Detecting Anomalies for High Performance Computing Resilience

Emmanuel Aviles, Shamir J. Quiñones, and Yael M. Camacho
Polytechnic University of Puerto Rico

Houssain Kettani, Faculty Advisor
Polytechnic University of Puerto Rico

George Ostrouchov, Mentor
Computer Science and Mathematics Division
Oak Ridge National Lab

Background

As high performance computing systems get larger and more complex, time between node failures decreases. Proactive measures that migrate application components to healthy nodes can increase time between failures and enable application progress. However, proactive measures require timely system information and an ability to predict where failures are likely to occur. Detecting anomalies is the first step to identifying failures and eventually developing a failure prediction capability. The main result of this project is a number of analysis tools for anomaly identification that are based on the R open source software environment for statistical computing and graphics and on the GGobi open source visualization program for exploring high-dimensional data. The tools we developed are for numerical data, they do not assume any specific set of system attributes, and they use only those attributes that display variability.

Clustering

Given a large collection of system attributes recorded at regular time intervals, the data are divided into node-periods and one set of derived attributes is constructed for each node-period. These node-periods are then clustered. Clustering is a method of unsupervised learning in which high-dimensional data sets are grouped together based on the relationships among them. A distance metric is used to measure the relationships among the data. We use the Euclidean distance on data that is scaled to have the same variability in each dimension so that results are invariant to units of attribute measurement. A single linkage hierarchical clustering algorithm constructs a dendrogram from the data. Anomalous behavior in single linkage clustering shows up as outliers or small outlying clusters.

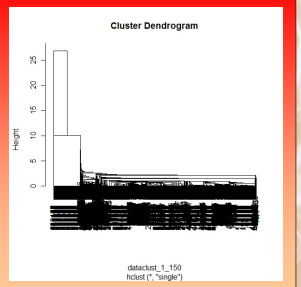

Figure 1: Hierarchical cluster dendrogram of data collected from xtorc.

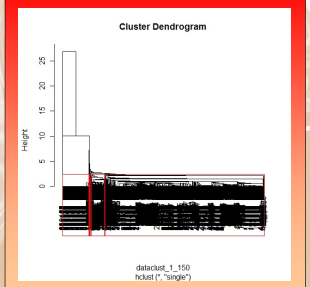

Figure 2: The same dendrogram cut at 20 clusters.

Anomaly Detection and Exploration

We compute an anomaly measure that is inversely proportional to cluster size. Node-periods are displayed by GGobi in the derived attribute high-dimensional space. Using R to control GGobi, we color node-periods by their anomaly measure ranging from blue as not anomalous, through yellow to red as highly anomalous. GGobi can then be used to explore the anomalies in any projection of the multivariate attribute space.

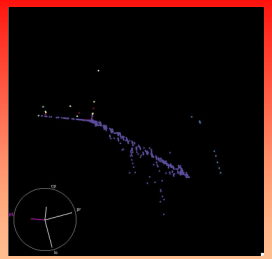

Figure 3: A GGobi projection of node-periods in a combination of four attributes emphasizing anomalies.

Figure 4: A GGobi projection of node-periods in a node by period arrangement. Faults, which show up as anomalies, were injected during data collection.

Displaying High-Dimensional Attribute Relationships

Using the same anomaly measure proportional to cluster size, we compute the contribution of each attribute to the distances between the largest cluster to the remaining clusters. The four attributes with the largest contribution are used to display the node-periods in a series of conditional scatter plots. The conditional scatter plots show a four-dimensional relationship. Typically the four-dimensional relationship of the displayed attributes provides an explanation of the anomalies and clues to why they are different from the rest of the node-periods.

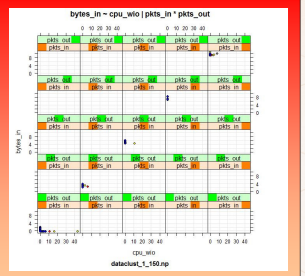

Figure 5: Conditional scatter plot series showing a four-dimensional explanation of the anomalies.

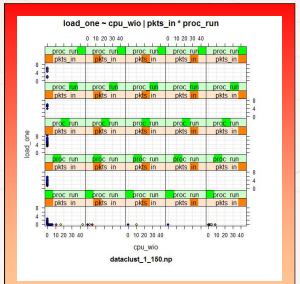

Figure 6: The same data showing another four-dimensional explanation discovered in GGobi.

Conclusion

We have developed a set of analysis tools that can be used in high performance computing operations to detect and analyze anomalous behavior. Identifying anomalies is the first step toward timely failure detection. The development of these tools required learning many new tools and concepts including the Unix operating system, the R environment for statistical computing and graphics, the GGobi visualization program, clustering techniques, high-dimensional data, and statistics. Overall, the experience was mainly educative.

Future Work

Detecting anomalies leads to identifying failures and eventually to the development of a failure prediction capability. Future work can include classification of anomalies into those that require attention and those that are only informative. A set of discovered anomalies can be used to collect data segments that contain predictive information about those anomalies. With appropriate data, failure predictive capabilities can be built.

Acknowledgements

We would like to thank the Department of Energy (DOE) for providing the funds for The Faculty and Student Teams (FaST) program, Oak Ridge National Laboratory (ORNL) for providing the facilities and tools for this appointment, and Oak Ridge Institute for Science and Education (ORISE) along with Oak Ridge Associated Universities (ORAU) which made the necessary accommodations to make sure we had a pleasant experience. Our mentors, Dr. George Ostrouchov from ORNL and Dr. Houssain Kettani from the Polytechnic University of Puerto Rico for providing guidance and support throughout the whole experience. We would also like to thank Hector J. Machin from University of Turabo for extracting data from xtorc, and Thomas Naughton from ORNL for providing guidance on using xtorc. This work was partially supported by the Department of Defense and the Extreme Scale Systems Center at the Oak Ridge National Laboratory.

References

- The R project for statistical computing, www.r-project.org
- GGobi data visualization system, www.ggobi.org

National Center for Computational Sciences
Polytechnic University Puerto Rico
United States Nuclear Regulatory Commission
ORISE
Office of Science U.S. Department of Energy
FaST
Oak Ridge National Laboratory