Name: Prashanth Mallyampatti
Unity ID: pmallya
Student ID: 200250501

CSC724-Advanced Distributed Systems
Paper Review

Correlating instrumentation data to system states: A building block for automated diagnosis and control

Summary:

This paper proposes a method to categorize, diagnose and evaluate system metrics to classify them as a type of failure condition and repair them. Large scale systems are prone to failure, which have to detected and classified and if possible, provide a corrective measure so as to prevent any further happenings. Various failure detection models – unsupervised and supervised learning models have been proposed with each having it's own pros and cons.

In this paper the author discusses about Tree-Augmented Bayesian Network model an automated learning model. This model is applied on SLO state, system metrics obtained from 3-tier web application with Java middleware component. The proposed approach is a classic classification problem rather than anomaly detection. TAN gives the probability that a state is S for an observed metric M, which considers metric relationships and thus forming an optimal tree on a Bayesian network. An additional feature of interpretability and modifiability has been defined with a mathematical equation which gives an insight into the causes of violation and sometimes how to repair it. This model was applied on synthetic, environment and testbed workloads to determine the accuracy of the classification. The workloads covered request rates, connection requests, database and app-server intensive requests for unbiased testing.

The results of the model were noticeably good with 87%- 94% accuracy. This range includes various workloads and different conditions. Metric aggregation and adaptation while testing have shown better results than usual classification. The accuracy of TAN for forecasting SLO violations is good on certain workloads and should be used accordingly. Since TAN uses no prior knowledge of the system the above results are very interesting and can be applied to a wide range of systems.

Strong Points:

1) Use of TAN model for classifying data is the main strength of this paper as Bayesian Networks have been proven to have better efficiency than most of the classifiers.
2) Aggregation of metrics is an amazing approach towards classifying and its results shows it has better accuracy.
3) The paper considers various types of workloads and tabulates their accuracies which provides the user a reference for better comparative studies.

Weak Points:

1) More mathematical proof or results of interpretability and modifiability could have been shown, as the paper discusses very briefly about it.
2) Use of external failure detectors can impose unnecessary problems and effect the accuracies achieved.
3) The overheads of using the model have not been discussed, which if not considered can hinder achieving greater accuracies.