文章编号:2096-1472(2024)03-0074-05
DOI:10.19644/j.cnki.issn2096-1472.2024.003.015

基于随机森林的硬盘故障率预测研究

张永强[1,4],孔君君[1],崔 摇[2],李向南[3,4]

(1.河北科技大学信息科学与工程学院,河北 石家庄 050018;
2.石家庄常宏智能科技有限公司,河北 石家庄 050004;
3.石家庄春晓互联网信息技术有限公司,河北 石家庄 050061;
4.河北省智能物联网技术创新中心,河北 石家庄 050018)

✉ zyq@hebust.edu.cn; kjunjun555@163.com; cuiyao@changhong.cc; xiangnan.li@chunxiao.net

摘 要:为了避免硬盘出现故障而造成大量数据丢失,文章提出一种基于随机森林的方法对硬盘的故障进行预测,降低其丢失数据的风险。首先,在数据预处理方面,对所采用的数据做特征映射预处理;其次,通过对决策树进行构建及选取等,构建随机森林预测模型,根据所选取的特征属性预测硬盘故障率所在的区间,并且特征属性的变化能反映出硬盘故障率的变化趋势;最后,对构建的随机森林模型参数进行调优,选取不同的 n_estimators 参数值进行测试和优化。实验结果表明,与 XGBoost(Extreme Gradient Boosting)、LSTM(Long Short-Term Memory)等方法相比,本文方法的 $F1$ 值(F-Measure)分别提高了 0.93%和 1.84%,并且对随机森林预测模型的参数值进行不同取值测试,最终准确率达到 98.18%,比默认值提高了 1.23%,证明该方法能更精确地预测硬盘故障率,反映出硬盘故障率基于特征属性的变化趋势。

关键词:随机森林;硬盘故障率;故障率预测;特征映射;S.M.A.R.T 属性
中图分类号:TP391 **文献标志码**:A

Research on Hard Disk Fault Rate Prediction Based on Random Forest

ZHANG Yongqiang[1,4], KONG Junjun[1], CUI Yao[2], LI Xiangnan[3,4]

(1.*School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang* 050018, *China*;
2.*Shijiazhuang Changhong Intelligent Technology Co., Ltd., Shijiazhuang* 050004, *China*;
3.*Shijiazhuang Chunxiao Internet Information Technology Co., Ltd., Shijiazhuang* 050061, *China*;
4. *Hebei Technology Innovation Center of Intelligent IoT, Shijiazhuang* 050018, *China*)

✉ zyq@hebust.edu.cn; kjunjun555@163.com; cuiyao@changhong.cc; xiangnan.li@chunxiao.net

Abstract: Aiming at hard disk faults which result in a large amount of data loss, this paper proposes a Random Forest-based method to predict hard disk faults and reduce the risk of data loss. Firstly, in terms of data processing, feature mapping preprocessing for the data used is performed. Secondly, by constructing and selecting Decision Trees, a Random Forest model is constructed to predict the range of hard disk fault rate based on the selected feature attributes, the changes of which reflect the changing trend of hard disk fault rate. Finally, the parameters of the constructed Random Forest model are optimized and tested with different n_estimators parameter values. The experimental results show that compared with methods such as XGBoost (Extreme Gradient Boosting) and LSTM (Long Short Term Memory), the $F1$ value (F-Measure) of the proposed method has increased by 0.93% and 1.84%, respectively. In addition, the parameter values of the Random Forest model are tested with different values, and the final accuracy reaches 98.18%, which is 1.23% higher than the default value. This proves that the proposed method can predict the hard disk fault rate more accurately and reflect the changing trend of the hard disk fault rate based on feature attributes.

Key words: Random Forest; hard disk fault rate; fault rate prediction; feature mapping; S.M.A.R.T attribute

0 引言(Introduction)

随着互联网、信息技术的发展,新时代数据中心也迅速发展,数据存储的数量呈指数增长,而硬盘驱动器是数据存储系统中常见的一种设备,大量的数据都存储在硬盘驱动器上[1]。

收稿日期:2023-10-27
基金项目:河北省自然科学基金(F2022208002);河北省高等学校科学技术研究重点项目(ZD2021048)

根据存储服务商 Backblaze 对硬盘使用情况的统计报告,硬盘经常出现故障,并且是最严重的一类硬件故障,它会导致大量的数据丢失,降低硬盘使用的可靠性。若能提前预测硬盘的寿命,进而对硬盘进行有针对性的维护,则会降低数据丢失的可能性。目前,自我监测、分析及报告技术(Self-Monitoring Analysis and Reporting Technology, S. M. A. R. T)可以对硬盘进行故障预测,它是一种自动的硬盘状态检测与预警系统和规范[2],但 S. M. A. R. T 阈值检测法仅能实现简单的磁盘故障评测,在达到 0.1%误判率(False Acceptance Rate, FAR)时,其故障检测率(False Discovery Rate, FDR)仅有 3%～10%,无法满足用户的实际需求。因此,本文提出一种随机森林预测模型,根据所选取的特征属性预测硬盘故障率所在的区间,并且根据特征属性的变化预测硬盘故障率的变化趋势,进而对硬盘故障率进行准确地预测。

1　相关工作(Related work)

近年来,众多的研究者利用机器学习和统计学对硬盘故障率预测问题开展了研究。姜少彬等[3]使用一种非监督对抗学习方法对硬盘进行故障预测,设计了一种基于长短期记忆神经网络的自编码器,并引入生成式对抗网络增强非监督学习。与传统监督和半监督的方法相比,尽管该模型在训练时不需要使用异常样本,避免了模型过拟合问题,但是该模型预测的准确率还有待提高。乔旭坤等[4]建立了基于机器学习的硬盘故障检测评估平台,在统一的实验平台中对随机森林、逻辑回归、多层感知神经网络、决策树、朴素贝叶斯、极端梯度提升树、梯度提升决策树和 AdaBoost 算法模型进行了故障预测性能比较,但实验中只针对同一公司的同一种型号的硬盘进行测试。李国等[5]根据精度和多样性值选取决策树并对其分配权重,组成变权重随机森林模型对硬盘进行故障预测,最终达到 93.12%的故障检测率和 0.008%的误报率。BASAK 等[6]讨论了所使用的长短期记忆神经网络的架构,描述了选择各种超参数的机制,并预测磁盘是否会在未来 10 天内发生故障,但是预测的精度不高。李顺等[7]提出一种基于深度学习长短期记忆神经网络(Long Short-Term Memory, LSTM)的硬盘剩余寿命预测方法。该方法相对于传统的机器学习方法能够捕获硬盘特征的序列信息,建立的 LSTM 模型可以在训练样本和测试样本上分别达到 0.27 和 1.85 的平均绝对误差(Mean Absolute Error, MAE),但是对硬盘寿命预测还没有达到更好的精度。XU 等[8]引入了一种基于递归神经网络(Recurrent Neural Network, RNN)的新方法,以基于逐渐变化的顺序 S. M. A. R. T 属性评估硬盘驱动器的健康状态。与简单的故障预测方法相比,健康状态评估在实践中更有价值,技术人员能够根据紧急程度安排不同硬盘驱动器的恢复。曹渝昆等[9]提出一种基于 LSTM 神经网络的风机齿轮带断裂故障预测方法,结合风电厂 SCADA(Supervisory Control and Data Acquisition)系统的风机运行状态监控数据,在随机森林算法的数据特征筛选基础之上,采用 LSTM 对齿轮带故障进行预测。刘雅卉等[10]对 ATM 机交易数据集进行交易特征提取,针对不同故障情景将数据分为正常-异常二分类,通过 Bootstrap 重抽样,建立多棵 CART (Classification and Regression Tree)决策树,形成随机森林模型,实现 ATM 机故障的诊断。基于上述研究,对硬盘故障的预测准确率还有待提高。可以根据硬盘 S. M. A. R. T 属性以外的属性进行预测,比如对硬盘的型号、使用天数等属性进行预测。此外,对随机森林模型的参数进行了调优,在原有的基础上可以提高预测的准确率,并且对所采用的数据做出特征映射[11]等预处理,通过本文方法可以更精确地预测硬盘故障率,反映出硬盘故障率基于特征属性的变化趋势。

2　随机森林预测模型(Random Forest prediction model)

随机森林模型是将许多棵决策树整合到一起成为森林并用来预测最终结果[12],根据所选取的特征属性预测硬盘故障率所在的区间,根据特征属性的变化预测硬盘故障率的变化趋势。随机森林预测模型包括决策树的构建、随机森林算法参数调优等内容,其模型图如图 1 所示。

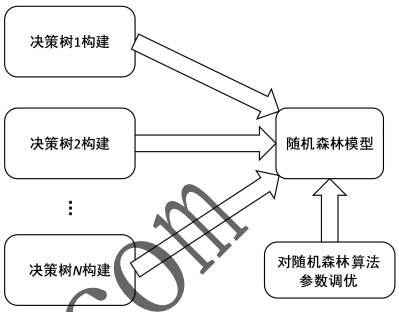

图 1　随机森林预测模型图

Fig. 1 Diagram of Random Forest prediction model

2.1　决策树的构建

使用随机森林模型预测硬盘故障率所在区间,主要根据所选取的特征属性进行分类预测。通过计算硬盘的品牌、型号、内存大小、使用数量、使用时间 5 个属性的基尼系数,通过计算对比选取决策树的根节点以及依次所选用的叶子节点。决策树的构建过程如图 2 所示。

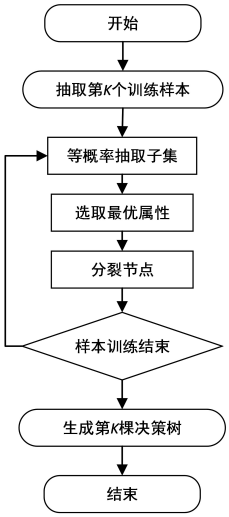

图 2　决策树的构建过程

Fig. 2 The construction process of Decision Tree

其中,对于决策树构建的节点,是通过计算所选属性的基尼系数选取的,基尼系数如公式(1)所示:

$$Gini(x) = 1 - \sum_{x=1}^{X} p(x)^2 \tag{1}$$

在构建决策树的过程中,分类算法计算所有可能的分裂带来的基尼系数,进而选择基尼系数小的分裂作为下一次的分

支。通过递归执行分支算法生成决策树模型，直到所有节点不满足分裂条件为止。

2.2 随机森林算法

随机森林模型是一种经典的 Bagging 模型，其弱学习器为决策树模型。随机森林模型在原始数据集中随机抽样，构成 n 个不同的样本数据集，然后根据这些数据集搭建 n 个不同的决策树模型，根据这些决策树的投票结果获得最终的分类结果[13]。

输入：训练数据 D，随机森林中决策树个数 N，选取的特征值个数 M。

输出：随机森林硬盘故障率预测模型。

建立的随机森林模型参数对预测结果很重要，为了更好地预测结果，本模型引用 GridSearchCV 对参数调优，预测结果的准确率在原有的基础上有一定程度的提高。关于随机森林算法参数调优的步骤如下。

(1)设置所有参数为默认值。建立基于随机森林的硬盘故障率预测模型，其中的参数默认设置值如表 1 所示。

表1 参数默认设置值

Tab.1 Parameter default setting value

参数	默认值	含义
n_estimators	10	随机森林预测模型分类器个数
random_state	None	代表一个随机种子
min_samples_split	2	内部节点再划分所需最小样本数
min_samples_leaf	1	叶子节点最小样本数
max_features	None	最大特征数
min_impurity_decrease	0	最小不纯度

最终的参数为默认值的预测结果，如表 2 所示，准确率为 96.95%，查准率为 91.10%，查全率为 91.08%，F1 值为 97.04%。

(2)创建 GridSearchCV 对象，并对参数 n_estimators 进行调优。对参数 n_estimators 调优不会增加模型的复杂度，对模型预测准确率的提升有帮助。

(3)对外层的 Bagging 框架进行参数调优，先对参数 n_estimators 进行调参，其他参数仍然设置为默认值。参数 n_estimators 的设置范围为 1～101，步长为 10。

(4)输出的较优准确率和最优参数与默认值对比如表 2 所示。

表2 参数调整后预测结果对比

Tab.2 Comparison of predicted results after parameter adjustment

n_estimators 值	准确率/%	查准率/%	查全率/%	*F*1 值/%
10	96.95	91.10	91.08	97.04
31	98.18	98.73	99.12	98.17

从对比结果来看，调参后，在原有准确率的基础上，预测的准确率有一定程度的提高。

2.3 硬盘故障率预测处理流程

硬盘故障率预测的处理流程如图 3 所示。根据数据集拥有的其他属性特征预测故障率特征的值，采用硬盘的品牌、型号、内存大小、使用数量、使用时间等特征值，构建随机森林预测模型，用于预测硬盘故障率。把数据集按照 8∶2 拆分成训练集与测试集，首先输入训练数据，其次使用经过预处理的数据形成训练模型进行故障率预测，最后输出预测结果。

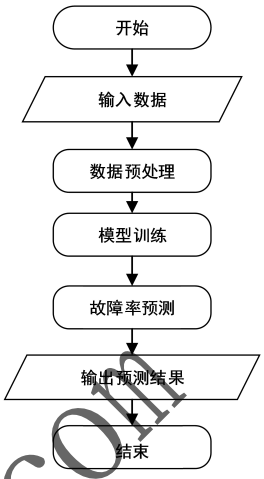

图 3 硬盘故障率预测的处理流程

Fig. 3 Process flow of hard disk fault rate prediction

3 实验及方法(Experiments and methods)

3.1 实验环境

操作系统：Windows 10；开发环境：PyCharm；程序语言：Python 3.9。

3.2 数据集与特征选择

实验数据选用 Backblaze 公开数据集，选取 2016—2021 年的汇总数据，数据集的特征属性包括硬盘的品牌、型号、数量、使用时间、内存大小、故障率。选取公开数据集用到的 6 个特征属性进行故障预测处理，数据集属性值对应表如表 3 所示。

表3 数据集属性值对应表

Tab.3 The corresponding table of dataset attribute value

Attributes Name	属性名
MFG	硬盘品牌
Model	硬盘型号
DriveSize	硬盘内存大小
DriveCount	硬盘数量
Drivehours	硬盘使用时间
Failures	硬盘故障率

3.3 实验数据处理

首先进行数据预处理，对原始数据中的缺失值进行填补，将其填充为 0；对于异常值，因为包含异常值的记录很少，所以可以直接删除包含异常值的记录。

其次进行特征映射，其中硬盘品牌转换规则如表 4 所示。硬盘故障率分为 13 个区间，其转换规则如表 5 所示。

表4　硬盘品牌转换规则

Tab.4 The conversion rules of hard disk brands

硬盘品牌	转换规则
HGST	1
Seagate	2
WDC	3
Toshiba	4

表 5　硬盘故障率转换规则

Tab.5 The conversion rules of hard disk faults

硬盘故障率区间	转换规则
[0,0.01)	1
[0.01,0.03)	3
[0.03,0.05)	5
[0.05,0.07)	7
[0.07,0.09)	9
[0.09,0.1)	10
[0.1,0.15)	13
[0.15,0.3)	25
[0.3,0.45)	35
[0.45,0.6)	55
[0.6,0.75)	70
[0.75,0.9)	85
[0.9,1)	95

3.4　评价指标

对算法进行评价,选择统一的评价指标,包括准确率(Accuracy)、查准率(Precision)、查全率(Recall)、查准率与查全率的调和平均值($F1$)、混淆矩阵[14]。

查准率也叫精度,简记为 P,表示预测为正例的样本中[TP(True Positive)+FP(False Positive)]有多少是真正的正样本(TP),其公式如下:

$$P=\frac{TP}{TP+FP} \tag{2}$$

查全率也叫召回率,简记为 R,表示在实际真正的正样本中[TP(True Positive)+FN(False Negative)]预测为正例的样本数(TP)所占的比例,其公式如下:

$$R=\frac{TP}{TP+FN} \tag{3}$$

$F1$ 值是查准率与查全率的调和平均值,它是一个广为接受的指标,其公式如下:

$$F1=\frac{2PR}{P+R} \tag{4}$$

混淆矩阵是对分类问题的预测结果的总结,是衡量分类型模型准确率中最基本、最直观且计算最简单的方法。使用计数值汇总正确和不正确预测的数量,并按照每个类别进行细分,这是混淆矩阵的关键。混淆矩阵显示了分类模型在进行预测时会对哪一部分产生混淆,不仅能了解分类模型所犯的错误,更重要的是可以了解错误的类型。正是这种对结果的分解,克服了仅使用分类准确率衡量本文模型带来的局限性。

本实验通过随机森林硬盘故障率预测模型运行的结果混淆矩阵和准确率如图 4 所示。从图 4 可以看出混淆矩阵统计的本文随机森林模型的正确预测和不正确预测的数量,并且不正确预测的数量较少。

```
[[ 88   4   1   0   0   0   0   0   0   0   0   0   0   0]
 [  2 115   0   0   0   0   0   0   0   0   0   0   0   0]
 [  0   0  46   0   0   0   0   0   0   0   0   0   0   0]
 [  0   1   0  65   1   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0  22   0   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0  13   0   0   0   0   0   0   0   0]
 [  0   0   0   0   0   1  19   0   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0  16   0   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0  71   0   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0  17   0   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0  25   0   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0  30   0   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   6   0]
 [  0   0   0   0   0   0   0   0   0   0   0   0   0   9]]
0.9818840579710145
```

图 4　随机森林模型预测结果

Fig. 4 Prediction results of Random Forest model

4　实验结果及分析(Experimental results and analysis)

本文随机森林模型与 XGBoost 算法[15]和长短期记忆神经网络(LSTM)模型[16]的评价指标对比如表 6 所示,从中可以看出,构造的随机森林方法 $F1$ 值达到 98.17%,相比 XGBoost 和 LSTM 算法效果较好。

表6　各模型的预测效果对比

Tab.6 Comparison of prediction effects of various models

算法模型	P/%	R/%	$F1$ 值/%
XGBoost	91.73	91.22	97.24
LSTM	96.42	96.33	96.33
随机森林(本文)	98.73	99.12	98.17

通过对特征值进行特征映射等预处理,并且对随机森林预测模型的参数进行调参,在设置默认参数的基础上,通过对参数的调优,提高了模型预测的准确率。取得的预测结果如表 7 所示,从中可以看出,对随机森林预测模型进行参数调整后,其预测的准确率达到 98.18%,相比之前明显提高。

表7　不同 n_estimators 值预测的准确率对比

Tab.7 Comparison of the prediction accuracy of different n_estimators values

数据集	算法模型	n_estimators 值	准确率/%
Backblaze 数据集	随机森林	10(默认值)	96.95
Backblaze 数据集	随机森林	91	96.53
Backblaze 数据集	随机森林	31(本文)	98.18

根据 ROC(Receiver Operating Characteristic)曲线衡量模型的预测效果。ROC 曲线与横轴围成的面积大小称为学习器的 AUC(Area Under ROC Curve),该值越接近于 1,说明这个模型的预测效果越好。随机森林预测模型的 ROC 曲线图如

图5所示，横坐标和纵坐标分别为反正例率(False Positive Rate，FPR)、真正例率(True Positive Rate，TPR)。从图5可以看出，随机森林预测模型的效果较好。本文方法是基于多分类的分类模型，由表5可知，预测结果的种类分为13种，所以ROC曲线图中“ROC curve of class”1～13对应的是13种分类结果。

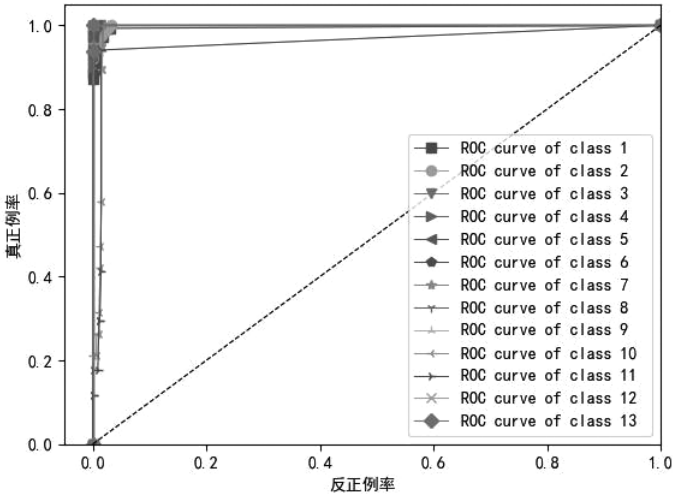

图5 ROC曲线图

Fig. 5 ROC curve

实验对基于随机森林的算法进行硬盘故障率预测，采用Backblaze公开数据集，选取2016—2021年的数据，硬盘的型号、数量、使用时间、内存大小等特征作为自变量，硬盘故障率作为因变量。在原有的预测结果是否为故障的基础上，通过基于随机森林模型的方法对硬盘故障率的所在区间进行预测，结果反映出特征属性变化后其硬盘故障率的变化，并且对随机森林预测模型参数进行调优后，最终预测准确率可达到98.18%。

5 结论(Conclusion)

数据是信息时代的宝贵资源，硬盘作为保存数据的主要部件，提高硬盘的可靠性对提高数据存储的安全具有重要意义。为了解决硬盘出现故障导致数据丢失的问题，提出了基于随机森林模型的方法对硬盘进行故障率预测。实验结果表明，与XGBoost、LSTM的方法相对比，基于随机森林的方法的精度提高了1%～2%，并且对模型的n_estimators参数进行调参后，最终预测准确率可达到98.18%，所以可以对硬盘的故障情况进行有效的预测。本研究可以对硬盘故障率的数据划分区间，并且对区间进行特征映射预处理，相比传统的方法，在对硬盘故障预测结果为是否故障的情况下，更容易反映出硬盘故障率的变化。

本实验虽然可以很好地对硬盘故障进行预测，但是对于故障率区间转换的细化程度还需进一步优化。下一步工作将致力于更准确地预测硬盘故障率具体值，并且提高模型的泛化能力和预测精度。

参考文献(References)

[1] LI J，STONES R J，WANG G，et al. Hard drive failure prediction using Decision Trees[J]. Reliability engineering & system safety，2017，164：55-65.

[2] 万成威，王霞，王猛. 基于SMART数据模式的HDD硬盘状态预测方法[J/OL]. 电讯技术，2022：1-6[2024-1-15]. https://kns.cnki.net/kcms/detail/51.1267.TN.20221118.1710.002.html.

[3] 姜少彬，杜春，陈浩，等. 一种硬盘故障预测的非监督对抗学习方法[J]. 西安电子科技大学学报，2020，47(2)：118-125.

[4] 乔旭坤，李顺，李君，等. 基于机器学习的硬盘故障预测研究[J]. 计算机技术与发展，2022，32(6)：215-220.

[5] 李国，常甜甜，李静. 基于变权重随机森林的硬盘故障预测方法[J]. 计算机工程与设计，2021，42(10)：2988-2994.

[6] BASAK S，SENGUPTA S，DUBEY A. Mechanisms for integrated feature normalization and remaining useful life estimation using LSTMs applied to hard-disks[C]//IEEE. Proceedings of the IEEE：2019 IEEE International Conference on Smart Computing(SMARTCOMP). Piscataway：IEEE，2019：208-216.

[7] 李顺，李君，吴鑫，等. 基于LSTM的硬盘剩余寿命预测[J]. 浙江万里学院学报，2020，33(4)：69-77.

[8] XU C，WANG G，LIU X G，et al. Health status assessment and failure prediction for hard drives with recurrent neural networks[J]. IEEE transactions on computers，2016，65(11)：3502-3508.

[9] 曹渝昆，巢俊乙，王晓飞. 基于LSTM神经网络的风机齿轮箱断裂故障预测[J]. 电气自动化，2019，41(4)：92-95.

[10] 刘雅卉，滕志霞. 基于随机森林的ATM机监测预警方法[J]. 电子技术与软件工程，2018(12)：162-164.

[11] 程森海，楼俏，王琼，等. 基于随机森林算法的配网抢修故障量预测方法[J]. 计算机系统应用，2016，25(9)：137-143.

[12] 吕红燕，冯倩. 随机森林算法研究综述[J]. 河北省科学院学报，2019，36(3)：37-41.

[13] 王梓杰，周新志，宁芊. 基于PCA和随机森林的故障趋势预测方法研究[J]. 计算机测量与控制，2018，26(2)：21-23，26.

[14] LI J，STONES R J，WANG G，et al. New metrics for disk failure prediction that go beyond prediction accuracy[J]. IEEE access，2018，6：76627-76639.

[15] 王陶，吴鑫，李君，等. 基于XGBoost算法的硬盘故障预测[J]. 数字技术与应用，2021，39(2)：123-126.

[16] 伍乙杰，黄文灏，赖仕达，等. 基于随机森林和双向长短期记忆网络的超短期负荷预测研究[J]. 电气自动化，2022，44(5)：35-37，40.

作者简介：

张永强(1981-)，男，博士，副教授。研究领域：人工智能，电磁防护理论与技术。

孔君君(2000-)，女，硕士生。研究领域：人工智能。

崔 摇(1982-)，男，工程师。研究领域：物联网，企业数字化管理，传感网。

李向南(1985-)，男，硕士。研究领域：物联网应用，边缘智能研究。