CLUSTERING SOFTWARE VULNERABILITIES USING SELF-ORGANIZING MAPS: OBSERVATIONS AND ANALYSIS

Khyati Panchal, Timothy Cain

BACKGROUND

- The National Vulnerability Database (NVD), a public data repository, provides valuable information to help mitigate future attacks by understanding trends and patterns in software vulnerabilities.
- The MITRE Corporation created a list of the Common Enumeration of Vulnerabilities (CVE) in order to categorize these vulnerabilities.
- The CVE list, stores information about security vulnerabilities and exposures to provide common names of publicly known security exploits.
- Another important community developed list from MITRE is the Common Weakness Enumeration (CWE). The CWE list is a group of common software and hardware weaknesses, developed by the Cyber Security community.
- We are using the NVD data set, transformed in a vector representation using natural language processing with the V2W-BERT framework.
- We used a Self Organizing Map (SOM) toolbox in MATLAB to obtain the visual representations and perform analysis on the vector dataset.

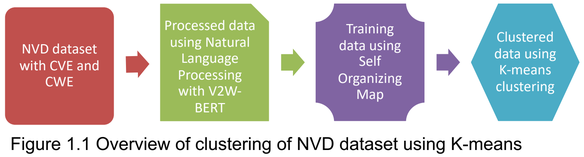

Figure 1.1 Overview of clustering of NVD dataset using K-means clustering with SOM

OVERVIEW OF CLUSTERING WITH SELF ORGANIZING MAP(SOM)

- SOM is a data visualization paradigm that clusters similar data and perform data-mining using topology preserving mapping and training through data distribution to generate high-dimensional data.
- As we can see in figure 1.1, data set processed with natural language processing using V2W-BERT model.
- After the processing we receive the data set in vector format that we use with SOM toolbox. SOM toolbox has inbuilt functions to perform various functionalities such as som_kmeans(), som_read_data(), etc.

CREATING VISUALIZATION USING SOM TOOLBOX: U-MATRIX

- Unified distance matrix also know as U-matrix is used for visualization of SOM to represent distance between each map unit to its adjacent neighbors by using different coloring scheme for adjacent nodes.
- U-matrix define structure based on computation between best matching unit with respect to height structure of U-matrix.

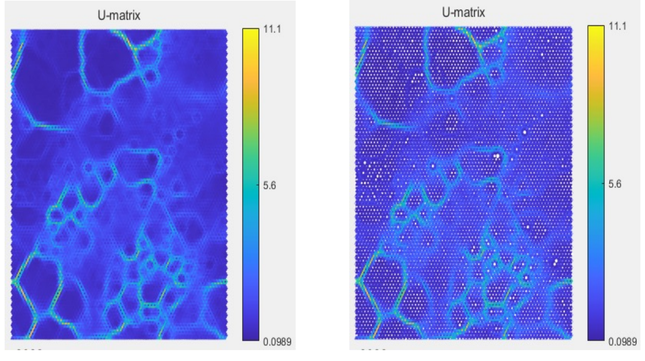

Figure 1.2 (a) U-matrix using SOM, (b) U-matrix with hits

FINDING BEST VALUE FOR CLUSTERS USING DAVIES-BOULDIN INDEX(DBI)

- Davies-Bouldin's validity index is utilized for partitioning with different number of clusters to choose the best clustering.
- Usually, when using DBI we try to reduce distance between points in a cluster, while also trying to increase the inner-cluster distance.

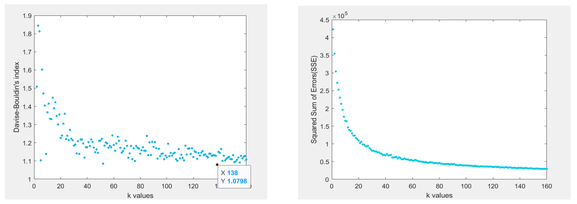

Figure 1.3 (a) K-means clustering using Davise-Bouldin's Index with an extended range search suggested best value for k, (b) Squared sum of errors for selecting best k

K-MEANS CLUSTERING WITH SOM

- Using SOM with the K-means algorithm, provides accuracy and consistency because it only works with local optimum with random initial center points. We used the batch K-means algorithm for clustering our data set.

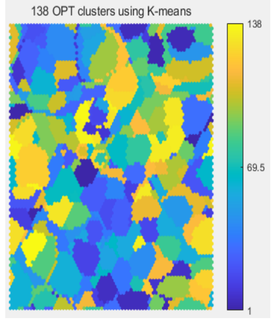

Figure 1.4 U-matrix using K-means with 138 clusters

CLUSTERING RESULTS AND ANALYSIS

- As a result of k-means clustering, we received a DBI index of 138 as the best value for k.
- Figure 1.5 shows the distribution of CVEs based on CWE labels in the first 10 clusters.

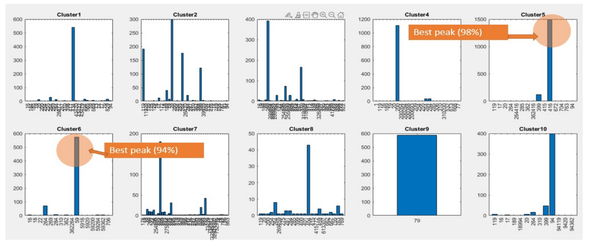

Figure 1.5 Bar graph for first 10 clusters

DETAIL ANALYSIS OF CLUSTER CWE-79 USING U-MATRIX

In the list of 2020's top 25 CWE's, CWE-79 scored highest with a score of 46.82 based on the Common Vulnerability Scoring System (CVSS).

CWE-79 is described as "Improper neutralization of input during web page generation" which is considered cross-site scripting.

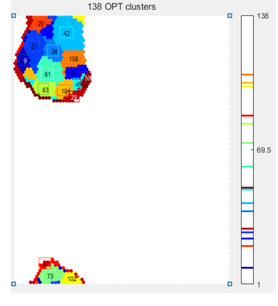

Figure 1.6 U-matrix of clusters only containing CWE-79 for prominent peak percentage more than 50%