Machine Learning for Storage System Reliability Prediction

Author's name: YANG zhi lin
the Chinese University of Hong Kong

supervisor's name: Patrick P. C. Lee

Introduction

Disk failure is a critical issue for large scale storage systems in big data centers. In this project, I tried to implement different machine learning based methods presented in 3 papers to predict disk errors or failures. Finally, I compared these methods in a comprehensive view. The objective of this research project is to enable scheduled system maintenance and proactive disk replacement in place of inefficient repair procedures. It is expected that an integrated solution combining the merits of these methods can be proposed for service availability improvement in future.

Experiment

1.paper title: 'Predicting Disk Replacement towards Reliable Data Centers'[1]
Procedures:
Preprocess data; Detect change points in time series; Compact time series; Downsample data; Train classifiers
Results:

	GBDT	SVM	DT	LR	RF	RGF
precision	0.912	0.334	0.875	0.723	0.905	0.934
recall	0.911	1.000	0.891	0.688	0.887	0.906
F1-score	0.911	0.500	0.883	0.699	0.896	0.919
Standard deviation	0.060	0.001	0.045	0.087	0.058	0.050

Table 1: performance of different machine learning models

2. paper title: 'Hard Drive Failure Prediction Using Classification and Regression Trees'[2]
Procedure: Data preprocessing; feature selection; training classifiers;
Results:

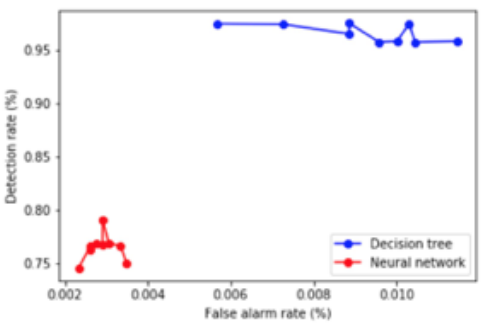

Figure 1: ROC curve of Decision tree and Neural Network

Model	Dataset	FAR (%)	FDR (%)
BP ANN	12 features	1.63	78.46
	19 features	2.52	83.72
	13 features	0.25	75.19
CT	12 features	0.46	91.54
	19 features	0.60	93.02
	13 features	0.60	93.02

Table 2: performance of CT and ANN with different feature sets

Acknowledgments:

Thanks for the help of my instructive Professor Patrick P. C. Lee and his PhD student Shujie Han.

3. Paper title: 'Improving Service Availability of Cloud Systems by Predicting Disk Error'[3]
Procedure: feature identification; feature selection; training classifiers
Results:
ROC curve and comparing different feature engineering methods:

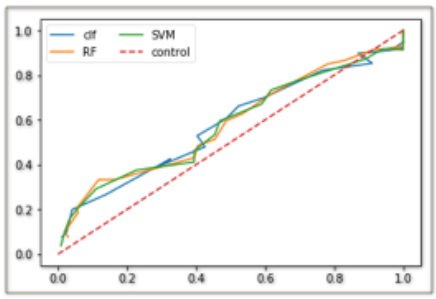

Figure 2: ROC curve of different machine learning models

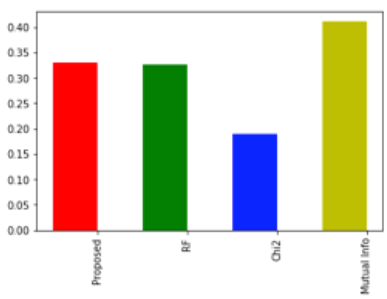

Figure 3: True Positive Rate (TPR) obtained using different feature engineering methods

Conclusions

1. Generally, tree-based machine learning methods are far superior to other types of machine learning algorithms in predicting the disk failure, especially those ensemble method such as Gradient Boosting Decision Tree and Gradient Boosting Regressor.
2. Apart from the choice of machine learning model, the feature engineering is also a crucial factor that will vastly affect the final results.
3. The quality of dataset is also an influential factor that determines the final result. Thus, it is worthwhile to study the transfer learning method in the future, which enables the use of existing trained prediction model on new disks model.
4. There are certain issues that should be considered when employing machine learning methods into the practical use, for example, the imbalance of disks classes and data noise, which has great impact on the prediction performance of machine learning models.

Reference:

[1] M. M. Botezatu, I. Giurgiu, J. Bogojeska, and D. Wiesmann, "Predicting Disk Replacement towards Reliable Data Centers," *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 16*, 2016.
[2] J. Li, X.P. Ji, Y.H. Jia, B.P. Zhu, G. Wang, Z.W. Li, X.G. Liu, 'Hard Drive Failure Prediction Using Classification and Regression Trees,' *2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, Atlanta, GA*, 2014, pp. 383-394.
[3] Y. Xu, K.X. Sui, R. Yao, H.Y. Zhang, Q.W. Lin, Y.N. Dang, P. Li, K.C. Jiang, W.C. Zhang, J.G. Lou, M. Chintalapati, D.M. Zhang, 'Improving Service Availability of Cloud Systems by Predicting Disk Error,' in *Proceedings of the 2018 USENIX Annual Technical Conference*, 2018.