带拒绝域的 ECOC 多类分类

雷 蕾[1],王晓丹[1],罗 玺[2],王 玮[3]
(1. 空军工程大学防空反导学院,陕西西安 710051;2. 空军工程大学信息与导航学院 陕西西安 710077;
3. 空军大连通信士官学校基础部,辽宁大连 116600)

摘 要: 针对纠错输出编码分解框架的自身特点、从降低误判风险出发,研究了带拒绝域的 ECOC 多类分类方法. 首先在二类划分过程中引入拒绝域,对不属于正负子类的待识别样本进行拒识;其次,在基分类器内部引入拒绝域,以最小化风险贝叶斯决策为目标,利用后验概率输出和代价矩阵寻找拒绝域阈值,对样本输出值落入拒绝域中的样本进行拒识;最后,研究了不同拒绝域输出的解码方法,并讨论了拒识码字个数和矩阵最小 Hamming 距离之间的关系. 实验结果表明基于二类划分构造的拒绝域能够提高分类正确率,而基于基分类器构造的拒绝域能够减小分类代价.

关键词: 多类分类; 纠错输出编码; 拒绝域; 支持向量数据描述; 贝叶斯决策
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2017)11-2779-08
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2017.11.027

Design of Reject Option for Multi-classification Based on ECOC

LEI Lei[1], WANG Xiao-dan[1], LUO Xi[2], WANG Wei[3]
(1. *The Air and Missile Defense Institute, Air Force Engineering University, Xi'an, Shaanxi* 710051, *China*;
2. *The Information and Navigation Institute, Air Force Engineering University, Xi'an, Shaanxi* 710077, *China*;
3. *The Fundamental Department, Dalian Air Force Communication NCO Academy, Dalian, Liaoning* 116600, *China*)

Abstract: Aiming at reducing misclassification costs, this paper studies the design of reject options for ECOC multi-classification based on its properties. The first level of reject option is constructed in the process of bipartitions to recognize an instance whose real labels does not belong to the meta-subclasses. Meanwhile, the second reject rule is presented in the dichotomizers based on posterior probabilities and cost matrix to make the minimum-risk Bayesian decision. Finally, different decoding strategies are analysed according to different reject output. The relationship between the number of rejected positions and the minimum Hamming distance of matrix is discussed. The two-stage reject rule makes the ECOC multi-classification with rejection come true and reduce the misclassification error and costs.

Key words: multi-classification; error-correcting output codes; rejection option; support vector domain description; Bayesian decision

1 引言

多类分类是机器学习领域的关键问题之一. 常用的方法是将多类分类问题分解为若干二类分类问题,直接利用二类分类方法的研究成果,通过结果融合实现多类分类. 作为一种广泛应用的分解框架,纠错输出编码[1,2]基于 $n \times L$ 的编码矩阵将 n 个类别分解为 L 个不同的二类划分,每个类别对应着长度为 L 的码字,初始数据根据编码矩阵对应的列重新划分构成正负子类,训练得到与该列对应的基分类器. 在测试阶段,对待识别样本 x,同时利用每个基分类器对其进行分类,通过某种解码规则对输出结果进行解码得到最终的分类结果. 目前,纠错输出编码的研究主要集中在如何构造有效的编码矩阵和解码策略,同时众多学者也研究发现纠错输出编码在优化偏差[3]、方差[4]和有效的概率估计[5]等方面效果很显著.

类似于其他分类机制,在实际应用领域,基于 ECOC 的多类分类也面临着分类错误的挑战,而这种分类错误有的时候会带来巨大的风险和损失,如医疗诊断、故障检测、弹道目标识别等,把一类样本误判为另一

收稿日期:2016-05-24;修回日期:2016-10-10;责任编辑:蓝红杰
基金项目:国家自然科学基金(No.61273275, No.61503407)

类的损失往往比相反的情况要高得多. 因此,对一不明确或误分代价很高的样本,拒绝对其进行识别(拒绝分类结果或做进一步处理)所带来的损失往往要小得多. 而目前基于 ECOC 的分类代价研究较少. 突出的有:Zhou 首次讨论了将拒识机制引入 ECOC 分类系统的可能性,并提高了测试结果的可信度[6]. P. Simeone 等针对 ECOC 的拒识问题,提出了两种拒识规则:一个是在解码阶段,通过系统输出的可靠性与阈值进行对比完成拒识功能;第二种方法是基于 ROC 曲线在基分类器内部直接构造拒绝域,并在解码阶段对拒识的码位进行了修正处理从而得到最终的预测输出. 不同于在基分类器内部直接构造拒绝域,基于 ECOC 的分类机制更需要在基分类器和二类划分两个层次完成拒识功能[7].

在实际应用的多类分类中,单纯地追求分类精度的提升已经不能满足问题需求,如何减小分类损失已成为研究的重点. 而纠错输出编码固有的结构特点决定了 ECOC 分类并不适用于以减小分类损失代价为目标的分类决策过程. 这是因为参与集成的基分类器对样本的输出仅限于正负类的二元输出,而不具备对样本拒绝分类的能力,从而不能实现选择性分类. 因此如何实现带拒识功能的 ECOC 多类分类成为本文研究的重点. 本文从如何减小分类代价出发,通过对 ECOC 拒识方法的研究,在二类划分和基分类器内部分别提出了拒识机制. 基于二类划分的拒识机制基于数据分布知识,对不属于二类划分的样本进行拒识,从而减小编码矩阵中码字零对解码的影响;在基分类器内部,重新构造拒绝域,对难分样本和错分代价大的样本进行选择性分类,从而实现分类代价的最小化. 在两层拒识完成后,对不同的拒识输出采取不同的解码改进方法,从而完成带拒绝域的 ECOC 多类分类任务.

2 纠错输出编码(ECOC)

ECOC 框架即用一种二元或三元的编码矩阵实现多类类别分解和基分类器集成[8]. “-1”代表一类(黑色),“+1”代表另一类(白色),“0”表示该码字位所对应的类不参与由该列所产生的基分类器的训练(灰色). 图 1 给出了四种常见的 ECOC 分类系统示意图.

3 带拒绝域的 ECOC 多类分类

3.1 基于二类划分的拒绝域

基于二类划分的拒绝域的产生绝大部分是因为原编码矩阵中 0 元素所对应的类别不参与训练所导致的. 当利用基分类器对不属于此二类划分的样本进行分类时,就有可能造成分类错误. 因此,我们将拒绝域引入二类划分,对不属于此列子类划分的样本进行拒识,从而使得基分类器更加关注符合自身分布的数据样本.

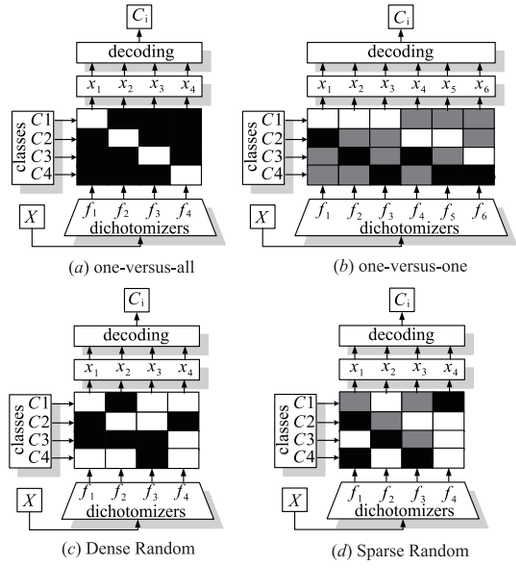

图 1 四种常见的ECOC

如何对不属于二类划分的样本进行识别是此层拒绝域构造的关键. 考虑到基于二类划分的拒绝域主要是识别不属于此二类划分子类的数据样本,因此,只需要判定一个样本是否属于此二类划分即可,属于则接受,不属于则拒绝. 则有:

$$v(\boldsymbol{x},t_b)=\begin{cases}\omega_k, & \text{if } d(S,\boldsymbol{x})<t_b\\ 0, & \text{if } d(S,\boldsymbol{x})>t_b\end{cases} \tag{1}$$

其中 $v(\boldsymbol{x},t_b)$ 为待识别样本 $\boldsymbol{x}$ 的判决输出,$d(S,\boldsymbol{x})$ 为 $\boldsymbol{x}$ 与二类划分子集 S 的距离,t_b 为阈值. 从式(1)可以看出,如果待识别样本与类别子集的距离小于阈值,则认为样本属于该二类划分,通过二类划分生成的基分类器对其进行进一步的识别;如果距离大于阈值,则认为样本不属于该二类划分,此时分类器给出的输出为 0. 从而将基分类器的二值输出在引入拒绝域后扩展到三值输出,同时尽量使该类输出值与其在编码矩阵中的类别码字相对应.

衡量待识别样本与二类划分子集的距离是一个属于与不属于的问题,即目标类与非目标类的关系. 支持向量数据描述(Support Vector Domain Description,SVDD)[9] 为解决此单类分类问题提供了强大的思路. 它通过在高维空间中构造一个超球体,使该球体尽可能地包含所有数据样本. 对未知样本而言,当它到超球体中心的距离小于或等于半径 r 时,则未知样本被判为目标类,否则为非目标类. 在基于二类划分的拒绝域构造中,将阈值 t_b 与半径 r 等同起来,小于或等于该半径,则认为样本属于该二类划分,参与基分类器训练和分类,反之亦然.

3.2 基于基分类器的拒绝域

基于二类划分的拒绝域着重解决的是待识别样本是否属于子类划分的问题,而基于基分类器的拒绝域就主要处理不确定样本和误分代价较高的样本. 基于

基分类器的拒绝域通常是假定基分类器的输出为概率输出$f_h(\boldsymbol{x})$,设定一对阈值 α、β,其中 $\alpha<\beta$,则拒识规则为:

$$v(f_h(\boldsymbol{x}),\alpha,\beta)=\begin{cases}+1, & \text{if } f_h(\boldsymbol{x})>\beta\\ -1, & \text{if } f_h(\boldsymbol{x})<\alpha\\ \text{reject}, & \text{if } f_h(\boldsymbol{x})\in[\alpha,\beta]\end{cases}\tag{2}$$

从式(2)中可以看出拒绝域构造的重点是确定阈值 α 和 β 的值[10].

由最小风险贝叶斯决策可知采取决策 a_i 情况下的条件期望损失 $R(a_i|\boldsymbol{x})$ 为

$$R(a_i|x)=E[\lambda(a_i,\omega_j)]=\sum_{j=1}^{c}\lambda(a_i,\omega_i)P(\omega_j|x)\tag{3}$$

而条件风险最小的决策 a_k,即 $R(a_k|x)=\min\limits_i R(a_i|\boldsymbol{x})$.

对于带拒绝域的基分类器而言,设有决策和代价风险矩阵:

	P	N	*reject*
P	c_{11}	c_{12}	c_{13}
N	c_{21}	c_{22}	c_{23}

(4)

其中 P、N、*reject* 分别表示将正负类样本划分为正类、负类和拒绝对其分类的决策. 损失代价 $c_{12}>c_{13}>c_{11}$,$c_{22}>c_{23}>c_{21}$. 基分类器输出为样本属于正子类的概率 $P(+1|\boldsymbol{x})$,负子类概率为 $1-P(+1|\boldsymbol{x})$. 则基分类器对待识别样本做出正类、负类和拒绝的风险为:

$$\begin{aligned}R(P|\boldsymbol{x})&=c_{11}P(+1|\boldsymbol{x})+c_{21}P(-1|\boldsymbol{x})\\ R(N|\boldsymbol{x})&=c_{12}P(+1|\boldsymbol{x})+c_{22}P(-1|\boldsymbol{x})\\ R(reject|\boldsymbol{x})&=c_{13}P(+1|\boldsymbol{x})+c_{23}P(-1|\boldsymbol{x})\end{aligned}\tag{5}$$

根据贝叶斯最小风险决策可知:

$$\begin{aligned}&P:\text{if } R(P|\boldsymbol{x})<R(N|\boldsymbol{x})\&\&R(P|\boldsymbol{x})<R(reject|\boldsymbol{x})\\ &\qquad\text{then } \boldsymbol{x}\in S_P\\ &N:\text{if } R(N|\boldsymbol{x})<R(P|\boldsymbol{x})\&\&R(N|\boldsymbol{x})<R(reject|\boldsymbol{x})\\ &\qquad\text{then } \boldsymbol{x}\in S_N\\ &reject:\text{if } R(reject|\boldsymbol{x})<R(P|\boldsymbol{x})\&\&(R(reject|\boldsymbol{x})\\ &\qquad<R(N|\boldsymbol{x}))\text{ then reject}\end{aligned}\tag{6}$$

对于 P 决策而言:

$$\begin{aligned}&R(P|\boldsymbol{x})<R(N|\boldsymbol{x})\Leftrightarrow\\ &c_{11}P(+1|\boldsymbol{x})+c_{21}P(-1|\boldsymbol{x})\\ &<c_{12}P(+1|\boldsymbol{x})+c_{22}P(-1|\boldsymbol{x})\Leftrightarrow\\ &c_{11}P(+1|\boldsymbol{x})+c_{21}(1-P(+1|\boldsymbol{x}))\\ &<c_{12}P(+1|\boldsymbol{x})+c_{22}(1-P(+1|\boldsymbol{x}))\Leftrightarrow\\ &P(+1|\boldsymbol{x})>\frac{c_{21}-c_{22}}{(c_{21}-c_{22})+(c_{12}-c_{11})}\end{aligned}\tag{7}$$

$$\begin{aligned}&R(P|\boldsymbol{x})<R(reject|\boldsymbol{x})\Leftrightarrow\\ &c_{11}P(+1|\boldsymbol{x})+c_{21}P(-1|\boldsymbol{x})\\ &<c_{13}P(+1|\boldsymbol{x})+c_{23}P(-1|\boldsymbol{x})\Leftrightarrow\\ &c_{11}P(+1|\boldsymbol{x})+c_{21}(1-P(+1|\boldsymbol{x}))\\ &<c_{13}P(+1|\boldsymbol{x})+c_{23}(1-P(+1|x))\Leftrightarrow\\ &P(+1|\boldsymbol{x})>\frac{c_{21}-c_{23}}{(c_{21}-c_{23})+(c_{13}-c_{11})}\end{aligned}\tag{8}$$

同理,对于 N 决策:

$$\begin{aligned}&R(N|\boldsymbol{x})<R(P|\boldsymbol{x})\Leftrightarrow P(+1|\boldsymbol{x})<\frac{c_{21}-c_{22}}{(c_{21}-c_{22})+(c_{12}-c_{11})}\\ &R(N|\boldsymbol{x})<R(reject|\boldsymbol{x})\Leftrightarrow(c_{12}P(+1|\boldsymbol{x})+c_{22}P(-1|\boldsymbol{x})\\ &\qquad<c_{13}P(+1|\boldsymbol{x})+c_{23}P(-1|\boldsymbol{x}))\Leftrightarrow\\ &\qquad P(+1|\boldsymbol{x})<\frac{c_{23}-c_{22}}{(c_{23}-c_{22})+(c_{12}-c_{13})}\end{aligned}\tag{9}$$

对于 *reject* 决策:

$$\begin{aligned}&R(reject|\boldsymbol{x})<R(P|\boldsymbol{x})\Leftrightarrow\\ &P(+1|\boldsymbol{x})<\frac{c_{21}-c_{23}}{(c_{21}-c_{23})+(c_{13}-c_{11})}\\ &R(reject|\boldsymbol{x})<R(N|\boldsymbol{x})\Leftrightarrow\\ &c_{13}P(+1|\boldsymbol{x})+c_{23}P(-1|\boldsymbol{x})<c_{12}P(+1|\boldsymbol{x})+c_{22}P(-1|\boldsymbol{x})\\ &\Leftrightarrow P(+1|\boldsymbol{x})>\frac{c_{23}-c_{22}}{(c_{23}-c_{22})+(c_{12}-c_{13})}\end{aligned}\tag{10}$$

令

$$\begin{cases}\alpha=\dfrac{c_{23}-c_{22}}{(c_{23}-c_{22})+(c_{12}-c_{13})}\\ \beta=\dfrac{c_{21}-c_{23}}{(c_{21}-c_{23})+(c_{13}-c_{11})}\\ \gamma=\dfrac{c_{21}-c_{22}}{(c_{21}-c_{22})+(c_{12}-c_{11})}\end{cases}\tag{11}$$

可得:

$$\begin{aligned}&P(+1|\boldsymbol{x})<\alpha &&\Leftrightarrow\boldsymbol{x}\in S_N\\ &\alpha<P(+1|\boldsymbol{x})<\beta &&\Leftrightarrow reject\\ &P(+1|\boldsymbol{x})>\beta &&\Leftrightarrow\boldsymbol{x}\in S_P\end{aligned}\tag{12}$$

其中 γ 为不带拒绝域时,属于正负子类的阈值. 至此,基于基分类器的拒绝域就得到了. 从式(11)可以看出,基于后验概率和最小风险贝叶斯准则的拒绝域构造方法原理简单、意义明确、求解方便,同时基分类器的拒绝域阈值只与代价矩阵有关系,这是由分类目的决定的. 根据损失代价最小风险决策,拒绝域随着基分类器的代价矩阵变化而改变,有的基分类器拒绝阈值就会相较于别的基分类器更严格或者宽松.

3.3　改进的解码方法

根据前面的分析,得到了 ECOC 多类分类的两层拒识机制,因为各层机制的拒绝输出有差异,所以在解码时对传统的解码方法得做适当的扩展,使其能更好的适用于带拒绝域的 ECOC 多类分类.

基于二类划分的拒绝域直接输出码字 0,即不参与基分类器训练的样本其在编码矩阵中的真实码字也就是 0,所以对于此层拒识的输出直接利用经典的 Hamming 或其他方法进行解码即可.

而基于基分类器的拒绝域输出的是 *reject* 的标识，其解码方法[11]如下：

Step1 将输出向量中被拒绝的码位全部用“-1”替换，并利用经典的汉明距离解码找出距离最小的类别码向量 $\boldsymbol{c}_{-1}$；

Step2 将输出向量中被拒绝的码位全部用“1”替换，并利用经典的汉明距离解码找出距离最小的类别码向量 $\boldsymbol{c}_1$；

Step3 在前两个步骤所得到的类别码向量 $\boldsymbol{c}_{-1}$ 和 $\boldsymbol{c}_1$ 中，在所有对应的非拒绝码位（即非“0”所标识的码位）中与输出向量最近的类别码向量（$\boldsymbol{c}_{-1}$ 或 $\boldsymbol{c}_1$）所对应的类别即为样本所属的最终类别.

在上述解码策略中，第一步目的是为了找出在拒绝域对应的位都假设为负类的情况下该输出向量最可能属于的类别 $\boldsymbol{c}_{-1}$，该类别为拒绝域都为负类的最大可能类别. 第二步目的是为了找出在拒绝域对应的位都假设为正类的情况下该输出向量最可能属于的类别 $\boldsymbol{c}_1$，该类别为拒绝域都为正类的最大可能类别. 由于拒绝域所占的长度一般不能超过总长度的一半（可通过代价参数矩阵控制），同时，采取拒绝策略后，码字之间的最小 Hamming 距离也相应地减小，这势必会影响到编码的纠错能力. Simeone 在文章[11]中指出编码时错误码字的个数 v 和拒识的码字个数 μ 存在如下的关系：

$$2v+\mu<d_{\min} \tag{13}$$

即想要纠正一个错误的码字比拒识一个码字困难得多. $d_{\min}$ 为编码矩阵之间的最小 Hamming 距离. 因此拒识的码字个数必须小于矩阵之间的最小 Hamming 矩阵，这也是要保证一定的拒识率，因为过高的拒识率在实际中对于分类机制而言没有太大的意义. 因此在解码之前，要判定拒识位个数与 $d_{\min}$ 直接的关系：如果 $\mu<d_{\min}$ 则按照该方法进行解码；否则，当 $\mu>d_{\min}$，就对该样本进行拒识. 这也是属于带拒绝域的 ECOC 机制，只是该拒识发生在最后的解码阶段.

由前面的分析可以得出，本文的带拒绝域的 ECOC 多类分类的大致框图如图 2 所示.

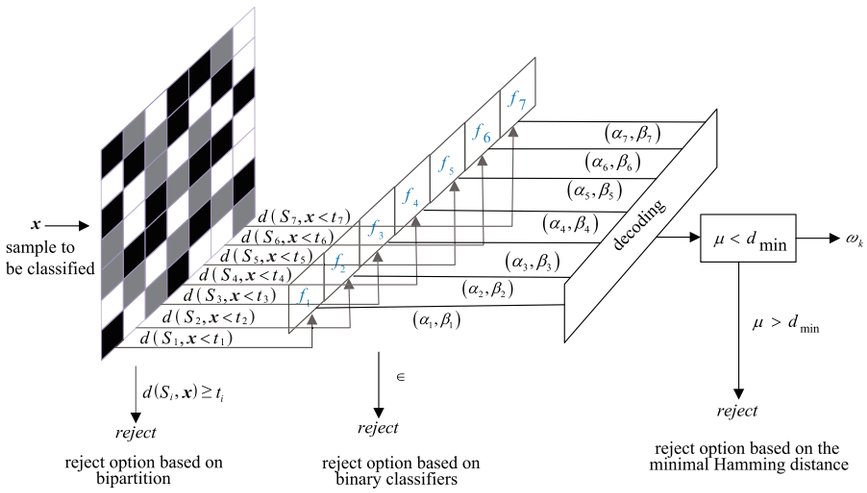

图2 带拒绝域的ECOC识别过程

4 实验

4.1 实验数据

实验中所用的 UCI 数据集及各类数据描述如表 1 所示，其处理及说明见文献[12].

4.2 实验设计

本文在拒绝域设计时，考虑到了二类划分中码字零元素的影响，因此实验中均采用三符号纠错输出编码一对一编码（one-versus-one）、稀疏随机编码（sparse random）、子类编码（SECOC）[13]、基于混淆矩阵的编码方法 CMECOC[14] 和基于 SVDD 的 HECOC 编码方法[12]. 选择支持向量机和决策树为基分类器，线性加权损失函数解码（LLW）和 Hamming 距离解码（ELW）为解码方法. 实验中所用到的各算法的参数设置如表 2 所示.

在训练基分类器并得到带拒绝域的二分器时，人为给定如式（4）的代价矩阵，为了不影响整个实验的可信度，该代价矩阵都被应用于所有 ECOC 集成的分类器. 该代价矩阵可以设为：

	P	N	*reject*
P	0	c_{12}	c_{13}
N	c_{21}	0	c_{23}

其中 $c_{12}>c_{13}$，$c_{21}>c_{23}$. 估计分类错误率时采用的交叉验证法和 t 检验法如文献[15]，计算公式如下：

$$\frac{|\bar{x}-\mu|}{\frac{\sigma}{\sqrt{n}}}\geqslant t_{0.025}(n-1) \tag{14}$$

表 1　UCI 数据集及数据描述(Features:C-continous,B-binary,N-nominal)

	Dataset	Cases	Classes	Atts	Features		
					C	B	N
(*a*)	Ecoli	336	8	7	7	–	0
(*b*)	Glass	214	6	10	9	–	– 1
(*c*)	Segment	2310	7	19	19	–	–
(*d*)	Soybean	306	18	35	–	35	–
(*e*)	vehicle	846	4	18	18	–	–
(*f*)	Vowel	990	11	13	13	–	
(*g*)	Wine	178	3	13	13	–	–
(*h*)	Yeast	1484	10	8	8	–	–
(*i*)	Zoo	101	7	16	1	15	–

表 2　各算法参数设置

Algorithm	Parameters
SVM_{poly}	$C = 1.0$ Tolerance parameter = 0.001 Epsilon = 1.0E-12 Kernel type = polynomial Fit logistic models = true
SVDD[9]	Fracrej = 0.05 Kernel function = RBF Sigma = 5
Treec	Maxcrit = purity Prune = 0 no pruning

μ,σ 分别表示 n 重交叉验证的均值和标准差,$t_{0.025}(4) = 2.7764$,$t_{0.025}(9) = 2.2622$. 实验中所用分类器均来自 PRTool(http://www.prtools.org)工具箱.

4.3　实验结果及分析

(1)基于二类划分的拒绝域分类效果

实验首先利用UCI 公共数据集来验证在二类划分阶段引入拒绝域后应用到不同编码矩阵中的分类效果. 其中 HD * 、LLW * 分别表示引入拒绝域后的 Hamming 距离解码和线性损失函数解码. 从表 3 可以看出,在二类划分阶段引入拒绝域的方法的分类精度都要优于对应的经典 ECOC 多类分类方法,从而应证了带拒绝域的 ECOC 分类在对不属于二类划分的样本数据进行处理时的效果比不包含拒绝域的经典的 ECOC 方法要好,这也正是引入拒绝域的原因所在. 基于二类划分的拒绝域基于样本数据,通过 SVDD 构造拒绝域,对部分不属于二类划分的样本进行拒绝识别,样本输出码字零,从而代替原始的正负类输出,将 2 值输出转化为 3 值输出,从而缩短了输出码字与目标码字的距离,提高分类精度.

表 3　基于二类划分的拒绝域应用到编码矩阵中的效果对比

Datasets		(*a*)	(*b*)	(*c*)	(*d*)	(*e*)	(*f*)	(*g*)	(*h*)	(*i*)
one-versus-one	HD	66.68	91.19	94.85	50.26	68.68	80.91	85.22	51.75	69.05
		11.70	3.41	2.02	6.14	2.58	3.05	12.74	0.60	2.43
	HD *	75.68	93.27	94.72	55.54	80.00	91.82	91.87	67.74	81.14
		11.88	1.86	1.27	4.07	3.39	3.52	6.88	2.06	2.97
	LLW	68.38	91.19	94.85	51.28	67.98	79.60	87.66	51.29	66.10
		8.94	3.41	2.02	6.89	6.31	2.99	3.76	1.78	8.55
	LLW *	73.33	93.76	94.72	53.91	79.05	92.83	93.34	65.22	87.31
		10.39	1.86	1.27	6.68	4.12	2.77	5.05	1.82	1.96
sparse random	HD	70.81	88.45	81.73	40.52	46.55	54.44	79.03	47.92	65.19
		13.55	6.03	1.92	6.96	4.72	3.21	8.82	2.29	14.28
	HD *	79.10	92.76	91.77	52.85	59.56	64.94	88.64	62.32	80.38
		12.61	1.86	2.37	6.11	5.56	3.10	7.73	0.50	10.81
	LLW	71.98	85.71	81.73	48.18	43.60	55.45	80.13	47.85	61.33
		12.10	7.58	1.90	3.74	3.56	3.93	10.28	2.10	9.53
	LLW *	78.90	92.76	87.53	60.77	57.81	65.86	90.24	62.78	79.83
		6.61	1.86	2.39	6.15	4.55	3.75	1.04	3.25	11.66
SECOC	HD	64.02	94.46	93.03	52.60	63.36	65.96	86.42	48.46	72.24
		8.73	5.63	2.37	3.65	5.46	2.30	8.79	3.51	9.58
	HD *	75.93	95.83	93.12	56.15	72.07	71.92	91.45	64.39	87.29
		7.22	3.33	0.99	5.59	4.46	1.81	7.18	2.79	15.97
	LLW	65.49	94.46	93.03	57.52	63.46	68.08	89.34	49.26	71.24
		6.20	5.63	2.37	10.44	5.46	1.36	5.64	1.99	10.40
	LLW *	73.16	95.83	93.12	59.11	71.32	73.49	91.58	64.79	86.33
		12.16	3.33	0.99	2.82	4.05	2.32	6.62	5.33	16.84

续表

Datasets		(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
CMECOC	HD	63.69	91.49	92.03	59.15	63.97	59.39	89.24	58.26	67.43
		9.49	7.04	0.90	3.41	5.22	2.45	10.13	2.84	8.38
	HD *	74.80	94.38	93.46	64.96	75.60	66.06	92.15	74.65	75.33
		9.90	5.32	1.51	5.92	5.61	3.69	7.61	3.28	8.96
	LLW	54.12	91.49	92.03	61.53	62.20	60.20	88.39	56.43	69.46
		7.49	7.04	0.90	1.92	4.95	2.10	11.57	2.28	11.00
	LLW *	63.69	94.38	93.46	65.41	73.16	65.05	91.00	70.07	76.43
		8.16	5.32	1.51	3.04	5.71	3.95	2.92	2.02	9.97
HECOC	HD	62.21	93.91	94.33	59.15	64.04	71.72	93.27	50.32	82.14
		9.18	5.73	1.19	5.39	5.61	2.32	1.83	3.29	16.83
	HD *	78.45	95.27	94.72	63.19	76.58	80.68	95.87	68.05	91.46
		5.17	4.67	1.39	2.68	5.29	3.30	7.44	2.84	15.52
	LLW	67.42	93.91	94.33	62.29	65.00	73.17	82.77	51.26	80.33
		5.09	5.73	1.19	2.96	3.60	3.23	6.06	0.99	9.53
	LLW *	75.97	95.27	94.72	64.54	76.07	81.38	92.20	64.80	87.64
		6.83	4.67	1.39	3.48	3.64	2.28	6.05	1.97	7.38

同时注意到基于二类划分的拒绝域方法在 HD 解码的基础上的分类效果的提升要高于 LLW 解码,这是因为 HD 解码为硬输出解码,基分类器的三值输出直接作用于预测码字和类别码字的最近识别,故拒绝域的作用效果更明显. 部分数据集在经典 ECOC 多类分类下的分类效果基本相同,例如数据集 Glass 和 segmentation. 从总体上看,在二类划分阶段引入拒绝域的方法在性能上还是要高于初始化的分类方法,这是因为拒绝域的引入能够使子类基分类器基于样本数据,有针对性的避免对不属于该二类划分的样本进行决策. 在这一过程中,基分类器对这些样本的输出由原始的正负类输出扩展到零输出,这与其在编码矩阵中的类别码字正好吻合,在解码时就能避免对非子类样本进行决策带来的误差,从而整体上大幅提高分类精度.

(2)拒绝率 ρ 对分类错误率的影响

考虑到纠错输出编码结构的特殊性,编码矩阵的列数决定了基分类器的个数,也就是拒绝域的个数. 例如对于 ecoli 数据集,在基于混淆矩阵的编码输出下,其编码矩阵大小为:8×9. 这样就需要对 9 个二类划分进行拒绝域的构造. 表 3 是在阈值 $t_b = r$ 的情况下产生的. 为体现拒绝率与整体分类错误率的关系,我们将调整阈值,使拒绝率 $\rho \in [0, 0.3]$,步长为 0.05. 当拒绝率过大,整个分类机制将失去意义.

从图 3 中可以看出,在不同的数据集上,拒识率与分类错误率的关系基于不同的基分类器走势大致相同. 随着拒识率的提高,分类错误率有所下降,这是因为对难分样本或者不属于该列二类划分的样本拒识的结果,从而提高了能正确分类样本的比例,拒识率的引入对 ECOC 多类分类的分类性能有促进作用.

5 基于基分类器的拒识分类

表 4 给出了基于基分类器的拒识方法在不同编码矩阵和 Hamming 距离解码中的分类代价,其中 HD * 表示引入拒绝域的 Hamming 距离解码方法. 在绝大部分情况下,在基分类器层引入拒识规则的最小平均风险都要小于与之对应的经典 ECOC 方法. 这反映了基于基分类器的拒识分类在引入代价矩阵和以最小风险作为评价准则的时候,其分类性能要优于经典的方法. 这也是带拒绝域的 ECOC 多类分类的出发点和优势所在. 其中基于 one-versus-one 编码矩阵的 ECOC 多类分类代价减小的幅度最大,这是因为在一对一编码矩阵中,各基分类器的正负类别各只有一种,引入拒绝域后,预测输出中码字零的比例增大,从而减少了将本是零码字类别进行错分的代价. 实验结果表明在基分类器内部引入拒绝域能够降低对困难样本和不属于子类划分样本的错分代价.

6 结论

经典的纠错输出编码拥有的固有结构并不适用于以减小分类损失代价为目标的分类决策过程,本文通过在二类划分和基分类器内部两层引入拒绝域,分别实现对不属于子类划分的样本和难分、易错分样本拒识,并对相应的解码策略进行了调整,实现了带拒绝域的 ECOC 多类分类. 在具体实现过程中初始化编码矩阵可以为任意的编码矩阵,从而提高带拒绝域 ECOC 方法的普适性. 在基分类器内部寻找最佳拒绝域时,提出了一种基于后验概率和最小风险贝叶斯准则的拒绝域构造方法,同时也从侧面证明拒绝域的形成与代价矩阵关系密切. 最后利用实验数据分别对其进行验证发现基于二类划分的拒绝域构造能够实现分类精度的提升,而基于基分类器构造

的拒绝域能够使 ECOC 分类获得最小风险. 如何构造代价敏感的 ECOC 多类分类是文章下一步研究重点.

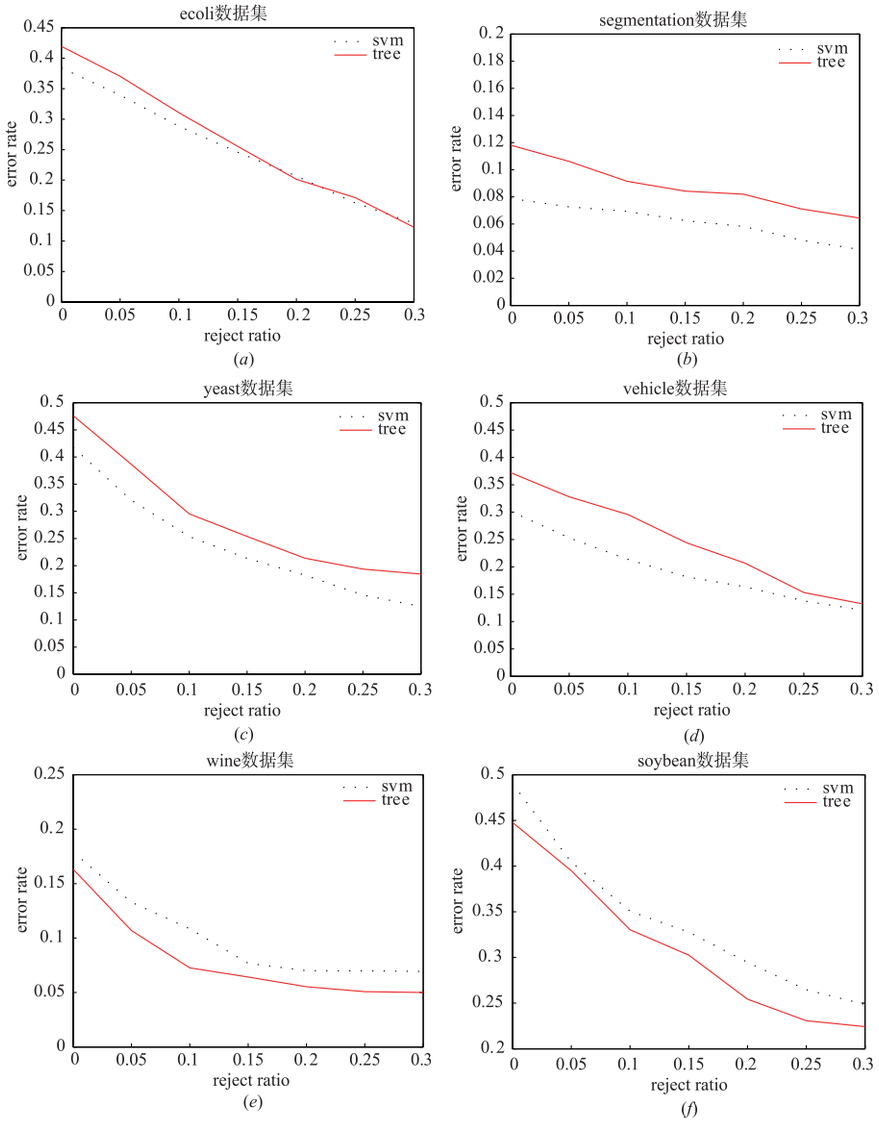

图3 拒识率与分类错误率关系

表 4　基于基分类器的拒绝域应用到编码矩阵中的效果对比

Datasets		(a)	(b)	(c)	(d)	(e)	(f)	(g)	(h)	(i)
one-versus-all	HD	34	10	50	57	140	48	22.5	8.5	16
	HD *	19.5	2.5	40.5	36	110	28	13	4	9.5
one-versus-one	HD	159	252	345	230	300	34	35	13	39.5
	HD *	123	198	168	115	227.5	14	20	7.5	12
sparse random	HD	73.5	58.5	85	299.5	245.5	30	17	9	25
	HD *	64.5	49.5	68.5	252	192.5	28	13	4.5	12
SECOC	HD	126	104	152	230	174	17	18.5	13	38
	HD *	115	92	123	178	103	13.5	14	8.5	27
CMECOC	HD	121	62	346	301	93	22.5	23	12	47.5
	HD *	98.5	53.5	339	269.5	68	11	19	9	35
HECOC	HD	89.5	45.5	85.5	283.5	129	31	24.5	10	28
	HD *	83	37.5	70	212	107	25	13.5	6.5	14

参考文献

[1] T G Dietterich, E Kong. Error correcting output codes corrects bias and variance[A]. Proc of the 21th International Conference on Machine Learning[C]. AAA Press, 1995. 313 – 321.

[2] T G Dietterich, G Bakiri. Solving multi-class learning problems via error-correcting output codes[J]. Journal of Artificial Intelligence Research, 1995, 34(2): 263 – 286.

[3] Mikel Galar, Alberto Fernández, Edurne Barrenechea, Francisco Herrera. DRCW-OVO: Distance-based relative competence weighting combination for one-vs-one strategy in multi-class problems[J]. Pattern Recognition, 2015, 48(1): 28 – 42.

[4] Francesco Ciompi, Oriol Pujol, Petia Radeva. ECOC-DRF: Discriminative random fields based on error correcting output codes[J]. Pattern Recognition, 2014, 47(6): 2193 – 2204.

[5] Zhou Jingdeng, Wang Xiaodan, et al. Research on the Unbiased Probability Estimation of Error-Correcting Output Coding[J]. Pattern Recognition, 2011, 44(7): 1552 – 1565.

[6] Jie Zhou, Hanchuan Peng, Ching Y. Suen. Data-driven decomposition for multi-class classification[J]. Pattern Recognition, 2008, 41(1): 67 – 76.

[7] Paolo Simeone, Claudio Marrocco, Francesco Tortorella. Two stage reject rule for ECOC classification systems[J]. LNCS, 2011, 6713(1): 217 – 226.

[8] 雷蕾, 王晓丹, 罗玺等. ECOC 多类分类研究综述[J]. 电子学报, 2014, 42(9): 1794 – 1800.
LEI lei, WANG Xiao-dan, LUO Xi, et. al. An overview of multi-classification based on error-correcting output codes[J]. Acta Electrinica Sinica, 2014, 42(9): 1794 – 1800. (in Chinese)

[9] David M J Tax. Support vector data description[J]. Machine Learning, 2004, 54(1): 45 – 66.

[10] Bing Zhou, Yiyu Uao, Jigang Luo. Cost-sensitive three-way email spam filtering[J]. Journal of Intelligent Information Systems, 2014, 42(1): 19 – 45.

[11] P Simeon, C Marrocco, F Tortorella. Design of reject rules for ECOC classification systems[J]. Pattern Recognition, 2012, 45(2): 863 – 875.

[12] Lei LEI, Wang Xiao-dan, et al. Hierarchical error-correcting output codes based on SVDD[J]. Pattern Analysis and Applications, 2016, 19(1): 163 – 171.

[13] S Escalera, David M J Tax, O Pujol, P Radeva, Robert P W Duin. Subclass problem-dependent design for error-correcting output codes[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2008, 30(6): 1041 – 1054.

[14] 周进登, 王晓丹. 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.
Zhou Jindeng, Wang Xiaodan. Multiclass classification of adaptive error-correcting output codes based on confusion matrix[J]. Systems Engineering and Electronics, 2012, 34(7): 220 – 226. (in Chinese)

[15] Jindeng Zhou, Xiaodan Wang, et al. Coding design for error-correcting output codes based on perception[J]. Optical Engineering, 2012, 51(5): 322 – 331.

作者简介

雷 蕾 女, 1988 年生于四川南充, 博士生. 研究方向为智能信息处理和目标识别.
E-mail: wendyandpaopao@163.com

王晓丹 女, 1966 年生于陕西汉中, 教授, 博士. 研究方向为模式识别, 机器学习等.
E-mail: afeu_wang@163.com

罗 玺 男, 1988 年生, 硕士, 讲师. 研究方向为智能信息处理.
E-mail: luoxi19887302@126.com

王 玮 男, 1985 年, 讲师. 研究方向为计算机应用.