CS-7641-ML Project Proposal : Hard Drive - Fail Me Not

Akarshit Wal, Gnanaguruparan Aishvaryaadevi,
Karthik Nama Anil, Parth Tamane, Vaishnavi Kannan

1 Introduction

Hard disk failures can be catastrophic in large scale data centers leading to potential loss of all data. To alleviate the impact of such failures, companies are actively looking at ways to predict disk failures and take preemptive measures. Traditional threshold-based algorithms were successful in predicting drive failures only 3-10% of the times[1]. Thus, we saw a shift to learning based algorithms that use Self-Monitoring, Analysis and Reporting Technology (S.M.A.R.T) attributes to make predictions. These attributes are different hard drive reliability indicators of an imminent failure. In recent times, people have applied insights and learnings obtained from analysing hard drives of one vendor to other vendor using tranfer learning techniques[2]. These models either used drives from specific vendors to achieve high F score[2] or used a subset of data and selected attributes[3]. In this project, we aim to explore unsupervised and supervised learning techniques to predict and analyze hard drive crashes.

The challenge lies in understanding how different attributes interact to contribute to failures, then selecting the most important attributes to train our models. Furthermore, the number of failed drives contribute to less than 10% of the data-set, it will be interesting to find means of handling the imbalanced data set where we may face the problem of curse of dimensionality.

2 Data Set

Backblaze owns and operates multiple data centers that have more than a million drives [4] and they regularly releases reports about the performance of these drives. The data set includes drive's serial number, model number, disk capacity, label indicating disk failure, and S.M.A.R.T stats. The 2019 data set reports 62 different statistics.

3 Methods

Given the plethora of features in our data set, we will first perform feature selection and dimension reduction using PCA to identify the most crucial features that contribute to disk failures. This is needed as not all drives report all stats and some stats can have different meaning across vendors. We then plan to use classification techniques (supervised learning) like Logistic Regression, XGBoost[5] and Random Forests[3] to label hard disks as failed or not. Additionally, we plan to incorporate clustering techniques(unsupervised learning) like DBSCAN and Hierarchical Clustering to identify hard disks with similar characteristics across vendors.

4 Results

The main goal is to identify disk failure with higher accuracy based on the given S.M.A.R.T statistics. With the help of simpler models like classification trees and regression, we hope to achieve more intuitive results. Also, we aim to draw correlation between S.M.A.R.T attributes of hard drive families making this model more practical.

5 Discussion

The current approaches are restricted to vendor specific hard disk failure analysis and uses only a small subset of features available in the data set. Through this project, we aim to use a holistic approach to determine the most relevant attributes to obtain a high failure prediction rate and low false alarm rate. This could be used to take preemptive measures to mitigate the affects of hard disk crashes and significantly improve the reliability of large scale storage systems in data centers.

6 Checkpoints

Data pre-processing and implementation of supervised learning methods for determination of hard disk failure as mid-term checkpoint. Application of unsupervised methods to cluster hard disk failure based on S.M.A.R.T attributes is the final checkpoint.

References

[1] C. Xu, G. Wang, X. Liu, D. Guo, and T. Liu. Health status assessment and failure prediction for hard drives with recurrent neural networks. *IEEE Transactions on Computers*, 65(11):3502–3508, Nov 2016.

[2] Mirela Madalina Botezatu, Ioana Giurgiu, Jasmina Bogojeska, and Dorothea Wiesmann. Predicting disk replacement towards reliable data centers. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[3] Jing Shen, Jian Wan, Se-Jung Lim, and Lifeng Yu. Random-forest-based failure prediction for hard disk drives. *International Journal of Distributed Sensor Networks*, 14(11):1550147718806480, 2018.

[4] Backblaze. Backblaze hard drive state, 2020.

[5] J. Li et al. Hard drive failure prediction using classification and regression trees. In *44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, Atlanta, GA, 2014*, 2014.