VCM: An Economic Value Cost Perspective for Hard Drive Failure Prediction

Tianming Jiang[†], Ping Huang[‡], Hao Xu[♭], Yongfeng Ji[♭], Wenjie Liu[†], and Ke Zhou[†]

[†]Wuhan National Lab for Optoelectronics, Key Laboratory of Information Storage System (School of Computer Science and Technology, Huazhong University of Science and Technology), Ministry of Education of China
[‡]Department of Computer and Information Sciences, Temple University, USA
[♭]Baidu, Inc., China

***Abstract*—Self-Monitoring, Analysis and Reporting Technology (SMART) is a technology in hard disk drive to predict the impending disk failure for data repair in advance. The prediction performance of the threshold-based method of SMART, however, is unsatisfactory. Recently, statistic and machine learning methods using SMART attributes as features have been explored to improve the prediction accuracy. Although some of the methods achieve satisfactory high failure detection rate and low false alarm rate, there is no unified metric to measure the overall revenues of these methods. In this paper, we introduce an economic value cost model to maximize the value brought by disk failure prediction. To evaluate the effectiveness of our method, we implement a multi-stage model integrating BP network, leaky bucket algorithm and value cost model. A real-world data set consisting of 94, 499 hard disk drives is used to train and test our hard drive failure prediction model, and the experimental results demonstrate the effectiveness of our model.**

I. Model Review

We use a three stages model, which takes real economic value cost and time series into account, to predict disk drive failures. As shown in Figure 1, we train a BP neural network to forecast the probability of disk failure in the first stage. The a threshold is selected according to the ROC curve produced by the BP prediction model. In the last stage, a leaky bucket is applied to decrease the false alarm rates while not impacting the failure detection rate significantly.

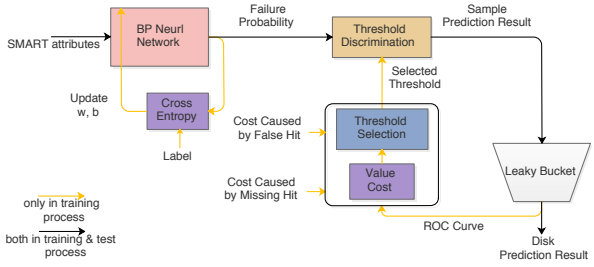

Fig. 1. Value cost model for disk failure prediction.

A. Data Collection

We collect a dataset containing 94,499 disk drives in a single running data center of Baidu Inc. [1]. Out of the total disks, there are 2,442 failed disks and 92,057 good disks.

B. Value Cost Model

The cost of erroneous predictions falls into two types, false hit cost and missing hit cost [2]. Applying to the storage systems in our scenarios, disk replacements mean higher storage price, and the remote data repair process consumes bandwidth requirement [3]. For convenience, the two prices are denoted as follows:

1)Storage price: S USD per bit,

2)Bandwidth price: B USD per bit.

Given the capacity of one disk is C bits and remote data repair of one disk consumes R bits, then the total value cost of the prediction model can be formulated as:

$$\begin{aligned} Cost &= B * R * FN + S * C * FP \\ &= B * R * N * FAR + S * C * P * (1 - FDR) \end{aligned}$$

where N and P represent good disks number and failed disks number of the total data set, respectively.

II. Experimental Results

In our value cost model, the economic cost caused by false hits and missing hits are taken into account to minimize the total cost. The benefit of the time series feature is presented in Figure 2. Especially, results of our model outperform the Logistical Regression model both with or without time series features.

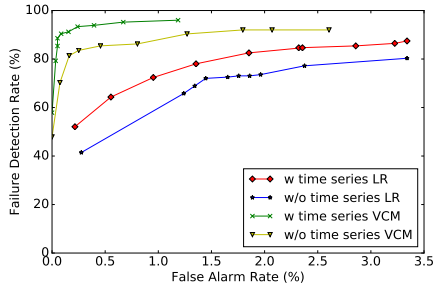

Fig. 2. Prediction performance of LR and VCM.

Acknowledgment

The research is supported in part by the National Natural Science Foundation of China (NSFC) under grant No.61232004, 61502189.

References

[1] W. Yang, D. Hu, Y. Liu, S. Wang, and T. Jiang, "Hard drive failure prediction using big data," in *Reliable Distributed Systems Workshop (SRDSW), 2015 IEEE 34th Symposium on*. IEEE, 2015, pp. 13–18.

[2] K. Zhou, Y. Liu, J. Song, L. Yan, F. Zou, and F. Shen, "Deep self-taught hashing for image retrieval," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1215–1218.

[3] Z. Huang, H. Jiang, K. Zhou, C. Wang, and Y. Zhao, "Xi-code: A family of practical lowest density mds array codes of distance 4," *IEEE Transactions on Communications*, vol. 64, no. 7, pp. 2707–2718, 2016.