International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-8, June 2020

Predicting Reliability of Storage Systems

Rahul Nandgave, Amar Buchade

Abstract*: Large Organizations have to make use of various storage devices like HDD and SDD to provide storage of information of their clients as well as themselves. These Storage devices are present in large numbers and are the basic building blocks that are used to store information and in case of failure occurs then replacing these devices can halt some services which can cause loss to the Organization in terms of money and time as well. To remediate this we can monitor each of the storage devices, as these storage devices come with a SMART (Self Monitoring and Reporting Technology) system that monitors and reports the stats back to the user. Thus with the help of these SMART Parameters we can train a machine learning model to predict if the hard disk will experience failure in the near future or not. In this study we did a survey of various techniques are based on various machine learning models and provide a brief overview of each of the techniques. Among these techniques we find that random forest and deep learning methods provide better results than the other methods discussed in various studies.

***Keywords* : Failure Detection, Machine Learning, Storage Devices, SMART Parameter..**

I. INTRODUCTION

These days, the technology provides a way for people to share content over the internet and communicate with each other with ease. Over the internet, people share their thoughts in the form of text as well as media. The security and integrity of this data is important as it belongs to the individuals. This data also needs to be stored somewhere safe. Organizations that provide such services also provide numerous ways of storing and handling such data. All of these ways basically boil down to making use of physical storage system such as hard disks, tape drives, solid State Drives etc. These Storage Devices vary based on the manufacturers and each manufacturer provides a wide range of models to choose from. All these storage devices provide SMART parameters which are also prone to vary based on the manufacturer. SMART stands for Self Monitoring and Reporting Technology. Almost all manufacturers provide some common set of these parameters, more or less there are some common set of these parameters. The SMART Parameters can be extracted by using disk utilities program that are provided by different operating systems.

Revised Manuscript Received on June 30, 2020.
***** Correspondence Author
Rahul Nandgave*, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India. Email: vedh.n513@gmail.com
Dr. A. R. Buchade, Department of Computer Engineering, Pune Institute of Computer Technology, Pune, India. Email: arbuchade@pict.edu

For example, in Linux we can install and use SMARTMON tools utility that, lets us extract the SMART parameters from the underlying disks [13].
Similarly in windows operating system we have several utilities that can help us find the SMART parameters. These parameters may vary a little from manufacturer to manufacturer but most of these are similar. Some of these parameters are shown in table 1 with SMART Parameters and meaning of each one.

Table 1: SMART Attributes

SMART Parameter	Meaning
SMART 1	Read Error Rate
SMART 3	Spin Up Time
SMART 4	START/STOP Count
SMART 5	Reallocated Sectors Count
SMART 7	Seek Error Rate
SMART 9	Power-On Hours
SMART 10	Spin Retry Count
SMART 12	Power Cycle Count
SMART 184	End-to-End error/ IOEDC
SMART 187	Reported Uncorrectable Errors
SMART 188	Command Timeout
SMART 189	High Fly Writes
SMART 190	Airflow Temperature
SMART 191	G-sense Error Rate
SMART 192	Power-off Retract Count
SMART 193	Load Cycle Count
SMART 194	Temperature
SMART 197	Current Pending Sector Count
SMART 198	Uncorrectable Sector Count
SMART 199	Ultra DMA CRC Error Count
SMART 240	Head Flying Hours
SMART 241	Total LBAs Written
SMART 242	Total LBAs Read

SMART parameters are important, because they help us identify status of health of the storage device. It can help in identifying the deteriorating health of a storage device and point it in advance if, the hard disk will experience a failure in the near future or not. In a large scale organization the number of such devices is exorbitant. To keep track of SMART parameters of each of these devices is not a simple task and if multiple failures are experienced then, it will affect the productivity of the organization. To make the task of monitoring these SMART parameters of individual storage devices we can make use of various machine learning models.

These models can train themselves to identify whether a hard disk is going to experience a failure or not based on these SMART parameters and hence automating the monitoring task. The machine learning models are able to justify only the instantaneous failure of the storage device, whereas, the deep learning models provide a better perspective because models like RNN (Recurrent Neural Networks) take into consideration the time aspect as one of the training parameter.

One of the challenges that surfaced after the survey was of unbalanced dataset. Storage devices hardly face any failure but when they do they need to be backed up and replaced, because of this the ratio of instances of storage devices experiencing failure to that to not experiencing failure is highly one sided. This raises issues for many machine learning models, in such cases we need to preprocess the dataset carefully and try to balance the dataset while training the model otherwise; one target class will heavily overwhelm the other class and thus providing defected results.

The next section provides an overview of machine learning techniques that are used to handle the problem of failure prediction which is then followed by conclusion.

II. LITERATURE SURVEY

The following literature explores various methods to recognize failures in the hard disk. The various method involve machine learning algorithms such as Decision trees, Random Forest, Recurrent Neural Networks etc.

Most of these Studies involve dataset from the servers of big organization and most common of these involve the blackbaze dataset[10]. The blackbaze organization release a yearly updated of dataset that contains the SMART parameters which are sampled each day for a year. This dataset involves with updated set of parameters each year and on every release some new parameters are introduced.

Ji Wang et. al. has proposed an attention augmented deep neural network that is able to focus on the history and then predict the failure of the hard disks [1]. There are many SMART parameters which can raise the issue of curse of dimensionality and finding the relevant parameters to train the model is an important challenge, which will help us to reduce the time. For finding these parameters they applied a z-test over the dataset to find if the parameters affects the performance of the hard disk or not and thus based on precision, recall and specificity they determined the overall score of their model.

Fernando Dione S. Lima et. al, have used a deep learning model known as recurrent neural network model, which has the capability to consider the history of the input dataset [2]. The model thus was able to predict that the hard disk will fail or not in long terms, with respect to time but for short term it was a difficult. Jing Li, et. al. in their work provided a prediction model using Decision Trees and Gradient Boosted Decision Trees, both models were able to reduce the false alarm rate and false detection rate and tree pruning was successfully applied at the required parameters [3]. From this study the authors recognized that the Decision Tree Model had better performance than the GBDT model. This model also provided some useful insight upon the relevance of parameters.

Farzaneh Mahdisoltani et. al, also explored various machine learning models over the black baze dataset of 2016 for HDD as well as SDD[12]. The training classifiers they used are classification and regression trees (CART), random forests, support vector machines, neural networks and logistic regression. Among these classifiers the random forest was the one which provided better results.

Carlos A. Rincon, et. al, have used three models, Decision Trees, Neural Networks and Logistic Regression [4]. Thus, while testing their model Decision Tree outperformed the other two. The models were tested on a homogeneous environment where SMART parameters of different models of different make and model were considered at the same time, these machine learning models can be trained over a homogeneous environment for model to increase the efficiency of the system.

Jiang Xioa, et. al, used a online random forest algorithm. The nature of this model is dynamic, the model can adapt to the new information [5]. Because of the adaptive nature the decision trees, generated during previous learning phase needs to be constantly replaced by the new one in the next learning phase. The model has the capability to learn and test at the same time in the deployment environment.

Fernando Dione S. Lima, et. al, provided a deep neural network architecture inspired by LSTM Networks [6][11]. These networks while training the model consider the long time series of the dependent previous history of the model and can predict the failure of hard disk in the long run. Thus, the model is not very good at predicting the short-term failures. In their previous work [2], they have used an RNN architecture and what they have concluded is that LSTM provided a better result than the RNN model, since LSTM is not bounded to a specific set of time period and is equipped with a gating mechanism which has the ability to take input only, the information which is relevant.

Ardeshir Raihanian Mashhadi et. al, did a survey on SMART parameters using a statistical model to compare and analyze the SMART parameters and find such parameters[7], which can help us predict the failure, before it happens for this instance. They used Random Forest, Decision Trees and Ensemble Trees techniques to recognize the patterns among the SMART parameters, among which the Random Forest algorithm overcame the problem of over-fitting and was better candidate among others.

In their study, Eduardo Pinheiro et. al, concludes that the SMART parameters for individual models are not sufficient but collective dataset might help to gain better perspective for better predictive model[9]. They focused on SMART parameters such as temperature, seek errors, power cycles, calibration retries, CRC errors, spin retries, etc.

Venkata Krishnan Mittinamalli Thandapani, has used an ensemble model which uses Random Forest, Feed Forward Neural Network with unsupervised K-means clustering algorithm [8]. The work determined that the Random Forest Model gave better accuracy as compared to the other two models. The work has foundation upon a limited number of data set inspired by only one model of hard disk.

The data set can be improved and involving a greater number of attributes that can be tried and tested in order to improve the efficiency of the model.

These studies revolve around various machine learning models and some explore the SMART parameters as a statistical model, in order to recognize the patterns between the parameters. Almost all of the work described, face the issue of unbalanced dataset problem along with the curse of dimensionality problem.

III. CONCLUSION

With the advent of new storage solutions and an exponential rise in the data, for which we make use of HDD and SDD. The SMART parameters that these storage devices provide can be used to find patterns which in return will help to predict the failure of these devices. This work provides an overview of studies, to identify the methodologies that can be applied to the binary classification problem of storage device failure detection.

The various machine learning models can learn with less error rate, but these models do not consider the time aspect while training and thus are good for the short-term predictions. On the other hand, Deep Learning Models such as LSTM networks, even though takes more time to train, take into consideration the remaining useful life of the storage devices as one of the relevant training parameters. These models thus can be trained to notify a failure within 'n' number of days (as per the user) instead of instantaneous failure notification.

Thus, in future, we will design an LSTM architecture which will be able to predict the failure in a hard disk before a given time and which will also reduce the false alarm rates of the system.

REFERENCES

1. Ji Wang, Weidong Bao, Lei Zheng, Xiaomin Zhu, Philips S. Yu, "An Attention-augmented Deep Architecture for Hard DriveStatus Monitoring in Large-scale Storage Systems", ACM Trans. Storage15, 3, Article21 (August 2019), 26 pages.
2. Fernando Dione S. Lima, Francisco Lucas F. Pereira, Iago C. Chaves, Joao Paulo P. Gomes, Javam C. Machado, "Evaluation of Recurrent Neural Networks for HardDisk Drives Failure Prediction", 7th Brazilian Conference on Intelligent Systems, pp. 85-90, 2018.
3. Jing Li, Rebecca J. Stones, Gang Wang, Xiaoguang Liu, Zhongwei Li, Ming Xu, "Hard Drive Failure Prediction using Decision Trees", Reliability Engineering and System Safety, March 2017.
4. Carlos A. Rincon, Jehan-Francois Paris, Ricardo Vilalta, Albert M. K. Cheng and Darell D. E. Long, "Disk Failure Prediction in Heterogeneous Environments", Society for Modelling and Simulation International, 2017.
5. Jiang Xiao, Zhuang Xiong, Song Wu, Yusheng Yi, Hai Jin, Kan Hu, "Disk Failure Prediction in Data Centers via Online Learning", InICPP2018: 47th International Conference on Parallel Processing, August 13–16, 2018,Eugene, OR, USA.ACM, New York, NY, USA, 10 pages.
6. Fernando Dione S. Lima, Gabriel M. R. Amaral, Lucas G. M. Leite, Joao Paulo P. Gomes, Javam C. Machado, "Predicting Failures in Hard Drives with LSTM Networks", Brazilian Conference of Intelligent Systems, pp. 222-227, 2017.
7. ArdeshirRaihanianMashhadi, Willie Cade, Sara Behdad, "Moving Towards Real-Time, Data-Driven Quality Monitoring: A Case Study of Hard Disk Drives", in 46th SME North American Manufacturing Research Conference, pp. 1107-1115, 2018.
8. Venkata Krishnan MittinamalliThandapani, "A Stable Model to Predict the Hard Disk Failure".
9. Eduardo Pinheiro, Wolf-Dietrich Weber and Luiz Andre Barroso, "Failure Trends in a Large Disk Drive Population", 5th USENIX Conference on File and Storage Technologies, pp. 17 – 29, 2007
10. Backblaze.com. (2020). *Backblaze Online Backup*. [online] Available at: https://www.backblaze.com/ [Accessed 17 Sep. 2019].
11. S. Hochreiter and J. Schmidhuber, "Long-short-term memory", Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.
12. Farzaneh Mahdisoltani Ioan Stefanovici Bianca Schroeder, "Improving Storage System Reliablility with Proactive Error Prediction", USENIX Annual Technical Conference, pp. 391-402, 2017.
13. Smartmontools.org. 2020. Smartmontools. [online] Available at: <https://www.smartmontools.org/> .

AUTHORS PROFILE

Rahul Nandgave, completed engineering from Kavikulguru Institute of Technology and Science, Ramtek under the university of Nagpur. Currently pursuing Masters in Computer Engineering from Pune Institute of Computer Technology under Savitri bai Phule University, Pune.

Dr Amar Buchade completed Ph. D. in Computer Engineering from College of Engineering, Pune. He received B.E. and M.E. in Computer Science and Engineering from Walchand College of Engineering, Sangle respectively. His research areas are distributed system, cloud computing and security.