Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 5.258

IJCSMC, Vol. 5, Issue. 7, July 2016, pg.177 – 185

Enhanced Rules Framework for Predicting Disk Drives Failures

Vineetha B.Y
M.Tech (CSE), New Horizon College of Engineering, Bengaluru
Vineetha.kumbar@gmail.com

Asha Borah
Assistant Professor (CSE), New Horizon College of Engineering, Bengaluru
Asha.borah@gmail.com

ABSTRACT: Disk Drive failure prediction and analysis is one of the important areas in the field of storage. The SMART [Self- Monitoring, Analysis and Reporting Technology] is the closest technology that can predict to an extent that which disk drives may go bad. For the enterprise cloud storage environment it needs to have an additional rules and framework which can ensure the impending drive failures are caught. This paper proposes disk drive prediction model based on statistical and machine learning methods using Maximum Likelihood rule induction algorithm for solving classification problems through probability distribution based on SMART attributes, which are evaluated by using the data provided by the real world data center called Backblaze and also based on the IO latency of each of the disks drives, we are predicting the failures of the disks drives.

Keywords: Disk Drive failure prediction, SMART, Enterprise cloud storage, Maximum Likelihood, Backblaze

I. INTRODUCTION

Predicting the impending failure of hard disks in the field accurately can help storage systems or data center administrators to take corrective actions before the failure to avoid loss of data and performance degradation. The main hard challenge here is for predicting the impending failure of hard disks which factors should be taken into consideration are not clear. Most of the hard drive vendors support self-monitoring, analysis and reporting technology (S.M.A.R.T.), which measures drive characteristics such as temperature, spin-up time, data error rates, etc. however, it has been found that S.M.A.R.T. parameters alone are not enough to reliably predict individual drive failures [5].

In addition to SMART attributes we have taken live disk performance parameters like message queues, read errors, write errors, busy also for consideration for predicting the impending failure of hard disks. We have taken the drives monitored data which consists of the SMART attribute values for respective failed drives and good drives, using these observed data we can use the proposed system for classifying the bad disks.

The next challenge is of classification. An algorithm is trained with data to recognize the characteristics of two classes – "good disks" and "failing disks". By using an application of an existing machine learning approach called Maximum Likelihood [ML] Rules algorithm [3], which is used on a rule induction principle for solving classification problems through probability estimation that can be effectively employed to detect impending disk failures. The MLRules algorithm generates an ensemble of classifiers to give an overall prediction of whether a disk will fail based on the hard disk events and warnings logged during the lifetime of the hard disk. It uses decision rules as its base classifier. A simple decision rule classifier contains a set of logical statements or conditions which, when satisfied, votes for a particular class. It is in the form: " if condition then response". It can be treated as a simple classifier that gives a constant response for the objects satisfying the condition part, and abstains from the response for all the other objects. The MLRules algorithm sequentially generates decision rules with an associated weight factor, by greedily minimizing the negative log-likelihood to estimate the conditional class probability distribution. The main advantage of decision rules is their simplicity and good

interpretability. We follow an approach in which a single decision rule is treated as a base classifier in an ensemble.

Rules can be learnt and calculated with low memory and computational requirements and can be used to develop an efficient stand-alone system for predicting hard disk failures with very good accuracy and low false alarm rate, unlike many other machine learning techniques. Rule-based techniques, since they are intuitively comprehensible, have special significance in case of hard disk failures.

II. RELATED WORK

Hard drive manufacturers have been developing self-monitoring technology in their products for predicting failures before it actually occurs, to allow users or storage systems enough time to back up their data. The Self-Monitoring and Reporting Technology (S.M.A.R.T.) system is used by nearly all hard drive manufacturers. It collects data such as temperature, spin-up time, data error rates, etc., during normal operation and uses them to set a failure prediction flag. The S.M.A.R.T. flag is a one-bit signal that is generated to warn users of impending drive failure. Nearly all hard drive manufacturers use a very naive threshold algorithm which triggers an S.M.A.R.T. flag when any attribute exceeds a predefined value. These thresholds are set so as to avoid false alarms at the expense of predictive accuracy.

Past research [5] has found very little correlation between failure rates and either elevated temperature or activity levels. They show that some S.M.A.R.T. parameters (scan errors, reallocation counts, offline reallocation counts, and probational counts) can be helpful in predicting disk failure. However, due to the lack of occurrence of predictive S.M.A.R.T. signals on a large fraction of failed drives, they concluded that S.M.A.R.T. alone cannot be used to form an accurate predictive failure model.

Bairavasundaram et al. [1, 2] found that SATA disks and their adapters develop checksum mismatches an order of magnitude more often than FC disks. They also observed that the probability of developing checksum mismatches varies significantly across different disk models even within the same disk class. Also, the fraction of disks with latent sector errors varies significantly across manufacturers and disk models. They observed that as disk size increases, the fraction of disks with latent sector errors increases across all disk models. We use these observations to partition our dataset by model and interface, for better rule discovery.

Another study on drive failure prediction was performed by Hughes et al. [6]. They used Wilcoxon rank-sum test to build prediction models. They proposed two different strategies: multivariate test and ORing single attribute test. Their methods were tested on 3;744 drives. The highest detection rate achievable was 60% with 0:5% FAR.

Murray et al. [7] compared the performance of SVM, unsupervised clustering, rank-sum test and reverse arrangements test. In their subsequent work [8], they developed a new algorithm termed multiple-instance naive Bayes (mi-NB). They found that, on the dataset concerning 369 drives, ranksum test outperformed SVM for certain small set of SMART attributes (28:1% failure detection at 0% FAR). When using all features, SVM achieved the best performance of 50:6 % detection with 0% FAR.

Hamerly and Elkan [4] employed two Bayesian approaches to predict hard drive failures based on SMART attributes. Firstly, they used a cluster-based model named NBEM. The second approach was a supervised naive Bayes classifier. Both algorithms were tested on a dataset from Quantum Inc. concerning 1;927 good hard drives and 9 failed drives. They achieved prediction accuracy of 35□40% for NBEM and 55% for naive Bayes classifier with about 1% FAR.

III. EXISTING SYSTEM

Most of the modern hard disk drives support Self-Monitoring, Analysis and Reporting Technology (SMART), which can monitor internal attributes of individual drives and predict impending drive failures by a thresholding method. As the prediction performance of the thresholding algorithm which triggers a S.M.A.R.T. flag when any attribute exceeds a predefined value. These thresholds are set so as to avoid false alarms at the expense of predictive accuracy.

IV. PROPOSED SYSTEM

Data Sets:

The data set consists of time series of SMART attributes values for every disk model. Data were collected from the Backblaze Company where they provide the drive data every year which consist of SMART attributes values for the failed and good disks.

These data sets consist of Disk Model name, Serial Number, Capacity, failures, and totally 200 SMART attributes as columns. In the failure column if the value is 0 then that disk is failed and

if it is 1 that disk is good. These data sets are used as training examples for the algorithm to learn the data and give the result.

Feature selection:

Among 200 SMART attributes we selected the below parameters given in a table as a feature from the data sets as they were common for all the disk vendors and which are important in predicting the drive failures, and also we have considered the drive live performance parameters with the help of storage layer which decides the performance of the disk by the IO latency using the last 4 features.

ID#	ATTRIBUTE_NAME
4	Start_Stop_Count
5	Reallocated_Sector_Count
7	Seek_Error_Rate
9	Power_On_Hours
12	Power_Cycle_Count
192	Power-Off_Retract_Count
193	Load_Cycle_Count
196	Reallocated_Event_Count
197	Current_Pending_Sector
	Message Queue
	Read Error
	Write Error
	Busy

Table 1: Selected Disk Drive parameters

Architecture of predicting the failures of Disk Drives:

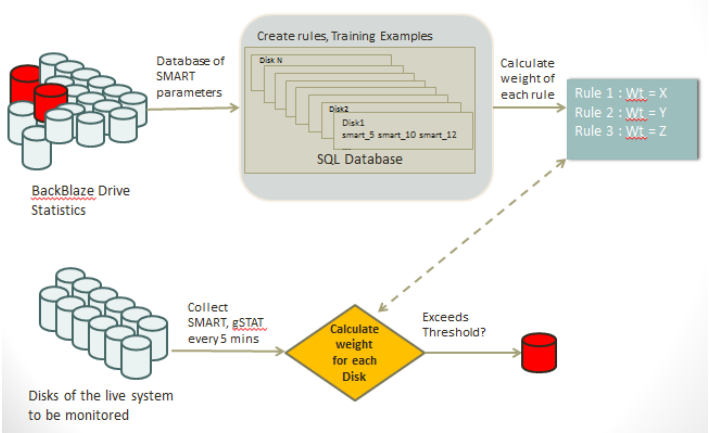

Figure 1: Architecture diagram for predicting disk failures.

The Figure 1 shows the design of predicting failures of disk drive system. Using Backblaze datasets with the help of SQL database rules are created and weights are created for each parameter, According to the created weights of selected parameters if the combined weight of the selected parameter in live system exceeds threshold value(fixed by system administrator) the drive is flagged as "BAD" and same disk is predicted for replacement.

Algorithm:

We use a new rule induction algorithm Maximum likelihood for solving classification problem through probability estimation. The purpose of maximum likelihood is to find the parameters of the model that best explain the data in the sense of yielding the largest probability or likelihood of explaining the data.

Probability estimation provides us with the conditional class distribution P(y/x), by which we can measure the prediction confidence. Moreover, all we need to obtain the Bayes classifier for any loss function is the conditional probability distribution. Here we consider the estimation of probabilities using the well-known maximum likelihood estimation (MLE) method. MLE can be stated as the empirical risk minimization by taking the negative logarithm of the conditional likelihood as the loss function (l) as shown in below:

$$\ell = \sum_{i=1}^{n} -\log P(y_i|\mathbf{x}_i). \qquad \ldots\ldots (1)$$

Algorithm for predicting failure of disk drives, given the Backblaze datasets:

Input: Set of n training examples given by Backblaze data
X = number of features selected
Output: Decision Rule

1. for x=1 to X
2. Find the likelihood value for the feature
3. Calculate the weight for the feature
4. Add a new feature to the existing one
5. Create a decision rule using the likelihood value and its respective weights

Generation of Decision Rule:

Decision rules are in the form of logical statements which are used for classifying the disk as bad or good, here single decision rule is treated as base classifier. Decision rule is formed by using the likelihood value calculated for each feature and weight associated with it. For Example

if feature1 > a_MLvalue

then feature1_weight = a

if feature2 > b_MLvalue

then feature2_weight = b

total_weight = feature1_weight + feature2_weight

Decision Rule :

if total_wgt > M

then class = Bad

Here M is a real time constant value decided upon the storage environment drive failure experience. M can vary according to the requirements of storage administrators.

V. RESULTS

The weights and likelihood value generated by using the Maximum likelihood Rule Algorithm with the Backblaze datasets for the above selected parameters are given below according to their parameters. According to these weights the system administrator can monitor their drives in order to predict the failure of disk drives.

Parameters	Weights	Likelihood value
Start_Stop_Count	0.01670	22
Reallocated_Sector_Count	0.59803	6452
Seek_Error_Rate	0.01280	7102252
Power_On_Hours	0.01213	16420
Power_Cycle_Count	0.01135	13
Power-Off_Retract_Count	0.02095	3045
Load_Cycle_Count	0.00774	22

Reallocated_Event_Count	0.00907	925
Current_Pending_Sector	0.01116	4896
Message Queue	0.0340	>90%
Read Error	0.10046	>10%
Write Error	0.13200	>10%
Busy	0.23012	>90%

Table 2: Calculated weights and likelihood value for the parameters

VI. CONCLUSION AND FUTURE WORK

In this paper we designed a system framework using the drive parameters and result of IO latency of each drive, which helps to predict the drives which are going to fail in the future so that storage administrators can replace the disks early before failing of disks. As the predicting failure of disks becomes very important for the performance of the storage read write performance in cloud storage.

This framework can be extended by adding more parameters of the drives causing the drive to fail. Can be considered for all different types of the disk drives where Backblaze has not used.

REFERENCES

[1] Lakshmi N. Bairavasundaram, Garth R. Goodson, Shankar Pasupathy and Jiri Schindler. An analysis of latent sector errors in disk drives. In Proceedings of the 2007 SIGMETRICS Conference on Measurement and Modeling of Computer Systems, 2007.

[2] Lakshmi N. Bairavasundaram, Garth R. Goodson, Bianca Schroeder, Andrea C. Arpacidusseau, and Remzi H. Arpaci-dusseau. An analysis of data corruption in the storage stack. In Proceedings of the 6th USENIX Symposium on File and Storage Technologies (FAST 08), 2008.

[3] Krzysztof Dembczynski, Wojciech Kotlowski, and Roman Slowinski. Maximum likelihood rule ensembles. In Proceedings of the 25th International Conference on Machine Learning, 2008.

[4] G. Hamerly and C. Elkan, "Bayesian approaches to failure prediction for disk drives," in Proceedings of the 18th International Conference on Machine Learning, San Francisco, CA, June 2001, pp. 202–209.

[5] Eduardo Pinheiro, Wolf Dietrich Weber, and Luiz Andr´e Barroso. Failure trends in a large disk drive population. In Proceedings of the 5th USENIX Symposium on File and Storage Technologies (FAST 07), 2007.

[6] G. F. Hughes, J. F. Murray, K. Kreutz-Delgado, and C. Elkan, "Improved disk-drive failure warnings," IEEE Transactions on Reliability, vol. 51, no. 3, pp. 350–357, Sep 2002.

[7] J. F. Murray, G. F. Hughes, and K. Kreutz-Delgado, "Hard drive failure prediction using non-parametric statistical methods," in Proceedings of the International Conference on Artificial Neural Networks, June 2003.

[8] Jerome H. Friedman and Bogdan E. Popescu. Predictive learning via rule ensembles. Annals of Applied Statistics, 2:916, 2008.

[9] Weihang Jiang, Chongfeng Hu, Yuanyuan Zhou, and Arkady Kanevsky. Are disks the dominant contributor for storage failures? a comprehensive study of failure characteristics. In Proceedings of the 6th USENIX Symposium on File and Storage Technologies (FAST 08), 2008.

[10] Joseph F. Murray, Gordon F. Hughes, and Kenneth Kreutz-Delgado. Machine learning methods for predicting failures in hard drives: A multiple-instance application. Journal of machine Learning research, volume 6, 2005.

[11] Jae Myung, Tutorial on maximum likelihood estimation, October 2002

[12] Vipul Agrawal, Chiranjib Bhattacharyya, Thirumale Niranjan, Sai Susarla, Prediction of Hard Drive Failures via Rule Discovery from
AutoSupport Data, 2009

