StreamDFP: A General Stream Mining Framework for Adaptive Disk Failure Prediction (Supplementary File)

Shujie Han, Patrick P. C. Lee, Zhirong Shen, Cheng He, Yi Liu, and Tao Huang

The following materials provide supplementary results to our main file.

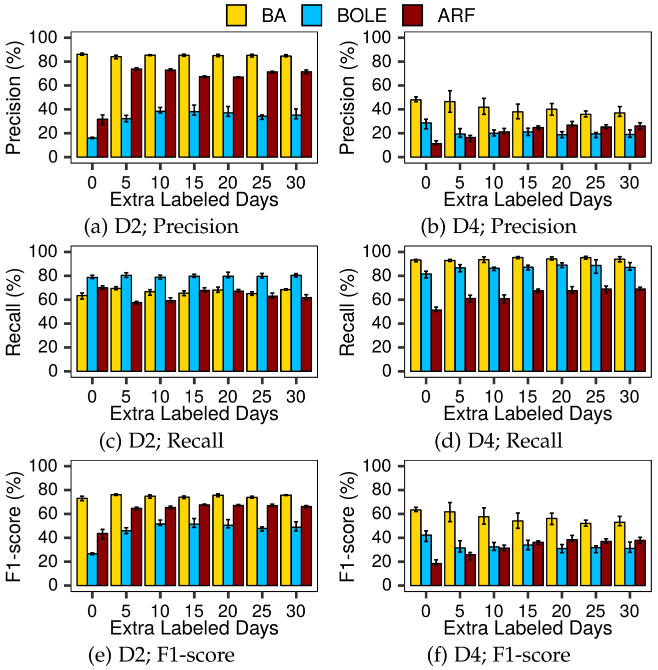
1 Impact of Parameters

In this section, we evaluate the impact of two parameters, including the extra labeled days D_L and the FPR threshold.

Exp#S1 (Sensitivity of extra labeled days). We study how the extra labeled days D_L affects the accuracy. We vary D_L from zero to 30 days; the zero days mean that we only label the samples as positive on the day when the failure occurs. Here, we focus on D2 and D4, which are derived from different disk manufacturers, and the three ensemble learning algorithms with concept-drift adaptation, i.e., BA, BOLE, and ARF.

Figure 1 shows the prediction accuracy versus D_L. In general, introducing extra labeled days (i.e., $D_L > 0$) increases the prediction accuracy especially for ARF in both D2 and D4 (by 23.4% and 19.9% F1-score, respectively) and BOLE in D2 (by 23.9% F1-score) when $D_L = 20$ days (our default setup). However, for D4, a smaller D_L for BOLE and BA achieves higher prediction accuracy. For example, the top two F1-scores are 63.4% and 61.8% for BA are on zero and five days, respectively. This implies that the optimal value of D_L varies across algorithms and datasets. Thus, StreamDFP opts to allow users to flexibly tune D_L based on production needs.

Exp#S2 (Impact of FPR thresholds). Machine learning models can be configured with a higher recall through increasing the FPR threshold (and vice versa). We study how different FPR thresholds affect the prediction accuracy, and examine if concept-drift adaptation still achieves accuracy gains. We focus on D2 as the representative dataset and bagging (including Bag and BA) as the representative algorithms.

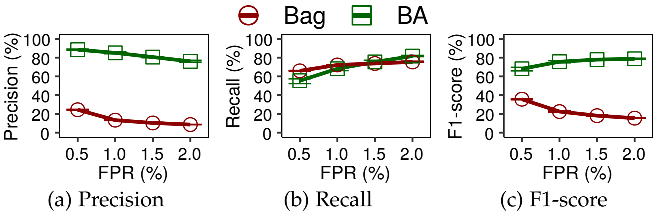


(a) D2; Precision (b) D4; Precision

(c) D2; Recall (d) D4; Recall

(e) D2; F1-score (f) D4; F1-score

Fig. 1: Exp#S1 (Sensitivity of extra label days). We focus on D2 and D4.



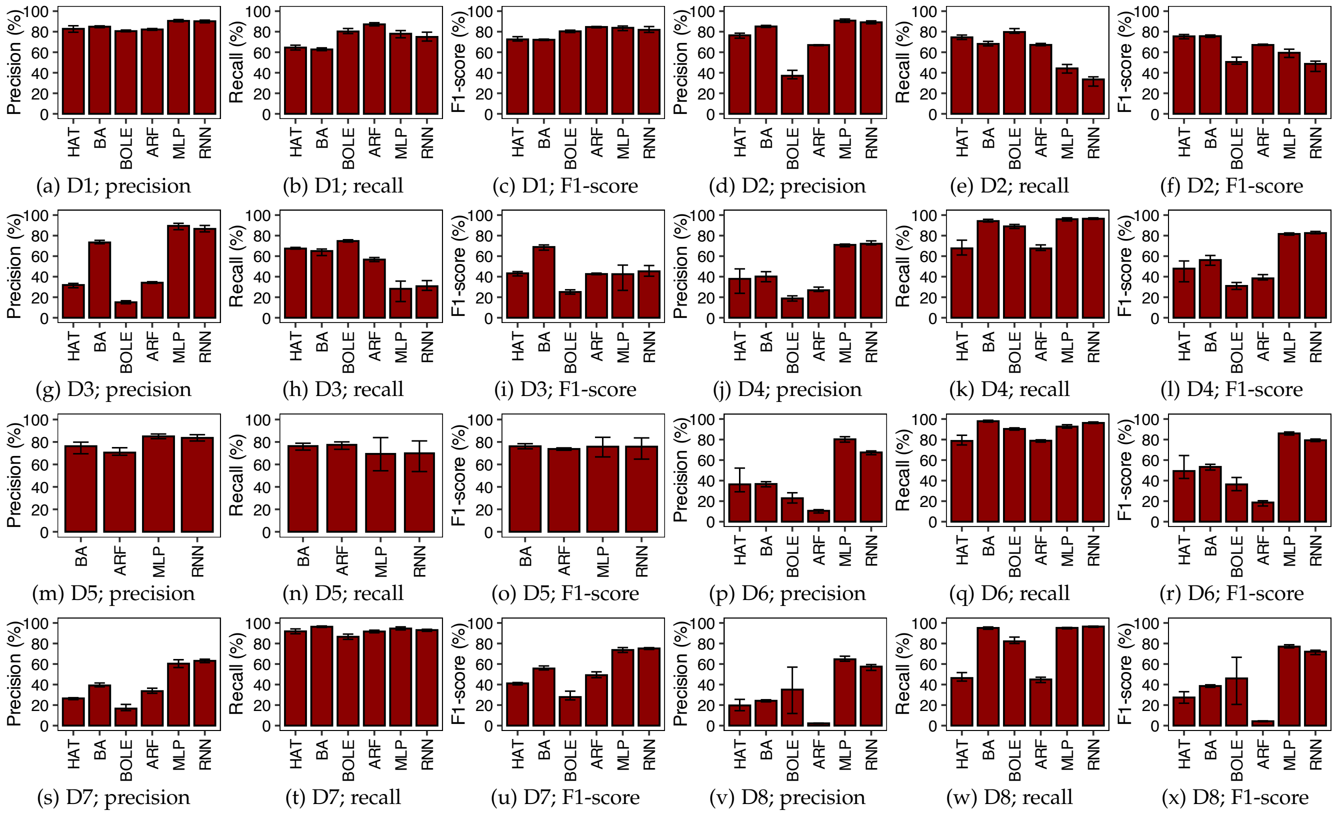
(a) Precision (b) Recall (c) F1-score

Fig. 2: Exp#S2 (Impact of FPR thresholds). We focus on Bag and BA for D2. Note that the error bars are not visible when the FPR is at least 1%.

Figure 2 shows the prediction accuracy of Bag and BA for D2 versus the FPR threshold (varied from 0.5% to 2.0%) and the FPR threshold of 1.0% is our default setup. BA improves the precision and F1-score of Bag by 64.0-71.8% and 32.5-63.3%, respectively, while its recall is less than Bag by 10.5% when the FPR is 0.5% but becomes higher than Bag by 6.3% when the FPR is 2.0%. It shows that BA improves the precision and F1-score significantly, while preserving the recall. We also emphasize that the relative differences between Bag and BA are consistent with our findings in Exp#1. We make similar observations for other datasets and algorithms under different FPR thresholds.

2 Recurrent Neural Network

In addition to MLP, we consider *Recurrent Neural Network (RNN)* [1], a state-of-the-art artificial neural network learning algorithm. Specifically, RNN comprises not only interconnected neurons like MLP, but also has recurrent connections



(a) D1; precision (b) D1; recall (c) D1; F1-score (d) D2; precision (e) D2; recall (f) D2; F1-score

(g) D3; precision (h) D3; recall (i) D3; F1-score (j) D4; precision (k) D4; recall (l) D4; F1-score

(m) D5; precision (n) D5; recall (o) D5; F1-score (p) D6; precision (q) D6; recall (r) D6; F1-score

(s) D7; precision (t) D7; recall (u) D7; F1-score (v) D8; precision (w) D8; recall (x) D8; F1-score

Fig. 3: Exp#S3 (Prediction accuracy of RNN).

between hidden neurons. These recurrent connections are used to store the past information and feed it together with the current inputs into the hidden layer. RNN forms a directed neural network along a temporal sequence, so it also exhibits temporally dynamic behavior. It can be also trained via stochastic gradient descent [2], and uses backpropagation through time (BPTT) [3] as concept-drift adaptation to propagate the errors between the predicted and true outputs to the neural network. As opposed to offline RNN learning that is used in disk failure prediction (e.g., [4]), we consider RNN with BPTT on stream mining and realize RNN with BPTT via stochastic gradient descent in StreamDFP.

Exp#S3 (Prediction accuracy of RNN). We evaluate the prediction accuracy of RNN on the datasets D1-D8. Figure 3 shows the prediction accuracy of RNN with BPTT compared with the other classification algorithms with concept-drift adaptation. We observe that RNN with BPTT achieves the highest F1-score on D4 and D7. The F1-scores of RNN are similar to those of MLP on the datasets with the absolute differences ranging from 0.0066% to 10.8%. The reason of small differences between RNN and MLP is that the correlations between the samples from different disks on each day may be limited, although RNN is used to capture the correlations for a sequence of time-series data. We point out that RNN is not always the "best" algorithm across datasets. This conforms to the main design goal of StreamDFP that it serves as a general framework to support various machine learning algorithms instead of a specific machine learning algorithm.

References

[1] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323:533–536, 1986.

[2] D. Saad. Online algorithms and stochastic approximations. *Online Learning*, 5:6–3, 1998.

[3] P. J. Werbos. Generalization of backpropagation with application to a recurrent gas market model. *Neural networks*, 1:339–356, 1988.

[4] C. Xu, G. Wang, X. Liu, D. Guo, and T.-Y. Liu. Health status assessment and failure prediction for hard drives with recurrent neural networks. *IEEE Trans. on Computers*, 65:3502–3508, 2016.