begin{cases}1, & c_i^j\ge \mathrm{fp}^*_\beta\\ 0, & \mathrm{fp}^*_\alpha< c_i^j<\mathrm{fp}^*_\beta\\ -1, & c_i^j\le \mathrm{fp}^*_\alpha\end{cases} \tag{7}$$

其中，c_i^j为利用编码矩阵第 j列对应的最优分类器对“0”标识的第 i类的分类正确率。表 2 给出了基于再编码的多类分类方法的具体实现步骤。初始编码矩阵经过带拒绝域的基分类器重新编码以后，矩阵中码元“0”的个数会有所下降，而基分类器对未知样本的输出不再只有两种输出，即正类输出和负类输出，而是出现第 3 种输出结果，即拒绝做出判决(用“0”标示)。对此我们可以直接利用经典的汉明距离进行解码。

下面将重点通过实验验证该方法所得的编码矩阵在分类中的应用效果。

4 实验

本节采用人工数据集和 UCI 数据集来验证本文方法的分类效果。

4.1 实验数据

实验中所用的第 1 类数据集为 5 类 2 维正态分布人工数据集[16]，各类别数据的先验概率相同且样本个数为 400，其类别分布参数如表 3 所示。

第 2 类数据集为 UCI 数据集，其各类数据描述如表 4 所示，针对部分 UCI 数据集，本文对其进行了归一化处理，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)进行了降维处理。

表 2 基于 ROC 三元再编码的多类分类方法

输入：初始编码矩阵$\boldsymbol{M}_{\mathrm{init}}\in\{-1,0,+1\}^{K\times L}$，训练样本集$\Im$

步骤1 从训练样本集$\Im$中随机抽取部分样本作为验证集$\Im_{\mathrm{validate}}$，更新训练样本集$\Im'=\Im-\Im_{\mathrm{validate}}$，令编码矩阵$\boldsymbol{M}=\boldsymbol{M}_{\mathrm{init}}$。

步骤2 基于编码矩阵$\boldsymbol{M}$和训练样本集$\Im'$训练得到 L 个基分类器$l_i(i=1,2,\cdots,L)$，利用每个基分类器对验证集$\Im_{\mathrm{validate}}$进行分类，保留分类结果。

步骤3 针对上一步每个基分类器画出相应的ROC曲线，同时利用3.1节所示的方法确定构造每个基分类器拒绝域的阈值对$(\mathrm{fp}^*_\alpha,\mathrm{fp}^*_\beta)$，得到编码矩阵各列所对应的最优基分类器$f_i(i=1,2,\cdots,L)$。

步骤4 利用f_i对编码矩阵$\boldsymbol{M}$中“0”对应的类别进行分类，根据式(7)对其进行重新编码，得到新的编码矩阵$\boldsymbol{M}_{\mathrm{new}}$。

步骤 5 利用f_i对未知样本进行分类，将输出向量与$\boldsymbol{M}_{\mathrm{new}}$进行 Hamming 距离解码。

输出：分类正确率

表 3 5 类人工数据集各类别分布参数

class	C_1	C_2	C_3	C_4	C_5
μ_i	(0,0)	(3,0)	(0,5)	(7,0)	(0,9)
σ_i^2	1	4	9	25	64

表 4 UCI 数据集及数据描述

数据集	样本书	类别数	维数
Balance	625	3	4
Ecoli	336	8	8
Glass	214	6	9
Iris	150	3	4
Satimag	6435	6	36
Segment	2310	7	19
Thyroid	215	3	5
vehicle	846	4	18
Vowel	990	11	10
Wine	178	3	13
Yeast	1484	10	8
Zoo	101	7	18

4.2 实验设计

实验比较了原始“一对一”编码、稀疏随机编码和基于两者的再编码矩阵的分类效果。在选择随机编码方法时，我们将分别从已产生的 2000 个稀疏随机编码阵集(对应各码元$p(-1)$=1/3，$p(0)$=1/3，$p(+1)=1/3$)中随机选择所需要的编码阵。基分类

器采用了线性逻辑分类器(LOGLC)和多项式核函数支持向量机($C=2$)。采用 Hamming 距离进行解码，输出最后分类决策。

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 时我们采用 10 重交叉验证，小于 500 时采用 5 重交叉验证，双边估计 t 检验法见文献[17]。

ROC 曲线的绘制采用算法 2 实现[18]，如图 5 所示，其中 p，n 分别为正负样本个数。根据 3.1 节的方法求出该基分类器对应的阈值对和最优分类器，用于下一步测试样本的分类。

4.3 实验结果及分析

4.3.1 人工数据集　图 6 展示了基于一对一编码矩阵的线性逻辑基分类器所绘制的 ROC 曲线。从图 6 中可以看出，将所有的点根据不同的决策阈值进行分类，得到一条从点(0,0)向点(1,1)方向延伸的 ROC 曲线，并最终趋于稳定。而对于有限数据集合而言，生成的 ROC 曲线是阶梯形状的，只有当样本数据趋近无穷时，才能得到理论上的真实 ROC 曲线。

对于人工数据集而言，正负类样本比例大致为 1 ： 1，所以得到的最优分类器的阈值为 (0.2754, 0.8751)，表 5 为验证集的分类正确率。其中 f_i 为基分类器，class_i 为类别。

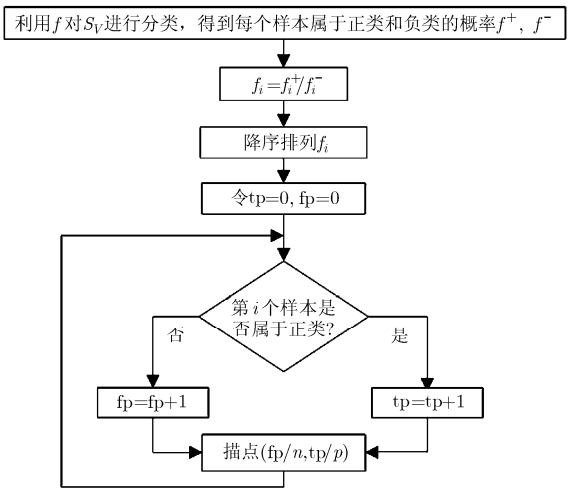

图 5　ROC 曲线的实现

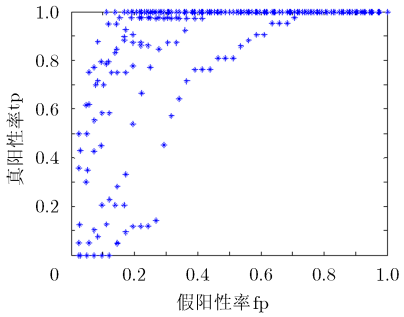

图 6　ROC 实验曲线

表 5 基于验证集的分类正确率

	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}
class_1	0	0	0	0	0.8802	1	1	0.9323	1	0.7188
class_2	0	0.9100	0.6050	0.9800	0	0	0	0.3200	1	0.9100
class_3	0.6963	0	0.9159	0.3785	0	0.9907	0.3411	0	0	0.1355
class_4	0.1256	0.7236	0	0.8543	0.8794	0	0.9095	0	0.9548	0
class_5	0.6443	0.2577	0.7577	0	0.2268	0.8247	0	0.8299	0	0

得到新的编码矩阵如表 6 所示。

表 7 所示结果为各编码类型在人工数据集下的分类正确率及其置信区间。ROC+编码方法为对应的基于 ROC 的再编码多类分类方法。

表 6 新编码矩阵行列

行	列									
	1	2	3	4	5	6	7	8	9	10
1	1	1	1	1	1	1	1	1	1	0
2	−1	1	0	1	1	1	1	0	1	1
3	0	−1	1	0	−1	1	0	1	1	−1
4	−1	0	−1	0	1	−1	1	−1	1	1
5	0	−1	0	−1	−1	0	−1	0	−1	−1

从表 7 可以看出，基于 ROC 的再编码方法是有效的，其分类得到的结果要优于原始编码方法，而基于一对一编码方法分类精度的提高要高于随机编码，这是因为在一对一编码中，码元“0”所占的比例高，对其进行再编码，能明显地减少“0”的数量，缩小测试样本输出向量与真实类码字的距离，从而提高分类正确率。同时再编码的复杂度也没有明显的增加，所需时间消耗在可接受范围内。

4.3.2 UCI 数据集　表 8 和表 9 列出了分别以 SVM 和 LOGLC 作为基分类器的不同编码方法的分类结果比较。其中 Re-为文献[10]所采用的再编码方法。在每张表中加粗的数据为最大分类正确率。

表 7 人工数据集分类正确率及置信水平为 0.95 的置信区间(%)

编码类型	基分类器			
	线性逻辑分类器		支持向量机	
	分类正确率&置信区间(%)	时间复杂度(s)	分类正确率&置信区间(%)	时间复杂度(s)
一对一	61.45±2.19	57	69.80±1.68	115.00
ROC 一对一	63.25±1.08	102	**72.85±4.79**	136.73
一对多	59.70±1.30	6	67.75±3.59	59.00
密集编码	55.20±0.12	18	60.35±0.12	156.00
稀疏编码	48.90±3.10	26	63.10±7.06	84.00
ROC 稀疏编码	**63.90±1.27**	50	64.45±14.57	117.54

表 8 基于 SVM 和 Hamming 距离解码的各数据集分类正确率及置信水平为 0.95 的置信区间(%)

	一对一	Re-一对一	ROC 一对一	一对多	密集编码	稀疏编码	Re-稀疏	ROC-稀疏
Balance	95.52±8.22	95.68±6.01	96.00±2.11	**96.96±10.23**	–	96.28±6.04	95.36±7.14	96.32±6.02
Ecoli	74.93±14.60	75.34±2.68	**75.88±5.86**	68.75±6.15	42.56±0.56	69.64±9.67	70.79±8.54	70.67±11.81
Glass	96.70±18.60	96.23±24.65	**97.69±17.26**	96.26±0.89	32.72±7.77	80.83±10.49	87.95±8.39	88.30±8.72
Iris	92.68±5.38	96.67±8.47	**97.33±16.94**	94.00±25.41	–	64.67±5.41	64.67±8.47	65.33±3.85
Satimag	77.54±3.03	80.40±4.59	**82.59±12.39**	62.00±8.19	67.80±6.54	72.80±3.10	71.09±13.59	71.63±1.98
Segment	93.46±3.85	**93.70±9.90**	93.68±9.35	92.81±14.30	54.29±8.17	82.34±2.86	83.85±14.85	87.14±11.55
Vehicle	74.94±9.76	75.53±12.78	**75.65±17.29**	73.52±3.80	53.52±0.80	58.87±5.28	59.57±10.23	60.73±3.54
Vowel	96.26±6.42	98.58±1.36	**98.71±2.89**	87.68±28.24	82.12±7.97	79.80±2.48	80.03±1.03	79.09±0.00
Wine	73.12±6.19	79.81±3.95	**80.35±2.19**	74.71±10.75	–	78.09±4.01	78.89±14.12	79.18±3.86
Yeast	56.33±5.67	65.73±2.01	66.61±14.90	**67.06±2.56**	46.44±0.84	51.21±10.47	52.97±8.41	53.16±22.10
Thyroid	87.93±5.89	92.09±6.38	**92.57±5.32**	91.62±12.31	83.72±6.87	87.44±6.65	88.75±4.78	89.10±17.26

表 9 基于 LOGLC 和 Hamming 距离解码的各数据集分类正确率及置信水平为 0.95 的置信区间(%)

	一对一	Re-一对一	ROC 一对一	一对多	密集编码	稀疏编码	Re-稀疏	ROC-稀疏
Balance	88.48±4.30	90.04±3.88	91.72±12.01	**96.96±10.23**	–	46.08±0.94	47.04±13.15	49.76±17.29
Ecoli	80.07±17.40	82.76±14.08	**83.62±12.01**	77.71±6.26	42.56±0.56	72.31±2.10	71.12±12.09	75.32±14.72
Glass	84.90±8.04	87.79±3.85	**89.70±14.32**	86.91±3.11	32.72±7.77	68.27±2.89	69.62±1.28	70.90±3.70
Iris	94.00±8.47	96.00±16.94	**97.33±16.94**	94.67±0.05	–	74.67±5.08	75.67±4.23	76.33±16.94
Satimag	86.53±4.07	86.15±0.51	**86.76±6.79**	79.39±3.43	63.82±0.06	79.78±0.71	80.20±2.07	80.26±2.25
Segment	93.11±17.05	**94.03±13.35**	93.03±11.00	88.01±2.69	74.29±14.29	71.77±3.30	74.16±2.80	74.50±3.52
Vehicle	79.31±3.21	77.42±3.83	**78.01±3.66**	75.65±12.75	23.52±0.80	71.76±4.27	68.56±17.08	69.86±3.60
Vowel	**77.17±5.13**	76.36±17.97	77.37±7.70	38.08±19.25	–	45.56±5.05	48.28±2.84	48.84±3.85
Wine	94.12±13.16	95.05±12.98	95.06±7.70	95.24±13.64	–	94.39±13.48	92.16±14.45	**95.48±29.20**
Yeast	56.73±14.21	57.69±4.68	**58.09±2.98**	53.22±12.85	46.44±0.84	55.53±6.28	53.70±6.66	54.05±13.86
Thyroid	95.82±5.65	95.81±17.98	**96.27±12.04**	95.83±14.12	93.97±14.01	94.41±12.15	93.96±5.55	92.10±17.26

从实验结果可以看出，基于 ROC 的再编码方法在大部分数据集上能获得较好的分类效果。尽管 Re-ECOC 有时能取得更好的分类效果，但它不能实现基分类器的选择性分类，对一些错分代价较大的样本有一定的局限性。同时基于“一对一”编码矩阵的 ECOC 多类分类方法在大部分数据集上的分类性能提升优于基于随机编码的 ECOC 分类。而对于 Glass 数据集而言，基于稀疏随机编码提高的幅

度较大，这是因为此时稀疏编码矩阵为

$$\begin{bmatrix} -1 & 0 & 0 & -1 & -1 & -1 & -1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 0 & 0 & 0 & -1 & -1 \\ -1 & 0 & -1 & -1 & 0 & 1 & -1 \\ -1 & 0 & -1 & -1 & 0 & 0 & 1 \end{bmatrix}$$

可以看出，此时的编码矩阵中“0”的数量较多，对其采取再编码策略能够明显地提高分类性能。

为了更直观地体现再编码对码元“0”的影响，本文引入拒分率来对编码前后的分类性能进行比较。

$$\rho = \frac{1}{l}\sum_{i=1}^{l}\frac{N_i^{\{0\}}}{N_i} \tag{8}$$

其中，N_i 为第 i 列编码对应的样本总数，$N_i^{\{0\}}$ 为其码元“0”所应对的样本个数，即拒分样本个数，l 为分类器个数。图 7 为不同数据集在原始编码和再编码情况下的拒分率比较。从图 7 中可以看出，再编码之后拒分率明显下降，且基于一对一编码的拒分率下降幅度大于随机编码。通过对“0”进行再编码能够使三元类码字更加具体，提高了解码速度和精度，从而提升分类效果。

5　结束语

在三符号 ECOC 多类分类中，码元“0”对应的类别样本将不参与此列基分类器的训练，因此利用该基分类器对未参与训练的类别样本进行分类可能会带来误差。本文提出的再编码 ECOC 方法在不进行二次训练的前提下，通过引入“拒绝域”来实现选择性分类，将基分类器的二值输出扩展为三值输出，从而实现编码矩阵中的码元“0”再编码，缩小预测输出与类别码字的距离，提高分类正确率。在为基分类器构造拒绝域时，一种基于 ROC 曲线的拒绝域构造方法被应用到再编码过程中。最后实验验证了本文方法的有效性。本文为实现多类别的选择性分类提供了新思路。如何从降低误判风险出发，研究有效的带拒绝域的 ECOC 多类分类方法是下一步的工作重点。

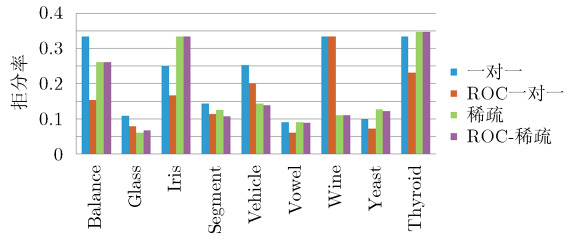

图 7 拒分率比较

参 考 文 献

[1] DIETTERICH T G and BAKIRI G. Solving multi-class learning problems via error-correcting output codes[J]. *Journal of Artificial Intelligence Research*, 1995, 34(2): 263–286. doi: 10.1613/jair.105.

[2] PHYO K S, JIAN G W, and EAM K T. Facial age range estimation with extreme learning machines[J]. *Neurocomputing*, 2015, 149A: 364–372. doi: 10.1016/j.neucom.2014.03.074.

[3] ELIF D Ü. ECG beats classification using multiclass support vector machines with error correcting output codes[J]. *Digital Signal Processing*, 2007, 45(17): 675–684. doi: 10.1016/j.dsp.2006.11.009.

[4] SERGIO E, DAVID M, ELOI P, *et al*. Online error correcting output codes[J]. *Pattern Recognition Letters*, 2011, 32(3): 458–467. doi: 10.1016/j.patrec.2010.11.005.

[5] ERIN L A, ROBERT E S, YORAM S, *et al*. Reducing multiclass to binary: a unifying approach for margin classifiers[J]. *Journal of Machine Learning Research*, 2000, 39(1): 113–141. doi: 10.1162/15324430152733133.

[6] SERGIO E, ORIOL P, and PETIA R. Separability of ternary error-correcting output codes[J]. *Pattern Recognition Letters*, 2009, 30(5): 285–297. doi: 10.1016/j.patrec.2008.10.002.

[7] SERGIO E, DAVID M J T, ORIOL P, *et al*. Subclass problem-dependent design for error-correcting output codes [J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008, 30(6): 1041–1054. doi: 10.1109/TPAMI.2008.38.

[8] 周进登, 王晓丹, 周红健. 基于混淆矩阵的自适应纠错输出编码多类分类方法[J]. 系统工程与电子技术, 2012, 34(7): 220–226. doi: 10.3969/j.issn.1001-506X.2012.07.38.
ZHOU Jindeng, WANG Xiaodan, and ZHOU Hongjian. Multiclass classification of adaptive error-correcting output codes based on confusion matrix[J]. *Systems Engineering and Electronics*, 2012, 34(7): 220–226. doi: 10.3969/j.issn.1001-506X.2012.07.38.

[9] WANG Y, CHEN S C, and XUE H. Can under-exploited structure of original-classes help ECOC-based multi-class classification?[J]. *Neurocomputing*, 2012, 89: 158–167. doi: 10.1016/j.neucom.2012.02.035.

[10] SERGIO E, ORIOL P, and PETIA R. Re-coding ECOCs without Re-training[J]. *Pattern Recognition Letters*, 2010, 31(7): 555–562. doi: 10.1016/j.patrec.2009.12.002.

[11] MIGUEL A B, SERGIO E, XAVIER B, *et al*. On the design of an ECOC-compliant genetic algorithm[J]. *Pattern*

Recognition, 2014, 47(2): 865–884. doi: 10.1016/j.patcog.2013.06.019.

[12] FRANCESCO C, ORIOL P, and PETIA R. ECOC-DRF: discriminative random fields based on error correcting output codes[J]. *Pattern Recognition*, 2014, 47(6): 2193–2204. doi: 10.1016/j.patcog.2013.12.007.

[13] MIKEL G, ALBERTO F, EDURNE B, *et al*. DRCW-OVO: distance-based relative competence weighting combination for one-vs-one strategy in multi-class problems[J]. *Pattern Recognition*, 2015, 48(1): 28–52. doi: 10.1016/j.patcog.2014.07.023.

[14] LEI L, WANG X D, LUO X, *et al*. Hierarchical error-correcting output codes based on SVDD[J]. *Pattern Analysis and Applications*, 19(1): 163–171. doi: 10.1007/s10044-015-0455-5.

[15] TADEUSZ P. On the use of ROC analysis for the optimization of abstaining classifiers[J]. *Machine Learning*, 2007, 68(2): 137–169. doi: 10.1007/s10994-007-5013-y.

[16] ZHOU J D, and WANG X D. Research on the unbiased probability estimation of error-correcting output coding[J]. *Pattern Recognition*, 2011, 44(7): 1552–1565. doi: 10.1016/j.patcog.2010.12.020.

[17] ZHOU J D, YUN Y，ZHANG J M, *et al*. Constructing ECOC based on confusion matrix for multiclass learning problems[J]. *Science China Information Sciences*, 2016, 59(1): 1–14. doi: 10.10071/s11432-015-5321-y.

[18] 邹洪侠, 秦峰. 二类分类器的 ROC 曲线生成算法[J]. 计算机技术与发展, 2009, 19(6): 109–112.
ZOU Hongxia and QIN Feng, Algorithm for generating ROC curve of two-classifier[J]. *Computer Technology and Development*, 2009, 19(6): 109–112.

雷 蕾： 女，1988 年生，博士生，研究方向为机器学习、多类分类.

王晓丹： 女，1966 年生，教授，博士生导师，研究方向为机器学习、计算机视觉处理、信息融合.

罗 玺： 男，1988 年生，硕士，讲师，研究方向为智能信息处理.