Report 2

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: 2017 Q1 – 2018 Q1

Model: ST4000DM000 (SgtA)

Serial_number	Percentage	Serial_number	Percentage
smart_242_raw	0.571197	smart_193_normalized	0.279935
smart_7_raw	0.554207	smart_1_raw	0.224919
smart_193_raw	0.529126	smart_5_raw	0.204693
smart_9_raw	0.505663	smart_3_normalized	0.135113
smart_240_raw	0.501618	smart_183_raw	0.127832
smart_190_normalized	0.487055	smart_183_normalized	0.127832
smart_190_raw	0.486246	smart_198_normalized	0.11246
smart_194_raw	0.486246	smart_197_normalized	0.11246
smart_194_normalized	0.486246	smart_5_normalized	0.064725
smart_9_normalized	0.467638	smart_192_raw	0.045307
smart_198_raw	0.443366	smart_189_raw	0.021036
smart_197_raw	0.443366	smart_189_normalized	0.020227
smart_241_raw	0.412621	smart_184_normalized	0.018608
smart_7_normalized	0.384304	smart_184_raw	0.018608
smart_187_normalized	0.337379	smart_188_raw	0.0089
smart_187_raw	0.337379	smart_199_raw	0.004854
smart_4_raw	0.308252	smart_4_normalized	0.000809
smart_12_raw	0.307443	smart_192_normalized	0.000809

Smart_5_raw: reallocated sectors count

smart_1_normalized	0.291262	smart_241_normalized	0
smart_3_raw	0	smart_10_raw	0
smart_191_normalized	0	smart_191_raw	0
smart_240_normalized	0	smart_188_normalized	0
smart_199_normalized	0	smart_242_normalized	0
smart_10_normalized	0	smart_12_normalized	0

Features Selected (percentage > 0.01)

'smart_242_raw', 'smart_7_raw', 'smart_193_raw', 'smart_9_raw', 'smart_240_raw', 'smart_190_normalized', 'smart_190_raw', 'smart_194_raw', 'smart_194_normalized', 'smart_9_normalized', 'smart_198_raw', 'smart_197_raw', 'smart_241_raw', 'smart_7_normalized', 'smart_187_normalized', 'smart_187_raw', 'smart_4_raw', 'smart_12_raw', 'smart_1_normalized', 'smart_193_normalized', 'smart_1_raw', 'smart_5_raw', 'smart_3_normalized', 'smart_183_raw', 'smart_183_normalized', 'smart_198_normalized', 'smart_197_normalized', 'smart_5_normalized', 'smart_192_raw', 'smart_189_raw', 'smart_189_normalized', 'smart_184_normalized', 'smart_184_raw'

Features not in paper:

3: Spin-Up Time (NA)

4: Start/Stop Count (not in)

9: Power-On Hours (not in)

12: Power Cycle Count (not in)

183: SATA Downshift Error Count or Runtime Bad Block (0.5%)

192: Power-off Retract Count, Emergency Retract Cycle Count (not in)

Smart_5_raw: reallocated sectors count

Precision, Recall, F-score, Deviation of different classifiers

	RGF	GBDT	RF	SVM	LR	DT
F1	0.988	0.990	0.986	0.990	0.897	0.991
Recall	0.986	0.986	0.992	1.0	0.870	0.989
Precision	0.990	0.993	0.980	0.980	0.948	0.993

Apply RGF model to *ST8000DM002* **without** transfer learning:

F1: 0.022

Recall: 0.991

Precision: 0.011

Conclusion: It nearly predicts all the test cases to be failed, which is far from the truth (only 114 truly failed in test cases, but in prediction 9999 failed). Hence, without transfer learning, the trained-model is not applicable to another disk model, even if they are from the same producer.

Smart_5_raw: reallocated sectors count