Yunchuan GUAN, Yu LIU, Ke ZHOU, Qiang LI, Tuanjie WANG, Hui LI, 2023. A disk failure prediction model for multiple issues. *Frontiers of Information Technology & Electronic Engineering*, 24(7):964-979.
https://doi.org/10.1631/FITEE.2200488

A disk failure prediction model for multiple issues

Key words: Storage system reliability; Disk failure prediction; Self-monitoring analysis and reporting technology (SMART); Machine learning

Corresponding author: Yu LIU
E-mail: liu_yu@hust.edu.cn
ORCID: https://orcid.org/0000-0002-1964-9278

Issues of disk failure prediction

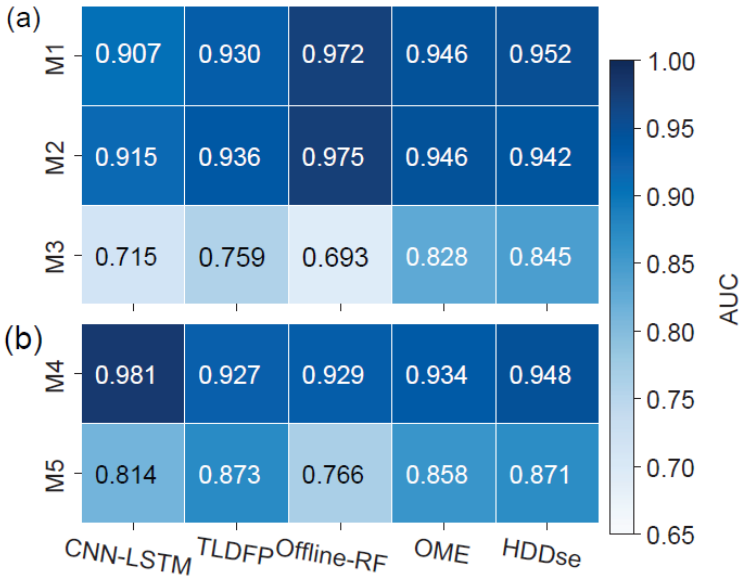
- **Heterogeneous disks**. This issue arises from differences in the distribution of model-specific SMART attributes. These distribution differences can introduce biases into the prediction model and reduce its prediction performance for each disk model.
- **Model aging**. The distribution of SMART attributes can exhibit differences over time, even for the same disk model. This phenomenon results in the model aging issue.

Issues of disk failure prediction

- **Environmental variation.** Recently, researchers found that location markers and the neighborhood disk information may affect the prediction accuracy of machine learning (ML) models. This implies that disks in different environments may manifest different distributions of SMART attributes.
- **Minority samples.** Using traditional ML algorithms with the training data of minority disks would dramatically increase the risk of over-fitting or poor generalization, which would weaken the performance of predictive models.

Motivation

- The issues of disk failure prediction exist in a mixed manner.
- Although there are effective learning methods for the issues of heterogeneous disks, model aging, environmental variation, and minority samples, each method claims to be good at solving only one or two issues.
- Experimental results show that none of the prediction models can be applicable to both datasets (a) and (b), as given in the figure.



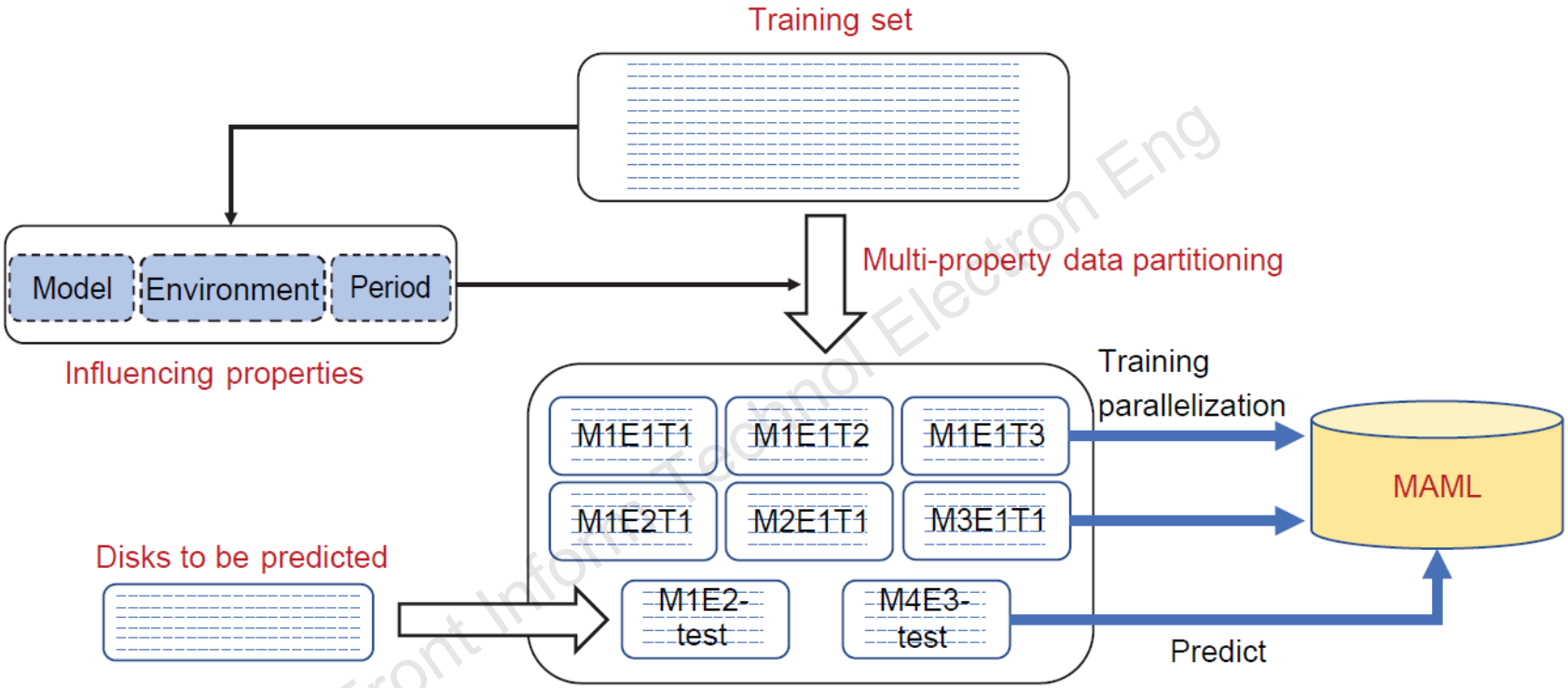
Main idea

- We find that the properties of the period, disk models, and environments have similar effects on the distribution of SMART data. As a result, we argue that the issues of heterogeneous disks, model aging, and environmental variation originate from the same problem, i.e., the difference in the distribution of SMART data, which we call data heterogeneity.
- We find that the issue of minority samples does not exist independently; it is caused by unplanned data partitioning and must therefore be considered in the context of the multi-issue scenario. In addition, we draw a qualitative conclusion; i.e., it is the number of failed disks rather than the number of disks that determines the minority sample status.

Main idea

- We find that the learning process of heterogeneous disks—data partitioning and transfer learning—can be applied directly to solve the problem of data heterogeneity.
- We determine the pattern for partitioning the data based on multiple issues, i.e., multi-property data partitioning (MDP).
- In terms of transfer learning, we treat the partitioned data as the tasks of multi-task learning. As a result, we introduce model-agnostic meta-learning (MAML), which can provide unified learning for multiple tasks.

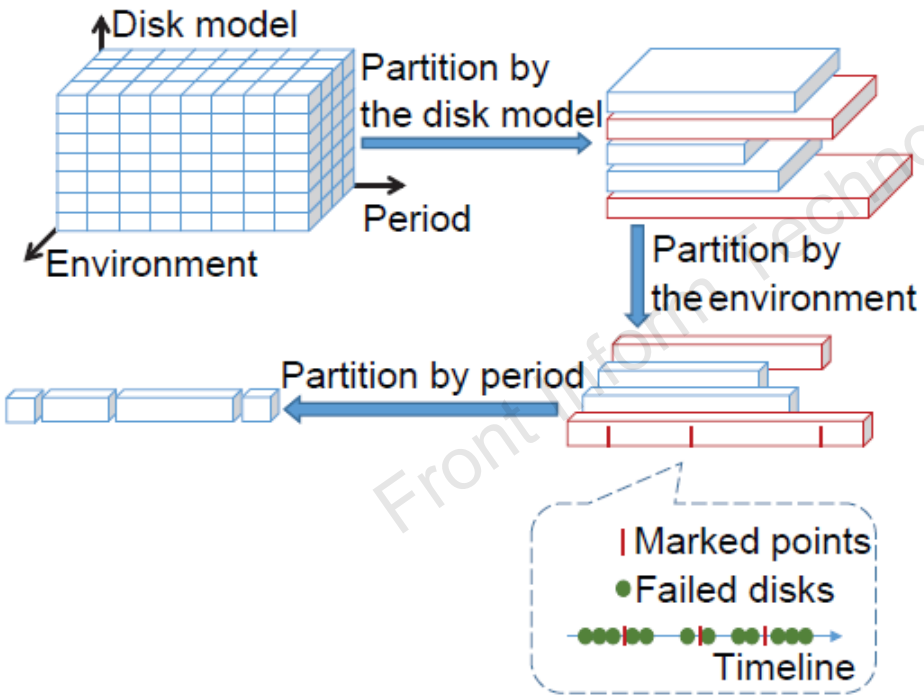
Framework



Overview of the model. By multi-property data partitioning (MDP), the training set is partitioned into multiple subsets by the properties corresponding to the issues. For example, M1E1T1 denotes the subset consisting of disk samples from a certain disk model, a certain environment, and a certain period. M denotes the disk model, E denotes the environment, and T denotes the time period.

Method

1. Multi-property data partitioning



By MDP, larger subsets are continuously selected for partitioning based on multiple properties, until the number of failed disks in each subset is smaller than a critical value. Note that the partitioned subsets consist of both healthy and failed disk samples, with the former accounting for the majority of the population.

Method

2. Multi-task learning algorithm—MAML

MAML optimizes the process of "finetuning on a task" on multiple tasks to achieve quick adaptation to unknown tasks. When learning on a task, MAML is concerned with the process of finetuning on that task, instead of the domain of that task.

Training process:

Algorithm 1 Training process of MAML

Initialize task set $\{\text{task}_i\}$ and meta-learner f_ϕ
Initialize learning rates α and η
while not done **do**
  for all task_i **do**
    $f_{\theta_i} = f_\phi$
    $f_{\theta_i'} = f_{\theta_i} - \eta \nabla f_{\theta_i}(S_i)$
    Calculate loss of fine-tuning $f_{\theta_i'}(Q_i)$
  end for
  $f_\phi = f_\phi - \alpha \sum_{i=1}^{K} \nabla f_{\theta_i'}(Q_i)$
end while

Prediction process:

The process of finetuning and evaluation.

$$f_{\phi^{*\prime}} = f_{\phi^*} - \eta \nabla f_{\phi^*}(S_{\text{test}})$$

$$\overline{y} = f_{\phi^{*\prime}}(Q_{\text{test}})$$

Major results

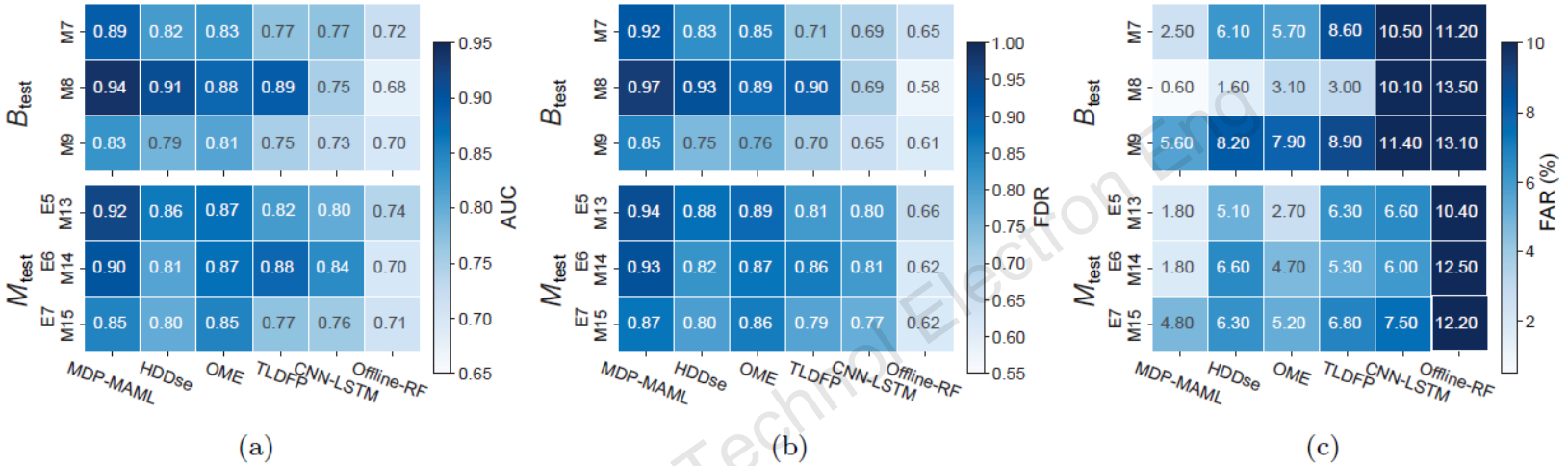


Fig. 10 Comparisons of performance with state-of-the-art disk failure prediction methods on the issues of heterogeneous disks and environmental variation: (a) AUC; (b) FDR; (c) FAR

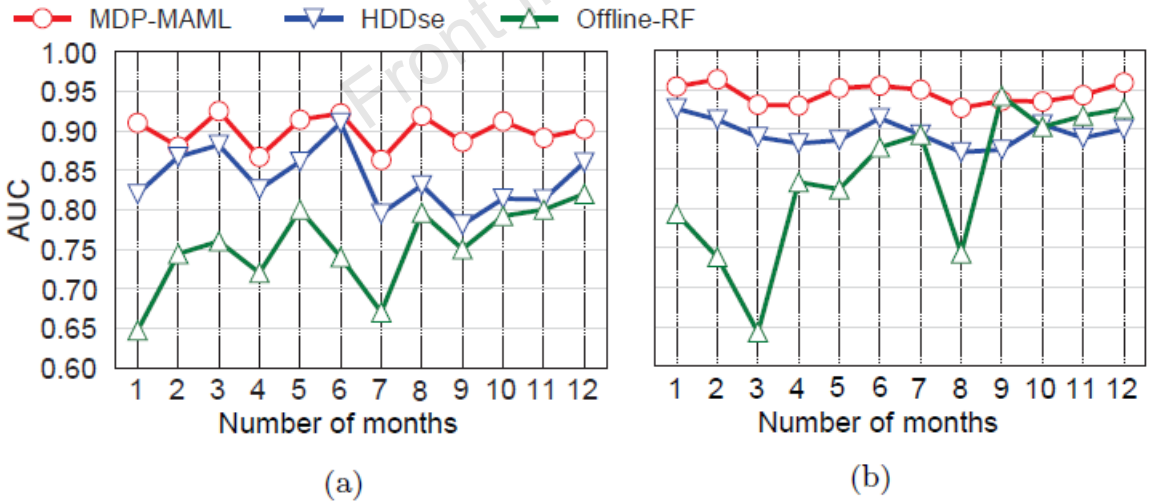


Fig. 11 AUC for the issue of model aging on M7 (a) and M8 (b)

Major results

Table 5 Average performance of the prediction model on the testing set

Model	AUC	FDR	FAR (%)
MDP-MAML	0.89	0.91	2.85
HDDse	0.83	0.84	5.65
OME	0.85	0.85	4.88
TLDFP	0.81	0.79	6.48
CNN-LSTM	0.77	0.74	8.68
Offline-RF	0.71	0.62	12.15

Table 7 Effectiveness of MDP

Property			AUC (mean)
Model	Period	Environment	
✓	✓	✓	0.912
	✓	✓	0.793
✓		✓	0.907
✓	✓		0.856

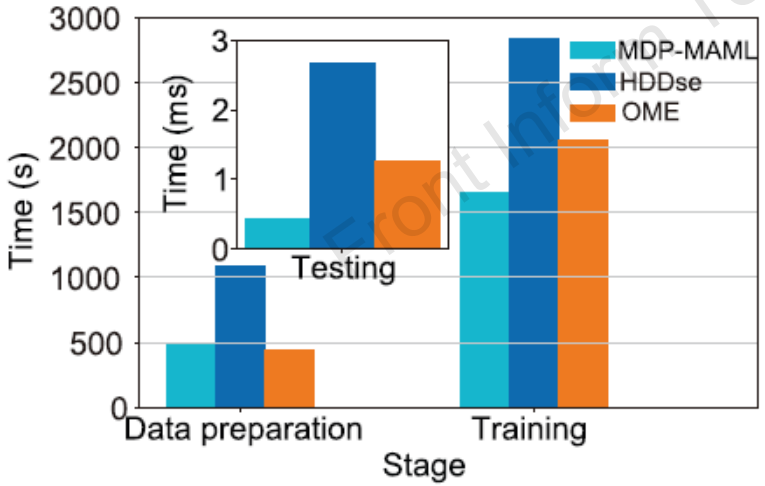


Fig. 12 Comparison in overhead

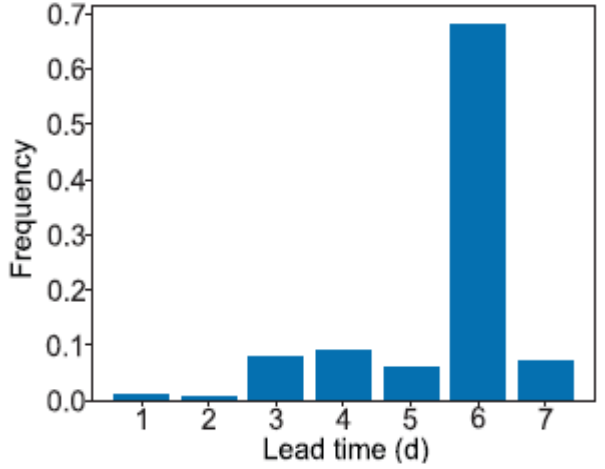


Fig. 14 Lead time of MDP-MAML

Conclusions

- We discover the common nature in the different issues of heterogeneous disks, model aging, and environmental variation, i.e., data heterogeneity.
- We summarize the commonality of different solutions to data heterogeneity and build a unified solution pattern for data heterogeneity, i.e., partitioning data and transfer learning.
- We propose MDP and introduce MAML to overcome the shortcomings of data partitioning and transfer learning methods used for disk failure prediction in a multi-issue scenario.

Yunchuan GUAN is a PhD candidate in WNLO, HUST, China. His main research interests include machine learning, disk reliability, and intelligent storage.

Yu LIU received his PhD degree from HUST, China, in 2017. He is an associate researcher of the School of Computer Science and Technology, HUST. He has more than 20 publications in international journals and conferences, including TOC, TDS, SIGMOD, DAC, IJCAI, ACM MM, ICME, ICMR, APWeb-WAIN, PR, and FGCS. His main research interests include similarity-hash-based smart storage, dark data, and AIOps for data and systems.

Ke ZHOU received his BE, ME, and PhD degrees in computer science and technology from HUST, China, in 1996, 1999, and 2003, respectively. He is a professor of the School of Computer Science and Technology and WNLO, HUST. He has more than 50 publications in international journals and conferences, including TPDS, PEVA, FAST, ATC, MSST, MM, INFOCOM, SYSTOR, MASCOTS, and ICC. He is a member of IEEE and USENIX. His main research interests include computer architecture, cloud storage, parallel I/O, and storage security.