Characterization of Single Port Time Series Via Hierarchical Cluster Analysis

Jeffrey L. Solka
NSWCDD, Code B10
17320 Dahlgren Rd.
Dahlgren, VA 22448-5100
jsolka@nswc.navy.mil

1. Abstract

This article discusses the application of hierarchical cluster analysis to single port machine time series. The characterization of these time series allow one to group a machine's daily single port activity time series. In this manner one can attempt to identify those days where the machines activity, on this particular, port may have been abnormal. These types of "abnormal days" can be caused by a multitude of different reasons including system compromise.

2. Port Activity Time Series

There are numerous types of network traffic that can be monitored and analyzed. One could focus attention on the activities at all of the ports on a group of machines. The analysis of packet counts at the various ports on a group of machines sans any sort of temporal information has been the focus of some of my previous work [5].

The traffic associated with a particular port on a group of systems viewed as a time series is the focus of the analysis discussed during this paper. Ports that are numbered less than 1000 typically are designated to run specific root processes although in practice which root owned service is running on which port is configurable. Even when one focuses attention on a particular port one still needs to decide the protocol(UDP, ICMP, or TCP) and the particular flags that are set within the packet.

The focus of attention for the discussions within are those packets with the syn ack flags set in them that orgininate on port 80 of the machines in question. These packets indicate that the machines in question are engaged in traffic that is usually associated with web pages during the time in question.

There still remains the question of what time granularity to use during the analysis. One hour is the finest time resolution studied in the paper. So the port 80 syn ack activity associated with a particular machine during a 24 hour time period constitutes a time series of length 24.

3. Cluster Analysis

Given a collection of port 80 syn ack time series associated with a group of machines, one needs a way of dividing the machines into homogeneous groups. The division of data into groups falls into the domain of statistical cluster analysis.

3..1 Hierarchical Agglomerative Clustering

In the absence of any information concerning the underlying number of groups in the data set one is well advised to employ hierarchical cluster analysis. Agglomerative hierarchical cluster analysis starts with each point contained in a singleton cluster and then proceeds through a series of iterations where at each iteration the two closest clusters are merged. The algorithm ultimately stops when all of the observations reside within a single cluster. One needs to provide a procedure to measure distances between clusters in order to implement the procedure.

So in stead of providing a particular clustering the agglomerative clustering methodology provides a continuum of clustering of a given data set from as many clusters as observations to all of the observations contained in a singleton cluster. The reader is referred to [1] for a thorough treatment of cluster analysis.

3..2 The Choice of the Number of Clusters

The choice of the number of clusters is a subject of continued research within the statistics community. This is due in part to the difficult nature of the process. In fact under the most general assumptions it is probably impossible to ascertain the "true" number of clusters and their inherent structure. The method of Mojena was used to ascertain the number of clusters in the port activity time series data [4].

Mojena suggested that one should select the number of groups corresponding to the stage in the dendrogram where

$$\alpha_{j+1} > \overline{\alpha} + k s_\alpha \tag{1}$$

where $\alpha_0, \alpha_1, \ldots, \alpha_{n-1}$ are the fusion levels corresponding to the stages with $n, n-1, \ldots, 1$ clusters. The terms $\overline{\alpha}$ and s_α are defined as the mean and unbiased standard deviation of the α-values and k is a constant. Mojena recommends that k should be chosen between 2.75 and 3.50. Milligan and Cooper [3] suggest a value of k of 1.25 and this is the value that is used in the results section.

4. Results

Example clustering results were produced based on an analysis of a small set of data collected at my home site facility.

4..1 Experimental Design and Data Collection

TCPDUMP was used to to produce data files for analysis. These data files characterize syn ack packet counts for each of the active ports on each of the machines in the machine set designated for study. Counts were recorded for each hour during a 6 month period from March 1,2000 till August 31,2000. A simple PERL script was written to extract those syn ack counts associated with port 80 for 5 test machines.

All statistical analysis on the data was performed using the R statistical processing language. R is a public domain implementation of the S language and it is available from //lib.stat.cmu.edu/R/CRAN. A good reference for the reader interested in the history and development of R is [2].

4..2 Examples

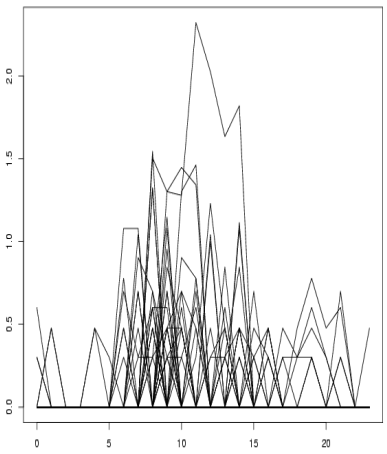

Figure 1: The plot presents the hourly port 80 syn ack time series for machine A during the period of time from March 1 of 2001 till August 31 of 2001. The x-axis designates hours past midnight and the y-axis denotes the log of the counts.

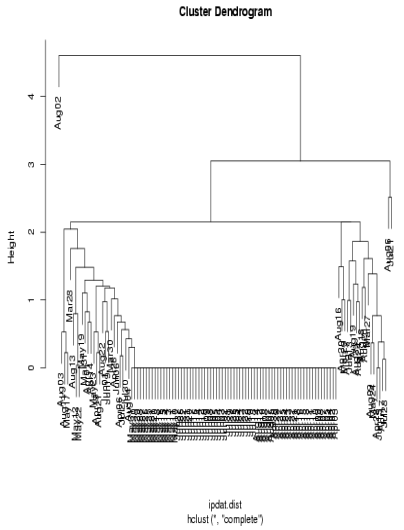

Figure 2: This plot presents a dendrogram of the port 80 syn ack daily time series from March 1, 2001 till August 31,2001 for machine A. The clustering was performed using a complete linkage clustering scheme with a Euclidean distance matrix.

Figure 1 presents a simple line plot representing the day by day time series associated with port 80 on this machine designated A. There is clearly one day whose time series seems to be an outlier as compared to the other series.

The time series were not scaled prior to processing. It was decided to use to standard Euclidean distance metric in order to compute the distance between the time series associated

with each day. Figure 2 presents a dendrogram plot representing the output of the agglomerative clustering procedure on the day by day time series associated with port 80 on this machine. The dendrogram clearly indicates the outlier role being played by the time series associated with the August 2 date.

The next step in the analysis process focused on determining the appropriate number of clusters. As discussed above the method of Mojeena was employed. Based on an analysis of the z-scores the data was broken into 13 clusters.

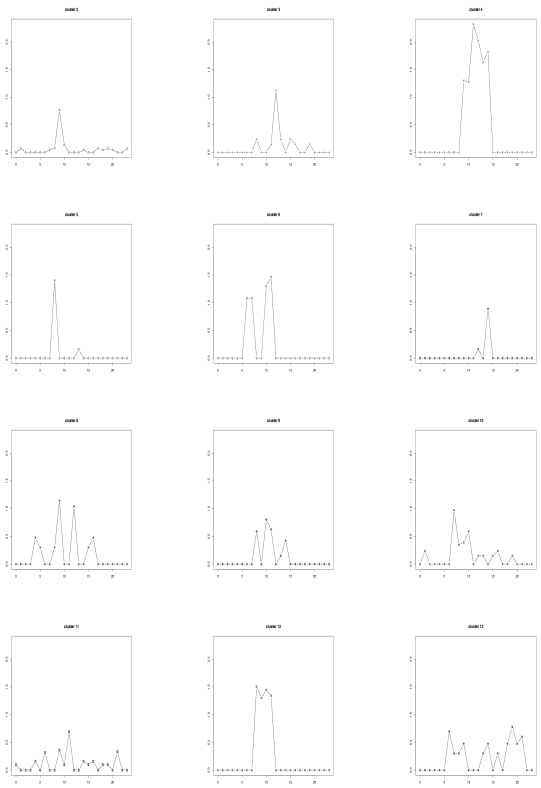

Figure 3: These plots present the mean expression levels associated with 12 of the 13 clusters. The omitted mean expression level plot corresponded to those machines that had very little activity during the time period.

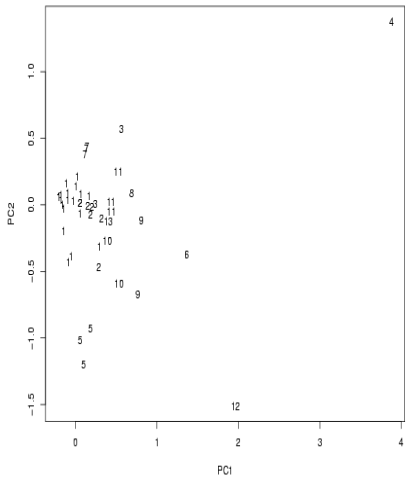

Figure 4: A scatter plot of the cluster structure for the time series data from March 1, 2001 till August 31,2001 for machine A. The plot has been rendered in the first two principal components. Each observation has been labeled with its respective cluster identifier.

Figure 3 presents plots of the mean expression levels for 12 of the 13 clusters of the syn ack data. The omitted plot corresponds to cluster number 1 where the port 80 syn ack activity on the machines in this cluster was very low. One can also examine the cluster structure via a projection into a lower dimensional space. One of the simplest ways to perform this projection is via the use of principal component analysis. Figure 4 contains a plot of our twenty four dimensional observations plotted in a space spanned by the first two principal components.The

outlier nature of cluster 4 which is a singleton cluster containing Aug 02 is very prominent in this plot.

5. Acknowledgments

Support for this effort is provided in part by the Office of Naval Research.

REFERENCES

[1] Brian S. Everitt. *Cluster Analysis*. John Wiley and Sons, New York, third edition, 1993.

[2] R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *IEEE Trans. Information Theory*, 5(3):299–314, 1996.

[3] G. W. Milligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *PPsychometrika*, 50:159–179, 1985.

[4] R. Mojena. Hierarchical grouping methods and stoppings rules an evlauation. *Computer Journal*, 20:359–363, 1977.

[5] J. L. Solka, D. J. Marchette, and B. C. Wallet. Statistical visualization methods in intrusion detection. *Proceedings of the Interface 2001*, 2000.

RESUME

Jeffrey L. Solka earned the B.S. degree in Mathematics and Chemistry from James Madison University in 1978, the M.S. in Mathematics from James Madison University in 1981, the M.S. in Physics from Virginia Polytechnic Institute and State University in 1989 and his Ph.D. in Computational Sciences and Informatics (Computational Statistics) at George Mason University, working under the direction of Prof. Edward J. Wegman, in May of 1995. Since 1984, Dr. Solka has been working in nonparametric estimation and statistical pattern recognition for the Naval Surface Warfare Center, Dahlgren, VA.