International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Predictive Maintenance of Storage Systems using LSTM Networks

Rahul Nandgave and Dr. A. R. Buchade

Department of Computer Engineering, PICT Pune

Received 10 Nov 2020, Accepted 10 Dec 2020, Available online 01 Feb 2021, **Special Issue-8 (Feb 2021)**

Abstract

There is an exorbitant amount of unstructured data which is available on the internet and is increasing at an exponential rate every day. The term 'Big Data' is used to represent such data. There is a need to handle such vast amounts of data efficiently and securely. RAID (Redundant Array of Independent Disks), NAS (Network Access Storage), SAN (Storage Access Network) etc. are some of the storage solutions that are available today and are supported by many companies that provide storage solutions. Significant loss of data as well as financial loss can be faced by companies in case of failures of storage solutions. Prediction of suchfailures at real time may help organizations for predictive maintenance and reducing the replacement downtime of such storage solutions. The LSTM Networks can be used to train our model which will predict the failure of storage devices based on data generated by S.M.A.R.T(Self-Monitoring and Reporting Technology) Parameters. Here we will provide an architecture inspired by an LSTM Network that will be able to predict the failure in a hard disk with lower false alarm rates and higher precision and recall.

Keywords: *LSTM, SMART Parameters, Predictive Maintenance, Failure Detection, Deep Neural Network,*

Introduction

With an explosion of world wide web, data on the internet is growing tremendously at an exponential rate. People communicate by sharing data and the amount shared between them is gigantic, this includes images, messages, videos, audios etc. This data is needed to be stored at some place securely and efficiently. Since advent of cloud, there is a heavy migration from basic storage architectures to cloud storage systems. This data is thus basically stored in the servers and these servers on raw level make use of either Solid-State Drives (SSD) or Hard Disk Drives (HDD). The manufacturers of such storage devices employ a system inbuilt inside them, known as the SMART system. SMART stands for Self-Monitoring and Reporting Technology. The SMART system collects information regarding various attributes related to storage devices. This information includes attributes such as read error rate, temperature, overall throughput, seek error rate etc. The standard for SMART was generalized in the 1990's and since then many attributes have been recognized and integrated as SMART attribute. Each of the manufacturer of such storage devices give their own version of some extra SMART attributes, thus variations in these SMART attributes can be found. In this study we will be only focusing on some common set of these SMART attributes.

The manufacturers of storage devices also provide raw attributes which are converted into normalized attributes which ranges between 0 and 100. Along with these values, we are also provided threshold values (for normalized values), these values decide whether if certain values exceed the threshold value then will it affect the storage device performance or not. The threshold values range of effect is usually decided by the manufacturer. For example, if we consider "READ_ERROR_RATE" then the threshold given is 63 and if at different time stamps if this value remains above the threshold then it's time to either keep taking backups at regular intervals or replace the hard disk. Keeping track of these values manually is possible if the number of storage devices is few but for large collection of such devices, which are usually found in servers, manually keeping track of such values is not possible. Some of these recognized and important attributes are Reallocated Sector Count(SMART 5), Reported Uncorrectable Errors(SMART 187), Command Timeout(SMART 188), Current Pending Sector Count(SMART 197), Uncorrectable Sector Count (SMART 198), Temperature (SMART 194) etc.[10]. Since there are numerous such attributes available it is difficult to identify and keep track of each attribute manually. Thus, we can make use of LSTM networks to predict the remaining useful life of the hard disk[9]. LSTM networks has the capability to predict in long term and hence perform better than other machine learning models. This will also help organizations to reduce their maintenance downtime, so that can make equipment replacement in time, thus increasing the reliability and fault tolerance of the system.

Literature survey

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. 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]. 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. Fernando Dione S. Lima, et. al, provided a deep neural network architecture inspired by LSTM Networks[6]. 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. ArdeshirRaihanianMashhadi, in their research work have analyzed the SMART parameters[7]. They used a statistical model to compare and analyze the SMART parameters and find such parameters, which can help us predict the failure, before it happens for this instance, they have considered a decision tree model. Venkata Krishnan MittinamalliThandapani, has used an ensemble model which uses Random Forest, Feed Forward Neural Network with unsupervised K-means clustering algorithm [8]. 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.

Proposed Methodology

The Dataset used is available from Blackbazefrom the year 2018 Q4 Quarter[10]. It consists of homogenous collection of various models of storage devices. The dataset is highly imbalanced and contains both the normalized as well as raw values of the SMART parameters. The dataset consists of Date, Serial Number, Model, Total Capacity in bytes, Failure (target variable) and the SMART stats (124 columns) For our purpose we will be using the normalized values of the SMART parameters. Normalized values are the transformed values derived from raw values of the smart attributes thus making the range of the values more compact. A Simple Neural Network takes into consideration only one instance of time, in such cases the previous history gets lost and if there is some temporal relation then this can lead to loss in accuracy. A Recurrent Neural Network (RNN) is a network which considers the sequential dependencies. In an RNN network a node is replaced with a memory cell, this allows the network to use the information related to previous time steps. This can be useful in area where there is a temporal relation between dataset. A Recurrent Neural Network's cell takes input the hidden state vector of previous time stamp as well as current time stamp input sequence. The RNN is then unfolded into desired number of layers which explores the temporal dependency between the input layers for a specific time period. Fig. 1. Provides a brief overview of RNN system where X is the input and abstract h is the hidden state layer, W_x, W_y and W_h are the weights of the system. At each instance the network takes in current time stamp input as well as hidden weights of the previous time stamp an provides output in Y. This limits the network to look back to certain length in time and thus limiting its capability to fully utilize from the past knowledge. As in case of a simple multilayer perceptron model, we make use of backpropagation algorithm to adjust their weights efficiently. Similarly, in RNN we make use of a modified version of this algorithm and since we are dealing with data with some temporal dependency between them, this algorithm is also known as backpropagation through time (BPTT). A BPTT works by taking the input and output pairs of the network and then accumulates error at each time step by unrolling the network at each individual time step. The weights are later updated by rolling up the network. These steps are repeated until error is minimized and network can not be trained any further or there is lack of data to train the model. A Long Short-Term Memory (LSTM) Network is an extension of RNN model. An independent Node in a LSTM network introduces a forget layer also known as a gating mechanism along with an internal memory. These two things help the network to retain the internal state information over a long period of time and eventually covering the long-term dependencies of the model. Another characteristic that makes LSTM network unique is the forget layer, this helps the network to decide which information to learn and which information to forget.

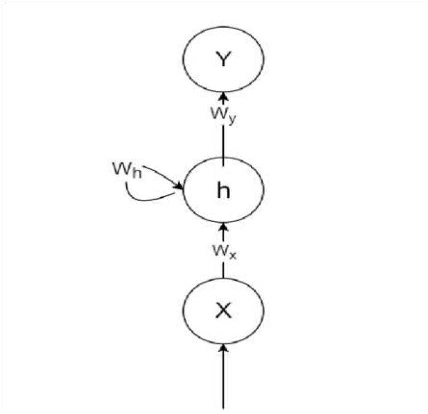

Fig 1: Recurrent Neural Network

The Algorithm for this work is described below- Step 1: Collect SMART parameters
Step 2: Refine SMART parameters (convert into normalized values)
Step 3: Replace null values with zeros
Step 4: Convert data according to time series form
Step 5: Split data into training and testing set in the ratio of 80% and 20 % respectively
Step 6: Train LSTM Model
Step 7: Check performance of the model
Step 8: Deploy model in test environment and reduce the false alarm rates of the overall system.

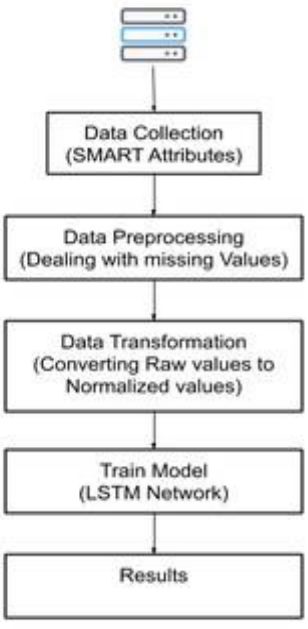

Fig. 2 Architecture of System

For our purpose we will be training our model with the Normalized values of SMART parameters which will be converted from Raw values of the same. Then we will train our model using k-fold cross validation technique in order to avoid overfitting of the model. In the end we will test our model and decide the efficiency of the result based on precision, recall and false alarm rate.

Result and Discussions

Our Objective is to determine whether a storage device is going to experience a failure or not in future. For this purpose, we will train our model to predict into three categories. These categories include if the storage device will fail, in next 5 days, if the storage device will fail in next 10 days or the storage device will fail in next 15 days. Thus the model will be capable of notifying the users beforehand, this will reduce the maintenance downtime and increase the reliability of the system also making the system make look like it is fault tolerant.

Conclusion

This work provides an overview of supervised deep learning algorithm i.e. a long short-term memory network, to identify the methodologies that can be applied to the binary classification problem of storage devicefailure detection. The variousmachine 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. Thus, we will design an architecture which will be able to predict the failure in a hard disk thus also reducing the false alarm rates of the system.

Acknowledgment

I am thankful to Dr. A. R. Buchade, for his indispensable support and guidelines for this project. I am thankful to Prof. M. S. Takalikar, HOD Computer Department, Pune Institute of Computer Technology I am grateful to Dr. P. T. Kulkarni, Principal, Pune Institute of Computer Technology. for his constant encouragement and for providing valuable resources when required.

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:

[6]. 47th International Conference on Parallel Processing, August 13–16, 2018,Eugene, OR, USA.ACM, New York, NY, USA, 10 pages.

[7]. 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.

[8]. 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.

[9]. Venkata Krishnan MittinamalliThandapani, "A Stable Model to Predict the Hard Disk Failure".

[10]. S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11]. Backblaze.com. (2020). *Backblaze Online Backup*. [online] Available at: https://www.backblaze.com/ [Accessed 17 Sep. 2019].