Analyze historical performance metrics to predict critical system errors before system degradation

Anthony Pagan

Abstract

The main goal of this paper is to build a proactive monitoring system model based on performance metric calculations that can be correlated to events and allow teams to build automated resolutions. Such a system would reduce support cost and increase efficiency while minimizing false non-actionable alerts.

1. Introduction

Many organizations rely on monitoring systems to monitor performance of their backend systems. The expectation is that the monitoring will notify support staff when critical systems are in a critical state. Unfortunately, many time support teams are only alerted after a system is degraded. In some cases, manual monitor tuning results in false alerts.

The main goal of this project is to build a proactive monitoring system model based on performance metric calculations that can be correlated to events and allow teams to build automated resolutions. Such a system would reduce support cost and increase efficiency while minimizing false non-actionable alerts.

2. Journal Reviews

There are several journals with different approaches to proactive monitoring that include some self-healing mechanisms resolve system performance issues. Some methods include existing algorithms and some attempt to build their own. The existing algorithms range from multiple linear regression, LSTM, RNN, Autoregression ARIMA models, ANN as well as others.

P Anjitha describes LSTM as a combination of a long and short-term memory RNN. A RNN or recurring Neural Network based on LSTM is called an LSTM network. LSTM has an input gate, output gate and a forget gate [1]. This journal attempts to describe a way to predict web performance with LSTM RNN using sequence data via user requests in web server logs.

The Monte Carlo approach [2] uses long term collects non-stationary data and uses residual short-term process to capture both the long-term linear combination of certain basis function with random application evolving with time, while short term is modeled as a traditional Auto regression process. Parameter for long-term and short-term data are estimated using sequential Monte Carlo (SMC).

In the Multi-resource MModel prediction they use autocorrelation with cross validation to achieve higher accuracy rates [4] . The 2 models used to predict resources are LAST which predicts next resource value from previous measurements and EWMA which predicts next resources based on weighted averages of previous predictions and the previous measurements.

To increase accuracy it uses autocorrelation and cross correlation that provides relationship between multiple resources. Since resources have a constant mean is it is convenient to treat them as a zero-mean series plus a constant, so Mmodel can always work on zero-mean series. EWMA algorithm is used to update estimation of mean for new measurements. ARM is used to determine benefits of each adaption technique by analyzing the resulting SSE and MSE. These techniques require N-step ahead predictions to provide a more accurate prediction.

The LSTM approach uses LSTM auto-encoder to capture the temporal structures from monitoring data and use deep neural network to determine reach rate or Qos of CDN service providers [5]. The method uses deep learning which uses algorithms that attempt to model high level abstractions in data by using artificial network architecture composed of multiple non-linear transformations. This method uses artificial neural network (ANN) instead of Multiple Linear regression (MLR) . ANN, which is non-linear, can detect complex relationship among features, compare learning techniques and can extract hierarchical level of features from raw data which can help build more accurate models with less time. They use correlation for all features and cluster highly correlated features together to eliminate redundant information.

Another approach to LSTM uses RNN, a LSTM many to one network with requests-to-vector. It applies RNN-LSTM network to predict the performance and workload of web servers then investigate the relation between user's requests sequence and web server performance [8]. This allows to create an analog workload of user-request sequence to be analyzed to predict performance. The LSTM network includes an input, output and forget gate. It includes R for recurring input, W for weight matrices for input, b for bias vectors of each gate formula and uses a sigmoid function as the activation function for each gate. A rectified linear function (ReLUs) is also used as a function in the formula which makes the network be trained several times faster. All these combined make the model learn long -term dependencies and make model train without hand generated features, unlike basic RNN. In addition, the solution uses cross-entropy as the loss function to achieve good performance. Workload comparison use cosine similarity.

In this next approach, the algorithms used include the k-means and gaussian clustering transformed with PCA [6]. First each metric is scaled to zero and a unit of variance, then principal component analysis is used to transform the metrics to get top 3 dimensions. K-means provides measure of distance of examined point to the center, if it exceeds some limit it is marked an anomaly. Gaussian clustering algorithm computes a per data point probability of each cluster. Cross validation with 3 folders picked up model hyperparameters making the best performing model the benchmark. Same approach was used to train the random forest model.

Another approach is a distributed target computation (DTC) for dynamically balancing load using information from neighboring entities [9]. Although this paper is for wireless network, the method can be applied to a range of applications for resource management. The prerequisites for this method include a way for each node to identify other nodes, current metric value for each node, a mechanism for each node to exchange target and metric values bilaterally and a mechanism to influence the choice between nodes to serve client demands. Methods used

were load mean/maximal scaling, mapping distance between actual and target loads to a given CRE range using sigmoid mapping and use 95% of cell capacity load metric from previous 2 methods.

In the decision tree approach the paper tries to use non-intrusive metric to passively measure traffic metrics and electric currents monitored by sensors attached to the power supply [10]. The method uses PHP scripts to collect metrics from 3 servers connected via cross over cables. A client program estimate bit from client to server via httperf and CPU utilization is capture via OS performance monitoring. The tests show the non-intrusive method for network metrics confirmed degradation of performance when compared to OS metrics, but power supply monitoring did not. The first method also had higher accuracy rate then both the network metrics and power supply methods combined.

The next approach main focus includes designing a multivariate time series prediction framework for PRESS, AGILE, ARIMA, NARNN, LSTM and BLSTM for resource usage prediction, analysis and comparison of performance of proposed methods [11]. It uses different method for selection of relevant subset of features and prediction and stability-based joint framework for selection of suitable feature selection techniques. Pearson correlation is used for recognizing the set of relevant features for forecasting and granger causality for multivariate time series and Mutual Information Criterion can be used for time series. Pearson and Spearman correlation select relevant features based on correlation of the candidate metric with desired resource metric. Granger causality select features based on causality and mutual information based on information shared between features.

This final approach is a custom built self-healing monitoring autonomic unit solution proposed by IBM [12]. It's a MAPE loop which is composed of monitor, analyzer, planner and executor. This method is based on files and their attributes and separate files when there are on hard disk H-State and memory M-state. H state are considered static files and M state are static and dynamic. Dynamic files have inner IA and outer OA attributes. IA are attributes representing nature of files in memory like efficiency responding time, running cost …etc. OA are attributes of communication like source component, target component ..etc. This monitor model solution shows it can improve performance of self-healing by reducing the gain of monitor and recovery, and deduce performance cost of following recovery so that it has a higher monitor performance and lower missed rate than single aspect monitor.

4. Data Selection

5. Review and Analysis

6. Testing and Performance Review

7. Conclusion

WORK CITED:

[1] Anjitha P, "Web server Performance Prediction using a Deep Recurrent network",International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 05 Issue: 09 | Sep 2018 www.irjet.net p-ISSN: 2395-0072

[2] T. Vercauteren, P. Aggarwal, X. Wang and T. Li, "Hierarchical Forecasting of Web Server Workload Using Sequential Monte Carlo Training," in IEEE Transactions on Signal Processing, vol. 55, no. 4, pp. 1286-1297, April 2007, doi: 10.1109/TSP.2006.889401.

[3] Chonggun Kim and H. Kameda, "An algorithm for optimal static load balancing in distributed computer systems," in IEEE Transactions on Computers, vol. 41, no. 3, pp. 381-384, March 1992, doi: 10.1109/12.127455.

[4] Jin Liang, K. Nahrstedt and Yuanyuan Zhou, "Adaptive multi-resource prediction in distributed resource sharing environment," IEEE International Symposium on Cluster Computing and the Grid, 2004. CCGrid 2004., Chicago, IL, USA, 2004, pp. 293-300, doi: 10.1109/CCGrid.2004.1336580.

[5] Ziyan Wua, Zhihui Lu a,*, Wei Zhang a, Jie Wua, Shalin Huang b, Patrick C.K. Hung c, "
A data-driven approach of performance evaluation for cache server
groups in content delivery network",ScienceDirect, J. Parallel Distrib. Comput. 119 (2018) 162–171

[6] Martin Adam1,2,, Luca Magnoni2, Martin Pilát3, and Dagmar Adamová1,"Detection of Erratic Behavior in Load Balanced Clusters of Servers Using a Machine Learning Based Method",
EPJ Web of Conferences 214, 08030 (2019) https://doi.org/10.1051/epjconf/201921408030
CHEP 2018

[7] Pavel Barca1*,Bojan Vujanić1,Nemanja Maček2,"MONITORING AND PREDICTING LINUX SERVER PERFORMANCE WITH LINEAR REGRESSION",SINTEZA 2018 INTERNATIONAL SCIENTIFIC CONFERENCE ON INFORMATION TECHNOLOGY AND DATA RELATED RESEARCH

[8] Zheng Huang,1,2 Jiajun Peng,1 Huijuan Lian,1 Jie Guo,1 and Weidong Qiu1,"Deep Recurrent Model for Server Load and Performance Prediction in Data Center" Wiley,Volume 2017, Article ID 8584252, 10 pages
https://doi.org/10.1155/2017/8584252

[9] Per Kreuger Rebecca Steinert Olof Görnerup Daniel Gillblad,"Distributed dynamic load balancing with applications in radio access networks",Wiley , DOI: 10.1002/nem.2014

[10] Satoru Ohta, "Obtaining the Knowledge of a Server Performance from NonIntrusively Measurable Metrics",International Journal of Engineering and Technology Innovation, vol. 6, no. 2, 2016, pp. 135 - 151

[11] Shaifu Gupta1 · A. D. Dileep1 · Timothy A. Gonsalves1, "A joint feature selection framework for multivariate resource usage prediction in cloud servers using stability and prediction performance", The Journal of Supercomputing (2018) 74:6033–6068 https://doi.org/10.1007/s11227-018-2510-7

[12] D. Xikun, W. Huiqiang and L. Hongwu, "A Comprehensive Monitor Model for Self-Healing Systems," 2010 International Conference on Multimedia Information Networking and Security, Nanjing, Jiangsu, 2010, pp. 751-756, doi: 10.1109/MINES.2010.159.