基于RB-XGBoost算法的智能电网调度控制系统健康度评价模型

谈林涛[1],李军良[2],任 昊[3],何 杨[3],高 欣[3],徐建航[2],黄晴晴[3]
(1. 国家电网华中电力调控分中心,湖北 武汉 430077;
2. 南瑞集团有限公司 国网电力科学研究院,北京 100192;3. 北京邮电大学 自动化学院,北京 100876)

摘要:针对智能电网调度控制系统(D5000系统)健康度评价,基于专家经验的传统评价方法存在主观性较大的问题,机器学习多分类方法是提高评价客观性的一种有效手段,但健康度各等级样本数目间存在的不平衡问题导致分类准确率较低,为此提出一种基于随机平衡和极端梯度提升(RB-XGBoost)算法的D5000系统健康度评价模型。首先,针对系统各评价等级样本数目严重不平衡的问题,提出一种自适应随机平衡(RB)的混合采样方法,分别以等级间样本数目的最大值、最小值作为采样区间的上、下限,生成多个随机数对各等级样本数据进行欠采样或过采样,增加训练数据的多样性并降低其不平衡程度;然后,训练平衡后的样本数据,建立极端梯度提升(XGBoost)算法子模型,考虑到各子模型重要度的一致性,提出采用硬投票方式集成所有子模型,得到与D5000系统各子模块对应的评价模型;最后,根据该系统指标层级关系,在评价过程中采用并、串行结合的计算方式,构建包含17个RB-XGBoost模型的D5000系统整体健康度评价模型。8组KEEL数据库中多类不平衡数据集的实验结果表明,与现有同类典型方法相比,所提方法的平均分类准确率最高提升了6.79%,平均提升了2.03%;某网省级D5000系统的实时采集数据验证了所提方法的有效性。

关键词:智能电网;D5000系统;健康度评价;多分类;自适应随机平衡采样;XGBoost算法
中图分类号:TM 734 **文献标志码**:A DOI:10.16081/j.epae.202001024

0 引言

智能电网调度控制系统(D5000系统)实现了电网调度业务的“横向集成、纵向贯通”,保障了电网的安全、稳定、经济、环保运行,在特大电网多级调度控制中发挥着重要的作用[1]。但随着电网规模的扩大、业务的升级,设备监视、运维压力大幅增加,为了保证D5000系统安全稳定运行,运维人员需实时掌握其健康程度,因此对其进行健康度评价变得尤为关键。

电力行业内相关系统的健康度分析大多采用基于专家经验的评价方法[2-6],典型方法有层次分析法(AHP)[3]、组合评价法[4]等,文献[5]通过AHP构建多层次指标体系结构模型,较全面地反映了火电厂的运行状态;文献[6]借鉴AHP的思想,实现了电力系统暂态模型的总体评价;文献[7-8]分别利用逼近理想解排序法(TOPSIS)-熵权法-AHP组合评价方法、主客观结合的熵权-网络层次分析组合权重分析法对智能电网管理进行相关评价;文献[9]利用相对熵组合赋权法对短路电流抑制方案进行综合评价。但上述方法在关键指标(指标权重、评价等级划分)的确定上依赖于专家经验,主观性较强;且一旦内部业务发生变动,该类方法无法快速进行相应的调整。

近年来,机器学习方法被广泛应用于解决电力系统的实际问题。例如,文献[10]提出基于进化思维的极限学习机分类算法,实现对航空发电机旋转整流器的故障分类;文献[11]针对二次设备缺陷的识别问题,利用Apriori算法对缺陷数据进行分析以建立识别模型。目前,机器学习方法也开始被应用于对象系统的评价,以提高评价结果的客观性。例如,文献[12]结合人工神经网络、信息融合技术对变压器进行状态评估;文献[13]利用模糊支持向量机算法实现了继电保护状态的在线评价。以上单一分类方法对特征空间中某些样本的分布区域有较好的学习效果,但对于由多类特征区域组成的整个样本空间而言,分类准确率有待提升。集成学习方法[14]将若干单一分类器相组合,实现各分类器的优势互补,其分类结果往往比单一分类器更加稳定、准确[15]。集成学习主要分为基于装袋算法(Bagging)、基于提升算法(Boosting)的2类方法[16-17],前者主要用于降低方差,后者主要用于降低偏差。随机森林(RF)算法[18]是Bagging方法的代表之一,通过Bootstrap方法生成多份不同的训练集来建立不同的子模型,增加多样性以提高整体的分类效果;陈天奇等人在2016年提出的极端梯度提升(XGBoost)算法[19]是Boosting方法的代表之一,其在梯度提升决策树(GBDT)[20]算法的基础上进行改进,实现了决策树的切分点查找算法的优化,通过正则化项解决了GBDT的过拟合问题[19],对于处理大规模数据的分类问题,具有准确度高、不易过拟合、可扩展性强的特点[21-22]。

收稿日期:2019-07-31;**修回日期**:2019-11-07

D5000系统的健康度评价等级分为完好、正常、注意、异常、严重[23]，其健康度评价问题可通过机器学习领域中的多分类方法来解决。该系统工作稳定，其健康程度长时间处于完好、正常状态，各个等级对应的样本数据分布存在较严重的不平衡现象，若直接利用XGBoost算法建立分类模型，将导致模型对注意、异常、严重等级对应样本的识别率偏低的问题。因此，所构建的系统健康度评价模型需能解决采集数据样本分布不均衡的多分类问题。此外，由于该系统的业务繁多、评价指标体系层级关系复杂，所有层的指标健康状态共同反映了整个系统的健康状态[23]；同时，该系统业务间具有“横向集成、纵向贯通”的特点。因此，还需建立一种多层级联的整体健康度评价模型，各层级间相互配合共同完成评价任务，所建立模型需满足并、串行工作的实时评价要求。

综上所述，针对D5000系统的健康度评价问题，本文提出一种基于随机平衡和极端梯度提升（RB-XGBoost）算法的D5000系统健康度评价模型，不仅解决了各等级样本分布不均衡的多分类问题，还建立了一种多层级联的整体健康度评价模型，各层级之间相互配合，共同完成评价任务以满足并、串行工作的要求。在构建单个评价子模型时，首先通过对所有评价等级的样本数据进行随机平衡（RB）处理，得到在各个评价等级下随机数目相同的平衡数据子集，然后通过XGBoost算法训练平衡数据子集，生成若干个子模型，最后通过硬投票方式得到最终的评价结果。为了验证所建模型的有效性，本文选取8组多类不平衡公开数据集以及该系统的实际运行数据进行实验对比验证。

1 基于RB-XGBoost算法的分类模型

1.1 XGBoost算法原理概述

XGBoost算法是一种基于Boosting思想的集成学习算法，其在GBDT算法的基础上进行改进，不仅在速度上得到了提升，而且为了提高精度，XGBoost算法对代价函数进行二阶泰勒展开，通过引入正则化项，避免产生过拟合现象。其集成树模型为：

$$\hat{y}_i=\theta(\boldsymbol{x}_i)=\sum_{k=1}^{K}f_k(\boldsymbol{x}_i)\quad f_k\in F \tag{1}$$

其中，K为子模型总数；$F=\{f(\boldsymbol{x})=\boldsymbol{\omega}_{q(\boldsymbol{x})}\}$为所有回归树集合，$\boldsymbol{\omega}_{q(\boldsymbol{x})}$为回归树所有叶子节点的权重组成的权重向量；$\hat{y}_i$为样本预测值；$\boldsymbol{x}_i$为样本输入特征；$f_k$为第$k$棵回归树，每棵回归树具有单独的叶子权重$\boldsymbol{\omega}$和树结构$q$。引入目标函数如式（2）所示。

$$O=l(y_i,\hat{y}_i)+\sum_{k=1}^{K}\Omega(f_k) \tag{2}$$

其中，$l(y_i,\hat{y}_i)$为损失函数，表示预测值$\hat{y}_i$与实际值y_i间的差值；Ω为正则化项，用于平滑最后的学习权重以避免过拟合。通过加法策略进行多次迭代，定义$\hat{y}_i^{(t)}$为第t次迭代中样本i的预测值，综合损失函数、正则化项，并结合式（1）和式（2），则目标函数可表示为：

$$O^{(t)}=\sum_{i=1}^{N}l\left(y_i,\hat{y}_i^{(t-1)}+g_if_t(\boldsymbol{x}_i)+\frac{1}{2}h_if_t^2(\boldsymbol{x}_i)\right)+\Omega(f_t) \tag{3}$$

$$g_i=\partial_{\hat{y}^{(t-1)}}l(y_i,\hat{y}^{(t-1)}),\ h_i=\partial^2_{\hat{y}^{(t-1)}}l(y_i,\hat{y}^{(t-1)})$$

其中，N为树的数量。主要实现过程如下：

（1）建立1棵新的决策树；

（2）根据式（3）所示的目标函数计算每个训练样本的g_i和h_i值，开始迭代；

（3）利用近似贪心算法寻找最佳分割点，得到决策树结构$f_t(\boldsymbol{x})$，其中t表示第t次迭代；

（4）将$f_t(\boldsymbol{x})$添加到集成树模型中；

（5）按照步骤（1）—（4）进行多次迭代得到最终的分类模型。

1.2 基于RB-XGBoost算法的分类模型

XGBoost算法通过残差不断优化下一个分类器，将弱分类器转换为强分类器，最后通过综合所有迭代分类器的输出结果得到待测样本的预测类别[19]。但是，D5000系统的健康度评价等级分为完好、正常、注意、异常、严重，由于系统的健康状况长期处于完好和正常这2个等级，因此注意、异常、严重这3个等级对应样本的数量远少于前2个等级的样本数量，产生了类别不平衡的现象，若直接采用原始数据训练模型，模型精度会受到很大的影响。因此，通过合理的采样方法平衡样本数量对提高模型的准确率至关重要。

机器学习中现有的采样方法主要分为欠采样技术和过采样技术[24]，典型的方法有随机欠采样和SMOTE（Synthetic Minority Oversampling TEchnique）过采样[25-26]。随机欠采样方法随机挑选部分多数类样本，旨在减少多数类样本以实现样本数量平衡，当不平衡程度较严重时，过度删除多数类样本会造成大部分有用信息的丢失；SMOTE过采样通过随机挑选少数类样本合成新样本，随着新样本数量的增加，易产生噪声，从而导致模型的过拟合。由于健康度数据集中各等级样本数量间存在差异，所以单一的采样方法无法全面处理D5000系统的不平衡数据。本文提出一种自适应随机平衡的混合采样方法以解决不同类别样本数量不平衡的问题。详细步骤如下。

（1）设置采样区间。将训练集分为n个类别，统计各类别的样本数量，以类别样本数量的最大值和最小值作为采样闭区间的两端点。

（2）自适应地选择采样方法。将n个类别单独划分为一个类别集，在采样区间内生成随机数，将各类别样本数量与随机数进行比较，若类别样本数量大于随机数，则采用随机欠采样方法；若类别样本数量不大于随机数，则采用SMOTE过采样方法。

（3）得到多份采样子集。综合各类别样本采样

子集，获得多份类别样本数量平衡的训练子集。

具体流程如图1所示。

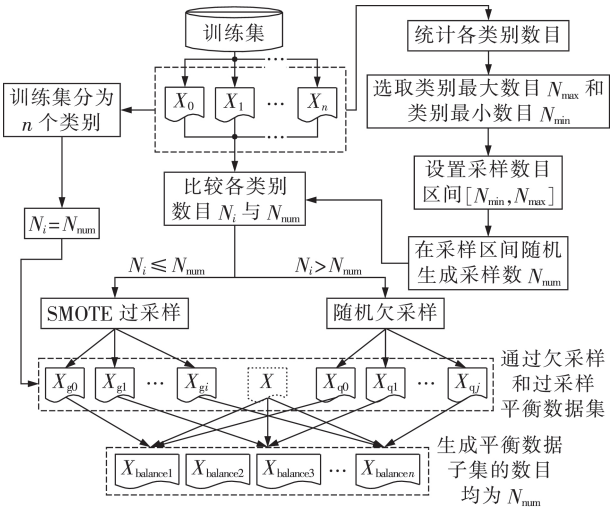

图1　随机平衡采样流程

Fig.1　Flowchart of RB sampling

考虑到将XGBoost算法（伪代码见附录A）直接用于训练不平衡数据在准确率上表现不佳的问题，本文结合随机平衡采样方法，提出RB-XGBoost算法训练分类模型。该模型中所有子模型包含的各类别样本数量相同，均为训练样本数量平衡的分类模型，其重要度具有一致性，并采用硬投票法集成所有子分类模型，得到最终的分类结果。主要的步骤如下：

（1）根据随机平衡方法获得平衡数据子集；

（2）考虑节点分裂时式（3）中的正则化项和目标函数，通过残差不断调整目标函数，训练平衡数据子集得到多个XGBoost树模型；

（3）将XGBoost树模型的分类结果集成，采用硬投票法得到最终的分类模型。

2　基于RB-XGBoost算法的D5000系统健康度评价模型

2.1　D5000系统的整体健康度评价模型

D5000系统的业务繁多，其业务主要包括fes应用、scada应用、public应用、data-server应用、数据库、消息总线、服务总线、公共服务等[23]，多种业务既相互配合保证电网调度的稳定运行，又单独负责实现各自的功能。该系统评价指标体系[23]的层级关系复杂，层级评价指标的优先级不同：上层评价指标受下层评价指标影响，即使是位于同一层的评价指标，其关联关系、重要程度也不同；所有层的评价指标共同反映该系统的健康度。

据此，本文提出基于RB-XGBoost算法的D5000系统整体健康度评价模型：考虑业务之间评价指标的异同，将评价指标相同部分构建通用的RB-XGBoost健康度子模型；根据评价指标的层级关系，将各层级RB-XGBoost子模块进行单独训练，然后汇总各子模块健康度等级进行最终评价。

D5000系统的整体健康度评价模型并行计算业务健康度、硬件健康度，首先，通过RB-XGBoost算法并行训练得到16个子模块的健康度评价模型，据此获得各子模块业务、硬件健康度等级，并将其保存在相应的数据库中，以便运维人员掌握各子模块的运行状况。然后，以各子模块健康度等级作为系统整体健康度评价模型的输入特征，通过RB-XGBoost算法训练各子模块健康度离散化等级数据，得到总模块健康度模型。并行训练子模块模型，串行训练子模块与总模块，构建包含17个RB-XGBoost模型的整体健康度评价模型，具体流程见附录B中图B1。

2.2　基于RB-XGBoost算法的D5000系统子模块评价模型

考虑到D5000系统的部分层级中所有子进程指标及其健康度的计算方法相同，因此集中不同应用下的进程数据，将其作为原始数据集训练RB-XGBoost子模块模型，不仅可增加原始数据集的数据量以降低样本分布不平衡程度，还可提高分类准确率。

D5000系统整体健康度评价模型包含17个RB-XGBoost模型，以数据库业务健康度为例，其模块示意图如附录B中图B2所示，将各个节点下数据库的进程指标数据和其他业务指标数据集中成6个特征属性。集合所有子进程的数据，将其作为RB-XGBoost数据库业务模型的原始数据集(X,Y)，其中X为子进程资源的占用情况，Y为子进程的健康度等级。RB-XGBoost数据库业务模型的生成步骤如下。

（1）将D5000系统中所有节点上的子进程资源占用情况数据作为RB-XGBoost数据库业务模型的原始数据集(X,Y)，样本含有进程占用内存、进程占用CPU、磁盘IO、网络IO、线程个数、网络连接数这6个输入特征。

（2）D5000系统的健康度等级分为完好、正常、注意、异常、严重5个等级，在构建RB-XGBoost数据库业务模型时将其离散化为0、1、2、3、4。

（3）分割原始数据集(X,Y)，其中的80% 作为训练集D，剩余的20% 作为测试集S。

（4）将训练集D分为X_0、X_1、…、X_4，分别表示完好、正常、注意、异常、严重这5个健康度等级数据构成的类别集合，统计各类别样本数量$N_i(i=0,1,\cdots,4)$，找出集合中样本数量的最大值和最小值，确定类别样本数量的区间$[N_{min},N_{max}]$。

（5）生成一个在区间$[N_{min},N_{max}]$内的随机数N_{num}，判断N_i与N_{num}的大小，若$N_i>N_{num}$，则采用随机欠采样方法；若$N_i<N_{num}$，则采用SMOTE过采样方法；若$N_i=N_{num}$，则该类别集无需处理。

（6）重复步骤（4）、（5），将欠采样类别集、过采样类别集、未处理类别集相组合，生成n个平衡数据子集。通过平衡数据子集训练得到n棵XGBoost分类树，其

对应n个分类结果，最后以少数服从多数的简单投票法确定最终的分类结果，从而建立RB-XGBoost数据库业务模型，具体训练过程见附录B中图B3。

3 实验结果与分析

3.1 实验环境及评价指标

实验环境为：Windows7操作系统，Inter Core i7-6700处理器，8 GB内存，采用python3.7.0编写程序。

对于数据集的不平衡程度，本文采用不平衡率γ进行度量，其定义如下：

$$\gamma = N_{\max}/N_{\min} \tag{4}$$

对于多分类问题，整体准确率指标无法全面评估分类器的性能，D5000系统的健康状况长期处于完好、正常这2个等级，造成各类别样本分布极不平衡，分类模型即使满足90%的整体准确率，也无法表明其对注意、异常、严重这3个等级具有较高的识别率。因此，本文以平均分类准确率τ_{macroacc}为评价标准，该指标为各类别赋予相同的权重，先单独计算各类别的准确率，然后求平均值获得最终的结果。相比于整体准确率指标，它能更有效地评估分类器对多类不平衡数据集的分类性能[27-29]，其定义如下：

$$\tau_{\text{macroacc}} = \frac{1}{N_{\text{d}}}\sum_{i=1}^{N_{\text{d}}}\rho_{\text{recall}\,i} \tag{5}$$

其中，N_{d}为数据集的类别总数；$\rho_{\text{recall}\,i}$为第i个类别中被正确分类的样本数量占该类别样本数量的比例。同时，为了更充分地将本文所提方法与其他各种算法的性能进行比较，在计算平均准确率的基础上，还使用了非参数统计检验方法中的Wilcoxon符号秩检验[28-30]进一步评估各分类算法的性能差异。

3.2 实验数据

为了验证本文所提方法的有效性，从机器学习权威KEEL数据库[16]中选取8组具有代表性的多类不平衡数据集进行对比实验分析，并选择D5000系统数据库业务数据进行验证，数据描述如表1所示，其中不平衡率按式(4)进行计算。

表1 数据描述

Table 1 Data description

数据集	样本总数	特征数	类别数	不平衡率
car	1728	6	4	18.61
zoo	101	16	7	10.25
glass	214	9	9	8.44
flare	1066	11	6	7.69
led7digit	500	7	10	1.54
page-blocks	5472	10	5	175.46
balance	625	4	3	5.87
shuttle	57999	9	7	4558
D5000系统	10000	6	5	21.38

从国家电网公司某网省级D5000系统中选取2016年12月1—20日每天10:00的数据库运行数据作为原始数据集，共10000条数据样本。由表1可看出，D5000系统的样本数据包含6个输入特征，共有5个类别，数据不平衡程度达21.38，是典型数据不平衡的多分类问题。具体数据描述见表2。

表2 D5000系统数据库数据输入属性与输出类别描述

Table 2 Description of data input attributes and output category for D5000 system database

输入属性名称	中文含义	类别	健康等级
DISK_IO	磁盘IO	0	完好
FILE_DISCRIPTOR_USED	网络连接数	1	正常
MEM_OCCUPY	进程占用内存	2	注意
THREAD_NUM	线程个数	3	异常
THREAD_NUM	网络IO	4	严重
CPU_PERCENT	进程占用CPU	—	—

3.3 模型实现及参数

为突出RB-XGBoost算法在数据不平衡多分类问题上的优势，选取RF[18]、XGBoost[19]、DES-MI[27]、RB-RF算法作为对比算法。其中RF、XGBoost算法分别为集成学习中基于Bagging和Boosting思想的代表方法，DES-MI算法是2018年提出的解决数据不平衡多分类问题的典型方法，RB-RF算法是将随机采样方法应用到RF算法中的多类不平衡学习方法。

使用python语言，导入RF和XGBoost等包，分别建立基于RF、XGBoost、DES-MI、RB-RF以及RB-XGBoost算法的分类模型。其中，RF算法的主要参数如下：决策树数量为1000棵，决策树分裂方式为gini，拆分树内部节点的最小样本数为2，叶子节点所需的最小样本数为1。XGBoost算法主要参数设置如下：迭代器为gbtree，学习速率为0.1，决策树数量为350棵，随机采样比例为0.8。DES-MI算法使用参考文献[27]中的默认参数。RB-RF算法中平衡子集个数设置为100，其他参数与RF算法的参数设置相同。RB-XGBoost算法中平衡子集个数设置为100，其余参数均与XGBoost算法相同。

采用5折交叉验证处理公开数据集，对比5种算法的平均分类准确率。为了突出RB-XGBoost算法在D5000系统健康度评价中的优势，以数据库业务健康度模块为例，结合5种算法各自的特点，分别按照附录B中图B3所示方式构建RF数据库业务模型、XGBoost数据库业务模型、DES-MI数据库业务模型、RB-RF数据库业务模型、RB-XGBoost数据库业务模型，采用5折交叉验证处理D5000系统数据，对比5种算法的平均分类准确率大小，结果见表3。

3.4 实验结果分析

3.4.1 公开数据集

将本文所提RB-XGBoost算法的实验结果与RF、XGBoost、DES-MI、RB-RF算法的结果进行对比，结果如表3、附录B中图B4和表B1所示。

从表3和附录B中图B4可以看出，通过在8组

表3　5种算法的平均分类准确率对比

Table 3　Comparison of average classification accuracy among five algorithms

公开数据集	$\tau_{macroacc}$ / %					RB-XGBoost 算法相比其他4种算法的提升比例（最小值）/ %
	RF	XGBoost	DES-MI	RB-RF	RB-XGBoost	
car	90.96	94.34	95.17	95.31	**96.57**	1.26
zoo	92.86	91.42	94.29	90.00	**95.71**	1.42
glass	70.15	77.69	78.01	78.35	**79.05**	0.70
flare	59.13	61.86	65.04	63.14	**71.83**	6.79
led7digit	70.80	72.35	72.18	70.31	**73.72**	1.37
page-blocks	84.34	84.43	91.34	92.09	**92.43**	0.34
balance	59.96	63.31	62.32	61.49	**66.51**	3.20
shuttle	89.10	93.91	94.21	91.60	**95.39**	1.18
平均值/%	77.16	79.91	81.57	80.29	**83.90**	2.03

公开数据集上进行实验验证，相较于RF、XGBoost、DES-MI、RB-RF算法，RB-XGBoost算法的$\tau_{macroacc}$均有不同程度的提高，相比于RF、XGBoost、DES-MI、RB-RF算法中表现最好的算法，RB-XGBoost算法的$\tau_{macroacc}$平均提升了2.03%，最高提升了6.79%，体现了RB-XGBoost算法可有效地应用于多类不平衡数据的分类问题。从附录B中表B1所示Wilcoxon符号秩检验结果也可看出，在满足显著性水平等于0.05的条件下，本文所提方法的$\tau_{macroacc}$均优于其他算法。

3.4.2　D5000数据库业务数据

采用5折交叉验证将RF数据库业务模型、XGBoost数据库业务模型、DES-MI数据库业务模型、RB-RF数据库业务模型、RB-XGBoost数据库业务模型进行对比，实验结果见表4、附录B中图B5和表B2。

表4　基于D5000子模块数据集的5种模型的平均分类准确率对比

Table 4　Comparison of average classification accuracy among five models based on D5000 submodule data set

数据集	$\tau_{macroacc}$ / %					RB-XGBoost模型相比其他4种模型的提升比例（最小值）/ %
	RF	XGBoost	DES-MI	RB-RF	RB-XGBoost	
D5000-D1	88.45	89.08	87.49	89.44	**93.97**	4.53
D5000-D2	90.03	91.05	88.48	91.23	**92.68**	1.45
D5000-D3	88.89	89.38	88.41	92.23	**93.21**	0.98
D5000-D4	88.58	89.21	86.93	90.19	**93.90**	3.71
D5000-D5	89.10	90.44	85.49	90.88	**93.34**	2.46
平均值 / %	89.01	89.83	87.36	90.79	**93.42**	2.63

从表4和附录B中图B5可看出，利用10 000条D5000系统数据库业务数据，分别在RF数据库业务模型、XGBoost数据库业务模型、DES-MI数据库业务模型、RB-XGBoost数据库业务模型上进行训练与测试，每折的$\tau_{macroacc}$均有不同程度的提高。相比于RF数据库业务模型、XGBoost数据库业务模型、DES-MI数据库业务模型、RB-RF数据库业务模型中表现最好的模型，RB-XGBoost数据库业务模型的$\tau_{macroacc}$平均提升了2.63%，最高提升了4.53%，体现了该方法可有效应用于D5000系统的健康度评价问题。相比于RF算法，XGBoost算法在D5000系统健康度评价问题上的分类效果更好，但由于各等级样本分布不均，平均分类准确率有待提升，RB-XGBoost算法有效解决了该问题。与DES-MI算法相比，RB-XGBoost算法保留所有集成树模型进行预测，增加了分类器的多样性以提高平均分类准确率。与RB-RF算法相比，RB-XGBoost算法采用相对预测精度更高的XGBoost集成树，有效提高了平均分类准确率。从附录B中表B2所示Wilcoxon符号秩检验结果也可看出，在满足显著性水平等于0.05的条件下，本文所提方法的平均准确率指标均优于其他算法。

4　结论

针对D5000系统的健康度评价需求，在考虑D5000系统数据不平衡特点的基础上，本文提出一种基于RB-XGBoost算法的健康度评价模型，该模型是机器学习在电力系统健康度评价中的一种应用，并将其与基于RF、XGBoost、DES-MI、RB-RF算法的模型进行对比以验证所提模型的有效性。同时，本文所提基于多层级联的RB-XGBoost整体健康度评价模型，通过不同业务特点、节点指标选择合适的健康度评价特征，然后训练各子模块的健康度评价模型，由下而上地进行多层综合评价，实现了D5000系统健康度评价功能。该模型具有良好的适应能力，一旦系统出现新业务，可重新获取数据进行学习，生成新的子模块健康度评价模型。在该模型的训练过程中，平衡子集的数目通过多次实验设置，而如何根据D5000系统各模块健康等级下样本数据的特点由采样方法自主对其进行调整将是后续研究工作的重点。

附录见本刊网络版（http://www.epae.cn）。

参考文献：

[1] 辛耀中，石俊杰，周京阳，等. 智能电网调度控制系统现状与技术展望[J]. 电力系统自动化，2015，39(1)：2-8.
XIN Yaozhong，SHI Junjie，ZHOU Jingyang，et al. Technology development trends of smart grid dispatching and control systems[J]. Automation of Electric Power Systems，2015，39(1)：2-8.

[2] 赵书强，汤善发. 基于改进层次分析法、CRITIC法与逼近理想解排序法的输电网规划方案综合评价[J]. 电力自动化设备，2019，39(3)：143-148，162.
ZHAO Shuqiang，TANG Shanfa. Comprehensive evaluation of transmission network planning scheme based on improved analytic hierarchy process，CRITIC method and TOPSIS[J]. Electric Power Automation Equipment，2019，39(3)：143-148，162.

[3] 杜栋，庞庆华，吴炎. 现代综合评价方法与案例精选[M]. 2版. 北京：清华大学出版社，2008：11-33.

[4] 张发明，刘志平. 组合评价方法研究综述[J]. 系统工程学报，2017，32(4)：557-569.
ZHANG Faming，LIU Zhiping. Combined evaluation methods：a literature review[J]. Journal of Systems Engineering，2017，32(4)：557-569.

[5] 李勇，何蕾，庞传军，等. 基于层次分析法的火电厂运行情况量化评价方法[J]. 电网技术，2015，39(2)：500-504.
LI Yong，HE Lei，PANG Chuanjun，et al. An analytical hierarchy process based quantitative method to evaluate operating condition of thermal power plant[J]. Power System Technology，2015，39(2)：500-504.

[6] 李娟，薛永端，徐丙垠，等. 基于层次分析的电力系统暂态模型评价方法[J]. 电网技术，2013，37(8)：2207-2211.
LI Juan，XUE Yongduan，XU Bingyin，et al. An analytic hierarchy process based assessment method for power system transient models[J]. Power System Technology，2013，37(8)：2207-2211.

[7] 孙蕾. 智能电网管理评价指标体系与多属性分析模型研究[D]. 北京：华北电力大学，2015.
SUN Lei. The study of evaluation system of smart grid management and multi-attribute analysis model[D]. Beijing：North China Electric Power University，2015.

[8] 孙强，葛旭波，刘林，等. 智能电网多属性网络层次组合评价法及其应用研究[J]. 电网技术，2012，36(10)：49-54.
SUN Qiang，GE Xubo，LIU Lin，et al. Multi-attribute network process comprehensive evaluation method for smart grid and its application[J]. Power System Technology，2012，36(10)：49-54.

[9] 李基康，滕欢，郭宁. 采用相对熵组合赋权法的500 kV电网短路电流抑制策略优化决策[J]. 电网技术，2016，40(6)：1814-1820.
LI Jikang，TENG Huan，GUO Ning. An optimal decision to restrain short circuit current for 500 kV power grids with combinatorial weighting method of relative entropy[J]. Power System Technology，2016，40(6)：1814-1820.

[10] 崔江，唐军祥，张卓然，等. 基于极限学习机的航空发电机旋转整流器快速故障分类方法研究[J]. 中国电机工程学报，2018，38(8)：2458-2466，2555.
CUI Jiang，TANG Junxiang，ZHANG Zhuoran，et al. Fast fault classification method research of aircraft generator rotating rectifier based on extreme learning machine[J]. Proceedings of the CSEE，2018，38(8)：2458-2466，2555.

[11] 张延旭，胡春潮，黄曙，等. 基于Apriori算法的二次设备缺陷数据挖掘与分析方法[J]. 电力系统自动化，2017，41(19)：147-151，163.
ZHANG Yanxu，HU Chunchao，HUANG Shu，et al. Apriori algorithm based data mining and analysis method for secondary device defects[J]. Automation of Electric Power Systems，2017，41(19)：147-151，163.

[12] 阮羚，谢齐家，高胜友，等. 人工神经网络和信息融合技术在变压器状态评估中的应用[J]. 高电压技术，2014，40(3)：822-828.
RUAN Ling，XIE Qijia，GAO Shengyou，et al. Application of artificial neural network and information fusion technology in power transformer condition assessment[J]. High Voltage Engineering，2014，40(3)：822-828.

[13] 曹海欧，张沛超，高翔. 基于模糊支持向量机的继电保护状态在线评价[J]. 电力系统保护与控制，2016，44(20)：70-74.
CAO Haiou，ZHANG Peichao，GAO Xiang. Online condition evaluation of relay protection based on fuzzy support vector machine[J]. Power System Protection and Control，2016，44(20)：70-74.

[14] 王德志，张孝顺，刘前进，等. 基于集成学习的孤岛微电网源-荷协同频率控制[J]. 电力系统自动化，2018，42(10)：46-52.
WANG Dezhi，ZHANG Xiaoshun，LIU Qianjin，et al. Ensemble learning for generation-consumption coordinated frequency control in an islanded microgrid[J]. Automation of Electric Power Systems，2018，42(10)：46-52.

[15] 龚越，罗小芹，王殿海，等. 基于梯度提升回归树的城市道路行程时间预测[J]. 浙江大学学报(工学版)，2018，52(3)：453-460.
GONG Yue，LUO Xiaoqin，WANG Dianhai，et al. Urban travel time prediction based on gradient Boosting regression tress[J]. Journal of Zhejiang University(Engineering Science)，2018，52(3)：453-460.

[16] 李诒靖，郭海湘，李亚楠，等. 一种基于Boosting的集成学习算法在不均衡数据中的分类[J]. 系统工程理论与实践，2016，36(1)：189-199.
LI Yijing，GUO Haixiang，LI Yanan，et al. A Boosting based ensemble learning algorithm in imbalanced data classification[J]. Systems Engineering-Theory & Practice，2016，36(1)：189-199.

[17] 黄新波，李文君子，宋桐，等. 采用遗传算法优化装袋分类回归树组合算法的变压器故障诊断[J]. 高电压技术，2016，42(5)：1617-1623.
HUANG Xinbo，LI Wenjunzi，SONG Tong，et al. Application of bagging-CART algorithm optimized by genetic algorithm in transformer fault diagnosis[J]. High Voltage Engineering，2016，42(5)：1617-1623.

[18] 王守相，周凯，苏运. 基于随机森林算法的台区合理线损率估计方法[J]. 电力自动化设备，2017，37(11)：39-45.
WANG Shouxiang，ZHOU Kai，SU Yun. Line loss rate estimation method of transformer district based on random forest algorithm[J]. Electric Power Automation Equipment，2017，37(11)：39-45.

[19] CHEN T G C. XGBoost：a scalable tree Boosting system[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco，USA：[s.n.]，2016：785-794.

[20] 毕云帆，撖奥洋，张智晟，等. 基于模糊Bagging-GBDT的短期负荷预测模型研究[J]. 电力系统及其自动化学报，2019，31(7)：51-56.
BI Yunfan，HAN Aoyang，ZHANG Zhisheng，et al. Study on short-term load forecasting model based on fuzzy Bagging-GBDT[J]. Proceedings of the CSU-EPSA，2019，31(7)：51-56.

[21] ZHANG D H，QIAN L Y，MAO B J，et al. A data-driven design for fault detection of wind turbines using random forests and XGboost[J]. IEEE Access，2018，6：21020-21031.

[22] 张晨宇，王慧芳，叶晓君. 基于XGBoost算法的电力系统暂态稳定评估[J]. 电力自动化设备，2019，39(3)：77-83，89.
ZHANG Chenyu，WANG Huifang，YE Xiaojun. Transient stability assessment of power system based on XGBoost algorithm

[J]. Electric Power Automation Equipment, 2019, 39(3): 77-83, 89.
[23] 李新鹏，邬小波，高欣，等. 业务与硬件融合的智能电网调度控制系统健康度综合评价[J]. 电力系统自动化，2017，41(20)：78-83，162.
LI Xinpeng, WU Xiaobo, GAO Xin, et al. Overall evaluation of health degree for smart grid dispatch and control systems based on integration of business and hardware [J]. Automation of Electric Power Systems, 2017, 41(20): 78-83, 162.
[24] CHEN L, BAO L J, LI J, et al. An aliasing artifacts reducing approach with random undersampling for spatiotemporally encoded single-shot MRI[J]. Journal of Magnetic Resonance, 2013, 237: 115-124.
[25] 毛文涛，田杨阳，王金婉，等. 面向贯序不均衡分类的粒度极限学习机[J]. 控制与决策，2016，31(12)：2147-2154.
MAO Wentao, TIAN Yangyang, WANG Jinwan, et al. Granular extreme learning machine for sequential imbalanced data [J]. Control and Decision, 2016, 31(12): 2147-2154.
[26] 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.
[27] GARCÍA S, ZHANG Z L, ALTALHI A, et al. Dynamic ensemble selection for multi-class imbalanced datasets [J]. Information Sciences, 2018, 445 / 446: 22-37.
[28] RAGHUWANSHI B S, SHUKLA S. Generalized class-specific kernelized extreme learning machine for multiclass imbalanced learning[J]. Expert Systems With Applications, 2019, 121: 244-255.
[29] ZHANG Z L, LUO X G, GONZÁLEZ S, et al. DRCW-ASEG: one-versus-one distance-based relative competence weighting with adaptive synthetic example generation for multi-class imbalanced datasets[J]. Neurocomputing, 2018, 285: 176-187.
[30] ARSHAD A, RIAZ S, JIAO L C. Semi-supervised deep fuzzy C-mean clustering for imbalanced multi-class classification [J]. IEEE Access, 2019, 7: 28100-28112.

作者简介：


谈林涛

谈林涛（1985—），男，湖北大悟人，工程师，硕士，研究方向为电力系统自动化、软件工程和人工智能（**E-mail**：tanlintao@cc.sgcc.com.cn）；

李军良（1981—），男，河北临漳人，高级工程师，硕士，研究方向为电力系统自动化、软件工程（**E-mail**：lijunliang_cn@163.com）；

任　昺（1995—），男，甘肃定西人，硕士研究生，研究方向为机器学习和电力系统自动化（**E-mail**：byrenbing@163.com）。

（**编辑**　陆丹）

Health evaluation model of smart grid dispatch and control system based on RB-XGBoost algorithm

TAN Lintao[1], LI Junliang[2], REN Bing[3], HE Yang[3], GAO Xin[3], XU Jianhang[2], HUANG Qingqing[3]
(1. Dispatching and Control Center, Central China Branch of State Grid Corporation of China, Wuhan 430077, China;
2. State Grid Electric Power Research Institute, NARI Group Corporation, Beijing 100192, China;
3. School of Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: For the health evaluation of smart grid dispatch and control system (D5000 system), the traditional expert experience-based evaluation methods have the problem of subjectivity. The multi-classification method based on machine learning is an effective way to improve the evaluation objectivity. However, the data imbalance among the samples of different health levels leads to a low classification accuracy. Therefore, a health evaluation model of D5000 system based on RB-XGBoost (Random Balance and eXtreme Gradient Boosting) algorithm is proposed. Firstly, in view of the serious imbalance of sample number among evaluation levels, a hybrid sampling method based on adaptive RB (Random Balance) is proposed. The maximum and minimum sample number among levels are used as the upper and lower limits of sampling interval respectively, and multiple random numbers are generated to undersample or oversample the data of each level, which increases the diversity and decreases the imbalance degree of training data. Then, after training the balanced sample data, sub-models of XGBoost (eXtreme Gradient Boosting) algorithm are established. Considering the consistency of each sub-model's importance, a hard voting method is proposed to integrate all sub-models, and then the evaluation model corresponding to each sub-module of D5000 system is obtained. Finally, according to the hierarchical relationship of the system indicators, a parallel-serial combined calculation method is adopted in the evaluation process. In this way, the overall health evaluation model of D5000 system including 17 RB-XGBoost models is constructed. The experimental results based on eight sets of KEEL multi-class imbalanced datasets show that the average classification accuracy of the proposed method increases by 6.79% at the most and 2.03% on average compared with the existing similar typical algorithms. And the effectiveness of the proposed method is verified by the actual sampling data of a provincial D5000 system.

Key words: smart grid; D5000 system; health evaluation; multi-classification; adaptive random balance sampling; XGBoost algorithm

附录 A

输入为训练数据集 D、测试集 S，训练集的分类标签数目为 T；输出为待分类样本的分类结果 l_{Label}。

步骤 1：基于随机平衡获得平衡数据子集。

（1）获得训练集分类标签数目 N_i（i=1，2，…，T），各分类标签训练子集 D_i（i=1，2，…，T）。

（2）选择 $N_{\max}=\max\{N_1,N_2,\cdots,N_T\}$、$N_{\min}=\min\{N_1,N_2,\cdots,N_T\}$ 作为闭区间的两端点。

（3）在区间 $[N_{\min},N_{\max}]$ 内选取随机数 N_{Num}。

（4）如果 $N_i<N_{\text{Num}}$，则 N_i 进行 SMOTE 过采样，得到 X_{gi}（i=1，2，…，i）；如果 $N_i>N_{\text{Num}}$，则 N_i 进行随机欠采样，得到 X_{qj}；否则，保留原始分类标签训练子集，得到 X_{yk}。其中 $i+j+k=T$。

（5）将 X_q、X_g、X_y 组合，获得平衡数据子集 X_t。

步骤 2：基于 XGBoost 算法训练平衡数据子集。

步骤 3：硬投票得出分类标签。

（1）测试集 S 通过 XGBoost 树获得标签 $l_{\text{Label}i}$（i=1，2，…，T）；

（2）统计获得标签的各类别数目 N_{si}（i=1，2，…，T）；

（3）输出 l_{Label}，对于 i=1，2，…，T，如果 $N_{si}=\max\{N_{s1},N_{s2},\cdots,N_{sT}\}$，则 $l_{\text{Label}i}=i$。

附录 B

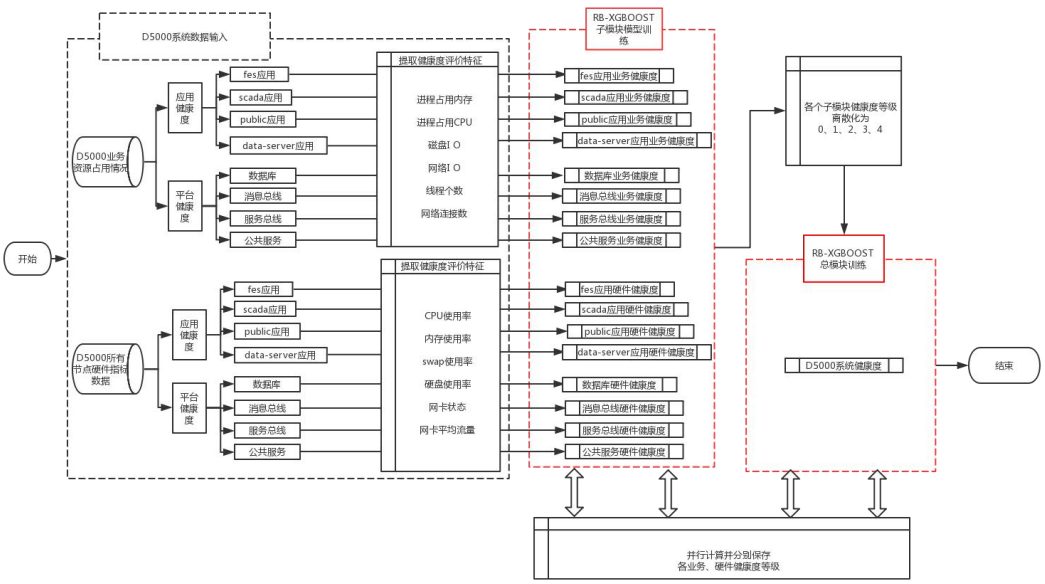

图 B1 D5000 系统整体健康度评价模型

Fig.B1 Overall health evaluation model of D5000 system

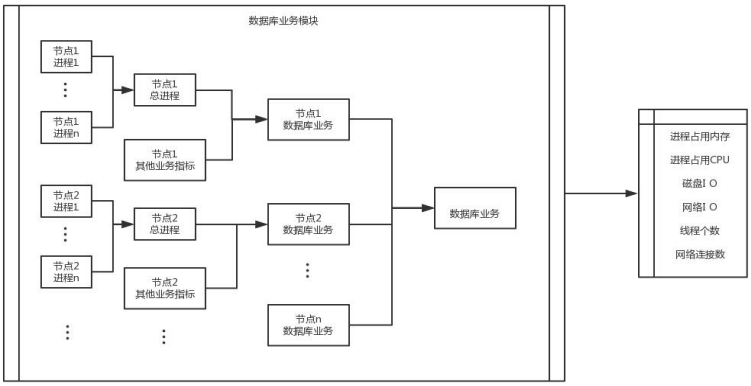

图 B2 D5000 数据库业务模块

Fig.B2 Business modules of D5000 database

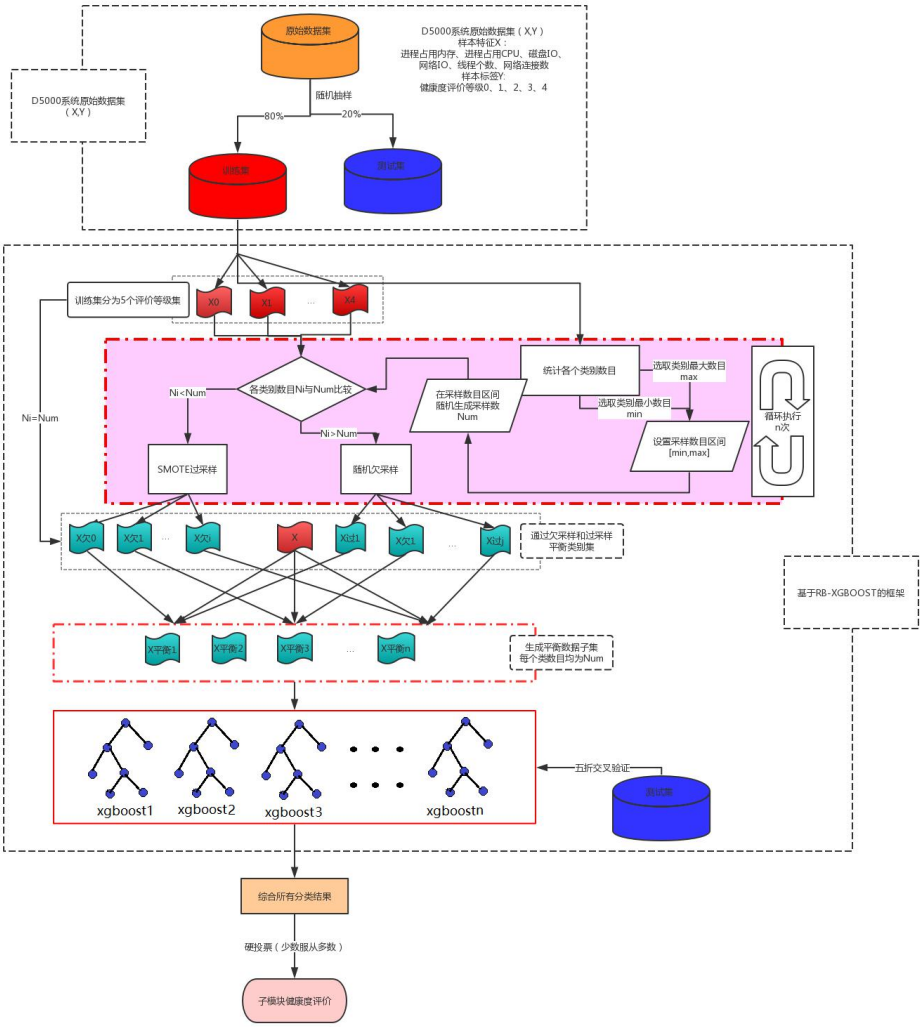

图 B3 基于 RB-XGBoost 算法的 D5000 系统数据库业务评价模型

Fig.B3 Business evaluation model of D5000 system database based on RB-XGBoost algorithm

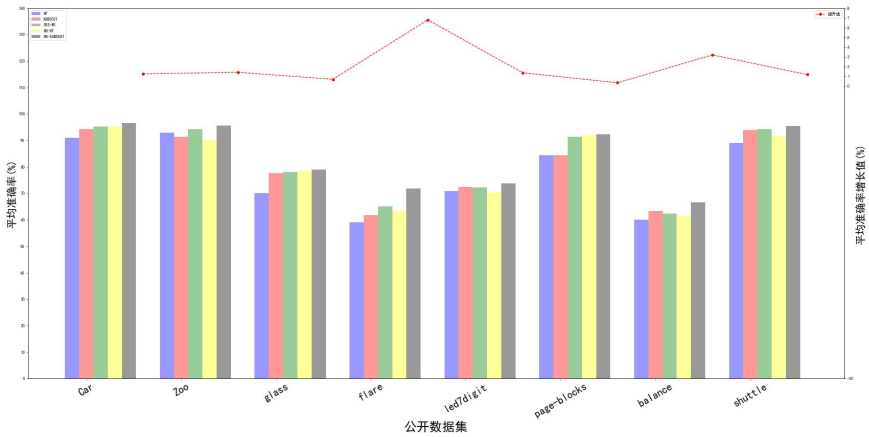

图 B4 RF、XGBoost、DES-MI、RB-RF、RB-XGBoost 算法的平均分类准确率对比

Fig.B4 Comparison of average classification accuracy among RF,XGBoost,DES-MI,RB-RF and RB-XGBoost algorithm

表 B1 在公开数据集上验证 RB-XGBOOST 算法与其他算法差异性的 Wilcoxon 符号秩检验结果($\alpha = 0.05$)

Table B1 Wilcoxon signed rank test results verifying difference between RB-XGBoost algorithm and other algorithms on public datasets($\alpha = 0.05$)

比较算法	R^+	R^-	p-value
RB-XGBoost、RB-RF	36	0	0.012
RB-XGBoost、DES-MI	36	0	0.012
RB-XGBoost、XGBoost	36	0	0.012
RB-XGBoost、RF	36	0	0.012

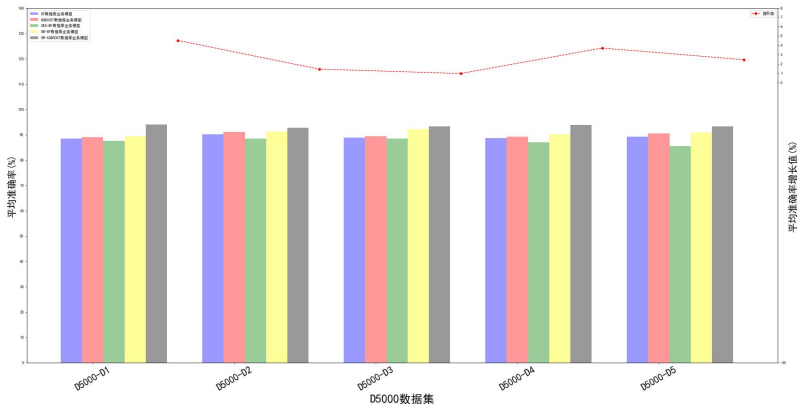

图 B5 在 D5000 子模块数据集上 5 种模型的平均分类准确率对比

Fig.B5 Comparison of average classification accuracy among five models on D5000 submodule dataset

表 B2 在 D5000 子模块数据集上验证 RB-XGBoost 与其他算法差异性的 Wilcoxon 符号秩检验结果($\alpha = 0.05$)

Table B2 Wilcoxon signed rank test results verifying difference between RB-XGBoost algorithm and other algorithms on D5000 submodule dataset($\alpha = 0.05$)

比较算法	R^+	R^-	p-value
RB-XGBoost、RB-RF	15	0	0.043
RB-XGBoost、DES-MI	15	0	0.043
RB-XGBoost、XGBoost	15	0	0.043
RB-XGBoost、RF	15	0	0.043