加权解码在解决纠错输出编码 Consistent-Diverse 平衡问题的应用

周进登[1,2],王晓丹[1],权 文[1],许 燕[3],姚 旭[1]
(1.空军工程大学计算机工程系,陕西三原 713800;2.空军 94719 部队,江西吉安 343000;
3.空军驻上海胶带有限公司军代室,上海 200235)

摘 要: 纠错输出编码作为解决多类分类问题的通用集成框架,能有效的把多类问题分解为二类问题从而使问题得以简化.然而在生成基分类器的过程中,经常面临提高基分类器之间的差异性和增加各基分类器与集成分类器学习的一致性的矛盾,称之为 consistent-diverse 平衡问题.在保证差异性的前提下减小由学习不一致性引起的分类错误率是解决该平衡问题的一个出发点,在此利用加权解码,通过对加权系数矩阵的再学习进而减弱和消除由基分类器学习不一致性产生的误差.实验利用人工数据集和 UCI 数据集分别加以验证,结果表明以集成分类器的分类错误率为适应度函数的遗传算法搜索出的最优加权系数矩阵相比其它方法产生的系数矩阵在解决 consistent-diverse 平衡问题更具有优越性.

关键词: 纠错输出码; 多类分类; 加权解码; 遗传算法
中图分类号: TN912.34 **文献标识码:** A **文章编号:** 0372-2112 (2011) 07-1514-09

Application of Weighted Decoding for the Consistent-Diverse Balance Problem of Error Correcting Output Codes

ZHOU Jin-deng[1,2], WANG Xiao-dan[1], QUAN Wen[1], XU Yan[3], YAO Xu[1]
(1. *Department of Computer Science, Air Force Engineering University, Sanyuan, Shaanxi* 713800, *China*;
2. *Air Force* 94719, *Ji'an, Jiangxi* 343000, *China*;
3. *Air Force Military Delegation Office for Shanghai Adhesive Tape Co, Ltd, Shanghai* 200235, *China*)

Abstract: Error-Correcting Output Codes as a unifying framework for studying the multiclass categorization problems can reduce them to multiple binary problems effectively, thus simplifying the problem. But when generating component classifiers, we usually need to face the contradiction between the diversity among the component classifiers and the consistency of learning between the component classifiers and the ensemble classifiers. We call this contradiction consistent-diverse balance problem. How to reduce the error ratio caused by the inconsistency under diversity big enough is the breakthrough of the balance problem. Using weighted decoding, we can reduce the classification error caused by the learning inconsistency through relearning for weight coefficient matrix. In the proposed algorithm, by using GA to learn the weight coefficient matrix and taking the final generalization error of the ensemble classifiers as the fitness function, we can get the weight coefficient matrix of which the error of the training samples is minimum. The experiments respectively on artificial data sets and UCI data sets have proved that the algorithm is better than others for the consistent-diverse balance problem.

Key words: error-correcting output codes; multiclass categorization; weighed decoding; genetic algorithms

1 引言

纠错输出编码(Error Correcting Output Codes, ECOC)作为一种利用二类分类方法解决多类分类问题的通用框架,能利用在传统二类分类领域取得的丰富成果用以解决多类分类问题,目前已成功应用于人脸识别[1]、文本识别[2]、手写数字分类[3]以及交通指示牌识别[4]等诸多领域,并取得了很好的识别效果.

自 1995 年 Dietterichand 和 Bakiri 首次提出应用 ECOC 解决多类分类问题以来[5],已有大量的相关文献针

收稿日期:2010-11-16;修回日期:2011-02-24
基金项目:国家自然科学基金(No.60975026)

对此方法进行了深入研究和扩展,这些文献的主要工作体现在对现有编码方法的改进和重新提出新的编码方法,此方面的工作有:BCH 编码[5,6]、无遗编码(Exhaustive Codes)[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(Dense Random Codes)和稀疏随机编码(Sparse Random Codes))[8],搜索编码(Searching Codes)[9]和 Hadamard 编码[23].此外根据文献[8]可知一些经典的多类分类方法如 one-versus-all[10]和 one-versus-one[7]也可以认为是 ECOC 框架下编码方法的一种.

解码规则是 ECOC 的另一研究热点,除汉明距离解码外,常用的有欧氏距离解码(Euclidean Decoding)[7],逆汉明距离解码(Inverse Hamming Decoding)[11],最大似然概率解码(Maximizing Likelihood Probability Decoding)[12,22],基于损失函数解码(Loss Based Decoding)[8]及加权损失函数解码(Loss Weighted Based Decoding)[13].式(1)、(2)为经典汉明距离解码公式,其中 $\boldsymbol{M}$ 为编码矩阵,h 为二分器输出结果.

$$d_H(\boldsymbol{M}(r),h(x)) = \sum_{s=1}^{L}\left(\frac{1-\operatorname{sign}((M(r,s)h_s(x)))}{2}\right) \tag{1}$$

$$\hat{y} = \operatorname{Arg}\min_r d_H(\boldsymbol{M}(r),h(x)) \tag{2}$$

在上述解码规则中,加权解码是最值得关注的一种解码方式,其核心是加权系数和解码方式的确定,虽然加权解码的提出往往与具体的解码方式一并出现,使得对解码方式的关注度远远超过对加权方式的研究,加权方式即加权系数向量或矩阵的确定方式,不同加权方式对应不同的系数向量或矩阵.加权方式的不同往往会导致不同的解码结果,其影响程度有时甚至超过解码方式本身.此外应注意传统意义的加权方式通常为一向量,权值的确定即为对该加权向量的求解,而针对 ECOC 解码的加权方式为一矩阵,因此其确定方法为对该权值矩阵的求解,故需做更深入的研究以期找到最适合的方法.另外促使我们对加权方式做深入研究的另一个重要原因是:基于 ECOC 集成框架中个体差异性与学习一致性平衡问题,为叙述方便下文将统一称此问题为 consistent-diverse 平衡问题,下一节对此有详细描述,此平衡问题在已有文献中未被提及过,本文将重点研究如何利用加权解码来解决此平衡问题.

2　consistent-diverse 平衡问题与加权解码

2.1　consistent-diverse 平衡问题

我们知道在一般的集成学习问题中,例如 AdaBoost 或 Bagging 学习器,每一个基分类器学习目标与总的分类器集成的学习目标都是一致的,因此我们称基分类器的学习具有一致性,而基于 ECOC 框架的基分类器学习却不具有一致性.为说明此情况,举例如下,如图 1(a)为 6 类分类问题,要准确的对这 6 类数据进行分类,最优分类器的决策边界应该由图 1(a)中的三条曲线组成,因此对任意学习方法其最终目标就是找出能产生最接近这三条曲线的决策边界的分类器,AdaBoost(Bagging)中每一个基分类器的学习都是以这三条曲线为学习目标并尽可能的产生差异性大的个体分类器,而在 ECOC 中由于每一个基分类器都是一个二分器,因此限制其学习目标只能是这三条曲线中的某一部分,图 1(b)把由这三条曲线组成的最优分类边界按相邻两类的类别号标注如下,假设 ECOC 中的编码矩阵中的某一列把{class1, class2, class3}$^+$划为正类,{class4, class5, class6}$^-$为负类,则该列对应的基分类器的学习目标为决策边界$\{B_{34}, B_{24a}, B_{24b}, B_{25}\}$,由于编码阵的每一列对应不同的二类划分,因此对应的学习目标也各不相同,故基于 ECOC 框架的集成学习不具备一致性.此问题带来的直接影响是即使各基分类器具有贝叶斯一致性*,最终的集成结果也可能是错误的.

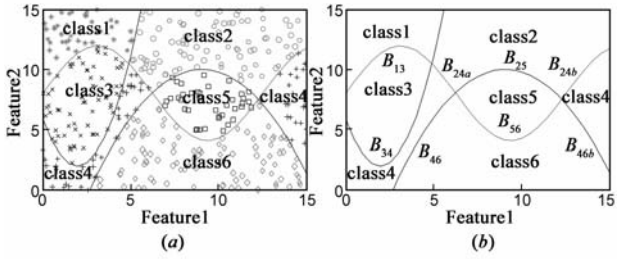

图1　六类均匀数据分布及其决策边界

为说明上述情况,假设有四类学习问题 $N_c = 4$,编码阵 $\boldsymbol{M}$ 已确定如图 2,黑色为正类,白色为负类.对测试样本 x,其类后验概率已知为(0.4,0.1,0.3,0.2),且基分类器(h_1, h_2, h_3)能提供准确的二类概率输出即具有贝叶斯一致性,则各基分类器正类概率输出结果为:$h_1: q_1 = p_3 + p_4 = 0.5$, $h_2: q_2 = p_1 + p_3 = 0.7$, $h_3: q_3 = p_1 = 0.4$.之所以能得到这样的概率输出是建立在各基分类器在学习过程中能得到最优决策边界,即基分类器在各自样本空间都是"最完美"的分类器.然而由于基分类器的学习目标各不相同(即图 1(b)中不同的目标边界),其最优成员个体的结果并不能对最终决策起促进作用,相反若我们利用汉明距离解码如图 2,则最终决策为 y_3,对应汉明距离最小为 1.2,而根据贝叶斯决策最终结果为 y_1,因为其对应的后验概率最大为 0.4,两者结果并不相同,为了排除解码对该问题的影响,我们再分别利用欧氏距离和指数损失函数的解码策略分别对其进行验证得到结果如图 2 所示.可以看出,在此两种解码规则下,该分类问题的最终决策仍为 y_3,因此

* 具有贝叶斯一致性,即该分类器与基于贝叶斯决策规则产生的结果相同,详细说明见文献[21]

可以说在基于 ECOC 框架的集成学习中确实存在不一致性学习给最终决策带来偏差甚至是错误的情况,这种部分与整体的学习不一致性产生的根本原因就是它们所学习的决策边界不同,因而产生的各基分类器虽然能对部分类别精确分类,但基分类器之间的决策"冗余"度却可能减少,而这"冗余"度的大小又直接影响着基于 ECOC 框架分类集成的最终效果,因此解决这种由部分与整体所产生的学习不一致性是利用 ECOC 框架进行多类分类所面临的一个重要问题.

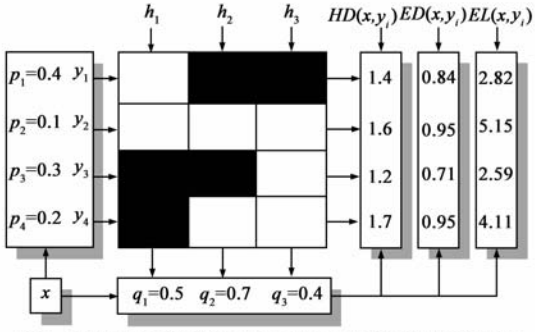

图2 基于贝叶斯决策与ECOC在三种不同解码规则下分类结果比较

值得注意的是文献[14]提到正是基于这种学习的不一致性,ECOC 才能产生差异性较大的基分类器,从而提高整个集成效果,文献[14]的实验结果也证实了这一点.然而与 AdaBoost 和 Bagging 等集成算法比较,ECOC 产生差异性大的基分类器的代价是学习的不一致性,可以说基于 ECOC 框架的集成学习中各基分类器学习不一致性是与生俱来的,因此解决好此问题将使基于 ECOC 集成学习更为有效.下面将介绍如何利用加权解码对该平衡问题进行求解.

2.2 加权解码

加权解码在基于 ECOC 框架的集成学习中应用非常普遍,但正如文章开头部分所述人们对加权解码的研究主要集中在解码方式的探讨,不同的解码方式会产生差异性较大的结果,如何根据实际情况选择不同的解码方式已被众多研究者关注,然而加权方式的选择却并没有引起足够的重视,特别是在 ECOC 框架下的加权方式更少被研究,目前仅有少数文献提到过此类问题[13,19],关于加权解码在 ECOC 下的作用,特别是在解决 consistent-diverse 平衡问题方面的能力在已有的文献中尚未被研究.

图 3 是基于 ECOC 加权解码的一般流程.假设编码矩阵已经确定,首先由训练样本根据编码矩阵进行重标记得到训练样本的 L 个二类划分,然后分别对每一个二类划分根据学习法则 F 进行训练得到 L 个基分类器($h_1(x),h_2(x),\cdots,h_L(x)$).在训练好基分类器后,经典的 ECOC 集成方法是选择一种解码方式,给定一个测试样本分别用这 L 个二分器对其进行判断得到一输出向量,最后根据选定的解码方式进行判定得出最终结果,其过程如图 3 虚线框内容所示.而对于具备再学习过程加权解码的 ECOC 集成方法,在选定好解码方式后,还需根据加权学习规则 H 对加权系数矩阵即加权方式进行再学习,学习样本为训练集全部样本,最后得到最优加权系数矩阵,正是由于拥有再学习过程,从而使先前由基分类器学习的不一致性带来的影响在此过程得到有效的修正,这也是本文提出利用加权解码来解决 consistent-diverse 平衡问题的出发点.为此文章提出利用在解码阶段以分类器集成的整体性能(本文以集成的最终分类错误率)为目标函数进行再学习的加权解码方法来解决不一致性学习带来的不利影响,而再学习的过程即最优加权方式的产生过程.

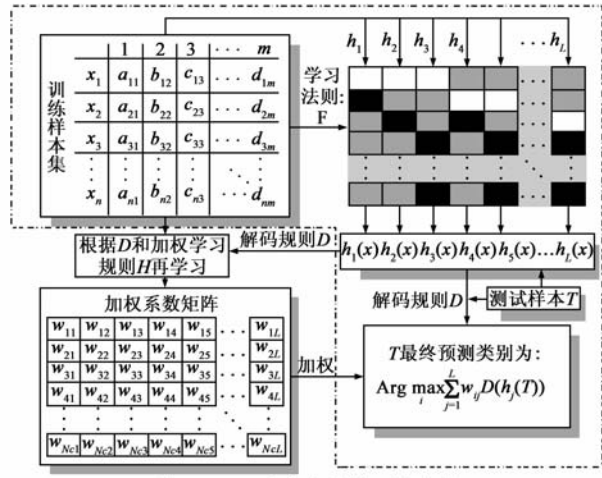

图3 ECOC加权解码的一般流程

3 三种不同加权方式加权解码在 consistent-diverse 平衡问题中的应用

由 2.2 节可知具有再学习的加权解码其再学习过程是加权方式的产生过程,是通过对训练样本再学习产生最优加权方式的过程,此学习过程也是减弱 ECOC 不一致性学习所带来偏差的关键部分,因此好的加权方式能在 consistent-diverse 平衡问题找到最有利的平衡点,从而有效地减少最终分类误差.为了突出基于再学习过程产生的加权方式确实能减弱不一致性学习对最终分类器性能产生的不利影响,下面介绍三种加权方式产生方法,并分别介绍其在加权解码中对 consistent-diverse 平衡问题所产生的影响.

3.1 错误率加权方式(EW)

该加权方式的产生来源于一种最直接的想法即:基分类器 h_s 对第 r 类样本分类正确率越高,则对应的加权系数 w_{rs} 越大[13].因此我们可以利用基分类器在各类样本的正确率作为加权方式产生的来源,此"学习"

过程仅仅是计算每个基分类器的错误率,因此并不具备真正意义上的学习.其计算错误率的公式为:

$$R(r,s)=\frac{1}{n_r}\sum_{k=1}^{n_r}\varphi(h_s),$$
$$\varphi(h_s)=\begin{cases}1,\text{if } h_s=M(r,s)\\0,\text{其他}\end{cases}\tag{3}$$

当 ECOC 为 Ternary-ECOC 时,由于编码阵存在不参与对应基分类器训练的类,即 $\boldsymbol{M}(r,s)=0$ 时所对应的 r 类在训练基分类器 h_s 时将被忽略,此时对应的加权系数矩阵中的权值 $\boldsymbol{W}(r,s)$ 也将被设置为 0.为了使加权系数矩阵被看作为一种离散概率密度分布,把系数矩阵的每一行都进行归一化处理:

$$\boldsymbol{W}(r,s)=\frac{R(r,s)}{\sum_{s=1}^{L}R(r,s)}\tag{4}$$

利用式(4)我们便可得最终的加权系数矩阵 $\boldsymbol{W}$.

可以看出该加权方式的产生方法的核心就是利用各基分类器的分类结果作为加权系数矩阵的权值,其特点是简单、高效(一次学习即可同时确定基分类器和加权方式)且不依赖具体的解码方式,因此我们可得算法 1 基于错误率加权方式产生算法(Weight based on Error Rate Algorithm,WERA).

算法 1　基于错误率加权方式产生算法(WERA)

输入:编码阵 $\boldsymbol{M}$

Step1　计算每个基分类器 h_s 对每一类的分类错误率:

$$R(r,s)=\frac{1}{n_r}\sum_{k=1}^{n_r}\varphi(h_s),$$
$$\varphi(h_s)=\begin{cases}1,\text{if } h_s=M(r,s)\\0,\text{其他}\end{cases}$$

Step2　把错误率矩阵 $\boldsymbol{R}$ 的每一行进行归一化处理得最终加权系数矩阵:

$$W(r,s)=\frac{R(r,s)}{\sum_{s=1}^{L}R(r,s)}$$

输出:加权系数矩阵 $\boldsymbol{W}$

然而我们注意到整个加权方式的产生过程并没有体现以分类器集成的最终分类误差作为目标函数对全部训练样本进行再学习的过程,因此可以说基于此加权方式的加权解码不具备解决 consistent-diverse 平衡问题的能力,其对分类性能的影响源自于各基分类器对每一类别数据分类性能优劣的加权,而由 2.1 节可知,单个基分类器的好坏并不能影响最终集成效果.

3.2　类可分性加权方式(SW)

在介绍此方法之前,首先定义如下函数:

$$C_s(x)=\begin{cases}1,\text{if } h_s(x)=M(r,s)\\0,\text{if } h_s(x)\neq M(r,s)\end{cases}\tag{5}$$

其含义为:当基分类器 h_s 对样本 x 的预测值 $h_s(x)$ 等于样本所属类 r 在编码阵第 s 列对应的值即 $M(r,s)$ 时,则该基分类器对应的度量函数 $C_s(x)$ 值为 1,若不等则为 0.同时我们定义其互补函数 $\overline{C}_s(x)=1-C_s(x)$,取值与 $C_s(x)$ 相反.继而给定样本类别标签 r 和基分类器标签 s,对应的类可分性可用如下公式表示[19]:

$$W_{rs}=\max\left\{0,\frac{1}{K_r}\left[\sum_{\substack{p\in class\ r\\q\notin class\ r}}C_s(p)C_s(q)-\sum_{\substack{p\in class\ r\\q\notin class\ r}}\overline{C}_s(p)\overline{C}_s(q)\right]\right\}\tag{6}$$

其中 p 和 q 都为训练样本元素,K_r 为归一化参数,用以确保权值矩阵中每一行元素之和为 1.式(6)的含义为:基于类可分性的加权系数矩阵权值为基分类器对每一类样本与其余类样本预测值正负相关性之差的和.算法 2 为基于类可分性加权方式产生算法(Weight based on Class Separability Algorithm,WCSA).

同基于错误率加权方式产生算法一样,类可分性加权方式产生算法并没有使基分类器向同一学习目标学习的趋势,因此该算法仍然没有直接关注 consistent-diverse 平衡问题,虽然在实验中观察到该加权方式也能对分类器集成性能产生积极作用,但本质上与上节所述加权方式相同.

算法 2　基于类可分性加权方式产生算法(WCSA)

输入:编码阵 $\boldsymbol{M}$,训练样本集 T

Step1 设 $r=1$,把样本集 T 划分成两类 $T_1\in class\ r$ 和 $T_2\notin class\ r$;

Step2 依次判断 $h_s(x)(s=1,2,\cdots,L)$ 对所有样本对 $\{(x_1,x_2)\mid x_1\in T_1,x_2\in T_2\}$,当 $h_s(x_1)=M(r,s)$ 且 $h_s(x_2)=M(r,s)$ 时(r' 为 x_2 的真实类标签),则 $w_{rs}=w_{rs}+1$;当 $h_s(x_1)\neq M(r,s)$ 且 $h_s(x_2)\neq M(r',s)$ 时,则 $w_{rs}=w_{rs}-1$;

Step3 若 $r<N_C$,则设 $r=r+1$ 并转到 Step1,否则取 $M(r,s)=\max(0,M(r,s))$ 并对 M 的每一行归一化处理.

输出:加权系数矩阵 $\boldsymbol{W}$

寻求通过再学习并直接以减小分类器集成的最终分类误差作为目标产生最优加权方式,并基于此加权方式进行加权解码,是我们提出用以解决 consistent-diverse平衡问题的一套方案,为验证该方案的可行性,我们提出利用遗传算法作为实现这一目标的工具,通过建立以集成最终分类错误率为目标函数,寻找出最优的加权方式.

3.3　基于遗传算法的加权方式(GW)产生方法研究

由 2.2 节分析可以看出利用加权解码解决 consistent-diverse 平衡问题的出发点是由于加权方式的产生是一个再学习过程,只要构造合适的再学习规则,则产生的加权方式就能在一定程度上缓解甚至是消除基分类器不一致性学习带来的影响,而考虑到该学习的不一致性是由各基分类器学习的目标各不相同所带来的,为此在本节中我们通过建立一个总的目标函数——集成最终分类错误率,并以此作为寻找最优加权方式的依据.

$$F = \frac{1}{n}\sum_{r=1}^{n} I(f_r \neq y_r)$$
$$f_r = \text{Arg}\max_{r}\sum_{s=1}^{L} W(r,s)L(h_s(x_r)M(r,s)) \quad (7)$$

其中,当 $f_r \neq y_r$ 时,$I(\cdot)=1$;当 $f_r = y_r$ 时 $I(\cdot)=0$. y_r 为样本 x_r 的真实类标签.

此目标函数的意义是找到最优加权系数矩阵,使加权之后的基分类器集成最大限度的减少训练样本错误率.在此再学习过程中需要不断优化的是各基分类器在每一类的加权系数,通过统一的目标函数来不断优化这些参数,最终使基分类器的不一致性学习带来的影响在其对应的加权参数中通过不断学习得以克服.图 4 为两种学习过程的示意图.在寻找最优加权方式时,利用遗传算法(Genetic Algorithms,GA)来对权值系数矩阵空间进行搜索,在具体的搜索过程中我们采用二进制编码对加权系数矩阵中的每一个元素进行编码,由于加权系数矩阵每一行元素都可以看作是对应类的概率密度函数,因此其元素值都是介于 0 和 1 之间的值,为此我们对每一加权系数矩阵中的元素采用 10 位二进制编码,对应保留 3 位小数因为 $2^9 < 1\times10^3 < 2^{10}$.这样每一个遗传个体占用 $\frac{N_c \times L \times 10}{8}$ 个字节,初始化时各加权系数矩阵元素值取 $1/L$,适应度函数为式(7),算法 3 为该算法的具体描述称之为基于遗传算法的加权方式产生算法(WGAA).

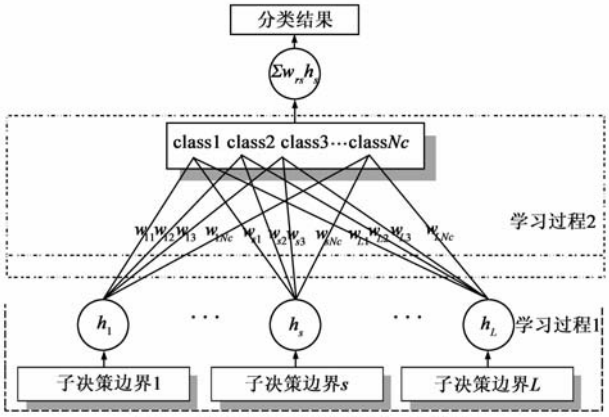

图4 再学习生成加权方式一般流程

算法 3 基于遗传算法的加权方式产生算法(WGAA)

输入:编码阵 $\boldsymbol{M}$,训练样本集 T,阈值 λ,迭代次数 N

Step1 利用二进制编码方法随机产生 m 个原始种群个体,每一个个体为加权系数矩阵所有元素编码的二进制串;

Step2 计算每一个体的适应度函数;

Step3 当迭代次数小于 N 时,若存在个体其适应度函数值小于 λ,则该个体为最优个体解码得到最优加权方式,否则进行交叉和变异产生新的种群个体并转到 Step2.

输出:加权系数矩阵 $\boldsymbol{W}$

该加权方式的学习与基分类器的学习起互补的作用,训练基分类器时不一致性学习能产生差异性较大的个体,而加权方式的再学习又使这种由不一致性学习带来的负面影响得到降低,因此可以说此加权方式能很好地解决 consistent-diverse 平衡问题,这也是直接以该平衡问题作为再学习目标加权方式产生的有益尝试.此方法将验证我们提出的"以分类器集成的整体性能为目标函数进行再学习过程产生的加权方式在加权解码中能对 consistent-diverse 平衡问题产生积极影响".

3.4 三种加权方式的比较

利用加权解码解决 consistent-diverse 平衡问题的重要手段是通过再学习产生最优加权方式,上述提出的三种加权方式产生方法,前两种加权方式的产生并没有体现对由基分类器学习不一致性带来的误差的再学习过程,其本质是对单个基分类器在样本中每一类数据分类性能的衡量,因此其加权解码并不能减小或消除这种误差,相反第三种加权方式的产生直接以分类器集成的最终分类误差为目标,并使其作为遗传算法的适应度函数,在加权系数矩阵空间进行搜索直到找到最优加权系数矩阵.因此可以说基于遗传算法的加权方式产生方法是本文提出利用再学习解决 consistent-diverse 平衡问题的最有效的体现.但同时我们注意到三种加权方式产生方法的时间复杂度各不相同,其中 WERA 算法加权系数矩阵的每一列都是通过对所有测试样本的计算得到,故其时间复杂度为 $O(n\times L)$,n 为测试样本数,L 为系数矩阵的列数. WCSA 算法加权系数矩阵的每一列中每一个权值为对应类和其它类的相对可分性之和,故其算法时间复杂度为 $O(n^2\times L)$,而 WGAA 算法计算加权系数矩阵时是对整个加权系数矩阵进行求解,每一次迭代过程都是对所有训练样本计算其适应度函数值,直到找到满足条件的系数矩阵使得适应度函数值达到设计要求或迭代次数达到最大为止,因此其时间复杂度最大为 $O(n\times L\times N)$,N 为最大迭代次数.由此可以看出在算法时间复杂度方面,WERA 最小,而 WCSA 与 WGAA 次之.

4 实验

通过实验我们将验证文章所提三种不同加权方式对解决基于 ECOC 框架集成学习中产生的 consistent-diverse 平衡问题所起的作用.

4.1 实验数据

在实验中我们通过对两种类型的数据分别进行验证.第一种是人工数据,我们将利用文献[14]提供的 6 类二维样本数据,之所以选择二维数据是为了更直观的显示出各类的分类间隔,以便更好的验证文章结论.第二种数据为公共数据集,选取 7 种 UCI 数据[15],数据

的属性如表 1 所示.

表 1　UCI 数据集各数据描述

Problem	# Train	# Attributes	# Classes
Yeast	1484	8	10
Segmentation	2310	19	7
Sat	6435	36	6
Glass	214	9	7
Vowel	990	13	11
Ecoli	336	8	8
Isolet	7797	617	26

4.2　实验设计

为了比较三种基于不同加权方式的解码策略在解决 consistent-diverse 平衡问题发挥的作用,实验中突出 consistent-diverse 平衡问题对分类的影响,我们构造 6 类完全可分的人工数据集如图 1(贝叶斯最小错误率为 0),且样本分布为均匀分布,同时为了尽可能降低由基分类器本身所产生的误差,我们选择支持向量机作为基分类器,其中核函数为径向基核函数(通过实验验证单个基于该类型核函数的 SVM 能使分类错误率保持较低水平,参数 $C=1, r=0.6$). 在选择输出编码矩阵时,各基分类器学习的不一致性通过其学习的决策边界差异性大小来体现. 如图 1 决策边界集为 $R=\{B12, B13, B24, B25, B34, B46, B56\}$,我们构造 4 种输出编码矩阵并规定编码长度都为 7,构造方法为:每一个基分类器从决策边界集 R 中任意取 $i(i=1,2,3,4)$ 个子边界学习,所取边界不能完全相同且保证每一个边界至少被取到 1 次,可得输出编码阵如图 5.

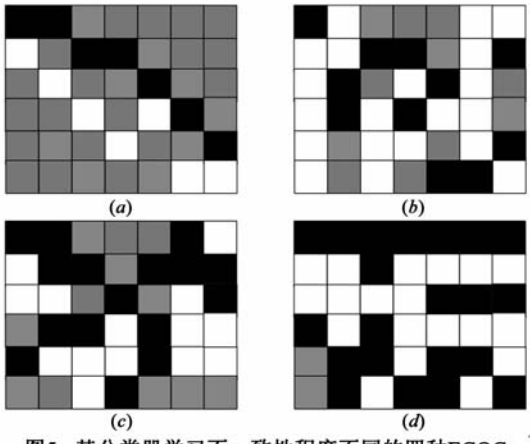

图5　基分类器学习不一致性程度不同的四种ECOC

在图 5 的 4 种编码阵中,编码阵 $\boldsymbol{M}_a$ 每个基分类器只学习 1 种子决策边界即 $subdecision_{Ma}(h_s)=1$,其它三种分别为:$subdecision_{Mb}(h_s)=2$、$subdecision_{Mc}(h_s)=3$ 和 $subdecision_{Md}(h_s)=4$,因此我们认为各编码阵不一致学习程度为:$\boldsymbol{M}_a>\boldsymbol{M}_b>\boldsymbol{M}_c>\boldsymbol{M}_d$. 实验中我们将对文章所提三种解码方式分别利用这四种不同输出编码进行测试和比较.

在 UCI 数据中我们选择 5 种不同的编码方法产生的编码矩阵,它们分别是:one-versus-all、one-versus-one、密集随机编码(Dense Random Codes)和稀疏随机编码(Sparse Random Codes),两种随机编码的码长分别取 $\lceil 10\log_2 Nc\rceil$ 和 $\lceil 15\log_2 Nc\rceil$. 另外我们还将选择 2 种不同的解码方式,它们分别是:汉明距离解码(HD)和指数损失函数解码(EL). 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 时我们采用 10 重交叉验证,小于 500 时采用 5 重交叉验证来进行,并利用双边估计 t 检验法来计算置信水平为 0.95 的分类错误率置信区间作为最终结果,计算公式如下:

$$\frac{|\bar{x}-u|}{\sigma/\sqrt{n}}\geqslant t_{0.025}(n-1) \tag{8}$$

u,σ 分别表示 n 重交叉验证的均值和标准差,$t_{0.025}(4)=2.7764, t_{0.025}(9)=2.2622$. 同时为了对实验结果进行统计分析,采用 Nemenyi 检验法对各算法之间的差异显著性进行检验.

4.3　实验结果和分析

4.3.1　人工数据集

在人工数据集中,四种编码的不一致性学习程度各不相同,由于存在 consistent-diverse 平衡问题,因此我们并不能判断这四种编码矩阵在处理该问题时的优劣,但这并不影响实验的有效性,在实验中我们主要验证的是不同加权方式对不一致性学习程度的影响,即利用不同加权方式加权是否能降低不一致性学习对总的分类器性能降低的不利影响,以此间接达到 consistent-diverse 平衡的目的. 表 2 为四种不一致性学习程度各不相同的编码阵在两种不同解码方式(HD:汉明距离解码,EL:指数损失函数解码)下的分类结果.

表 2　三种加权方式在两种不同解码方式下 ECOC 集成分类错误率

	$\boldsymbol{M}1$	$\boldsymbol{M}2$	$\boldsymbol{M}3$	$\boldsymbol{M}4$
HD	44.98 ± 1.45	14.09 ± 0.90	26.32 ± 1.89	4.09 ± 1.19
*SW*HD	49.51 ± 1.35	14.38 ± 1.04	9.18 ± 1.31	3.79 ± 1.14
*EW*HD	49.51 ± 1.35	14.28 ± 1.07	9.52 ± 2.08	3.39 ± 1.18
*GW*HD	37.59 ± 1.12	6.47 ± 1.67	9.49 ± 1.47	3.39 ± 1.35
EL	45.19 ± 1.99	14.61 ± 1.05	9.62 ± 1.77	4.00 ± 1.65
*SW*EL	47.99 ± 3.39	14.19 ± 0.87	9.91 ± 2.16	4.50 ± 1.19
*EW*EL	49.57 ± 1.43	14.10 ± 0.52	9.07 ± 1.51	4.09 ± 1.06
*GW*EL	37.59 ± 1.07	6.81 ± 1.25	10.17 ± 1.77	3.58 ± 1.19

从表 2 中的结果可以看出当输出编码阵为 $\boldsymbol{M}1$ 和 $\boldsymbol{M}2$ 时,由 3.2 节可知 $\boldsymbol{M}1$ 和 $\boldsymbol{M}2$ 的不一致性学习程度是比较大的($\boldsymbol{M}1$ 各基分类器只学习 1 个子决策边界,$\boldsymbol{M}2$ 各基分类器学习 2 个子决策边界),在两种不同解码方式下,通过再学习过程产生的加权方式(GWHD 和 GWEL)在加权解码中错误率是最小的. SWHD 和 EWHD 在此两种编码阵下其分类错误率相比无加权的汉明距

离解码(HD)不降反升.SWEL 和 EWEL 在与基本的基于指数损失函数解码(EL)比较时,当编码阵为 $M1$ 时错误率增加,$M2$ 时降低,但降低的程度非常小.相比之下 GWHD 和 GWEL 的分类错误率较 HD 和 EL 的降幅是非常大的,GWHD 较 HD 的错误率减少程度分别为 16.4% 和 54.1%,GWEL 较 EL 的错误率减少程度分别为 16.8%和 53.4%.由此可以看出基于这种通过再学习产生加权方式的加权解码(GWHD 和 GWEL)对基分类器学习不一致性程度大的 ECOC 集成确实能起到减少其不利影响的效果.

注意到实验结果中当输出编码为 $M3$ 时,基于 SWHD、EWHD 和 GWHD 的错误率要远小于基于 HD 的错误率,但 EL 和 SWEL、EWEL 及 GWEL 的错误率基本相同,且与除 HD 外的其它三种不同加权方式的加权解码错误率也基本一致,说明不同的解码方式(例如 HD 和 EL)可能会产生差异性大的结果,此结论在文献[8]也有证明,此外通过加权解码可以克服非加权解码带来的解码误差.

4.3.2 UCI 数据集

从图 6 和图 7 中可以看出基于遗传算法加权方式(GW)解码的 ECOC 集成分类错误率普遍要小于基于类可分性加权方式(SW)和错误率加权方式(EW)的解码错误率.在总共进行的 56 次实验中,基于 GW 的集成分类错误率最小的次数有 34 次,SW 为 13 次,EW 为 9 次.而在与无加权的汉明距离解码和指数损失函数解码相比较时,基于 GW 加权的集成分类错误率要小的次数有 43 次,SW 为 30 次,EW 为 23 次.由此可见以集成最终分类误差为目标通过再学习产生的加权方式在对基于 ECOC 框架的集成进行加权解码时确实能对 ECOC 的集成效果产生积极影响.

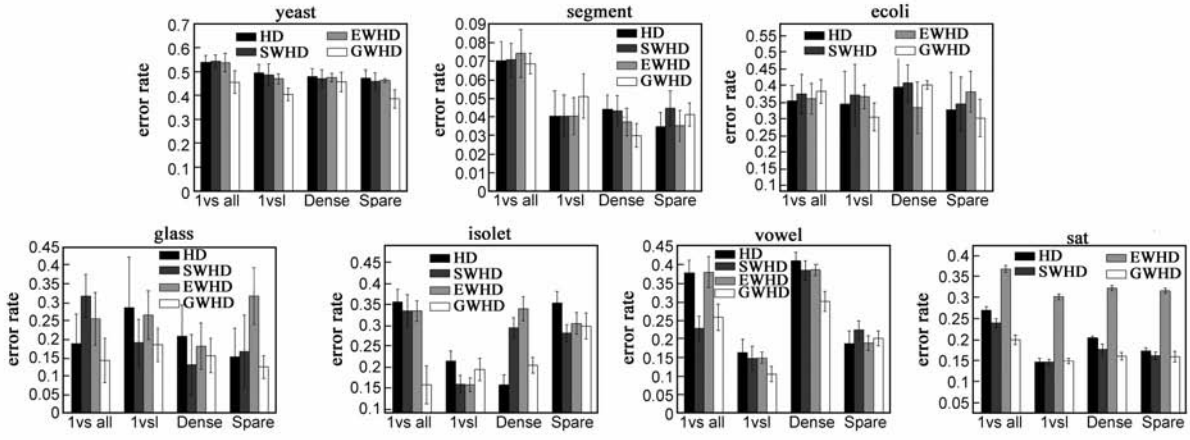

图6 基于汉明距离解码的UCI数据实验结果

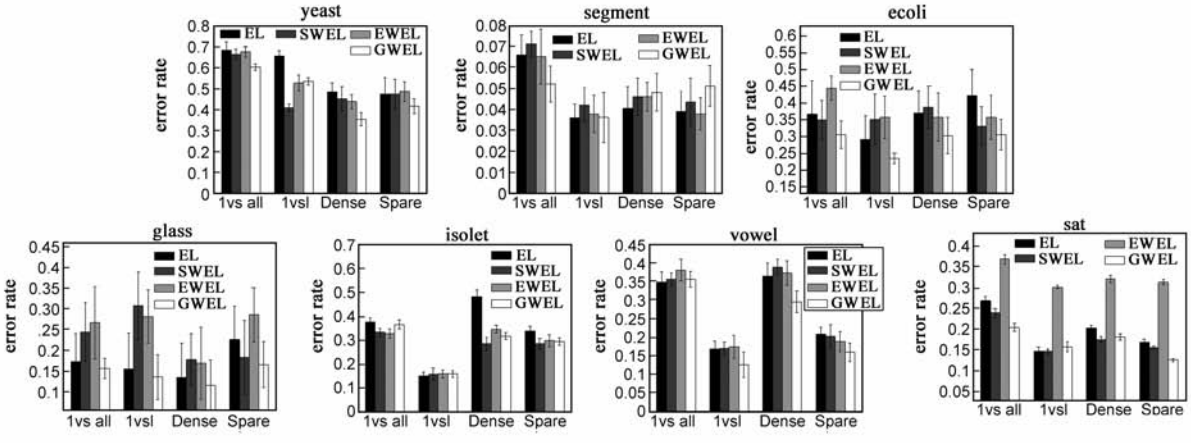

图7 基于指数损失函数解码的UCI数据实验结果

为了得到具有统计意义的实验结论,我们首先利用秩和检验法对上述结果进行分析,其中秩水平计算如下:

$$R_j = \frac{1}{J}\sum_i r_i^j \tag{9}$$

r_i^j 为每种加权方式在第 i 类问题中用基于第 j 种编码阵所得到的秩大小,J 为每种方法所进行的实验次数,在本次实验中 J 为(4 种编码阵 × 7 种 UCI 数据).表 3 为各方法所对应的秩和平均数,其中加粗部分为基于各编码阵中秩和平均最小值即对应分类错误率最小的加权方式.

表 3　各加权方式秩和平均数比较

Coding	HD	SWHD	EWHD	GWHD	EL	SWEL	EWEL	GWEL
1vs all	4.71	4.71	5.42	2.42	5.00	4.57	5.85	3.14
1vs 1	5.00	4.28	4.85	4.00	3.42	4.85	5.57	3.42
dense	5.85	4.42	5.28	3.00	4.85	4.57	4.85	3.00
sparse	3.85	4.57	5.57	3.00	6.28	4.42	5.28	2.85
global Rank	4.85	4.50	5.28	3.10	4.89	4.60	5.39	3.10

从表 3 中可以看出基于 GW 解码的 ECOC(GWHD 和 GWEL)秩和平均数最小都为 3.107,SW 次之,EW 最大.为了验证这三种不同加权方式对 ECOC 集成分类错误率的影响具有统计意义上的显著差别,我们利用 Nemenyi 检验方法—即两种方法具有显著性差异当此两种方法的秩和平均差大于临界值 *CD*(Critical Difference value)[20]

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6J}} \tag{10}$$

其中 q_α 可通过查询"The Studentized Range Statistic"表得到,k 为所要验证的方法数,J 为每次实验的次数.在本实验中我们比较了 8 种方法在置信水平为 $\alpha = 0.05$ 下的分类效果如表 3 所示,即 $k = 8$, $q_{0.05} = 1.860$,代入式(10)可得差异临界值(*CD*)为 1.218.观察表 3 可知基于 GW 的秩平均数比 SW 和 EW 的秩平均数都要小且差值都大于差异临界值,因此我们可以说基于 GW 的 ECOC 加权解码在 95% 的置信区间都要好于其它两种方法.

5　结论

Consistent-diverse 平衡问题是利用 ECOC 集成方法解决多类分类所必须面临的平衡问题.解决这种由集成个体之间的差异性带来的学习不一致性问题在已有文献中很少被关注,本文在对该平衡问题产生的原因进行分析和总结后提出利用加权解码,通过对加权方式的再学习达到消除由各基分类器学习不一致性带来的误差.在学习过程中利用遗传算法以集成最终分类误差为适应度函数进行搜索进而得到最优加权系数矩阵(即最优加权方式),最后利用此最优加权系数矩阵进行加权解码使基于 ECOC 集成面临的 consistent-diverse 平衡问题得以有效解决.如何在 ECOC 集成学习中对 consistent-diverse 问题找出最合适的平衡点是提高 ECOC 集成性能的重要手段,本文讨论了在给定输出编码阵的前提下怎样尽可能降低由 consistent-diverse 问题引起的不利影响即由基分类器的差异性导致的学习的不一致性问题,然而如果能在对多类问题确定输出编码阵的同时考虑该平衡问题,那么该平衡问题的不利影响将在编码阶段就被克服,此类问题即文章开头所述的基于数据的编码.为此如何在基于数据编码方法的研究中考虑 consistent-diverse 平衡问题将作为本文的下一步研究重点.

参考文献

[1] Windeatt T, Smith RS, Dias K. Weighted decoding ECOC for facial action unit classification[A]. 18th European Conference on Artificial Intelligence (ECAI)[C]. Patras, Greece, 2008. 26 – 30.

[2] R. Ghani. Combining labeled and unlabeled data for text classification with a large number of categories[A]. Proc Int'l Conf Data Mining[C]. IEEE Press, 2001. 597 – 598.

[3] J Zhou, C Suen. Unconstrained numeral pair recognition using enhanced error correcting output coding: a holistic approach [A]. Proc Int'l Conf Document Analysis and Recognition[C]. IEEE Press, 2005. 484 – 488.

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

[5] TG. Dietterich, G. Bakiri. Solving multiclass learning problems via error correcting output codes[J]. J Artificial Intelligence Research, 1995, 2(1): 263 – 286.

[6] W W Peterson, JR Weldon. Error-Correcting Codes[M]. MIT Press, Cambridge, MA, 1972.

[7] T Hastie, R Tibshirani. Classification by pairwise coupling[J]. The Annals of Statistics, 1998, 26(2). 451 – 471.

[8] EL Allwein, R E Shapire, Y Singer. Reducing multiclass to binary: a unifying approach for margin classifiers[J]. J Machine Learning Research, 2000, 1(1): 113 – 141.

[9] Jiang Yan-huang, , Zhao Qiang-li, Yang Xue-jun. A search coding method and its application in supervised classification [J]. Journal of Software, 2005, 16(6): 1081 – 1088.

[10] N J Nilsson. Learning Machines[M]. McGraw-Hill, 1965.

[11] T. Windeatt , R Ghaderi. Coding and decoding for multi-class learning problems[J]. Information Fusion, 2003, 4(1): 11 – 21.

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

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

[14] EB Kong, TG Diettrich. Probability estimation via error correcting output coding[A]. International Conference of Artificial Intelligence and Soft Computing[C]. Banff, Canada, 1997. 1 – 4.

[15] A Asuncion, D Newman. UCI Machine Learning Repository [D]. School of Information and Computer Sciences, Univ of California, Irvine, 2007.

[16] Cortes, C,& Vapnik, V. Support-vector networks[J]. Machine

Learning,1995,20(3):273 – 297.
[17] Friedman,J H. On Bias,Variance,0/1-Loss,and the Curse of Dimensionality[R]. Department of Statistics,Stanford University,1996.
[18] Domingos,P. A unified bias-variance decomposition and its applications[A]. Proceedings of the 17th International Conference on Machine Learning[C]. Bari,Italy: Morgan Kaufmann,2000. 123 – 131.
[19] Smith RS,Windeatt T. Class-separability weighting and bootstrapping in error correcting output code ensembles[A]. Benediktsson,JA,Kittler,J,Roli,F (eds.) MCS 2010. LNCS[C]. Springer,Heidelberg ,2010. 185 – 194.
[20] J Demsar. Statistical comparisons of classifiers over multiple data sets[J]. J Machine Learning Research,2006,7(1):1 – 30.
[21] GM James. Majority Vote Classifiers:Theory and Applications[D]. Department of Statistics,University of Stanford California, 1998.
[22] Jindeng Zhou,xiaodan Wang,Heng Song. Research on the unbiased probabilty estimation of error-correcting output Coding[J]. Pattern Recognition,2011,44(7):1552 – 1565.
[23] 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 An-rong, Xie Xiang, Kuang Jing-ming. Application of Hadamard ECOC in multi-class problems based on SVM[J]. Acta Electronica Sinica,2008,36(1):122 – 126.(in Chinese)

作者简介

周进登 男,1984 年生于江西鹰潭.博士生.研究方向为模式识别,智能信息处理和多传感器数据融合.
E-mail:zhoujin198417@yahoo.com.cn

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,智能信息处理和机器学习等.
E-mail:afeu_w@yahoo.com.cn