DOI:10.16451/j.cnki.issn1003-6059.2014.02.013

多类类别不平衡学习算法: EasyEnsemble. M*

李倩倩[1] 刘胥影[1,2]

[1](东南大学 计算机科学与工程学院 计算机网络和信息集成教育部重点实验室 南京 211189)
[2](南京大学 计算机软件新技术国家重点实验室 南京 211189)

摘 要 随机欠采样方法忽略潜在有用的大类样本信息,在面对多类分类问题时更为突出. 文中提出多类类别不平衡学习算法: EasyEnsemble. M. 该算法通过多次针对大类样本随机采样,充分利用被随机欠采样方法忽略的潜在有用的大类样本,学习多个子分类器,利用混合的集成技术最终得到性能较优的强分类器. 实验结果表明,与常用的多类类别不平衡学习算法相比,EasyEnsemble. M 可有效提高分类器的 *G-mean* 值.

关键词 机器学习, 类别不平衡学习,欠采样, 集成
中图法分类号 TP 391

EasyEnsemble. M for Multiclass Imbalance Problem

LI Qian-Qian[1], LIU Xu-Ying[1,2]

[1](*Key Laboratory of Computer Network and Information Integration*, *Ministry of Education*, *School of Computer Science and Engineering*, *Southeast University*, *Nanjing* 211189)
[2](*National Key Laboratory for Novel Software Technology*, *Nanjing University*, *Nanjing* 211189)

ABSTRACT

The potential useful information in the majority class is ignored by stochastic under-sampling. When under-sampling is applied to multi-class imbalance problem, this situation becomes even worse. In this paper, EasyEnsemble. M for multi-class imbalance problem is proposed. The potential useful information contained in the majority classes which is ignored is explored by stochastic sampling the majority classes for multiple times. Then, sub-classifiers are learned and a strong classifier is obtained by using hybrid ensemble techniques. Experimental results show that EasyEnsemble. M is superior to other frequently used multi-class imbalance learning methods when *G-mean* is used as performance measure.

Key Words Machine Learning, Class-Imbalance Learning, Under-Sampling, Ensemble

* 国家自然科学基金青年基金项目(No. 61105046)、教育部高等学校博士学科点专项科研基金项目(No. 20110092120029)、南京大学软件新技术国家重点实验室开放课题项目(No. KFKT2011B01)资助
收稿日期: 2013 - 05 - 13
作者简介 李倩倩,女,1989 年生,硕士研究生,主要研究方向为机器学习、数据挖掘. E-mail: liqianqianseu@gmail.com.
刘胥影(通讯作者),女,1981 年生,博士,讲师,主要研究方向为机器学习、数据挖掘. E-mail: liuxy@seu.edu.cn.

1 引 言

类别不平衡问题是指在分类问题中各类别的样本数量分布不平衡,即某些类别的样本数远小于其他类别[1].通常称样本数少的类别为小类(Minority Class),样本数多的类别为大类(Majority Class).在这类问题中,小类样本是关注的重点,其错分代价相比大类样本较大.如在某些医学诊断中,训练样本中大多数人是正常人,只有少数是患者,而患者是诊断目标.若一个健康人被误诊,只会消耗较多的检查费用,但若一个癌症患者被漏诊,代价很大,甚至会危及病人生命[2].以最大化正确率为目标的传统分类算法易忽略小类样本的正确分类,因此在类别不平衡问题中常失效[3].实际上,以正确率作为评价标准,隐含假设每个样本的重要性是相同的(即被错分的代价是相同的),而在类别不平衡问题中,小类样本更重要,因此正确率不再适合作为评价标准.目前常用的衡量标准主要有 F-measure[4],AUC[5],G-mean[6] 等.好的类别不平衡学习方法希望在确保大类样本正确率的前提下,提高小类样本的识别率[7].

现实生活中存在大量的多类类别不平衡问题,如蛋白质折叠问题[8-10],焊缝缺陷检测问题[11] 等.该类问题相比于两类问题,概念复杂度更高,类别间的数据分布更多样化,甚至某些类别样本数稀少不足以有效地从中学习到相应概念,这使得从多类数据中进行类别不平衡学习更困难[12].

目前,多类类别不平衡问题的方法大致可分为分解法和非分解法.

分解法主要是将多类问题分解为多个两类问题,然后采用两类类别不平衡分类算法进行学习.一对多方法(One-vs-All, OVA)[13] 和一对一方法(One-vs-One, OVO)[14] 是两种常用的分解方法.OVA 每次任选一个类别作为正类,其余类别作为负类,然后进行学习.OVO 方法为任意两个类别对学习一个二分类器.这两种方法均采用投票方式对样本进行分类.文献[11] 采用 OVO 和 OVA 分别将蛋白质折叠问题分解成多个两类问题,然后采用基于规则的分类器来提高小类样本的正确率.Liao[11] 提出将 OVA 和改进的过采样和欠采样方法相结合解决焊缝缺陷检测问题.Alejo 等[15] 将类别不平衡比例因子引入到误差函数中,采用代价敏感学习方法实现两类不平衡子问题的学习.虽然分解方法简化学习问题,但其主要缺点如下[16]:1) 需学习的分类器个数增加;2) 会导致模糊区域,即类别不确定的区域,使得某些样本类别无法判定;3) OVA 会加剧小类样本的不平衡程度.

通过对多类的非分解学习算法改进从而实现非分解方法.如文献[16] 采用遗传算法为每个类寻找最优代价,然后使用代价敏感的 AdaBoost.M1[17] 学习.王硕等[7] 将 AdaBoost.NC[18] 扩展成多类分类器,结合随机过采样解决多类类别不平衡问题.Hones 等[19] 将针对类别不平衡问题设计的两类决策树分支标准 Hellinger Distance[20] 扩展到多类分类问题中,提出 Multiclass Hellinger Distance 决策树.

随机欠采样方法可有效解决两类的类别不平衡问题.Drummond 等[6] 指出,随机欠采样方法优于随机上采样方法.该方法使得学习算法集中于不同的类别分布的同时,提高训练样本中小类样本的比例,从而可获得较鲁棒的分类器[10].但该方法的缺点也很明显,它忽略很多潜在有用的大类样本信息.

EasyEnsemble[21] 针对两类类别不平衡问题,利用集成技术充分挖掘随机欠采样方法所忽略的潜在有用的大类样本信息.该方法为本文工作提供重要的研究基础.

多类问题中,各类别样本只占训练集的一部分,且概念复杂度更高,直接使用欠采样方法会减少本来就不多的样本集,学习效果不理想[10].尤其是当小类样本数非常少的情况,随机欠采样方法忽略潜在有用信息的缺点会带来更严重的问题.为使随机欠采样方法适用于解决多类类别不平衡问题,本文提出 EasyEnsemble.M.该算法利用混合的集成技术,充分挖掘大类样本中潜在有用的信息,避免随机欠采样方法在解决多类类别不平衡问题时的不足.

2 EasyEnsemble.M

集成学习方法在机器学习领域占有重要地位.在集成学习中弱分类器的多样性是提高泛化性的重要因素,对于解决类别不平衡分类问题有重要作用[22].为使随机欠采样方法适用于多类的类别不平衡问题,本文提出 EasyEnsemble.M,利用集成技术充分挖掘被随机欠采样方法忽略的潜在有用信息.其基本思想如下:1) 采用随机欠采样方法减少各大类中的样本得到平衡的数据集,学习一个多类的 AdaBoost.M1 分类器;2) 多次独立地重复以上采样学习过程,得到多个 AdaBoost.M1 分类器;3) 将每个 AdaBoost.M1 分类器的所有弱分类集成,对测试样本进行预测.

为在下文中不失一般性,将样本数目最少的类

别称为小类,其余类别称为大类.

给定 K 类的多类类别不平衡分类问题,类别按样本数目从少到多排序: $C_1, C_2, \cdots, C_K$. C_1 类样本数目最少,记做 P. 在第 t 轮中,从其余大类中独立同分布的随机下采样得到若干子集 N_{it}, $|N_{it}| = |P|$. 每个大类的子集和小类一起组成平衡的多类训练集,在此训练集上,用 AdaBoost. M1[17] 学习多类分类器 H_t. 以上过程独立重复 T 次,得到 T 个多类 AdaBoost. M1 分类器$\{H_t\}_{t=1}^{T}$. 最终将所有 H_t 的所有弱分类器$\{h_{dt}\}$, $d = 1, 2, \cdots Si$; $t = 1, 2, \cdots, T$,集成.

AdaBoost. M1 是两类集成学习算法 AdaBoost[17] 的多类扩展. 两者主要区别在于两类问题中的误差计算 $|h_t(x_i) - y_i|$ 替换为 $\|h_t(x_i) \neq y_i\|$,其中,如果 π 成立,则 $\|\pi\|$ 为 1,否则为 0. 对测试样本的类别即为使得所有弱分类器的加权投票之和最大的类别. AdaBoost. M1 步骤如下所示.

算法 1　AdaBoost. M1

输入　训练集$\{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$,
$y_i \in Y = \{1, 2, \cdots, K\}$,迭代次数 T

输出　$H(x) = \arg\max\limits_{y \in Y} \sum\limits_{t: h_t(x) = y} \log_2 \dfrac{1}{\beta_t}$

初始化权重

$$D_1(x_i) = \frac{1}{m}$$

for $t = 1, 2, \cdots, T$

根据 D_t 学习弱分类器 h_t, $h_t: X \to Y$,

计算弱分类器 h_t 的错误率

$$\varepsilon_t = \sum_{i: h_t(x_i) \neq y_i} D_t(x_i)$$

若 $\varepsilon_t > \dfrac{1}{2}$,则 $T = t - 1$,结束 for 循环

计算

$$\beta_t = \frac{\varepsilon_t}{1 - \varepsilon_t}$$

更新权值,计算

$$D_{t+1}(x_i) = \begin{cases} \dfrac{D_t(x_i)}{Z_t}\beta_t, & h_t(x_i) = y_i \\ \dfrac{D_t(x_i)}{Z_t}, & h_t(x_i) \neq y_i \end{cases}$$

其中 Z_t 是归一化常数

end

值得注意的是,EasyEnsemble. M 将每个多类的 AdaBoost. M1 分类器的所有弱分类器进行集成,即 $H(x) = \arg\max\limits_{y \in Y} \sum\limits_{t=1}^{T} \sum\limits_{d=1}^{si} \left(\log_2 \dfrac{1}{\beta_{td}}\right) \|h_{td}(x) = y\|$,而不是将 AdaBoost. M1 分类器本身的输出集成,即

$$H(x) = \arg\max_{y \in Y} \sum_{t=1}^{T} \left\| \arg\max_{y' \in Y} \left(\sum_{d=1}^{si} \left(\log_2 \frac{1}{\beta_{td}}\right) \cdot \|h_{td}(x) = y'\| \right) = y \right\|.$$

通过上述方式,在最终分类器中摒弃传统的投票方式预测样本类别. 每个子问题包含不同的大类样本信息,利用这些样本信息得到的各分类器更具多样性,集成之后形成分类器 H 性能更优. EasyEnsemble. M 如下所示.

算法 2　EasyEnsemble. M

输入　训练集$\{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$,
$y_i \in Y = \{1, 2, \cdots, K\}$.
样本数最少的类别是 C_1,其样本集用 P 表示,
其余类别样本集用 $N_i (2 \leqslant i \leqslant K)$,
迭代次数 T,
在第 i 轮迭代中,AdaBoost. M1 的迭代次数为 si

输出　$H(x) = \arg\max\limits_{y \in Y} \sum\limits_{t=1}^{T} \sum\limits_{d=1}^{si} \left(\log_2 \dfrac{1}{\beta_{td}}\right) \|h_{td}(x) = y\|$

for $t = 1, 2, \cdots, T$

$D_t = P$

for $i = 2, 3, \cdots, K$

从 N_i 类中随机下采样得到样本子集 N_{it},

$$|N_{it}| = |P|$$

$D_t = D_t \cup N_{it}$

end

用 D_t 训练 AdaBoost. M1 分类器 H_t,其中,h_{td} 是 H_t 的第 d 轮弱分类器

$$H_t(x) = \arg\max_{y \in Y} \sum_{d=1}^{si} \left(\log_2 \frac{1}{\beta_{td}}\right) \|h_{td}(x) = y\|$$

end

EasyEnsemble. M 继承 EasyEnsemble 的特点,将原问题转化为 T 个平衡的子问题. 不同之处在于转化和分解方法,EasyEnsemble. M 学习到的每个 AdaBoost. M1 分类器,可对样本进行多类分类.

3　实验及结果分析

3.1　实验设置

实验采用 G-mean 作为性能衡量标准. 定义 n_i 为属于类别 C_i 的样本总数,$cm(i, j)$ 为类别为 C_i 的样本被判别为类别 C_j 的个数,则类别 C_i 的查全率和查准率可定义[18] 为

$$P_i = \frac{cm(i, i)}{\sum\limits_{j=1}^{k} cm(j, i)}, \quad R_i = \frac{cm(i, i)}{n_i},$$

G-mean 定义为

$$G\text{-}mean = \left(\prod_{i=1}^{k} R_i\right)^{\frac{1}{k}}.$$

本文使用 8 个 UCI 数据集进行实验,见表 1.

表 1 数据集信息

Table 1 Information of datasets

数据集	类别	样本个数	属性	最大不平衡比例/%
abalone	5	595 (6/14/58/126/391)	1(离散属性) 7(连续属性)	65.2
balance	3	625 (49/288/288)	4(离散属性)	5.9
Ecoli	5	327 (20/35/52/77/143)	1(离散属性) 6(连续属性)	7.2
glass	6	214 (9/13/17/29/70/76)	9(连续属性)	8.4
segment	7	669 (10/20/20/41/83/165/330)	19(连续属性)	33
satimage	5	775 (25/50/100/200/400)	19(连续属性)	16
vowel	11	530 (8/16/24/32/40/48/56/64/72/80/90)	10(连续属性)	11.3
yeast	10	1484 (5/20/30/35/44/51/163/244/429/463)	8(连续属性)	92.6

实验采用 10 次 5 折交叉验证. 交叉验证重复 10 次,最终结果记录 10 次的平均值.

本文采用的对比算法及参数设置如下.

1) 分类回归树(Classification and Regression Trees, CART)[23]. 训练时利用全部的训练样本.

2) AdaBoost. M1(AdaM1)[17]. AdaBoost. M1 使用全部的训练集. CART 作为其弱分类器,迭代 40 次.

3) Over-sampling + AdaBoost. M1(OverAdaM1). 除最大类样本集外,所有类别样本采用随机过采样方法有放回的从该类别中随机抽取样本构成新的该类样本集,与最大类样本集构成新的训练集参与 AdaBoost. M1 分类器的学习. CART 作为其弱分类器,迭代 40 次.

4) Under-sampling + AdaBoost. M1(UnderAdaM1). 除最小类 C_1 样本集 S_1 之外,从类别 $C_i(2\leqslant i\leqslant K)$ 样本集 S_i 中随机抽取一个样本子集 S_i', $|S_i'| = |S_1|$. AdaBoost. M1 采用 $S_1 \cup S_i'(i=2,3,\cdots,K)$ 作为训练集进行学习. 以 CART 为弱分类,迭代 40 次.

5) SMOTE + AdaBoost. M1(SMOTEAdaM1). 先采用少数样本合成过采样技术(Synthetic Minority Over-Sampling Technique, SMOTE)为类别 $C_i(i=1,2,\cdots,K-1)$ 合成新的样本集 S_i',使得

$$|S_i| + |S_i'| = |S_K|.$$

新的训练集由新合成的样本集合和原始样本集合构成进行 AdaBoost. M1 的学习. 弱分类器采用 CART,迭代 40 次. SMOTE 中 KNN 采用的是 5(数据集 yeast 最小类样本数为 5,所以该数据集采用是 3,数据集 abalone 采用的是 4).

6) Over-sampling + AdaBoost. NC(OVNC9)[7]. 文献[7]将多类的 AdaBoost. NC 结合随机过采样解决多类类别不平衡问题. AdaBoost. NC[18]是在 AdaBoost 的基础上结合负相关学习理论实现的多类分类算法. 实验中参数 λ 的设置与文献[7]中一致,取 $\lambda=9$. 弱分类器采用 CART,迭代 40 次.

3.2 实验结果

表 2 中记录 EasyEnsemble. M($T=4$)和所有对比算法的 *G-mean* 值,粗体表示最优方法的结果. 从表中可看出如下结论.

1) EasyEnsemble. M 在其中 5 个数据集上取得最高的 *G-mean* 值.

2) EasyEnsemble. M 优于将随机欠采样方法直接应用到多类类别不平衡问题的 UnderAdaM1. 这说明 EasyEnsemble. M 有效利用被随机欠采样忽略的潜在有用的大类样本信息.

3) 在较易分的数据集上,如 segment,AdaM1 的 *G-mean* 值高于除 SOMTEAdaM1 外的所有的类别不平衡分类算法. 而在某些难分类的数据集上,大部分

表 2 各算法 10 次 5 折交叉验证 *G-mean* 值对比

Table 2 *G-mean* values comparison of all algorithms based on ten-times 5-fold cross validation

数据集	CART	AdaM1	OverAdaM1	UnderAdaM1	SOMTEAdaM1	OVNC9	Ensemble. M
abalone	0.000±0.000	0.000±0.000	0.000±0.000	0.002±0.005	0.001±0.004	0.033±0.014	**0.087±0.022**
balance	0.067±0.065	0.022±0.018	0.020±0.011	0.009±0.011	0.013±0.014	0.398±0.027	**0.688±0.020**
ecoli	0.699±0.040	0.790±0.014	0.792±0.010	0.791±0.008	0.795±0.015	0.717±0.013	**0.812±0.007**
glass	0.349±0.127	0.556±0.095	0.565±0.137	0.569±0.104	0.620±0.111	0.488±0.074	**0.637±0.028**
segment	0.894±0.016	**0.930±0.016**	0.915±0.014	0.921±0.026	**0.930±0.020**	0.855±0.024	0.899±0.006
satimage	0.332±0.112	0.121±0.029	0.093±0.032	0.128±0.047	0.445±0.074	0.456±0.063	**0.761±0.020**
vowel	0.369±0.086	0.635±0.103	0.655±0.064	0.658±0.104	**0.753±0.089**	0.402±0.068	0.425±0.062
yeast	0.000±0.000	0.062±0.030	0.026±0.015	0.061±0.031	**0.143±0.062**	0.129±0.038	0.085±0.020
平均值	0.339±0.326	0.390±0.380	0.383±0.387	0.392±0.382	0.463±0.370	0.435±0.272	**0.549±0.318**

的类别不平衡分类算法的 *G-mean* 值高于 AdaM1，包括 OverAdaM1，UnderAdaM1，SMOTEAdaM1，EasyEnsemble. M.

4) 对于小类样本数目绝对稀少，且不平衡程度较高的数据集，如 abalone、segment 和 yeast，EasyEnsemble. M 在后 2 个数据集上效果不理想. 这大概是由于随机欠采样方法固有缺陷造成的.

为考察不同 T 值对实验结果的影响，表 3 记录 EasyEnsemble. M 在设定不同迭代次数 T(即大类样本子集个数) 后对于数据集 *G-mean* 值的影响. 图 1 中显示数据集 balance、glass 和 vowel 在取不同 T 值时 *G-mean* 值的变化情况. T 分别取 4，6，8，10，由表 3 和图 1 可看出，T 值对于 *G-mean* 值的影响较大，不同数据集有适用于其本身的 T 值. 同时可看出，T 值的选取不是越大越好，大部分数据集在 $T=8$ 时可取得最高的 *G-mean* 值，balance 和 satimage 的最佳 T 值分别为 4 和 6，在 $T=10$ 处，*G-mean* 值下降幅度较大.

表 3　迭代次数对实验结果的影响

Table 3　Influence of the number of iterations on results

数据集	$T=4$	$T=6$	$T=8$	$T=10$
abalone	0.087 (±0.022)	**0.089 (±0.039)**	**0.089 (±0.036)**	0.077 (±0.022)
balance	**0.688 (±0.020)**	0.601 (±0.021)	0.611 (±0.019)	0.586 (±0.023)
ecoli	0.812 (±0.007)	0.814 (±0.007)	**0.817 (±0.009)**	0.810 (±0.008)
glass	0.637 (±0.028)	0.633 (±0.043)	**0.653 (±0.030)**	0.622 (±0.038)
segment	0.899 (±0.006)	0.905 (±0.005)	**0.908 (±0.006)**	0.900 (±0.006)
satimage	0.761 (±0.020)	**0.765 (±0.011)**	0.759 (±0.026)	0.758 (±0.017)
vowel	0.425 (±0.062)	0.437 (±0.084)	**0.470 (±0.082)**	0.410 (±0.077)
yeast	0.085 (±0.020)	0.103 (±0.032)	**0.110 (±0.015)**	0.084 (±0.013)

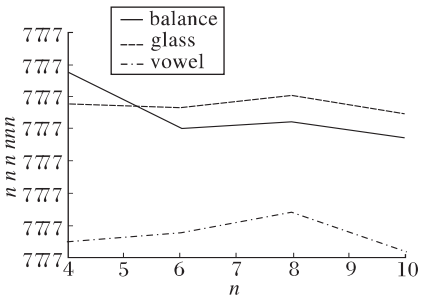

图 1　不同数据集上迭代次数 T 对 *G-mean* 值的影响

Fig. 1　Influence of the number of iterations on *G-mean* results on different datasets

本文通过实验衡量 Easy Ensemble. M 的 ROC 曲线下面积的多类扩展(Multiclass Area under the ROC Curve，MAUC) 值[24]，从整体来看，本文算法与其他算法的 *MAUC* 值相差不大，所以在此没有列出其结果.

综上所述，EasyEnsemble. M 有效利用被随机欠采样忽略的潜在有用的大类样本信息，可有效解决多类类别不平衡问题，同时针对不同数据集选取合适的 T 值可达到更好的实验结果.

4　结 束 语

随机欠采样方法可有效解决两类类别不平衡问题，并提高学习效率，但该方法忽略潜在有用的大类样本信息. 由于多类分类问题本身的特点，直接应用会使该方法的缺点加倍放大. 本文提出多类类别不平衡学习算法: EasyEnsemble. M，该算法利用混合的集成技术充分挖掘被随机欠采样方法忽略的潜在有用的大类样本信息. 实验结果表明，在以 G-mean 为性能评价准则时，EasyEnsemble. M 优于将随机欠采样方法直接应用于多类类别不平衡问题的简单扩展，及其他 5 个常用的多类类别不平衡学习方法. 本文实验表明不同 T 值对实验结果的影响在不同数据集上是不同的，下一步的工作可考虑如何确定更有效的子集个数，及如何获取更优的样本子集. 此外，针对样本绝对稀少的问题，随机欠采样有其固有缺点，应设计更有效的采样方法解决此类问题.

参　考　文　献

[1] Ye Z F, Wen Y M, Lü B L. A Survey of Imbalanced Pattern Classification Problems. CAAI Trans on Intelligent Systems, 2009, 4(2): 148 - 156 (in Chinese)
(叶志飞，文益民，吕宝粮. 不平衡分类问题研究综述. 智能系统学报，2009，4(2)：148 - 156)

[2] Dong Y J. Random-SMOTE Method for Imbalanced Data Sets. Master Dissertation. Dalian, China: Dalian University of Technology, 2009 (in Chinese)
(董燕杰. 不平衡数据集分类的 Random-SMOTE 方法研究. 硕士学位论文. 大连: 大连理工大学，2009)

[3] Chawla N V. Data Mining for Imbalanced Datasets: An Overview [EB/OL]. [2013 - 03 - 10]. http://link.springer.com/chapter/10.1007%2F0 - 387 - 25465 - X_40#page - 1

[4] Davenport M. Introduction to Modern Information Retrieval. Journal of the Medical Library Association, 2012. DOI: 10.3163/1536 - 5050.100

[5] Bradley A P. The Use of the Area under the ROC Curve in the Eval-

uation of Machine Learning Algorithms. Pattern Recognition, 1997, 30(7): 1145 - 1159

[6] Drummond C, Holte R C. C4.5, Class Imbalance, and Cost Sensitivity: Why Under-Sampling Beats Over-Sampling [EB/OL]. [2013-03-10]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.68.6858&repl&type=pdf

[7] Wang S, Yao X. Multiclass Imbalance Problems: Analysis and Potential Solutions. IEEE Trans on Systems, Man and Cybernetics, 2012, 42(4): 1119 - 1130

[8] Zhao X M, Li X, Chen L N, *et al*. Protein Classification with Imbalanced Data. Proteins: Structure, Function and Bioinformatics, 2008, 70(4): 1125 - 1132

[9] Chen K, Lü B L, Kwok J T. Efficient Classification of Multi-label and Imbalanced Data Using Min-Max Modular Classifiers // Proc of the International Joint Conference on Neural Networks. Vancouver, Canada, 2006: 1170 - 1775

[10] Tan A C, Gilbert D, Deville Y. Multi-class Protein Fold Classification Using a New Ensemble Machine Learning Approach. Genome Information, 2003, 14: 206 - 217

[11] Liao T W. Classification of Weld Flaws with Imbalanced Class Data. Expert Systems with Applications, 2008, 35(3): 1041 - 1052

[12] Zhou Z H, Liu X Y. Training Cost-Sensitive Neural Network with Methods Addressing the Class Imbalance Problem. IEEE Trans on Knowledge Data Engineering, 2006, 18(1): 63 - 77

[13] Rifkin R, Klautau A. In Defense of One-vs-All Classification. The Journal of Machine Learning Research, 2004, 5: 101 - 141

[14] Hastie T, Tibshirani R. Classification by Pairwise Coupling. The Annals of Statistics, 1998, 26(2): 451 - 471

[15] Alejo R, Sotoca J, Valdovinos R, *et al*. The Multi-class Imbalance Problem: Cost Functions with Modular and Non-Modular Neural Networks // Proc of the 6th International Symposium on Neural Networks. Berlin, Germany: Springer, 2009: 421 - 431

[16] Sun Y, Kamel M S, Wang Y. Boosting for Learning Multiple Classes with Imbalanced Class Distribution // Proc of the 6th IEEE Industrial Conference on Data Ming. Hong Kong, China, 2006: 592 - 602

[17] Freund Y, Schapire R. A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting. Journal of Computer and System Sciences, 1997, 55(1): 119 - 139

[18] Wang S, Chen H H, Yao X. Negative Correlation Learning for Classification Ensembles // Proc of the International Joint Conference on Neural Networks. Barcelona, Spain, 2010: 1 - 8

[19] Hoens T, Qian Q, Chawla N, *et al*. Building Decision Trees for the Multi-class Imbalance Problem. Lecture Notes in Computer Science, 2012. DOI: 10.1007/978 - 3 - 642 - 30217 - 6_11

[20] Cieslak D A, Chawla N V. Learning Decision Trees for Unbalanced Data // Proc of the European Conference on Machine Learning and Knowledge Discovery in Databases. Antwerp, Belgium, 2008, I: 241 - 256

[21] Liu X Y, Wu J X, Zhou Z H. Exploratory Undersampling for Class-Imbalance Learning. IEEE Trans on Systems, Man and Cybernetics, 2009, 39(2): 539 - 550

[22] Wang S, Yao X. Theoretical Study of the Relationship between Diversity and Single-Class Measures for Class Imbalance Learning // Proc of the IEEE International Conference on Data Mining. Miami, USA, 2009: 76 - 81

[23] Breiman L, Friedman J, Stone C J, *et al*. Classification and Regression Trees. London, UK: Chapman & Hall, 1984

[24] Hand D J, Till R J. A Simple Generalisation of the Area under the ROC Curve for Multiple Class Classification Problems. Machine Learning, 2001, 45(2): 171 - 186