Report 3

LI, Ruizhe | 1155076990

Github: https://github.com/rzli6/ML-Storage.git (private)

Target Paper: Predicting Disk Replacement towards Reliable Data Centers

Goal: Realize the method descripted in the paper.

Dataset: 2015 Q1 – 2015 Q4

Apply RGF model trained on ST4000DM000 (SgtA) on SgtA:

F1: 0.9948348265944833

Recall: 0.9897725626539187

Precision: 1.0

Apply RGF model to *ST31500541AS* **without** transfer learning:

F1: 0.12097669256381798

Recall: 1.0

Precision: 0.064

Note: It nearly predicts all the test cases to be failed, which is far from the truth.

Transfer Learning:

Knowing that the original f1 score is so low it might be meaningless to do transfer learning, I still tried the procedure described in the paper. I merged the SgtA and SgtB datasets together, and trained a LR model to classify the two models. There are 1693 disks in SgtB and 1586 disks in SgtA, totally 3279 disks in the dataset. The trained model can reach a high precision of 99%. Usually we expect a higher score, but under this situation, a high score means the two disks are so dissimilar to each other that the model can easily tell the difference. Therefore, I couldn't go any further to select similar cases for SgtA model.

Apparently, I failed on transfer learning. And the conclusion is that, we cannot carry it out until we know how to get a higher mark when applying the original learner on the new model. Until then, we could train a classifier to help us select samples similar to SgtB.

Smart_5_raw: reallocated sectors count

Target Paper: Improving Storage System Reliability with Proactive Error Prediction

Goal: Realize the method descripted in the paper.

Dataset: 2015 Q1 – 2015 Q4

Dataset Overview:

Drive model	Capacity (TB)	#Drives	Drives (Drive days) affected by:			
			SMART 5	SMART 187	SMART 196	SMART 197
Seagate ST4000DM000	4	36368	1.19% (0.02%)	2.33% (0.01%)	N/A	3.37% (0.02%)
Hitachi HDS5C3030ALA630	3	4664	3.58% (0.05%)	N/A	2.55% (0.04%)	2.72% (0.01%)
HGST HMS5C4040ALE640	3	7168	0.91% (0.03%)	N/A	0.91% (0.03%)	0.59% (0.002%)
Hitachi HDS722020ALA330	2	4774	11.84% (0.12%)	N/A	9.76% (0.08%)	6.47% (0.03%)
HGST HMS5C4040BLE640	4	9426	0.24% (0.003%)	N/A	0.24% (0.003%)	0.32% (0.002%)
Hitachi HDS5C4040ALE630	4	2719	2.54% (0.03%)	N/A	1.62% (0.02%)	1.95% (0.005%)
Seagate ST3000DM001	3	4707	25.15% (1.77%)	30.59% (0.31%)	N/A	35.33% (0.29%)

Figure in the paper

	smart_5_raw	smart_187_raw	smart_196_raw	smart_197_raw	Capacity (TB)	# Drives
ST4000DM000	0.00738119	0.0154702	0	0.0150657	4	29670
ST3000DM001	0.120719	0.177226	0	0.0702055	3	1168
Hitachi HDS5C3030ALA630	0.0323491	0	0.032132	0.0123752	3	4606
Hitachi HDS722020ALA330	0.133675	0	0.133675	0.0365151	2	4683
Hitachi HDS5C4040ALE630	0.0150376	0	0.0154135	0.0075188	4	2660
HGST HMS5C4040ALE640	0.0060317	0	0.0060317	0.00266517	4	7129
HGST HMS5C4040BLE640	0.000966806	0	0.000966806	0.00257815	4	3103

My result

Questions:

1. What we are going to predict? (Smart_5_raw or the increasement of it, or something else?)
2. Should we include the smart_5_raw as a feature? As shown in the paper?

Following is my result upon my understanding.

Label: 1 if smart_5_raw increases in the next week, else 0

Features (20 in total):

```
['y',
 'smart_1_raw', 'smart_4_raw', 'smart_5_raw',
 'smart_7_raw', 'smart_9_raw', 'smart_12_raw',
 'smart_187_raw', 'smart_193_raw', 'smart_194_raw',
 'smart_197_raw', 'smart_199_raw', 'smart_4_raw_increase',
 'smart_5_raw_increase', 'smart_7_raw_increase',
 'smart_9_raw_increase', 'smart_12_raw_increase',
 'smart_187_raw_increase', 'smart_193_raw_increase',
```


Smart_5_raw: reallocated sectors count

```
'smart_197_raw_increase', 'smart_199_raw_increase']
```

Note that, after data preprocessing I only got 135 cases with a label 1.

Down sampling to 500 health samples with KMeans.

Then feed the 635 training data into learner, I got:

	CART	SVM	NN	LR	RF
f1_score	0.8607239195419769	0.8643282687083159	0.7610262806215362	0.885173232083531	0.8489834139745188
recall	0.837037037037037	0.9259259259259259	0.7777777777777778	0.962962962962963	0.8296296296296296
precision	0.921951219512195	0.8551617873651771	0.892	0.8551617873651771	0.8970315398886829

Smart_5_raw: reallocated sectors count