AN ADVANCED HARD DISK DRIVE FAILURE PREDICTION

Mr. B. B. K. PRASAD[1], P. Sravani[2], S. Bhavya[3], SD. Mushraf[4]

#1(Associate Professor) Dept. Of INFORMATION TECHNOLOGY, NRI INSTITUTE OF TECHNOLOGY, AP, India-521212

#1(UG Scholar) Dept. Of INFORMATION TECHNOLOGY, NRI INSTITUTE OF TECHNOLOGY, AP, India-521212

ABSTRACT

Failures or unexpected events are inevitable in critical and complex systems. Proactive failure detection is an approach that aims to detect such events in advance so that preventative or recovery measures can be planned, thus improving system availability. Machine learning techniques have been successfully applied to learn patterns from available datasets and to classify or predict to which class a new instance of data belongs. In this paper, we evaluate and compare the performance of 21 machine learning algorithms by using the m for proactive hard disk drive failure detection. For this comparison, we use WEKA as an experimentation platform and benchmark publicly available datasets of hard disk drives that are used to predict imminent failures before the actual failures occur. This project implementation of Random forest, the results show that different algorithms are suitable for different applications based on the desired prediction quality and the tolerated training and prediction time.

keywords:

Failure prediction, random forest, Clustering algorithm, Hard disk drives

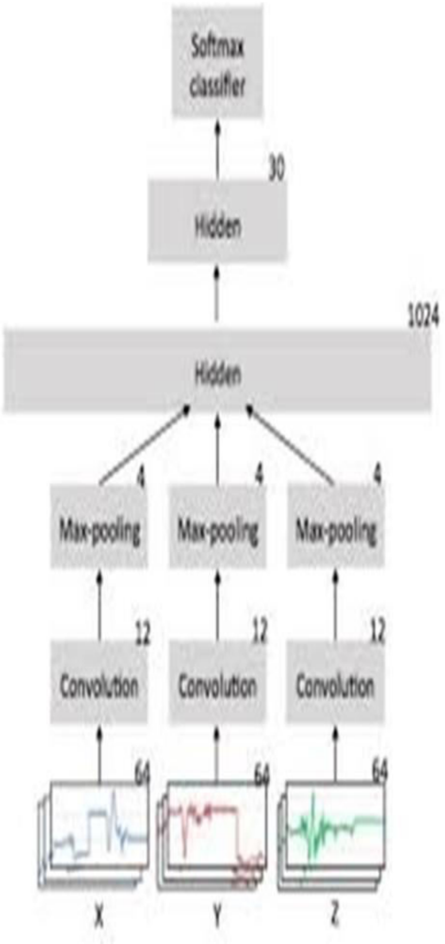
INTRODUCTION

Reliability is one of the most important factors of critical systems to maintain its functionalities and provide services without disruption. In complex systems, most components are communicating with each other and a failure of one component may lead to a failure of another component. If the problem of a component persists and cannot be resolved, it might propagate to other parts of the system and cause a total failure. The traditional approach is to prevent system failures in a reactive manner: when an internal misbehaviour is detected, a monitoring agent triggers a recovery procedure—to avoid or alleviate the problem—and a human operator maybe informed. This method, however, is performed after a misbehaviour has occurred, which may require some additional time until it is detected. This implies that when the recovery procedure starts, the problem may already have caused some damage to the system. Proactive failure detection [39], on the other hand, aims to foresee an imminent problem by detecting early signs instead of detecting the problem itself. These signs include unusual behaviours of system parameters, such as, system load, CPU utilization, memory usage, network traffic, and hardware temperature. When a failure can be detected in advance, one or even more recovery actions can be carefully planned, analysed, and evaluated to find the best solution for failure prevention [35]. Furthermore, in an extreme case when a failure cannot be avoided, other solutions, e.g. warming up spare components, can be initiated earlier to prepare for the service outage. Machine learning techniques have been used in many studies (see, e.g., [3, 4, 31, 40]) to analyze the signs of a failure and to make a prediction whether the system is likely to fail in the near future. In order to apply machine learning techniques, the prediction problem is formulated into a binary classification problem where the signs obtained from system monitoring are learned and used as reference Model is to classify new runtime observations and concludes whether the system is about to fail. In this paper, we compare the performance of machine learning techniques in terms of prediction quality, as well as time needed for training the models and for making predictions by applying it to a concrete dataset. We evaluate 21 compatible machine learning algorithms supported by WEKA [21], an open-source data mining tool, for the task of proactive failure detection in hard disk drives and make suggestions about which algorithm is most suitable under specific constraints. The problem of the hard disk failure prediction was proposed by Hemery and Elkan [22], who used supervised naïve Bayes and mixture models of unsupervised naive Bayes trained with the expectation-maximization algorithm to predict failures. Hughes et al. [26] applied a rank sum test, which is a statistical hypothesis test to predict which drives will fail. The null hypothesis of the test is constructed by using the data from good drives. At runtime, when the parameter of the tested drive deviates significantly from the null hypothesis, a failure warning is issued. The dataset used in this paper was originally used by Murray et al. [36] to compare a newly developed classifier called the multi-instance naive Bayesian classifier with a set of traditional algorithms, including support vector machines, unsupervised clustering, the rank sum test, and the reverse-arrangement test. The data set is thus also suitable to benchmark different machine learning algorithms in this study

Proposed system:

Failure prediction for hard disk drives is a typical and effective approach to improve the reliability of storage systems. The experimental results show that our proposed method can achieve a better prediction accuracy than state- of- the- art methods.

SYSTEM ARCHITECTURE:



IMPLEMENTATION:

MODULES

The activity recognition process is described, containing four main stages.

1. Data Collection: The first step is to collect multivariate time series data from the phone’s and the watch’s sensors. The sensors are sampled with a constant frequency of 30 Hz. After that, the sliding window approach is utilized for segmentation, where the time series is divided into subsequent windows of fixed duration without interwindow gaps (Banos et al., 2014). The sliding window approach does not require pre-processing of the time series, and is therefore ideally suited to real-time applications.

2. Pre-processing: Filtering is performed afterwards to remove noisy values and outliers from the accelerometer time series data, so that it will be appropriate for the feature extraction stage. There are two basic types of filters that are usually used in this step: average filter (Sharma et al., 2008) or median filter (Thiemjarus, 2010). Since the type of noise dealt with here is similar to the salt and pepper noise found in images, that is, extreme acceleration values that occur in single snapshots scattered throughout the time series. Therefore, a median filter of order 3 (window size) is applied to remove this kind of noise.

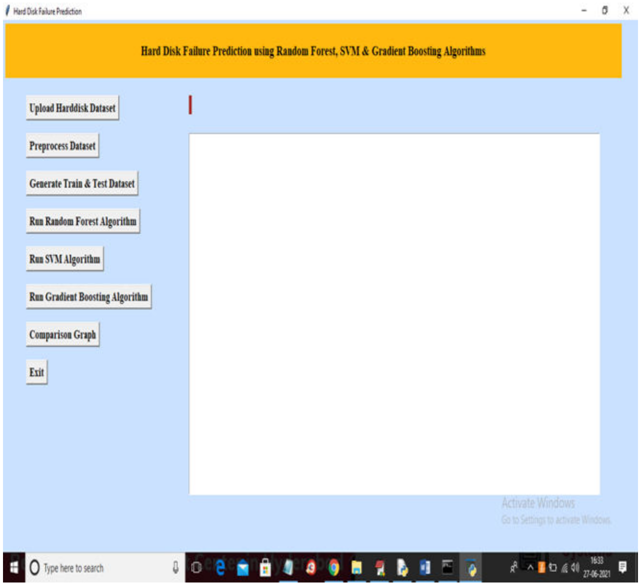
3. Feature Extraction: Here, each resulting segment will be summarized by a fixed number of features, i.e., one feature vector per segment. The used features are extracted from both time and frequency domains. Since, many activities have a repetitive nature, i.e., they consist of a set of movements that are done periodicallylike walking and running. This frequency of repetition, also known as dominant frequency, is a descriptive feature and thus, it has been taken into consideration.

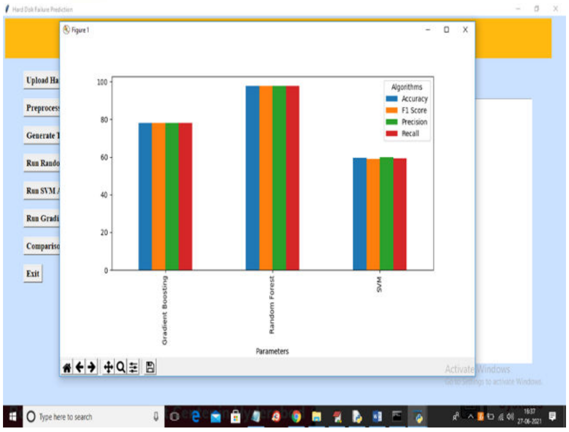
4. Standardization: Since, the time domain features are measured in (m/s 2), while the frequency ones in (Hz), therefore, all features should have the same scale for a fair comparison between them, as some classification algorithms use distance metrics. In this step, Z-Score standardization is used, which will transform the attributes to have zero mean and unit variance, and is defined as

$$xnew = (x-\mu)/\ \sigma$$

where μ and σ are the attribute’s mean and standard deviation respectively (Gyllensten, 2010).

Results:


Hard Disk Failure Prediction
Hard Disk Failure Prediction using Random Forest, SVM & Gradient Boosting Algorithms
Upload Harddisk Dataset
Preprocess Dataset
Generate Train & Test Dataset
Run Random Forest Algorithm
Run SVM Algorithm
Run Gradient Boosting Algorithm
Comparison Graph
Exit
Activate Windows
Go to Settings to activate Windows.
Type here to search
16:33
27-06-2021



:

CONCLUSION In this article, we presented a novel drive failure prediction method based on the part-voting random forest to improve the detection accuracy of soon-to-fail HDDs. Considering the different characteristics of each drive failure type, our method differentiates prediction of HDD failures in a coarse-grained manner by part-voting and similarity between health samples, because it is hard to classify drive failures accurately. The trees in random forest are specialized in classifying certain groups of samples rather than all samples. Therefore, the method uses a part of the decision trees that are suitable for voting on the classification result of a certain sample to improve the accuracy of the prediction. We tested the method with real-life data. The experimental results show that the random-forest-based method outperforms the other methods mentioned in section "Background and related work." Our prediction method can achieve an FDR of 97.67% with an FAR of 0.017% for family "B," an FDR of 100% with an FAR of 1.764% for family "S," and an FDR of 94.89% with an FAR of 0.44% for family "T". There are several aspects that need to be improved. The use of SMART data to indicate impending failures is limited.34 We will thus attempt to extract the workload and input/output performance of drives from system log data, and combine these data with SMART data along the time axis to facilitate classification of HDD failures and health status assessments. Moreover, we will utilize deep learning to detect the HDD failures, which can improve the accuracy of differential prediction and help users to take effective measures as early as possible.

References

1.B. Erçahin, Ö. Aktaş, D. Kilinç, and C. Akyol, ''Twitter fake account detection,'' in Proc. Int. Conf. Comput. Sci. Eng. (UBMK), Oct. 2017, pp. 388–392.

2. F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, ''Detecting spammers on Twitter,'' in Proc. Collaboration, Electron. Messaging, AntiAbuse Spam Conf. (CEAS), vol. 6, Jul. 2010, p. 12.

3. S. Gharge, and M. Chavan, ''An integrated approach for malicious tweets detection using NLP,'' in Proc. Int. Conf. Inventive Commun. Comput. Technol. (ICICCT), Mar. 2017, pp. 435–438.

4. T. Wu, S. Wen, Y. Xiang, and W. Zhou, ''Twitter spam detection: Survey of new approaches and comparative study,'' Comput. Secur., vol. 76, pp. 265–284, Jul. 2018.

5. S. J. Soman, ''A survey on behaviors exhibited by spammers in popular social media networks,'' in Proc. Int. Conf. Circuit, Power Comput. Technol. (ICCPCT), Mar. 2016, pp. 1–6.

6. A. Gupta, H. Lamba, and P. Kumaraguru, ''1.00 per RT #BostonMarathon # prayforboston: Analyzing fake content on Twitter,'' in Proc. eCrime Researchers Summit (eCRS), 2013, pp. 1–12.

7. F. Concone, A. De Paola, G. Lo Re, and M. Morana, ''Twitter analysis for real-

time malware discovery,'' in Proc. AEIT Int. Annu. Conf., Sep. 2017, pp. 1–6.

8. N. Eshraqi, M. Jalali, and M. H. Moattar, ''Detecting spam tweets in Twitter using a data stream clustering algorithm,'' in Proc. Int. Congr. Technol., Commun. Knowl. (ICTCK), Nov. 2015, pp. 347–351.

9. C. Chen, Y. Wang, J. Zhang, Y. Xiang, W. Zhou, and G. Min, ''Statistical features-based real-time detection of drifted Twitter spam,'' IEEE Trans. Inf. Forensics Security, vol. 12, no. 4, pp. 914–925, Apr. 2017.

10. C. Buntain and J. Golbeck, ''Automatically identifying fake news in popular Twitter threads,'' in Proc. IEEE Int. Conf. Smart Cloud (SmartCloud), Nov. 2017, pp. 208–215.

11. C. Chen, J. Zhang, Y. Xie, Y. Xiang, W. Zhou, M. M. Hassan, A. AlElaiwi, and M. Alrubaian, ''A performance evaluation of machine learning-based streaming spam tweets detection,'' IEEE Trans. Comput. Social Syst., vol. 2, no. 3, pp. 65–76, Sep. 2015.

12. G. Stafford and L. L. Yu, ''An evaluation of the effect of spam on Twitter trending topics,'' in Proc. Int. Conf. Social Comput., Sep. 2013, pp. 373–378.