Progress Report

Project Title: Machine Learning for Storage System Reliability Prediction

YANG zhi lin -1155092151

Project Name:	Machine Learning for Storage System Reliability Prediction
Supervisor:	Patrick PC Lee
Project Period:	June 1, 2018 – August 24, 2018
Reporting period:	June 1, 2018 – July 1, 2018
Report Submission date:	July 2, 2018

Section One: Summary

Hardware reliability is one of critical issues in managing storage system. In this project, we are exploring a suite of different machine learning methods to predict disk failures for data centers. We are studying several related works and reproducing the results described in these papers. By comparing their performances, we intend to make an overall analysis and propose some further improvements on the models.

Section Two: Activities and Progress

Task 1: Reproduce the paper: Botezatu et al., "Predicting Disk Replacement towards Reliable Data Centers"

Completed: Has done the preliminary implementation, including the following steps:
1.data preprocessing:
1) Change points detection
2) Exponential smoothing
3) Downsampling via K-Means Clustering

2. Model training, including: Logistic Regression, Random Forest, Support Vector Machine, Decision Tree, Gradient Boosting Decision Tree and Regularised Greedy Forest

3. Performance evaluation: Regularized Greedy Forest achieved the best performance as presented in the paper

Work to be done:
1) There are still some differences between my experiment results and the results in the paper, and further improvement on my results is expected.
2) The transfer learning method is expected to be tried.
3) Different disk models will be evaluated for better assessment of the method

Task 2: Reproduce the paper: Li et al., "Hard Drive Failure Prediction Using Classification and Regression Trees"

Completed: Has done the preliminary implementation of the methods on disk models in Blackblaze dataset, including the following steps:
1.data preprocessing

2. Model training, including: Classification Tree and BP artificial neural network

3. Performance evaluation: Since the dataset I used was different from the dataset in the paper, the performance is not as satisfactory as expected

Work to be done:
1) The dataset will be changed to the one used in the paper
2) The Regression Tree method will be tried to assess the healthy degree of disks

3) Different parameters for the methods will be tried

Task 3: Reproduce the paper: Mahdisoltani et al., "Proactive error prediction to improve storage system reliability"

Task 4: Reproduce the paper: Wang et al., "Storage device performance prediction with CART models"

Task 5: Reproduce the paper: Li et al., "Being Accurate is Not Enough: New Metrics for Disk Failure Prediction"

Section Three: Outputs and Deliverables

Detailed description of the experiments are summarised in 4 weekly reports. Please find the reports and the implementation code in the following link: https://github.com/yangzhilinAndy/Machine-Learning-Research-Project.git

Experiment result for Task1:

Healthy

	RGF	GBDT	RF	SVM	LR	DT
Precision	0.916 +/- 0.018	0.907 +/- 0.012	0.913 +/- 0.015	0.686 +/- 0.002	0.686 +/- 0.002	0.925 +/- 0.013
Recall	0.985 +/- 0.010	0.981 +/- 0.012	0.974 +/- 0.014	1.000 +/- 0.000	1.000 +/- 0.000	0.916 +/- 0.031
F-score	0.949 +/- 0.009	0.942 +/- 0.009	0.942 +/- 0.010	0.814 +/- 0.001	0.814 +/- 0.001	0.920 +/- 0.017

Replaced

	RGF	GBDT	RF	SVM	LR	DT
Precision	0.982 +/- 0.026	0.950 +/- 0.029	0.936 +/- 0.036	0.000 +/- 0.000	0.000 +/- 0.000	0.824 +/- 0.054
Recall	0.801 +/- 0.046	0.760 +/- 0.031	0.797 +/- 0.036	0.000 +/- 0.000	0.000 +/- 0.000	0.838 +/- 0.032
F-score	0.873 +/- 0.026	0.856 +/- 0.022	0.860 +/- 0.024	0.000 +/- 0.000	0.000 +/- 0.000	0.830 +/- 0.032

Experiment result for Task2:

Classification Tree:
False Alarm Rate = 38.8% Failure Detection Rate = 77.3%
BP artificial neural network:
Accuracy = 26%

Section Four: Self Assessment

From my personal perspective, I found my research performance in this month was not as satisfactory as I had expected. The reason was as follows: as a beginner and also a self-learner in machine learning, I am not very familiar with ML tools, thus have much difficulty in doing the implementation and has made quite a lot of mistakes. But with more and more practice, I expect myself to improve my skills and achieve better results.

To sum up, I have learned quite a lot in this month, including basic machine learning concepts and techniques. I also learned how to read paper and implement their results. I expect that I can finish all the planned tasks and actually deliver by the end of the research period.