不平衡数据集下基于自适应加权 Bagging-GBDT 算法的磁盘故障预测模型

李新鹏 1,2，高欣 2，何杨 2，阎博 3，孙汉旭 2，李军良 4，徐建航 4，刘震宇 5，庞　博 5

(1 国家电网有限公司，北京 100031；2 北京邮电大学 自动化学院，北京 100876；3 国网冀北电力有限公司，北京 100054；4 南瑞集团（国网电力科学研究院）有限公司，北京 100192；5 国网冀北电力有限公司 承德供电公司，河北 承德 067000)

摘　要： 针对磁盘数据集中正负样本数目严重不平衡导致基于机器学习的分类算法易出现故障预测准确率低的问题，本文提出一种基于自适应加权 Bagging-GBDT 算法的磁盘故障预测模型.首先，提出基于聚类的分层欠采样方法对健康磁盘样本进行多次抽样，解决随机欠采样方法易丢弃潜在有用样本的问题；其次，将每次采样后样本与全部故障磁盘样本组合得到多个样本子集，通过训练这些子集建立多个预测精度较高的 GBDT 子分类模型；最后，根据待测点邻域样本类别自适应确定各子模型权重，据此通过加权硬投票集成最终的磁盘故障预测模型.在 8 组 KEEL 不平衡数据集上实验结果表明，与现有典型不平衡学习算法相比，少数类的召回率平均提升了 9.46%；同时在磁盘公开数据集和某调度系统磁盘数据上对比验证了该方法在故障预测率上的先进性.

关键词： 磁盘故障预测；不平衡数据集；分层欠采样；Bagging-GBDT；自适应加权

Prediction model of disk failure based on adaptive weighted bagging-GBDT algorithm under imbalanced dataset

LI Xin-peng1,2, GAO Xin2, HE Yang2, YAN Bo3, SUN Han-xu2,LI Jun-liang4, XU Jian-hang4, LIU Zhen-yu5, PANG Bo5

(1 State Grid Corporation of China, Beijing 100031, China;2 College of Automation, Beijing University of Posts and Telecommunications, Beijing 100876, China;3 State Grid Jibei Electric Power Company Limited, Beijing 100054, China;4 Nari Group (State Grid Electric Power Research Institute) Corporation, Beijing 100192, China;5 State Grid Jibei Electric Power Company Limited Chengde Power Supply Company,Chengde 067000, China)

Abstract： Aiming at the problem that the classification algorithm based on machine learning is prone to low accuracy of fault prediction due to the serious imbalance between the number of positive and negative samples in the disk dataset, this paper proposes a disk fault prediction model based on adaptive weighted Bagging-GBDT algorithm. Firstly, a hierarchical under-sampling method based on clustering algorithm is proposed to sample healthy disk samples several times to solve the problem that the random undersampling method is easy to discard potentially useful samples. Secondly, each sample after sampling is combined with all the failed disk samples to obtain several subsets. By training these subsets, a number of GBDT sub-classification models with higher prediction accuracy are established. Finally, the weights of each sub-model are adaptively determined through the neighborhood sample label of the test sample, and the final disk failure prediction model is integrated by weighted hard voting. The experimental results on 8 sets of KEEL imbalanced datasets show that the recall of the negative is increased by an average of 9.46% compared with the existing typical imbalanced learning algorithm. At the same time, the advancement of the method in the fault prediction rate is verified on disk public datasets and the disk data of a scheduling system.

Key words： prediction of disk failure; imbalanced dataset; hierarchical under-sampling; Bagging-GBDT; adaptive weighted

作者简介：

李新鹏　男，(1989-)，博士研究生，高级工程师.研究方向为机器学习、电力系统自动化.

高　欣（通讯作者）　男，(1974-)，博士，副教授.研究方向为机器学习、数据挖掘；Email:xlhhh74@bupt.edu.cn.

何　杨　女，(1996-)，硕士研究生.研究方向为机器学习、数据挖掘.

阎　博　男，(1985-)，博士，高级工程师.研究方向为机器学习、电力系统自动化.

孙汉旭　男，(1960-)，博士，教授.研究方向为智能系统的分析与设计研究.

李军良　男，(1981-)，硕士研究生，高级工程师.研究方向为电力系统自动化、软件工程研究.

徐建航　男，(1988-).研究方向为电力系统自动化.

刘震宇　男，(1976-).研究方向为电力系统自动化.

庞　博　男，(1975-).研究方向为电力系统自动化.