WEEKLY REPORT 7

YANG zhi lin

1. Paper: **'Predicting Disk Replacement towards Reliable Data Centers'**
Main work: Refer to Rui Zhe's code and fixed bugs in my code.

1) Dataset: Blackblaze 2015 Q1-Q4
Model name: ST4000DM000
Total number of disks: 29670
Total number of failed disks: 586

2) ML result:

stat

	GBDT	SVM	DT	LR	RF	RGF
P	0.912	0.334	0.875	0.723	0.905	0.934
R	0.911	1.000	0.891	0.688	0.887	0.906
F	0.911	0.500	0.883	0.699	0.896	0.919
Sd	0.060	0.001	0.045	0.087	0.058	0.050

SVM still has the problem of low precision and high recall.

2. Paper: **'Hard Drive Failure Prediction Using Classification and Regression Trees'**

1) Vary different number of voters and apply a voting rule: 'When detecting a drive, we check the last N consecutive samples (voters) before a time point, and predict the drive is going to fail if more than N/2 samples are classified as failed, and the next time point is tested otherwise. '

Result in paper:

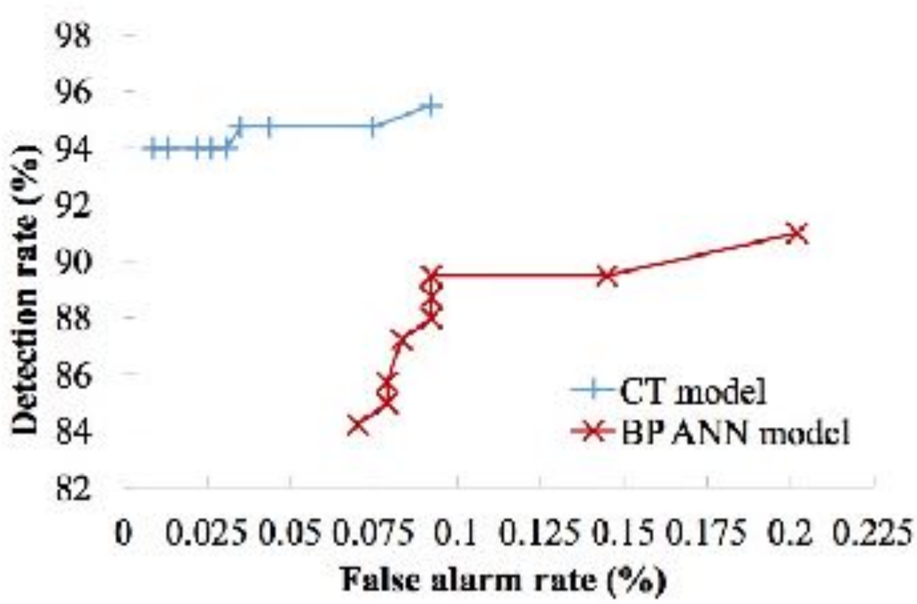
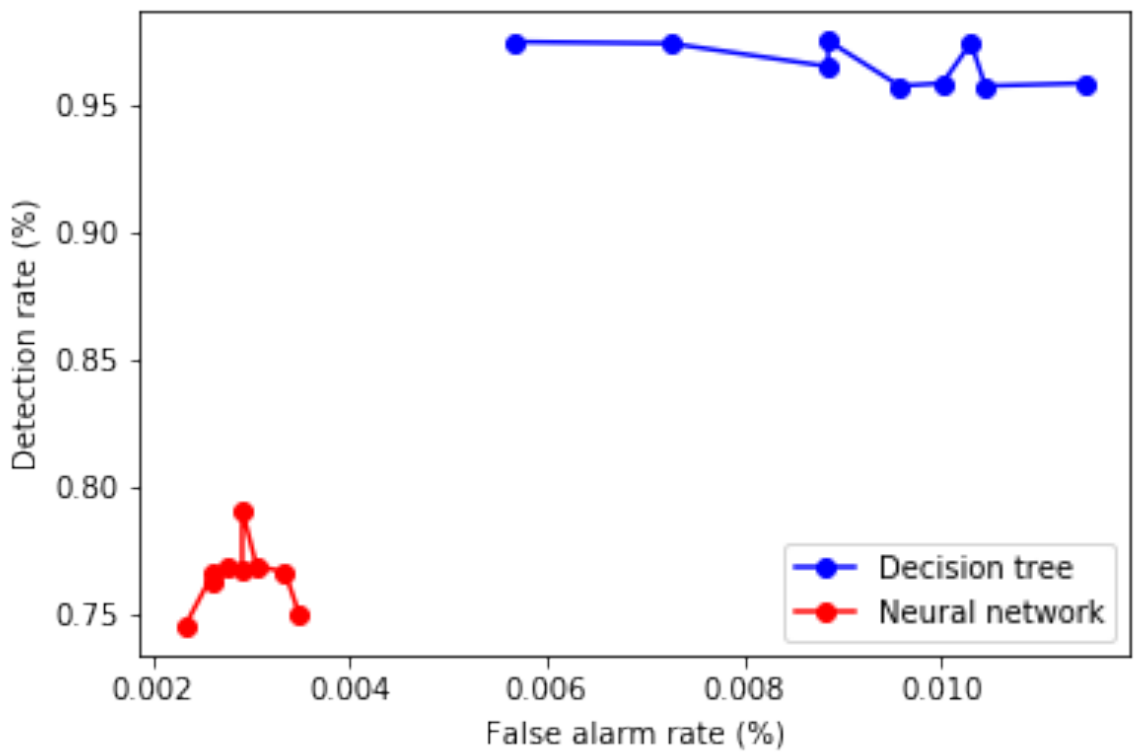


Figure 2. Impact of voting-based detection method on prediction performance. The points on each curve are obtained by the number of voters $N = 1, 3, 5, 7, 9, 11, 15, 17$, and 27 from right to left.

My result:



2) Calculate ‘time in advance’:
Result in paper:

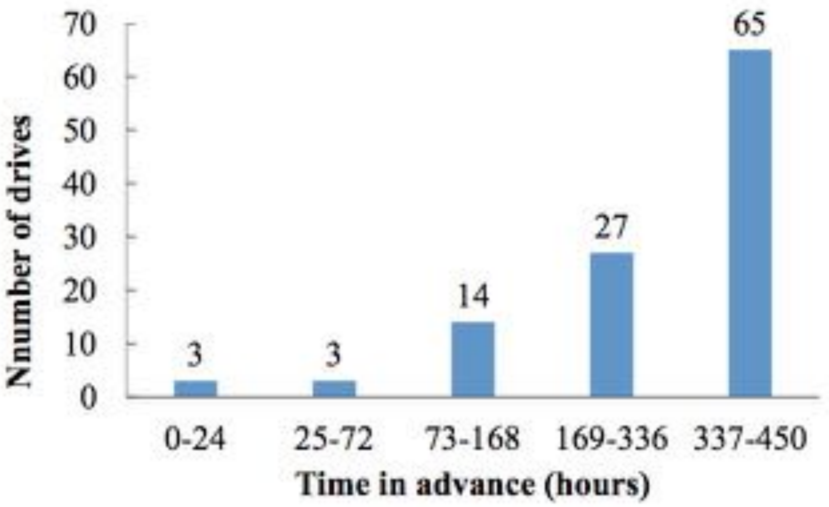


Figure 3. Distribution of time in advance of BP ANN model.

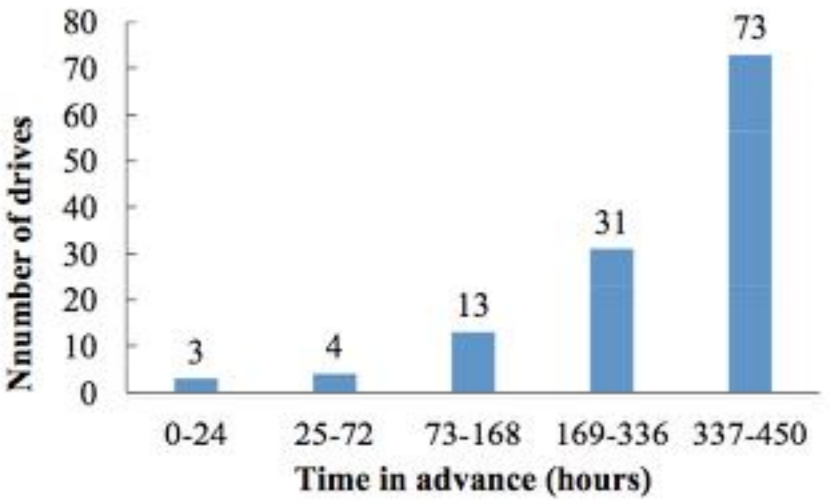
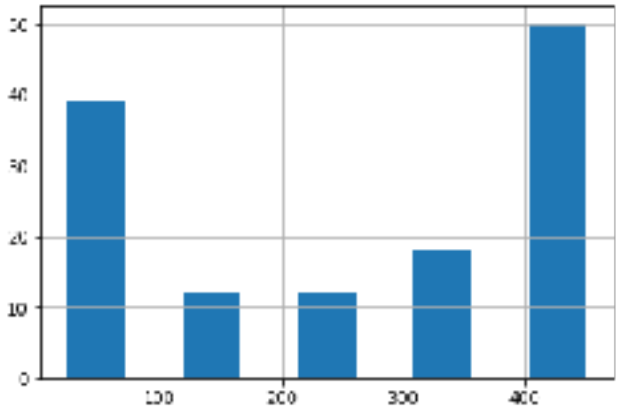
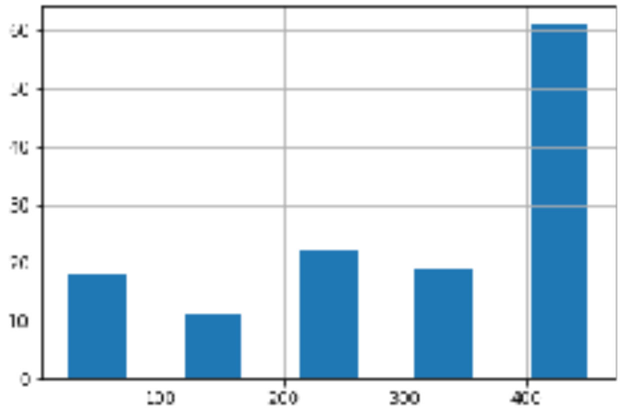


Figure 4. Distribution of time in advance of CT model.



Distribution of time in advance of BP ANN model.



Distribution of time in advance of CT model.

3) predict on smaller dataset: 10%, 25%, 50% and 75%
Result in the paper:

Model	Dataset	FAR (%)	FDR (%)
BP ANN	A	2.93	88.24
	B	1.10	90.63
	C	0.16	84.38
	D	0.03	81.82
CT	A	0.22	82.35
	B	0.07	90.63
	C	0.11	90.63
	D	0.09	91.82

My result:

Model	Dataset	FAR (%)	FDR (%)
BP ANN	A	0.377	57.627
	B	0.232	76.271
	C	0.015	66.949
	D	0.421	72.034
CT	A	0.842	84.034
	B	0.290	82.353
	C	0.682	89.076
	D	0.624	91.597

3. 'Improving Service Availability of Cloud Systems by Predicting Disk Error'

1) Dataset overview:

Dataset: Blackblaze 2017 Q4

Model name: ST4000DM000

2) Data preprocessing:

[1] create labels: use SMART 5 'Reallocated Sectors Count' as the error indicator. Labels are 'the number of days between the data is collected and the first error is detected'.

My question: How to handle those disks without any error?

[2]Dataset statistics:

Total number of disks that have SMART 5 error: 232

Total number of samples: 17284

[3] Feature identification:

All SMART attribute(46 features) + 3 kinds of statistical features:

Diff, Sigma and Bin

Totally 135 features

My questions: Should we consider the difference between cumulative features and noncumulative features?

[4] Feature selection:

An iterative algorithm to prune away non-predictive features.

Issue: zero accuracy, since it uses labels of higher range to predict labels of lower range.