Research Article

Literati Journals
Empowering Researchers

Asset Data Linkage USING Artificial Intelligence Method of Machine Learning

Krishna M V[1]*

Abstract: As organizations grow so do their IT system, while this is positive phenomenon it brings with itself numerous challenges. One of the challenges that are resultant of such growth is the management of vast amount of asset data (metadata) in a meaningful and relevant fashion. In situations where there is lack of adequate measures to manage asset data of key organizational systems it paves way to a potential issue that can impede management of IT operations and hence quality IT service. Effective and efficient management of asset data is an often ignored but an important part of the IT operations management. The problems of management of IT asset data described here is about identifying the hosts and linking them together based on their architectural relationship, this information is useful for a variety of operational uses such as asset change impact assessment, decommissioning of servers, etc. Unsupervised Machine learning (ML) methods of clustering can be applied to the IP traffic observed across the network to group hosts into sets that have similar communicating partners. Self-Organizing Map (SOM) an Artificial Neural Network trains itself on IP logs entries and forms clusters of those communication patterns.

Asian Journal of Engineering and Technology Innovation
Volume 4, Issue 7
Published on: 7/05/2016

Cite this article as: Krishna M V. Asset Data Linkage USING Artificial Intelligence Method of Machine Learning. Asian Journal of Engineering and Technology Innovation, Vol 4(7): 72-75, 2016.

INTRODUCTION

In most organizations asset data are managed by a top down approach where there are processes defined to collect, update and maintain the asset data in repositories known as configuration management databases (CMDB).

The integrity and consistency of the CMDB in such a top down approach is dependent on the capability of process maturity. Practical observation however shows that these processes by themselves are not adequate where there are macro organizational factors at play that may impact the way in which CMDB is managed and maintained.

Some of such factors that impact the management of CMDB are:

1. Organization mergers and acquisition: vast amount of IT asset data having different details from merged organization may not necessarily be congenial for integration.
2. Continually evolving IT landscape over decades having produced legacy assets, topologies, distributed asset ownership and demographics.
3. Lack of tools and manually managed asset registers give way to data inconsistencies owing to large churn of commissioning and decommissioning of IT assets.

[1]Reva Institute of Technology and Management, Rukmini Knowledge Park, Kattigenahalli, Yelahanka, Near Border Security Bustop, Bengaluru, Karnataka-560064, India.
E-mail: ashwin@revainstitution.org
*Corresponding author

The CMDB typically contains vast amount of information related to the IT asset such as server name, server IP address, related servers, dependent servers, software installed, license, etc.

The ad-hoc managed asset data renders inadequate when assessing impact of changes planned to IT systems due to the shared nature of many asset types such as network devices and virtual platforms. The IT asset data are stored in separate configurations with very less or missing linking data elements that can be used to easily correlate or determine end to end dependencies of IT systems and services, add to it the vast amount of information that needs to be processed. The decision making as a result tends to be manual, less accurate and introduces risk in the management of IT systems.

IT Asset linkage is the process of identifying pairs of host IP records that have a pattern. In most cases, each record refers to an IT asset category, and the challenge is to identify which records refer to the same IT system.

Many different applications can have the same physical server, and many different virtual servers share the same physical server, or can have other fields that agree. So although fields tend to agree for matching pairs of records (i.e., records that refer to the same physical server), it is also possible for them to disagree on matching pairs and it is certainly possible for some fields to agree on differing pairs of records (i.e., on records that do not refer to the same physical servers). In the absence of robust record keeping, one of the approaches that can aid the process of managing the current

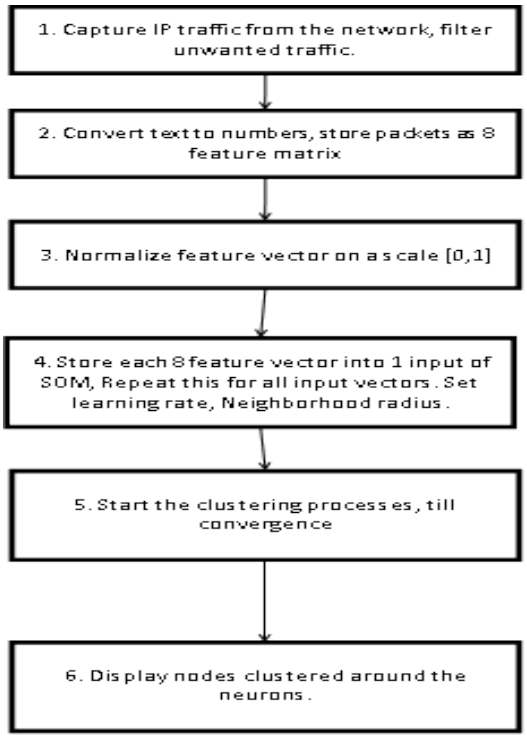

Figure 1: Block diagram of the approach

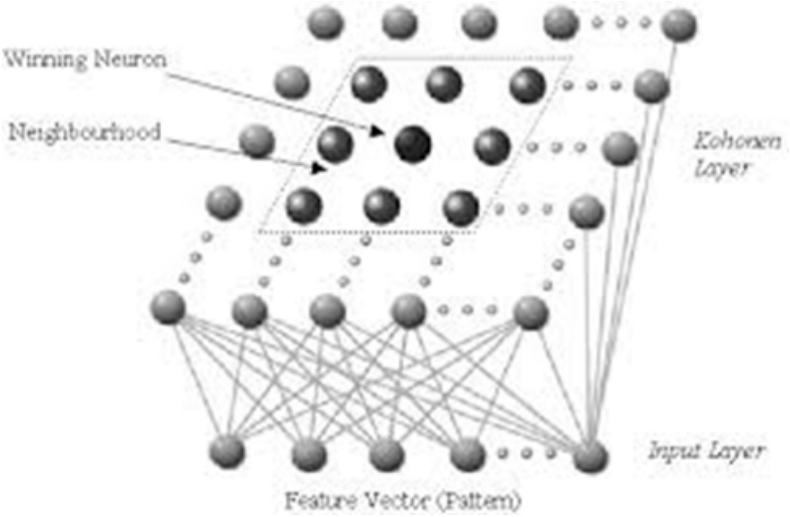

Figure 2: Kohonen's network (SOM)

and accurate information about IT asset is a bottom up approach. In this approach the communication between the asset hosts as captured from the organizations network (IP address pairs) can be used build a map that shows relationship between these hosts. Once this asset map is generated with basic information of the assets (IP address) other information can be fetched from CMDB and populated thereby obtaining an integrated and consistent asset data linkage.

The approaches to perform the above task can be grouped under as statistical, graphical and ML. Some of these techniques are used for Social Network Analysis (SNA). Given the vast amount of data to be traversed and analyzed Machine Learning methods are known to be efficient.

In the following papers [1] & [3], a machine learning method that uses Artificial Neural Network known as Self-Organizing Map (SOM) to cluster the IP traffic intrusion detection & clustering the relation data to identify clusters among book category and football teams, respectively.

This paper proposes IP traffic clustering (using SOM) for detection of the asset linkage groups (communities). We use the term node(s) to refer to the IT asset(s) here onwards.

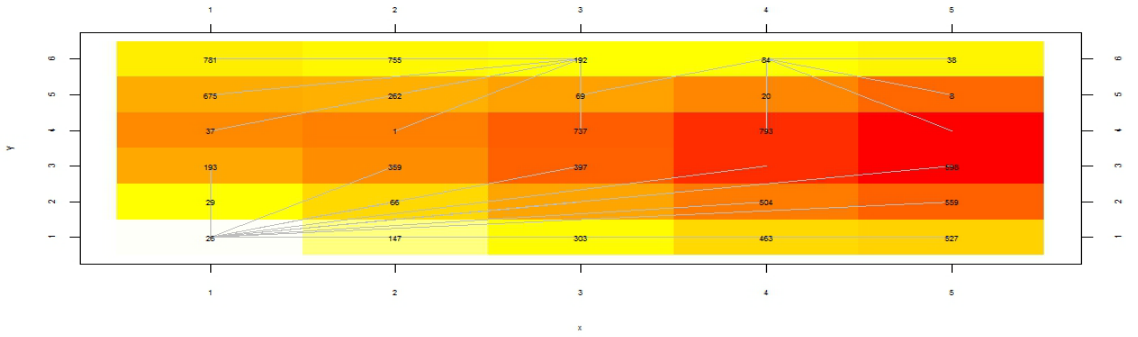

Figure 3: IP Clustersdetected after1.5 million iterations (SOM)

RELATED WORK

Graphical Method

The most natural way to imagine a node to node linkage is through network map (graph), this technique is very popular with SNA. In this method the network of nodes are represented using graphs and matrices.

There are different properties such as density, degree, group, reachability, centrality, etc. of a network that are useful in mining network data.The networks of nodes are detected for connectivity based on the path existence between nodes. There are various measures to detect the strength of the connection between the nodes. Breadth first search and depth first search is used to detect the connections.

In this method, the combined data pre-processing requirement is greater than the relation detection between the networks of nodes.

Computational complexity of O(|V||E|log(|V|2 /|E|)) to implement a network flow problem using Goldberg and Tarjan algorithm[2] prohibits larger scale implementation.

Statistical Method

Statistical method of solving the connectivity of nodes is also a popular.

There are methods such as:

1. Co-efficient clustering: It measures the degree to nodes that tend to cluster in a network together.
2. Probabilistic network model: they simulate the random graphs that tend to perform equivalent of real networks. Power Law networks, Erd¨os-Renyi's random graphs are some of the types of probabilistic approaches.

Machine Learning

Machine learning comprises of mathematical, statistical, and computational-techniques that have greater capability in detecting patterns from the given historical data. They are capable of better performance even when situations are not certain.

Due to high volume of data (IP traffic) classification techniques are not suitable, instead ML clustering methods are preferred. Clustering is an un-supervised learning method.

The seminal work presented here [3] show that SOM have a higher precision of clustering nodes in comparison.

METHODOLOGY

Unsupervised learning method SOM provides for simple and efficient way of clustering the data. SOM also preserves the topological character between the nodes.

The relationship between nodes is an important aspect that needs to be preserved to ensure that the generated cluster model can be populated with actual asset labels for slated human readability.

Data Collection and Preprocessing

Capture the IP traffic data from the organizations network using network capture tool at intervals of 10 mins several times during the day over few days.

The network from where the data is capture is typically within data center and across data centers.

The IP traffic data captured contains the communication information between pairs of IP addresses (client and server) along with details of the application hosted on the server, protocol, port, average number of bytes, bitrate, etc.

In this approach the IP payload information will not be touched to determine any relation details between the nodes. The IP address data in the network captured is multi-nominal, the port is numeric, protocol is nominal, etc. The IP traffic data is filtered to remove any unwanted communication trace that may introduce noise, in this case. Any common service related communication such as authentication, virus definition updates, PC based communications are filtered through.

For SOM learning the input vector data have to be selected. Here only IP addresses combination is selected for the learning purpose as only the relationship between the pair of nodes have to be clustered.

Each packet captured {
- Extract Client IP address
- Extract Destination IP address

}

Data Normalization and Scaling

The feature vector that represents the IP traffic is contains the Client and Destination IP addresses in full. Each IP address consists of 4 octets separated by period (.),(e.g. 201.58.54.9) these 4 octets of both IPs combination are normalized between the range of [0,1], to form 8 features input to SOM input vector.

The standard normalization is used here

$$\text{Normalized}(X) = (X - X_{min}) / (X_{max} - X_{min})$$

SOM Configuration

In this experiment, the SOM was configured as per following: Rectangular structure of neighbor detection was chosen. R representing number of neurons to be updated = 1. The structure had 30 output neurons arranged in a 5x6 matrix plane.

Kohonen's net was chosen for SOM implementation, the winning vector is the one that has the shortest distance in relation to the input vector. The learning rate was chosen as 0.6 this value is decremented inversely (0.5) in every epoch.

During the experiment the IP traffic of various sizes were filtered, normalized and input to SOM. The clustering process was run till convergence is reached. In other words, the loops continue till there are no changes to the neuron weights between successive loops, further the learning rate and radius are adjusted and the looping continues till termination condition is reached.

OBSERVATION AND FUTURE WORK

The experiment output shows promising result of clustering of IPs that has common communication partners.The mapping of the output of the SOM is as per below

Future work on the SOM with rate of convergence can be performed to determine the efficiency of clustering.

REFERENCES AND NOTES

1. Published in Langin, C., Zhou, H., Rahimi, S., Zargham, M., & Gupta, B. (2009). A self-organizing map and its modeling for discovering malignant network traffic. IEEE Symposium on Computational Intelligence in Cyber Security, 2009. CICS '09, 122-129.
2. Goldberg, A.V., Tarjan, R.E.: A new approach to the maximum-flow problem. Journal of the ACM (JACM) 35(4), 921–940 (1988)
3. SOMSN: An Effective Self Organizing Map for Clustering of Social Networks, International Journal of Computer Applications (0975 – 8887) Volume 84 – No5, December 2013
4. "Beyond probabilistic record linkage: Using neural networks and complex features to improve genealogical record linkage," in Neural Networks (IJCNN), The 2011 International Joint Conference on , vol., no., pp.9-14, July 31 2011-Aug. 5 2011
5. Fillegi, I. P. and Sunter, A. B. "A Theory for Record Linkage," Journal of the American Statistical Association, vol. 64, 1183-1210, 1969.
6. Introduction to Machine Learning Second Edition 2010 – EthemAlpaydin. M. Young, *TheTechincal Writers Handbook.* Mill Valley, CA: University Science, 1989.
7. The Third International Symposium on Optimization and Systems Biology (OSB'09)Zhangjiajie, China, September 20–22, 200.