doi:10.11835/j.issn.1000-582X.2022.05.012

一种处理不均衡多分类问题的特征选择集成方法

宿 晨,徐 华,崔 鑫,王玲娣
(江南大学 物联网工程学院,江苏 无锡 214122)

摘要:为解决不均衡多分类问题,提出一种特征选择和 AdaBoost 的集成方法。首先,数据进行预处理。利用 WSPSO 算法进行特征选择,根据特征重要性选取初始粒子构建初始种群,使得算法初期就可以沿着正确的搜索方向开展,减少不相关特征的影响。其次,利用 AdaBoost 算法对于样本权重较敏感的特点,增强对小类样本的关注度。并且利用 AUCarea 作为评价标准,相对于其他评价标准,AUCarea 具有可视化的优点且对较差 AUC 更加敏感。最后,与其他几种不均衡分类算法在不平衡数据集上进行对比,结果证明该算法可有效处理不均衡多分类问题。

关键词:不平衡数据;集成学习;AdaBoost;特征选择;多分类
中图分类号:TP181 **文献标志码**:A **文章编号**:1000-582X(2022)05-125-10

An ensemble learning algorithm for feature selection based on solution to multi-class imbalance data classification

SU Chen, *XU Hua*, *CUI Xin*, *WANG Lingdi*
(School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, Jiangsu, P. R. China)

Abstract: In order to solve the problem of unbalanced multi-classification, a feature selection and AdaBoost integration method is proposed. First, the data is preprocessed. The WSPSO algorithm is used to select features, and the initial population is constructed according to the importance of the feature. The initial algorithm can be carried out along the correct search direction to reduce the influence of incoherent features. Secondly, the AdaBoost algorithm is more sensitive to sample weights, and the attention to small samples is enhanced. And using AUCare is used, as the evaluation standard, because compared with other evaluation criteria, AUCare has the advantage of visualization and is more sensitive to poor AUC. Finally, compared with several other unbalanced classification algorithms on the unbalanced data set, the algorithm can effectively deal with the unbalanced multi-classification problem.

Keywords: imbalanced data; ensemble learning; AdaBoost; feature selection; multi-class classification

分类问题是数据挖掘中的难点。绝大多数分类算法只是在平衡数据集分类效果显著,而在不均衡数据集上分类效果欠佳。但现实生活中的分类问题往往是类别不均衡的,例如银行欺诈的检测、垃圾邮件

收稿日期:2020-12-25
基金项目:教育部-新华三集团"云数融合"基金资助项目(2017A13055)。
Supported by Ministry of Education Xinhua Three Group "Cloud Data Integration" Fund Project (2017A13055).
作者简介:宿晨(1993—),男,硕士研究生,主要从事机器学习、数据挖掘方向研究;徐华(1978—),女,副教授,博士,主要从事计算智能、车间调度、大数据方向研究,(E-mail)joanxh2003@163.com。

的检测、车辆识别、疾病诊断等。不均衡数据分类问题已经成为机器学习、数据挖掘等领域的重要研究方向之一[1-2]。在实际的生活中,不均衡分类问题大多数是多分类问题,因此更具有研究意义。而多数类的不均衡分类问题与二分类的分类问题相比,多数类对于分类模型要求更高,获取少数类的代价也更大,类间数据的分布也更多样化,也更难被分类。针对不均衡数据分类的研究主要集中在数据层面与算法层面的改进研究。数据层面的改进主要集中在对数据集的改进,增加少数类数据或减少多数类数据,使得原本的数据集相对均衡,主要的改进方法是过采样与欠采样。早期,Chawla 等[3-4],提出了一种 SMOTEBoost(synthetic minority over-sampling technique and boost)的方法,将 SMOTE 采样算法与集成算法 Boost 相结合,加强了对小类样本的关注度;武森等[5]将聚类算法运用到采样中,先利用聚类欠采样方法将数据集均衡化,然后利用 AdaBoost 算法对新生成的数据集进行分类操作;2013 年,Krawczyk 等[6]使用 PUSBE(pruned under-sampling balanced ensemble)方法,该方法有效运用了特征选择技术;2014 年,Krawczyk 等[7]又提出了 CS-MCS(cost-sensitive multiple classifier systems)集成方法,运用随机欠采样结合遗传算法相结合的方式。在低维数据集上效果明显,但是高维数据集上效果欠佳。文献[8]指出采样方法虽然可以提升小类样本的识别率,但是容易引入噪声,丢失有用信息,分类器对小类样本过分的关注也易使得算法陷入局部最优。对此,TAO 等[9]提出了一种新的过采样技术,该技术使用实值否定选择(RNS)过程来生成人工少数类数据,而无需实际的少数类数据。生成的少数类数据(如果有的话)会与实际的少数类数据一起使用,并与多数类数据相结合,作为二分类学习方法的输入,并且在实验中证明了其有效性。

从算法的角度来看,改变概率密度,单类学习分类,集成学习,代价敏感学习,核方法等五种主要方法来解决数据分类不平衡问题[10]。国际机器学习界的权威 Dietterich 已经将集成学习列为机器学习 4 大研究之首[11]。TAO 等[12]提出了一种新的基于自适应权重的支持向量机成本敏感集成方法,用于不平衡数据分类,还创新性的提出了一种自适应的顺序错误分类权重确定方法。该方法可以基于在提升过程中先前获得的分类器,在每次迭代时自适应地考虑少数实例对 SVM 分类器的不同贡献,这可以使其产生不同的分类器,从而提高泛化性能。随后,Tao 等[13]又提出了一种新的基于亲和度和类别概率的模糊支持向量机技术(ACFSVM)。多数类样本的亲和力是根据支持向量描述域(SVDD)模型计算的,该模型仅由给定的多数类训练样本在内核空间中进行训练,类似于 FSVM 学习所使用的模型。针对噪声样本的处理,Tao 等[13]采用核 k 最临近法来确定与以前相同的核空间中多数类别样本的类别概率。具有较低分类概率的样本更有可能是噪音,并且通过将相似度和分类概率结合起来构成的低隶属度,减少了噪声样本的影响。张苗燕等[14]结合细菌觅食算法的思想,提出了一种新的算法 AdAdaboost,并对加权系数进行了改进,全局优化最佳弱分类器,改善了 AdaBoost 算法误检率的同时得到了较好的检测性能;Guo 等[15]将 AdaBoost.M1 算法与特征选择结合起来,提出了一种新的集成方法 BAK(BPSO-AdaBoost-KNN),使用 KNN 作为基分类器,但 KNN 的缺点是不能直接处理带权数据,需要借助 re-samplingd 的方法转化数据集后使用,而且 AdaBoost.M1 针对于基分类器的要求过于严苛,错误率不能超过 50%。对此,将胡旺[16]等提出的 SPSO(simple particle swarm optimization)算法进行改进,并与 Zhu 等[17]提出的 SAMME.R 版本的 AdaBoost 算法相结合,提出了 WSPSO-SAMME.R-DT 算法,用以解决不平衡多分类问题。与 AdaBoost.M1 算法所不同的,SAMME.R 使用决策树作为基分类器,避免在训练样本上花费时间,降低对基分类器的要求。为了降低基分类器的相关性,引入了随机化的方法。使用 AUCarea 作为性能度量指标,并将其作为适应度值,优化特征选择。提升了小类样本的识别率。

1　算法介绍

1.1　SAMME.R AdaBoost 算法

AdaBoost 算法是一个迭代过程,弱分类器的生成是串行的。在 AdaBoost 的训练过程中,分类器的重心将转移到那些更难分类的样本上,即多次错误分类的样本。随后的训练也会偏重于这些样本,这是通过在算法运行期间为训练样本分配权重来实现的。样本权重最初都是一致的,后续过程中每轮都会对样本权重进行更新,最终得到一组弱分类器,将所有弱分类器加权组合成一个强分类器。

AdaBoost 算法适用于二分类问题, AdaBoost.M1 可用于解决多分类问题。但是 AdaBoost.M1 的前提条件是基分类器的错误率小于 50%,这一要求过于严格,易导致训练失败。针对以上不足,笔者选择 Zhu 等[17]提出的 SAMME.R 版本的 AdaBoost 算法,降低了对基分类器过于严苛的要求,仅比随机猜测略好即可。同时,使用分类器的类别估计概率值来对样本权重进行更新。

在该算法中,获得加权类概率估计的公式为:

$$p(t)_k(x)=\mathrm{Prob}_w(h=k\mid x),k=1,\cdots,K, \tag{1}$$

其中:t 为迭代次数,k 为类标签,Prob 函数是返回区域中的数值落在指定区间内的对应概率。获得加权类概率估计后,利用拉格朗日定理对称约束优化得到 $h(t)_k(x)$

$$h(t)_k(x)=(K-1)(\log\ p(t)_k(x)-\frac{1}{k}\sum_{k'=1}^{K}\log\ p(t)_{k'}(x)), \tag{2}$$

更新样本权重 w_i:

$$w_i=w_i\cdot\exp\left(-\frac{K-1}{K}\boldsymbol{y}T_i\log\ p(t)(x_i)\right),i=1,\cdots,m, \tag{3}$$

其中:$\boldsymbol{y}=(\boldsymbol{y}_1,\cdots,\boldsymbol{y}_m)^T$

$$y_k=\begin{cases}1, & \text{if}\quad \boldsymbol{y}=k\\ -\dfrac{1}{K-1}, & \text{if}\quad \boldsymbol{y}\neq k\end{cases}$$

1.2　性能度量指标 AUCarea

对于不平衡二分类问题来说,经常使用 ROC 曲线来度量分类中的不平衡性,ROC 是接受者操作特性曲线(receiver operating characteristic),利用 ROC 曲线下的面积(area under the curve)作为算法的评价标准,理想中分类器的 AUC 为 1.0,随机猜测的分类器 AUC 为 0.5。

AUC 评价标准无法直接应用与多分类问题,需要对其进行拓展。最常用的 2 种扩展方法分为[18]:1)一对一方法;2)一对多方法。为了更加清晰的对比这 2 种方法,令 $Y=\{y_1,y_2,\cdots,y_k\}$,Y 表示的是数据的类标签的集合。在一对一的方法中,计算所有类的两两组合$(y_i,y_j)(i\neq j)$ 的 AUC 值。一对多的方法中,先定义成二分类问题,令 $y_i\in Y$,属于 y_i 的样本定义为正类,剩余的样本为负类,然后计算定义后的 AUC 值。由此将会得到一组 AUC 值$\{r_1,r_2,\cdots,r_n\}$,最后取平均值,记作 avgAUC,作为性能度量值使用。以上 2 种方法都可简单的实现,但是无法做到可视化。因为当多个 AUC 都变化时,avgAUC 的值可能没有任何变化。例如,当 r_i 变为 $r_i+\sigma$,r_j 变为 $r_j-\sigma$,$(i,j)\in\{1,2,\cdots,n\}$), 最终其 avgAUC 的值没有改变,也无法进行分类模型调整的评价。

采用 Hand DJ[19]提出的一种具有可视化优点的度量指标方法 AUCarea。AUCarea 会将所有的 AUC 的值在极坐标上绘制出来。如图 1 所示,黄色虚线三角形代表的就一个是三分类的 AUCarea 极坐标图示,黄色虚线所覆盖的面积就是最终度量值。AUCarea 的计算公式如下

$$\mathrm{AUCarea}=\frac{1}{2}\sin\left(\frac{2\pi}{n}\right)\left(\left(\sum_{i=1}^{n-1}r_i\times r_j\right)+(r_n\times r_1)\right), \tag{4}$$

其中:n 为 AUC 的总数;r 为每对类组合$(y_i,y_j)(i\neq j)$的 AUC 的值。

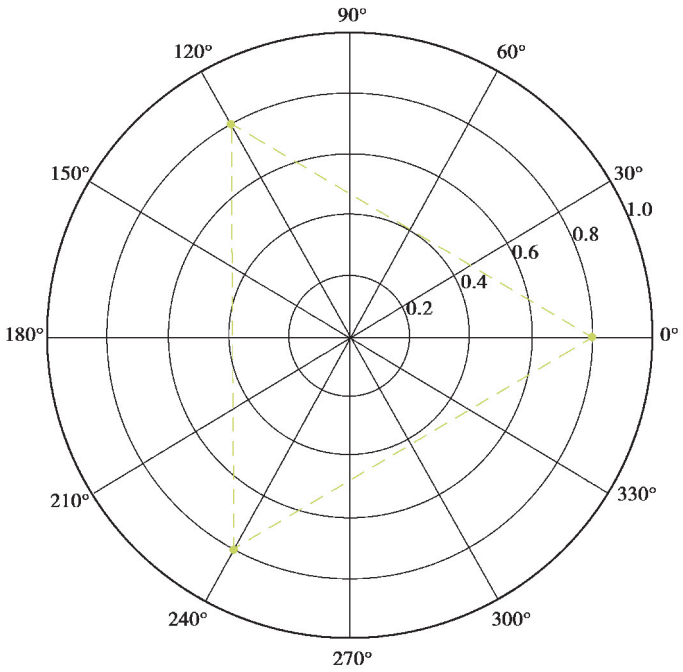


图1 三分类的AUCarea极坐标图示

Fig. 1 Three-category AUCare polar plot

当所有AUC的值为1时，就达到了理想中的最优状态，即AUCarea的最大值，如公式(5)所示

$$\mathrm{AUCarea\_max}=\left(\frac{n}{2}\right)\sin\left(\frac{2\pi}{n}\right), \tag{5}$$

计算归一化的为公式(6)

$$\mathrm{AUCarea\_nor}=\frac{\mathrm{AUCarea}}{\mathrm{AUCarea\_max}}=\frac{\left(\sum_{i=1}^{n-1}r_i\times r_{i+1}\right)+(r_n+r_1)}{n}, \tag{6}$$

使用归一化公式(6)所得值记为AUCarea，AUCarea除了可视化的优点之外，也对单个差的AUC较为敏感。

2 基于WSPSO特征选择方法

虽然AdaBoost算法可通过增加小样本的权重来增强对小样本的关注，但它仍然使用正确率作为优化目标，并且容易引起过拟合。因此，将特征选择方法与SAMME.R AdaBoost算法结合。SAMME.R算法中加入特征选择是基于以下考虑：去除不相干特征，减少时间与空间的浪费，加强对特征和特征值之间的联系，从而更好的进行分类[20]。特征选择算法主要有三类：嵌入式(embedded)、过滤式(filter)和封装式(wrapper)。嵌入式算法的思路是学习器自身自动进行选择，虽然效果较好，但是对于参数的设置需要较高的知识背景；过滤式算法的思路是先对各特征的相关性或发散性进行评估排序，根据设置的阈值来选择。但是对特征之间的相关性难以评估，会造成部分有用信息的遗失；封装式算法的思路是利用学习算法来评估特征的优劣，相对于嵌入式算法与过滤式算法，虽然需要巨大的搜索空间，执行时间稍长，但不需要过多的背景知识，可直接面向算法优化，并且特征间的组合效应也得到了充分的挖掘。综上，选择封装式算法来进行特征选择。

粒子群优化(PSO, particle swarm optimization)[21]算法，具有易实现、结构简单、没有复杂变异交叉操作的优点，可运用于特征选择优化问题在。文献[16]在证明PSO进化过程与粒子速度无关后提出了简化版粒子群优化(SPSO, simple particle swarm optimization)算法，去掉了速度选项，SPSO的进化公式为

$$xt+1_{id}=\omega xt_{id}+c_1r_1(p_{id}-xt_{id})+c_2r_2(p_{gd}-xt_{id}), \tag{7}$$

其中：$xt+1_{id}$表示的是第t代第i个粒子的第d维分量；ω是惯性权重因子；c_1和c_2是学习因子常数；r_1和

r_2 是随机数,服从 $U(0,1)$;p_{id} 表示第 i 个粒子个体极值的第 d 维,而 p_{gd} 表示全局最优解的第 d 维分量。

在粒子群算法中,惯性权重是重要的参数之一。其主要功能是平衡整个粒子群的全局搜索能力和局部搜索能力,从而显著的提高算法的整体收敛速度。而在标准 SPSO 算法中,ω 是固定的数值,无法改变。当惯性权重较小时,如果最优解在初始搜索空间中,则粒子群算法可以快速找到全局最优解,反之则无法正确找到。而惯性权值较大时,粒子群算法更像是全局搜索算法,总会探索新的区域。这意味着需要更多的迭代来寻找全局最优,并且更有可能在找不到最优解同时算法的时间复杂度也会增加。因此,ω 应该在算法的初期设置为较大值,在算法的后期设置为较小值。这样设置的优点在于:初始阶段的全局寻优能力会得到增强,有利于避免局部最优;而在算法的后期,可增强算法在局部的搜索能力,同时提高算法收敛速度。因此,借鉴文献[22]运用的一种线性递减动态获取惯性权重 ω 的方法,即

$$\omega = \frac{(\omega_{\text{ini}} - \omega_{\text{end}})(T - t)}{T} + \omega_{\text{end}}, \tag{8}$$

其中的参数取值为:$\omega_{\text{ini}}=0.9$;$\omega_{\text{end}}=0.4$ 时效果较好。t =当前迭代次数,T=最大迭代次数。

特性选择可看成 0～1 组合优化问题,Kennedy[23] 等最早提出了二进制粒子群优化(BPSO, binary particle swarm optimization)算法将 PSO 算法扩展到了离散二进制空间,针对 PSO 在特征选择应用很多都是建立在 BPSO 的基础上的,但其缺点是离散的 PSO 丧失了一些连续 PSO 的特性。在此情况下,选择在特征选择过程中,将特征选择问题转换为一个向量,由(0,1)来表示。$F=(\boldsymbol{ft}_{i1}, \boldsymbol{ft}_{i2}, \cdots, \boldsymbol{ft}_{id})$,$\boldsymbol{ft}_{id}$ 等于 1 时,该维特征被选中,等于 0 时,则未被选中。设定一个阈值 δ 来判断是否被选中,如公式(9)所示

$$\boldsymbol{ft}_{id} = \begin{cases} 1, & xt_{id} \geqslant \delta, \\ 0, & \text{other}, \end{cases} \tag{9}$$

δ 是随机数,取值范围 $U(0.2,0.8)$。根据公式(9)粒子的位置向量在连续空间域与离散问题域中完成特征向量转换。

在标准的 PSO 算法中,假设初始种群中存在先验近似最优粒子,则可确定整体的搜索方向,这将大幅度地缩短 WSPSO 的进化时间。所以,需要对数据预处理,得到特征的重要性。Brieman[24] 提出了一种确定特征重要性的方法,其主要思想是:每次选择特征时,随机替换特征的值,并记录预测精度的变化,预测的准确性越高说明该特征的重要性越高。这里所提的特征重要性也就是对预测结果贡献的百分比。因此可以得到占比最高的粒子,加入初始的种群。

2.1　WSPSO-SAMME.R-DT 算法

给出基于封装式特征选择的 WSPSO-SAMME.R-DT 算法的具体步骤。其中 DT 代表基分类器决策树。为了增加集成学习中基分类器间的多样性,将随机选择决策树中的最佳分割点。

算法 2:WSPSO-SAMME.R-DT 算法

输入:训练集 $\{(x_i,y_i)|i=1,2,\cdots,n\}$,最大迭代次数 T,种群大小 m。

1)初始化种群。依据特征重要性,选择重要性最高的一个粒子作为初始粒子,剩余的 $m-1$ 个粒子以随机的方式生成。这 m 个粒子的各维分量都是 $U(0,1)$ 的随机数,将所有粒子进行组合,完成初始种群的构建;

2)判断是否满足条件 $t \leqslant T$ and p_g 的适应度小于 1。若成立继续下一步,不成立跳出循环;

3)对于粒子 $i=1,2,\cdots,m$;

4)根据公式(9)将粒子 x_i 转化为特征向量,基于特征向量从训练集中选取训练子集。然后根据算法 SAMME.R,训练出一个强分类器 H;

5)根据强分类器 H 得到的预测结果,计算每对类别组合的 AUC,然后按照公式(6)计算 AUCarea,作为 x_i 粒子的适应度值;

6)根据得到的 AUCarea 的值来更新个体最优 p_i 和全局最优 p_g;

7)根据公式(7)、(8)更新粒子位置;

8)根据公式(9)将 p_g 转化为特征向量;

输出:最优特征子集、强分类器 H。

3 实验仿真与分析

实验机器配置为:Window7,内存6 GB,CPU2.50 GHz,算法基于Python3.6.2实现。实验所用的10组数据集来自KEEL官网与UCI数据集,数据都源于实际中的应用领域,表1给出具体信息。不均衡比IR(imbalance ratio)是最大多数类别的样本数与最小少数类别的样本数之比。学习因子 $c_1=c_2=2$,经实验显示,种群粒子数 $m=100$ 时效果最佳,初始设置迭代次数 T 为300次,图2给出了每代最优个体适应度值的曲线,随着迭代次数增加,可以看出迭代次数 $T=50$ 时个体适应度值已趋于稳定,所以最终选择的迭代次数为 $T=50$。

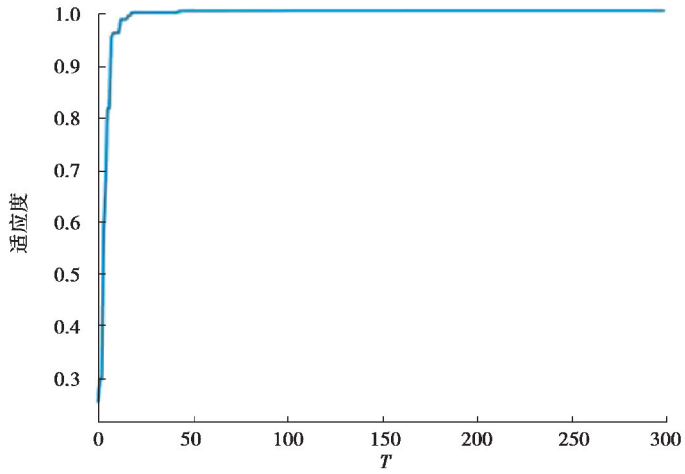


图2 最优个体适应度值的曲线

Fig. 2 Curve of optimal individual fitness value

实验结果利用AUCarea以及另一个被广泛应用的不平衡分类指标GM(G-Mean)[25]对算法进行评价。GM定义如下

$$\mathrm{GM}=\sqrt{\frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}}\times\sqrt{\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}},\tag{10}$$

其中TP,FP,FN,TN分别表示:小类正确分类的数量,预测为小类但是真实为大类,预测为大类但是真实为小类,大类正确分类的数量。

表1 数据集信息

Table 1 Dataset information

数据集	样本数/特征	类别分布	IR
Balance	625/4	(49/288/288)	5.88
New_thyroid	215/4	(30/35/150)	5.0
Hayes-Roth	132/5	(30/51/51)	1.7
Contraceptive	1437/9	(333/511/629)	1.89
Dermatology	358/34	(20/48/48/60/71/111)	5.55
Thyroid	720/21	(17/37/666)	39.18
Wine	178/13	(48/59/71)	1.48
Glass	214/9	(70/76/17/13/9/29)	8.44
Zoo	101/16	(41/20/5/13/4/8/9)	10.25
Autos	159/25	(2/14/33/32/20/9)	16.5

文献[6][7][15]各自提出了 PUSBE、CS-MCS、BAK 算法,与提出的 WSPSO-SAMME.R-DT 进行对比。采取一对一方法将 PUSBE 和 CS-MCS 扩充到多分类问题上,结果见表 2、3。

表 2　4 种算法的 AUCarea 值对比

Table 2　Comparison of AUCarea values of four algorithms

数据集	AUCarea			
	PUSBE	CS-MCS	BAK	WSPSO-SAMME.R-DT
Balance	0.631	0.651	0.792	0.981
New_thyroid	0.975	0.932	0.974	1.000
Hayes-Roth	0.872	0.784	0.767	0.972
Contraceptive	0.544	0.502	0.632	0.596
Dermatology	0.987	0.975	0.931	0.994
Thyroid	0.882	0.921	0.936	0.992
Wine	0.978	0.988	0.905	1.000
Glass	0.858	0.807	0.872	0.887
Zoo	1.000	0.923	1.000	1.000
Autos	0.804	0.832	0.927	0.931
平均值	0.853	0.832	0.888	0.935

表 3　四种算法的 GM 值对比

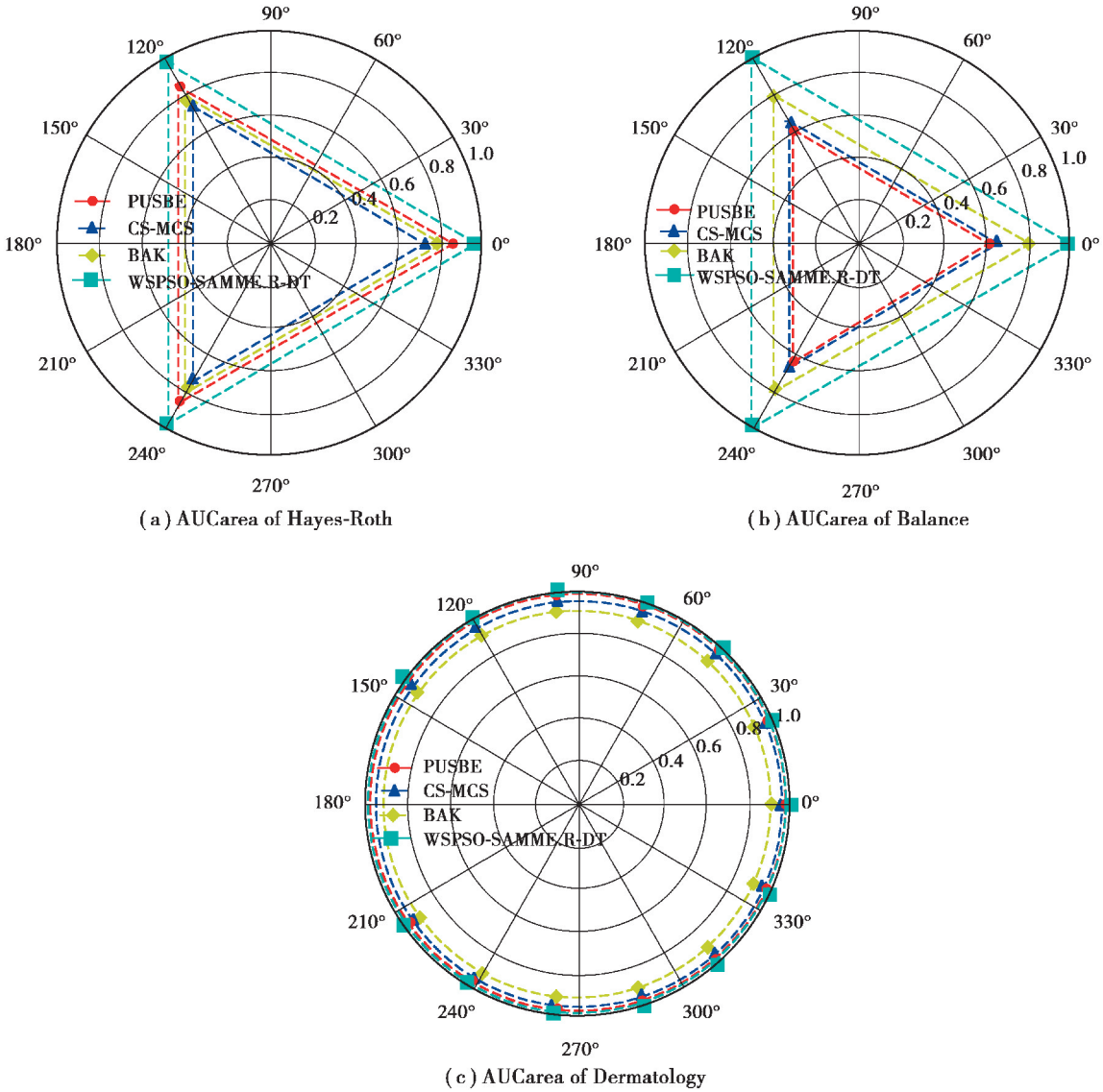
Table 3　Comparison of GM values of four algorithms

数据集	GM			
	PUSBE	CS-MCS	BAK	WSPSO-SAMME.R-DT
Balance	0.861	0.738	0.899	0.971
New_thyroid	0.970	0.939	0.981	1.000
Hayes-Roth	0.887	0.841	0.916	0.916
Contraceptive	0.639	0.614	0.530	0.459
Dermatology	0.997	0.921	0.957	0.991
Thyroid	0.987	0.971	0.931	0.996
Wine	0.989	0.994	0.834	1.000
Glass	0.899	0.865	0.992	0.993
Zoo	1.000	0.948	1.000	1.000
Autos	0.798	0.915	0.945	0.953
平均值	0.903	0.875	0.899	0.928

根据表 2 与表 3 可以得到如下结论:提出的算法 WSPSO-SAMME.R-DT 总体性能略优于要其他 3 种算法,尤其是在 New_thyroid、Wine 与 Zoo 数据集上,AUCarea 与 GM 的值都达到了 100%。除了 Contraceptive 数据集外,在其他数据集上,WSPSO-SAMME.R-DT 也略好于其他 3 种算法。其中 CS-MCS

在 AUCarea 上的平均值为 0.832，比 PUSBE 的平均值低了 2.1%；比 BAK 的平均值低了 5.6%；比提出的 WSPSO-SAMME.R-DT 的平均值低了 10.3%。而在 GM 值上 CS-MCS 的平均值为 0.875，比 PUSBE 的平均值低了 2.8%；比 BAK 的平均值低了 2.4%；比提出的 WSPSO-SAMME.R-DT 的平均值低了 5.3%。由此可看出算法总体性能相对较差，这是因为 CS-MCS 算法并没有跟其他 3 种算法一样采用特征选择技术，这也证明了特征选择可有效的应用于不均衡多分类的问题。

为了更直观的对比 4 种算法的分类效果，图 3 给出了 4 种算法 AUCarea 值的部分 polar 图。根据图 3 所示，图中红色虚线代表 PUSBE；蓝色虚线代表 CS-MCS；黄色虚线代表 BAK；青色虚线代表 WSPSO-SAMME.R-DT。他们所围成的面积就是其对应的 AUCarea 的值。从图中可以看出 WSPSO-SAMME.R-DT 在 Hayes-Roth 与 Balance 数据集中面积最大，意味着在这两种数据集下，WSPSO-SAMME.R-DT 优于其他 3 种算法，在 Dermatology 算法中排名第二，但是与该数据集下的最优算法 PUSBE 所产生的面积相差不多。



(a) AUCarea of Hayes-Roth

(b) AUCarea of Balance

(c) AUCarea of Dermatology

图 3 4 种算法的 AUCarea 极坐标图示比较

Fig. 3 Comparison of AUCarea polar coordinates of four algorithms

4 结　语

结合特征选择与集成学习方法提出了 WSPSO-SAMME.R-DT 算法，在 10 组不均衡数据集上对本算法进行实验测试，实验结果验证了该算法的有效性。WSPSO-SAMME.R-DT 使用了 WSPSO 算法并且以 AUCarea 作为适应度值，来优化特征选择。其中，AUCarea 具可视化的优点，并且对较差的 AUC 值更加敏感。笔者并没有采用采样技术对初始数据集进行数据层面的改进，避免了丢失重要信息、引入噪声等情况。WSPSO-SAMME.R-DT 可直接应用于多分类算法且并不需要进行拓展。

参考文献：

[1] Napierala K, Stefanowski J. Types of minority class examples and their influence on learning classifiers from imbalanced data[J]. Journal of Intelligent Information Systems, 2016, 46(3): 563-597.

[2] Glauner P, Boechat A, Dolberg L, et al. Large-scale detection of non-technical losses in imbalanced data sets[C]// 2016 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference. September 6-9, 2016, Minneapolis, MN, USA. IEEE, 2016: 1-5.

[3] Chawla N V, Lazarevic A, Hall L O, et al. SMOTEBoost: Improving prediction of the minority class in boosting[C]// European conference on principles of data mining and knowledge discovery. Berlin, Heidelberg: Springer, 2003: 107-119.

[4] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321-357.

[5] 武森，刘露，卢丹. 基于聚类欠采样的集成不均衡数据分类算法[J]. 工程科学学报，2017，39(8)：1244-1253.
Wu S, Liu L, Lu D. Imbalanced data ensemble classification based on cluster-based under-sampling algorithm[J]. Chinese Journal of Engineering, 2017, 39(8): 1244-1253. (in Chinese)

[6] Krawczyk B, Schaefer G. An improved ensemble approach for imbalanced classification problems[C]// 2013 IEEE 8th International Symposium on Applied Computational Intelligence and Informatics. May 23-25, 2013. Timisoara, Romania: IEEE, 2013: 423-426.

[7] Krawczyk B, Wo zniak M, Herrera F. Weighted one-class classification for different types of minority class examples in imbalanced data[C]// 2014 IEEE Symposium on Computational Intelligence and Data Mining. December 9-12, 2014. Orlando, FL, USA: IEEE, 2014: 337-344.

[8] 王莉莉，付忠良，陶攀，等. 基于主动学习不平衡多分类 AdaBoost 算法的心脏病分类[J]. 计算机应用，2017，37(7)：1994-1998.
Wang L L, Fu Z L, Tao P, et al. Heart disease classification based on active imbalance multi-class Ada Boost algorithm[J]. Journal of Computer Applications, 2017, 37(7): 1994-1998. (in Chinese)

[9] Tao X M, Li Q, Guo W J, et al. Self-adaptive cost weights-based support vector machine cost-sensitive ensemble for imbalanced data classification[J]. Information Sciences, 2019, 487: 31-56.

[10] 陶新民，郝思媛，张冬雪，等. 不均衡数据分类算法的综述[J]. 重庆邮电大学学报(自然科学版)，2013，25(1)：101-110，121.
Tao X M, Hao S Y, Zhang D X, et al. Overview of classification algorithms for unbalanced data[J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2013, 25(1): 101-110, 121. (in Chinese)

[11] Dietterich T G. Ensemble methods in machine learning[C]// International workshop on multiple classifier systems. Berlin, Heidelberg: Springer, 2000: 1-15.

[12] Tao X M, Li Q, Ren C, et al. Real-value negative selection over-sampling for imbalanced data set learning[J]. Expert Systems With Applications, 2019, 129: 118-134.

[13] Tao X M, Li Q, Ren C, et al. Affinity and class probability-based fuzzy support vector machine for imbalanced data sets[J]. Neural Networks, 2020, 122: 289-307.

[14] 张苗燕，王登飞，魏宗寿. 一种改进的 AdaBoost 快速训练算法[J]. 西北工业大学学报，2017，35(6)：1119-1124.
Zhang M Y, Wang D F, Wei Z S. An improved Ada boost training algorithm[J]. Journal of Northwestern Polytechnical University, 2017, 35(6): 1119-1124. (in Chinese)

[15] Guo H X, Li Y J, Li Y N, et al. BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification[J]. Engineering Applications of Artificial Intelligence, 2016, 49: 176-193.

[16] 胡旺, 李志蜀. 一种更简化而高效的粒子群优化算法[J]. 软件学报, 2007, 18(4): 861-868.
Hu W, Li Z S. A simpler and more effective particle swarm optimization algorithm[J]. Journal of Software, 2007, 18(4): 861-868. (in Chinese)

[17] Hastie T, Rosset S, Zhu J, et al. Multi-class AdaBoost[J]. Statistics and Its Interface, 2009, 2(3): 349-360.

[18] Nakas C T, Yiannoutsos C T. Ordered multiple-class ROC analysis with continuous measurements[J]. Statistics in Medicine, 2004, 23(22): 3437-3449.

[19] Hand D J, Till R J. A simple generalisation of the area under the ROC curve for multiple class classification problems[J]. Machine Learning, 2001, 45(2): 171-186.

[20] Qu Y, Fang Y, Yan F Q. Feature selection algorithm based on association rules[J]. Journal of Physics: Conference Series, 2019, 1168: 052012.

[21] Bratton D, Kennedy J. Defining a standard for particle swarm optimization[C]// 2007 IEEE Swarm Intelligence Symposium. April 1-5, 2007. Honolulu, HI, USA: IEEE, 2007: 120-127.

[22] 行鸿彦, 郭敏, 张兰, 等. 基于改进 SPSO-BP 神经网络的温度传感器湿度补偿[J]. 传感技术学报, 2018, 31(3): 380-385.
Xing H Y, Guo M, Zhang L, et al. The humidity compensation for temperature sensor based on improved SPSO-BP neural network[J]. Chinese Journal of Sensors and Actuators, 2018, 31(3): 380-385. (in Chinese)

[23] Kennedy J, Eberhart R C. A discrete binary version of the particle swarm algorithm[C]// 1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation. October 12-15, 1997. Orlando, FL, USA: IEEE, 1997: 4104-4108.

[24] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123-140.

[25] Galar M, Fernández A, Barrenechea E, et al. EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling[J]. Pattern Recognition, 2013, 46(12): 3460-3471.

(编辑 侯湘)