Int. J. Nonlinear Anal. Appl. 15 (2024) 7, 28–32
ISSN: 2008-6822 (electronic)
http://dx.doi.org/10.22075/ijnaa.2023.31039.4551

A summary of approaches to identify hard disk failure through the utilization of machine learning algorithms

Somayeh Askarpour, Maryam Saberi Anari

Department of Computer Engineering, Technical and Vocational University (TVU),Tehran, Iran

(Communicated by Seyyed Mohammad Reza Hashemi)

Abstract

This article delves into the techniques employed for identifying failures in hard disks through the utilization of machine learning algorithms. Hard disks serve as essential components within computer systems, and as they age and undergo repetitive usage, they may manifest indications of failure or inadequate performance, culminating in data loss and system malfunction. Consequently, the early detection and anticipation of hard disk failures are of utmost significance. Recent advancements in machine learning methods have enabled the precise detection of hard disk failures within a short timeframe. Within this investigation, we explore the foundational concepts pertaining to hard disks and their failures. We scrutinize various machine learning algorithms employed for the detection of hard disk failures. Furthermore, we introduce performance evaluation metrics for failure detection models. The challenges and limitations in the detection of hard disk failures are discussed, along with potential strategies for enhancing system performance and accuracy.

Keywords: Hard drives, Failures, Identification, Machine learning, Deep neural networks
2020 MSC: 68T07

1 Introduction

A hard disk is one of the important components of a computer that can become damaged due to continuous use. Machine learning algorithms [2] are used to detect hard disk failures. Despite technologies like SMART (Self-Monitoring, Analysis and Reporting Technology) that have been developed to predict the reliability of hard disk drives, the possibility of predicting and increasing reliability in hard disks is enhanced. These technologies monitor the status of the hard disk through sensors and data collection, indicating the current failure, but no intelligent analysis is performed by combining multiple sensor values. However, predictive analysis of sensor values can increase the security of storage systems by replacing hard drives before they fail.

SMART records the performance values of the hard disk in 62 attributes and sets a specific threshold for each attribute. If a feature exceeds its threshold value, the drive is recognized as faulty. Hard disk failures can be classified into two categories: predictable and unpredictable. Predictable failures can be detected over time before they occur, but unpredictable failures happen suddenly and rapidly and cannot be detected using SMART features. The percentage of predictable drive failures is 60%.

Extensive research has been conducted on detecting hard disk failures using machine learning algorithms. Currently, several research papers exist in the field of hard disk drive failure detection, where their results are evaluated using

Email addresses: somayeh.askarpour@gmail.com (Somayeh Askarpour), saberi-m@tvu.ac.ir (Maryam Saberi Anari)

Received: March 2023 *Accepted:* June 2023

binary classification to predict whether the hard drive will fail or not. This article examines some of these relevant studies conducted in recent years, aiming to identify the most accurate and widely used algorithms for detecting hard disk failures using machine learning algorithms.

The structure of this work is divided into four main sections. Section 2 describes the previous research studies and analyzes their results. In Section 3, the studied and analyzed results from the previous section are discussed, along with existing opinions and interpretations. Finally, Section 4 provides a conclusion from the literature review and presents recommendations for future research.

2 An overview of prior research

Hamerly and colleagues in their article [3] have examined the failures of computer hard drives and have pointed out that hard disk failures can incur significant costs in system support. In this study, two Bayesian methods for predicting drive failures using internal drive conditions measurements have been investigated. Initially, the problem was examined from the perspective of anomaly detection. A mixture model of naive Bayes submodels using the expectation-maximization algorithm was introduced.

The second method is a naive Bayes classifier that has a supervised learning approach. Both methods were tested using real data related to 1936 drives. The prediction accuracy of both algorithms is much higher than the accuracy of threshold-based methods used in the hard drive industry. Their data was provided by Quantum Inc. and includes 1927 healthy disks and 9 faulty disks.

iIn calculating the rates, equation (1) has been used to identify the correct disk hard failure.They achieved a fault detection rate of approximately 35% to 40% for the NBEM method and around 55% for the simple Bayesian classifier with a 1% false alarm rate.

$$tpr = \frac{tp}{tp + fn} = \frac{tp}{9},\ fpr = \frac{fp}{fp + tn} = \frac{fp}{1934}. \tag{1}$$

The presented failure prediction methods here outperform the current standard industry methods and are sufficiently efficient to be useful in practice. The current standard methods in the disk drive industry are estimated to operate with a high actual positive rate at the drive level (0.04 to 0.11) and a false positive rate of 0.002 [5].

At the same false positive rate, NBEM achieves an actual positive rate of 0.30, which is nearly three times better. In this study, the authors also examine their results from another perspective and assume that the drives are an array of inexpensive drives in a high-stress environment with 128 drives in the Quantum dataset. For anomaly detection, NBEM uses three features and issues an alert for 0.05% unusual readings, issuing an alert once every 16 hours. With this method, a real failure alert is predicted with a probability of 0.41, and there will be a false alert with a probability of 0.59. The false alarm rate at the drive level is acceptable due to the good design of the inexpensive drive array.

Jiang Xiao and his colleagues in their article [9] have addressed the challenges related to hard disk drives in data centers and storage system development. They emphasize that research on disk failure prediction using machine learning methods focuses on SMART features. However, most of these studies rely on offline training and face the ”model aging” problem. To solve this problem, this study introduces an innovative model based on Online Random Forests (ORFs). The formulas used in calculating the forest of equations are (2) and (3).

$$G(D) = p_0(1 - p_0) + p_1(1 - p_1) \tag{2}$$

$$\Delta G(D, s) = G(D) - \frac{|D_{ls}|}{|D|}G(D_{ls}) - \frac{|D_{rs}|}{|D|}G(D_{rs}). \tag{3}$$

This model has the capability to adapt to changes in the SMART distribution over time and provides better prediction performance compared to offline models. Experiments on real-world data demonstrate that the ORF model quickly converges to offline random models and achieves a low false positive rate while maintaining a high failure detection rate. Additionally, it shows the ability to preserve long-term prediction performance for use in data centers.

Shen and his colleagues in their research [6] state that in large-scale data center environments, various types of drives and models with different input/output patterns are used for diverse applications. This creates significant differences in the types of drive failures, which in turn is an important challenge in hard disk failure prediction. In this article, a failure prediction method for hard disk drives is presented based on a random forest model with partial voting, which separates failure predictions. In this method, the partitioning of nodes in the tree is performed using a small number of features that are randomly selected. Each node is divided based on the best feature (f_b) and the best splitting point (t) that maximizes the expected information gain. The expected information gain is calculated using equation (4).

$$\Delta_{info} = -\frac{S_{i1}}{S_i} info(S_{i1}) - \frac{S_{i2}}{S_i} info(S_{i2}). \tag{4}$$

Validation experiments are conducted on two real-world datasets containing 64,193 SMART drive data. The results of the experiments show that the proposed method has better prediction capability compared to current advanced methods.

The prediction method in this article has been tested using real-world data, and empirical results demonstrate that the random forest-based approach outperforms other methods. For the ”B” family, this method achieves a false detection rate (FDR) of 97.67% with a false alarm rate (FAR) of 0.017%. For the ”S” family, it achieves a 100% detection rate with a false alarm rate of 1.764%, and for the ”T” family, it achieves a 94.89% detection rate with a false alarm rate of 0.44%.

Hughes and his colleagues in their research [4] have focused on studying improved SMART algorithms. These enhanced algorithms aim to predict failures in individual hard disk drives as closely as possible to prevent data loss. Experimental tests have shown that SMART alone provides moderate accuracy in predicting failures at low false alarm rates. The proposed improvements in these algorithms have increased the prediction accuracy for unexpectedly failing drives by 3-4 times. However, the highest achievable accuracy is approximately in the range of 40%-60%.

Murray and his colleagues in their article [5] have utilized internal monitoring features of each drive. In fact, this research examines the detection of rare events in a time series of noisy and non-parametric data.

A new algorithm based on the multi-instance learning framework and naive Bayes classifier has been developed, which has shown good performance in reducing high error rates. The results of this research indicate that non-parametric statistical tests should be considered in learning problems involving the detection of rare events.

In this research, the model was tested with an experiment on 3780 drives with a failure rate of 0.9%. The test results showed an acceptable performance with a detection rate of 60% and a false reporting rate of 0.5%. The researchers have used several methods, including the sum rank test, support vector machine (SVM), and unsupervised clustering. The dataset in this research demonstrated the best performance with SVM, achieving an FDR of 50.6% and FAR of 0%.

In their research [7], Wang, Y, and their colleagues mention the important point that online monitoring of hard disk health can provide valuable information about the hardware failure trend.

In their research, they have developed an approach for detecting hardware drive anomalies using the Mahalanobis distance (MD) equation (5).

$$Z_{ij} = \frac{(x_{ij} - \overline{X}_j)}{S_j}, i = 1, 2, \ldots, m, j = 1, 2, \ldots, n. \tag{5}$$

By employing Feature Mechanisms, Modes, and Effects Analysis (FMMEA), and with the help of the minimum Redundancy Maximum Relevance (mRMR) method, vital parameters have been selected. The results of this research show that approximately 67% of failures in failed drives can be detected with a zero error rate, providing users with at least 20 hours of time to back up their data.

Wang et al. in their paper [8] have presented a two-stage parametric approach for predicting sudden HDD failures using statistical models. The proposed two-stage solution includes anomaly detection and failure prediction. Initially, the Mahalanobis distance is used to aggregate the monitored variables, which are then transformed into Gaussian variables using the Box-Cox transformation.

By determining the threshold, anomalies in the HDD are identified. In the second step, a relative in-place test is proposed to monitor the progress of anomalies in the HDD. In this study, a new cost function has been derived to adjust the prediction rate. This is crucial as it can balance the failure detection rate and false alarm rate, providing advanced warning to users about HDD failures so that they can back up their data in a timely manner. This method has been applied to both synthetic and real HDD datasets, and the results of this research demonstrate that this approach has performed better than other advanced methods with a 68% failure detection accuracy and 0% false alarm rate.

Chang Xu et al. in their study [10] point out that most of these studies have focused on predicting hard disk failures and determining the status of a hard disk as ”healthy” or ”unhealthy” by labeling it. This article presents a new method based on recurrent neural networks [1] (RNN) equations (6) and (7) to evaluate the health status of hard disk drives. The method utilizes the gradual changes in sequential SMART features to assess the disk drive's health.

$$U(t+1) = U(t) + i(t)e_h(t)^T\alpha - U(t)\beta \tag{6}$$

$$R(t+1) = R(t) + h(t-1)e_h(t)^T\alpha - R(t)\beta. \tag{7}$$

The important point is that in real storage systems, hard disks gradually deteriorate until sudden failure.

3 Conclusion

Machine learning algorithms are widely used in many computer vision and machine learning applications for detecting hard disk failures due to their high efficiency and prominent results. However, most existing algorithms are offline, which limits their usability for practical purposes.

Consequently, the design of online learning algorithms for detecting hard disk failures is highly beneficial because they can work with sequential training data or dynamic base distributions. These algorithms combine ideas from online learning, extremely random forests, and online decision tree growth methods, and they incorporate time-weighting techniques to add and discard new trees in the algorithm. By using these algorithms, excellent results can be achieved in detecting hard disk failures, and they can be applied in various practical applications.

In comparison to offline algorithms, online learning algorithms achieve better accuracy and performance due to their dynamics in working with temporal, sequential, or adaptable data. Additionally, these algorithms can be used in various applications such as visual tracking, interactive segmentation, and more.

By using these algorithms, the status of a hard disk can be continuously and in real-time monitored, and preventive measures can be taken against any potential risks. Furthermore, online learning algorithms for detecting hard disk failures are effective in reducing the costs associated with hard disk failures.

By detecting hard disk failures early on, one can quickly proceed with repair or replacement, preventing data loss and high expenses resulting from hard disk failures. Considering that hard disks are essential components of computer systems and widely used in many applications, the development of online learning algorithms for detecting hard disk failures is crucial and vital.

References

[1] M. Fadavi Amiri, M. Hosseinzadeh, and S.M.R. Hashemi, *Improving image segmentation using artificial neural networks and evolutionary algorithms*, Int. J. Nonlinear Anal. Appl. Article in Press, doi: 10.22075/IJNAA.2023.30232.4371

[2] A. Ghanbari Sorkhi, M. Iranpour Mobarakeh, S.M.R. Hashemi, and M Faridpour, *Predicting drug-target interaction based on bilateral local models using a decision tree-based hybrid support vector machine*, Int. J. Nonlinear Anal. Appl. **12** (2021), no. 2, 135–144.

[3] G. Hamerly and C. Elkan, *Bayesian approaches to failure prediction for disk drives*, ICML. **1** (2001), 202–209.

[4] G.F. Hughes, J.F. Murray, K. Kreutz-Delgado, and C. Elkan, *Improved disk-drive failure warnings*, IEEE Trans. Reliab. **51** (2002), no. 3, 350–357.

[5] J.F. Murray, G.F. Hughes and K. Kreutz-Delgado, *Machine learning methods for predicting failures in hard drives: A multiple-instance application*, J. Mach. Learn. Res. **6** (2005), 783–816.

[6] J. Shen, J. Wan, S.J. Lim, and L. Yu, *Random-forest-based failure prediction for hard disk drives*, Int. J. Distrib. Sensor Netw. **14** (2018), no. 11.

[7] Y. Wang, Q. Miao, E.W. Ma, K.L. Tsui and M.G. Pecht, *Online anomaly detection for hard disk drives based on Mahalanobis distance*, IEEE Trans. Reliab. **62** (2013), no. 1, 136–145.

[8] Y. Wang, E.W. Ma, T.W. Chow, and K.L. Tsui, *A two-step parametric method for failure prediction in hard disk drives*, IEEE Trans. Industr. Inf. **10** (2013), no. 1, 419–430.

[9] J. Xiao, Z. Xiong, S. Wu, Y. Yi, H. Jin and K. Hu, Disk failure prediction in data centers via online learning, Proc. 47th Int. Conf. Parallel Process., 2018, pp. 1–10.

[10] C. Xu, G. Wang, X. Liu, D. Guo, and T.Y. Liu, *Health status assessment and failure prediction for hard drives with recurrent neural networks*, IEEE Trans. Comput. **65** (2016), no. 11, 3502–3508.