Report 4

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.

In this week, I applied the method to every model described in the paper, to see whether this predicting model is equally applicable to other disk models. The followings are the result:

1. Overview

	# total	# failure	# failure percentage
ST4000DM000	29908	824	0.03
ST31500541AS	1970	271	0.14
Hitachi HDS722020ALA330	4737	159	0.03
Hitachi HDS5C3030ALA630	4634	74	0.02

Figure 1: the overview statistics of each disk model

Before tapping into the training procedure and predicting result, we should first take a look at the data set we are currently working on. Based on the figure 1 shown above, the ST4000DM000 (Sgt A) model has the biggest total number of disks, and it has a failure proportion of 3%. And following is the Hitachi HDS722020ALA330 (Hit A), which has 4737 disks in total, and 159 failed disks. Keeping this information in mind could help us later construe the discrepancies among different predicting results when applying the same method on dissimilar models.

Smart_5_raw: reallocated sectors count

2. Apply to different model

model	Metrics	GBDT	SVM	DT	LR	RF	RGF
ST4000DM000	P	0.99	0.98	0.92	0.60	0.99	0.99
	R	0.95	0.96	0.87	0.61	0.95	0.94
	F	0.97	0.97	0.89	0.60	0.97	0.97
	Sd	0.02	0.01	0.03	0.30	0.02	0.02
ST31500541AS	P	1.00	1.00	0.90	0.34	1.00	1.00
	R	0.92	0.92	0.76	0.97	0.92	0.92
	F	0.96	0.96	0.82	0.50	0.96	0.96
	Sd	0.01	0.01	0.07	0.02	0.01	0.01
Hitachi HDS722020ALA330	P	0.99	0.93	0.89	0.71	0.99	0.99
	R	0.86	0.87	0.86	0.85	0.85	0.85
	F	0.92	0.89	0.86	0.77	0.91	0.91
	Sd	0.04	0.07	0.09	0.08	0.04	0.04
Hitachi HDS5C3030ALA630	P	1.00	0.97	0.63	0.46	1.00	1.00
	R	0.71	0.71	0.74	0.56	0.71	0.69
	F	0.83	0.82	0.67	0.50	0.83	0.81
	Sd	0.06	0.07	0.05	0.05	0.06	0.04

Figure 2: the precision, recall, f1 score, and Sd of each model, obtained from experiment

		RGF		GBDT		RF		SVM		LR		DT	
		SgtA	HitA	SgtA	HitA	SgtA	HitA	SgtA	HitA	SgtA	HitA	SgtA	HitA
Replaced	P	0.98	0.84	0.97	0.82	0.93	0.82	0.93	0.72	0.73	0.72	0.89	0.74
	R	0.98	0.79	0.96	0.78	0.94	0.76	0.95	0.65	0.81	0.59	0.87	0.61
	F	**0.98**	**0.81**	**0.96**	**0.80**	**0.94**	**0.79**	**0.94**	**0.68**	**0.77**	**0.65**	**0.88**	**0.67**
	Sd	**0.01**	**0.02**	0.01	0.04	0.05	0.08	0.02	0.05	0.07	0.1	0.04	0.03
Healthy	P	0.99	0.93	0.98	0.92	0.97	0.92	0.97	0.87	0.89	0.85	0.94	0.86
	R	0.98	0.95	0.98	0.94	0.96	0.93	0.96	0.90	0.85	0.90	0.95	0.91
	F	**0.98**	**0.94**	0.98	0.93	0.97	0.92	0.96	0.88	0.87	0.87	0.94	0.88
	Sd	**0.01**	**0.02**	0.02	0.03	0.04	0.05	0.02	0.04	0.08	0.05	0.02	0.02

Table 3: Precision, Recall, F-score, Deviation of different classifiers - median on 100 runs , each of which using randomly-drawn training and test data points

Figure 3: the precision, recall, f1 score, and Sd of each model, stated in the paper

When I tried to apply the same machine-learning method on different disk models, the result is shown in Figure 2. Basically, this method can reach a predicting accuracy like 80% high. And different machine learning methods actually got different result. For example, in general the predicting accuracy of LR (Logistic Regression) model is way less than the RGF model, which is the most preferred method according to the paper author.

Smart_5_raw: reallocated sectors count

2.1. Similarities

My result has several similarities comparing to the one got from paper. Take Sgt A as an example. The orders of accuracy of different machine learning model are **similar: RGF >= GBDT >= RF >= SVM >= DT >= LR.** These consistency means RGF, GBDT, RF models do have some strengths on answering this kind of questions. The algorithms behind "tree structure" can properly interpret relationship between health status and the selected SMART attributes of each disk. And the result also indicates some differences among the accuracies of several disk models. If we focus on the RGF model, order of accuracies of each model is Sgt A > Sgt B > Hit A > Hit B, which is consistent with the descending order of the number of failed disks, and this order seems not related to the total number of disks and failed percentage. Down sampling could be one of key reasons for this relationship. In the experiment, the data is sampled to be commensurate to failed number. Therefore, the size of training set is not determined by the whole data set, but the number of failed disks.

2.2. Differences

However, there are also some discrepancies. For example, in the paper, RGF is a dominant machine learning method. It shows a conspicuous advantage over other method. On the contrary, this huge gap did not show up in my result. GBDT, SVM, RF, and RGF can all reach a F1 score around 97%, in Sgt A model. This could be interpreted with different attributes we selected. I selected 6 more attributes than the paper. And based on my understanding, why these attributes were not shown in the paper is because either some attributes were not accessible at that time, or the author thought they were irrelevant. And the bigger number of attributes may lead to the differences in which model my accuracy is higher. In addition to attributes, another difference between the author and I is our datasets was not completely same. He used a larger dataset, from 2013 to 2015 excluding several months in between, totally 27 months' data. However, my dataset only contained the data from 2015 whole year, 12 months' data. Thus, the accuracy dropped in my result could be explained by the decrement of dataset size.

Smart_5_raw: reallocated sectors count

3. Feature summarize of each model.

attributes	get_percent	get_median	get_mean
smart_193_raw	0.4429999887943268	112.0	157.0
smart_197_raw	0.4300000071525574	9.0	43.0
smart_198_raw	0.4169999957084656	9.0	40.0
smart_194_raw	0.41100001335144043	131.0	179.0
smart_7_raw	0.4050000011920929	123.5	137.0
smart_9_raw	0.3959999978542328	139.5	150.0
smart_194_normalized	0.3880000114440918	134.0	181.0
smart_190_raw	0.3880000114440918	134.0	181.0
smart_190_normalized	0.3880000114440918	134.0	181.0
smart_187_raw	0.37299999594688416	4.0	44.0
smart_187_normalized	0.36899998784065247	4.0	45.0
smart_242_raw	0.36000001430511475	148.0	168.0
smart_240_raw	0.32600000500679016	68.0	103.0
smart_193_normalized	0.32499998807907104	111.0	156.0
smart_9_normalized	0.3050000071525574	144.0	158.0
smart_7_normalized	0.2919999957084656	163.0	180.0
smart_1_normalized	0.29100000858306885	32.0	98.0
smart_241_raw	0.28999999165534973	172.0	184.0
smart_12_raw	0.27399998903274536	171.0	208.0
smart_4_raw	0.2720000147819519	169.0	207.0
smart_1_raw	0.25999999046325684	53.5	116.0
smart_3_normalized	0.22300000488758087	170.0	212.0
smart_5_raw	0.20000000298023224	8.0	53.0
smart_192_raw	0.11900000274181366	176.0	227.0
smart_189_normalized	0.10899999737739563	211.5	226.0
smart_189_raw	0.10899999737739563	211.5	226.0
smart_5_normalized	0.09700000286102295	9.0	38.0
smart_198_normalized	0.0949999988079071	2.0	12.0
smart_197_normalized	0.0949999988079071	2.0	12.0
smart_188_raw	0.07400000095367432	68.0	136.0
smart_183_raw	0.04100000113248825	5.0	50.0
smart_183_normalized	0.04100000113248825	5.0	50.0
smart_184_raw	0.03500000014901161	8.0	83.0
smart_184_normalized	0.03500000014901161	8.0	83.0
smart_199_raw	0.01899999938905239	35.5	136.0
smart_188_normalized	0.004000000189989805	1.0	3.0
smart_4_normalized	0.0020000000949949026	1.0	1.0
smart_12_normalized	0.0010000000474974513	2.0	2.0
smart_199_normalized	0.0010000000474974513	3.0	3.0
smart_240_normalized	0.0010000000474974513	2.0	2.0
smart_192_normalized	0.0	nan	nan
smart_242_normalized	0.0	nan	nan
smart_241_normalized	0.0	nan	nan
smart_3_raw	0.0	nan	nan
smart_10_raw	0.0	nan	nan
smart_191_normalized	0.0	nan	nan
smart_191_raw	0.0	nan	nan
smart_10_normalized	0.0	nan	nan

Sgt A

Smart_5_raw: reallocated sectors count

attributes	get_percent	get_median	get_mean
smart_195_normalized	0.527999997138977	68.0	98.0
smart_5_raw	0.5090000033378601	30.0	53.0
smart_241_raw	0.47200000286102295	95.0	128.0
smart_5_normalized	0.4429999887943268	27.0	50.0
smart_242_raw	0.4350000023841858	56.5	82.0
smart_1_normalized	0.3869999945163727	48.0	106.0
smart_12_raw	0.3799999952316284	105.0	124.0
smart_4_raw	0.37599998712539673	105.0	124.0
smart_7_raw	0.37599998712539673	118.0	127.0
smart_194_raw	0.37299999594688416	144.0	168.0
smart_9_normalized	0.335999995470047	125.0	132.0
smart_9_raw	0.32100000977516174	135.5	139.0
smart_194_normalized	0.32100000977516174	155.0	175.0
smart_190_raw	0.32100000977516174	155.0	175.0
smart_190_normalized	0.32100000977516174	155.0	175.0
smart_197_raw	0.31700000166893005	15.0	48.0
smart_1_raw	0.3140000104904175	64.0	130.0
smart_198_raw	0.2879999876022339	15.0	49.0
smart_187_normalized	0.2840000092983246	9.0	46.0
smart_187_raw	0.2840000092983246	9.0	46.0
smart_195_raw	0.26899999380111694	88.0	135.0
smart_7_normalized	0.2549999952316284	138.5	153.0
smart_189_raw	0.17299999296665192	32.0	56.0
smart_240_raw	0.17000000178813934	70.5	135.0
smart_189_normalized	0.16200000047683716	26.0	52.0
smart_188_raw	0.1289999932050705	65.0	124.0
smart_197_normalized	0.054999999701976776	6.0	11.0
smart_198_normalized	0.054999999701976776	6.0	11.0
smart_183_raw	0.029999999329447746	22.5	24.0
smart_183_normalized	0.029999999329447746	22.5	24.0
smart_188_normalized	0.017999999225139618	2.0	3.0
smart_199_raw	0.014999999664723873	81.0	84.0
smart_12_normalized	0.0	nan	nan
smart_3_raw	0.0	nan	nan
smart_184_normalized	0.0	nan	nan
smart_184_raw	0.0	nan	nan
smart_10_normalized	0.0	nan	nan
smart_199_normalized	0.0	nan	nan
smart_240_normalized	0.0	nan	nan
smart_4_normalized	0.0	nan	nan
smart_241_normalized	0.0	nan	nan
smart_3_normalized	0.0	nan	nan
smart_242_normalized	0.0	nan	nan
smart_10_raw	0.0	nan	nan

Sgt B

Smart_5_raw: reallocated sectors count

attributes	get_percent	get_median	get_mean
smart_1_normalized	0.36500000953674316	81.0	145.0
smart_1_raw	0.35199999809265137	95.5	162.0
smart_196_raw	0.3330000042915344	136.0	153.0
smart_5_raw	0.3330000042915344	138.5	157.0
smart_197_raw	0.32100000977516174	104.0	158.0
smart_194_raw	0.29600000381469727	215.0	216.0
smart_194_normalized	0.289000004529953	204.5	207.0
smart_4_raw	0.22599999606609344	241.0	245.0
smart_12_raw	0.22599999606609344	241.0	245.0
smart_193_raw	0.18199999630451202	212.0	190.0
smart_192_raw	0.18199999630451202	212.0	190.0
smart_9_raw	0.17599999904632568	232.0	203.0
smart_3_raw	0.1379999965429306	226.5	229.0
smart_3_normalized	0.13199999928474426	216.0	218.0
smart_196_normalized	0.13199999928474426	19.0	97.0
smart_5_normalized	0.12600000202655792	29.5	116.0
smart_9_normalized	0.11900000274181366	222.5	221.0
smart_193_normalized	0.07500000298023224	56.0	112.0
smart_192_normalized	0.07500000298023224	56.0	112.0
smart_199_raw	0.05000000074505806	45.5	138.0
smart_10_normalized	0.02500000037252903	37.0	31.0
smart_10_raw	0.02500000037252903	37.0	31.0
smart_7_raw	0.01899999938905239	37.0	28.0
smart_7_normalized	0.01899999938905239	37.0	28.0
smart_2_raw	0.01899999938905239	198.0	199.0
smart_197_normalized	0.01899999938905239	2.0	3.0
smart_8_raw	0.013000000268220901	209.0	180.0
smart_8_normalized	0.013000000268220901	209.0	180.0
smart_2_normalized	0.006000000052154064	123.0	123.0
smart_198_raw	0.006000000052154064	6.0	6.0
smart_199_normalized	0.006000000052154064	11.0	11.0
smart_12_normalized	0.0	nan	nan
smart_4_normalized	0.0	nan	nan
smart_198_normalized	0.0	nan	nan

Hit A

Smart_5_raw: reallocated sectors count

attributes	get_percent	get_median	get_mean
smart_1_normalized	0.5410000085830688	26.0	89.0
smart_1_raw	0.5	29.0	115.0
smart_5_raw	0.3779999911785126	33.0	100.0
smart_197_raw	0.36500000953674316	44.0	99.0
smart_196_raw	0.35100001096725464	28.0	82.0
smart_194_raw	0.2840000092983246	224.0	227.0
smart_194_normalized	0.25699999928474426	230.5	221.0
smart_192_raw	0.24300000071525574	220.5	214.0
smart_193_raw	0.24300000071525574	220.5	214.0
smart_196_normalized	0.2160000056028366	20.5	80.0
smart_5_normalized	0.2160000056028366	23.0	80.0
smart_9_raw	0.17599999904632568	224.5	198.0
smart_4_raw	0.14900000393390656	263.5	275.0
smart_9_normalized	0.13500000536441803	198.5	191.0
smart_12_raw	0.13500000536441803	284.0	285.0
smart_8_normalized	0.1080000028014183	274.0	261.0
smart_8_raw	0.1080000028014183	274.0	261.0
smart_3_raw	0.0949999988079071	260.5	247.0
smart_192_normalized	0.0949999988079071	43.5	59.0
smart_193_normalized	0.0949999988079071	43.5	59.0
smart_2_raw	0.08100000023841858	287.5	262.0
smart_3_normalized	0.06800000369548798	312.5	258.0
smart_2_normalized	0.05400000140070915	277.0	246.0
smart_199_raw	0.04100000113248825	72.0	108.0
smart_7_raw	0.014000000432133675	185.5	186.0
smart_7_normalized	0.014000000432133675	185.5	186.0
smart_197_normalized	0.014000000432133675	20.0	20.0
smart_198_raw	0.014000000432133675	6.0	6.0
smart_12_normalized	0.0	nan	nan
smart_10_normalized	0.0	nan	nan
smart_4_normalized	0.0	nan	nan
smart_198_normalized	0.0	nan	nan
smart_199_normalized	0.0	nan	nan
smart_10_raw	0.0	nan	nan

Hit B

Smart_5_raw: reallocated sectors count