Predictive Fingerprinting of System State

Helen Cunningham
VMware, Inc.

Abstract. This work explores visual fingerprinting for pre-emptive identification of performance problems in computing systems. Simple statistical methods provide a system administrator with a 2-dimensional **visual fingerprint** of each computing system, suggesting whether the system is normal or abnormal. The method also has a statistical solution. Tests show that detectable fingerprint differences emerge early in an operational day and predict performance problems later in the day.

Background

Performance problems in a datacenter are unpredictable, appearing suddenly and disappearing just as quickly. Adaptive thresholding of individual metrics is sensitive to current problems, but no *single* metric is a good predictor of future state. Therefore we have developed a multivariate approach.

This method is fast and computationally lightweight, trading deterministic precision for sensitivity to faint signals. Correlation reduces a time series (with 10s of 1000s of data points) down to a single value (the correlation coefficient). Multiple correlation characterizes any number (m) of metrics with an m x m correlation matrix.

In a slight departure from the standard use of "fingerprinting" in computer science, (e.g., Tadayoshi Kohno, Broido & Claffy 2005, Conti & Abdullah 2004, "fingerprint" here refers to a data reduction that characterizes a dataset with respect to one or more target variables, but need not uniquely identify the dataset.

Defining normal behavior

The data for this study came from UNIX metrics on blade-configured servers running in a thin client network. Using data sets from known normal and abnormal machines, multiple correlation was used to generate a surface-colored fingerprint. K-means clustering was used to order variables in the 2-D fingerprint so as to optimize visual analysis.

hcunningham@vmware.com

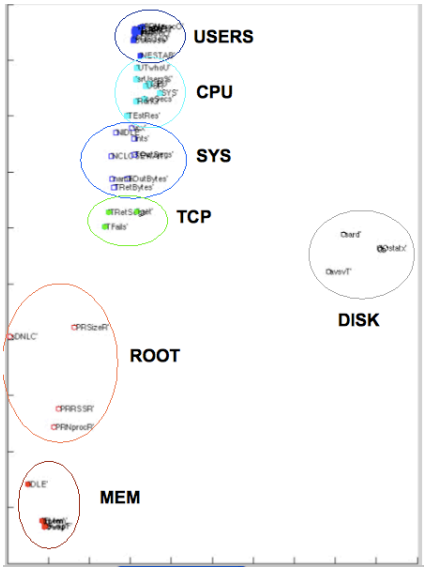
IEEE Conference on Visual Analytics Science and Technology 2012
October 14 - 19, Seattle, WA, USA
978-1-4673-4753-2/12/$31.00 ©2012 IEEE

The bayesian information criterion (BIC) was used to determine optimal number of clusters. The BIC is:

$$\mathbf{BIC} = RSS + k\ log(n)$$

RSS is residual sum-of-squares of the cluster analysis
n is the number of samples
k is the number of clusters in the analysis
Optimal value from the BIC was 7, and Figure 1 shows the result for 7 clusters. Cluster centroids correspond nicely with known UNIX system metrics.

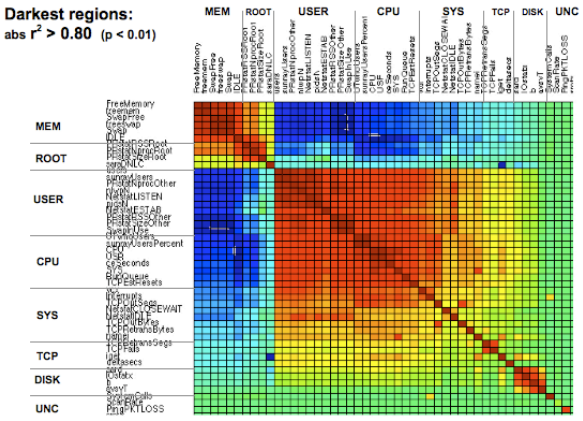
Figure 1: K-Means Clusters of UNIX metric time series



Fingerprints of normal machines are highly similar (mean correlation 0.93) so their correlation matrices were combined by cellwise averaging to form a normal template. In Figure 2, dark red cells indicate high positive correlation, dark blue indicates high negative correlation, and pale green indicates low correlation.

"Abnormal" machines produce dissimilar fingerprints (mean correlation 0.72) and so are not combined. Figures 3-5 show how normal correlation structure can break down: 1) increases in zero correlations: 2) banding corresponding to particular variables. Arrows mark areas of visual interest.

Figure 2 - The 'normal' fingerprint



Abnormal fingerprints

Figure 3

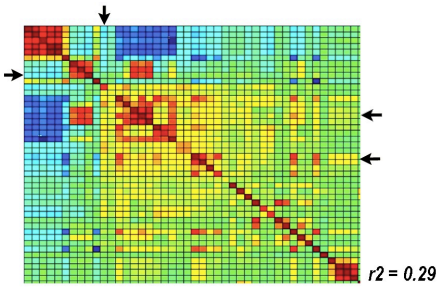


Figure 4

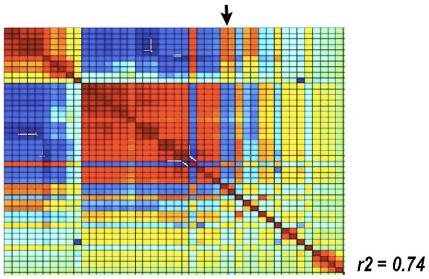
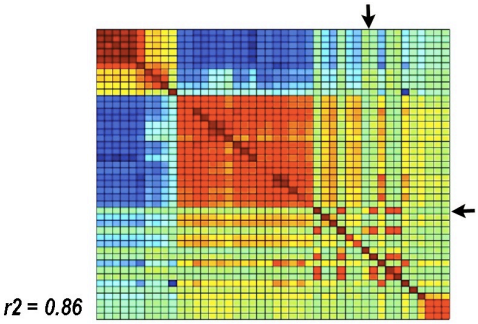


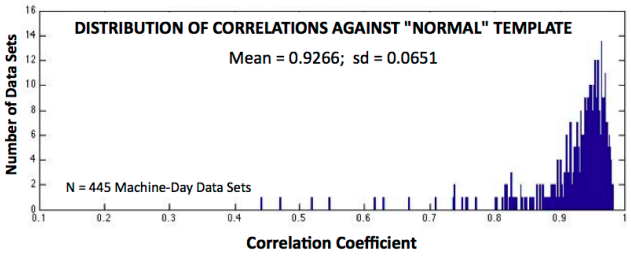
Figure 5



Statistical Approach

Distance from normality can be computed by correlating a candidate machine's correlation matrix with the normal template (after unraveling the matrix into a vector). The resulting distribution of correlations against the template is truncated normal (max correlation =1.0) and provides a statistical basis for identifying abnormal machines based on outliers.

Figure7 – Distribution of Correlations



The statistical approach was tested by using it to identify "suspect" systems early in the day and sending them to system administrators for further analysis. The admin examined and monitored the machine for the rest of the day and then classed each "suspect" as either normal or abnormal. Table 1 shows the number of cases in which the admins agreed or disagreed with each prediction.

	Suspects (54)	Non-suspects (232)
Sys Admin: Trouble Found	41	7
Sys Admn: No Trouble Found	13	225

Assuming a rate of abnormal hosts consistent with the number of suspects generated, then a randomly drawn sample of 54 suspects should have netted only 23% with discoverable problems (12-13 hosts). In fact our method contained 76% with discoverable problems (41 hosts). The observed proportion is thus statistically significant (<0.0001 probability of getting this result by chance).

References

Tadayoshi Kohno, Andre Broido, K.C. Claffy, "Remote Physical Device Fingerprinting," *IEEE Transactions on Dependable and Secure Computing*, vol. 2, no. 2, pp. 93-108, April-June 2005, doi:10.1109/TDSC.2005.26.

Gregory Conti & Kulsoom Abdullah, "Passive visual fingerprinting of network attack tools", *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*
Pages 45 – 54.