Modeling of a Software Vulnerability Identification Method

Diako Doffou jerome[1†] Behou Gérard N'Guessan[2†] , ACHIEPO Odilon Yapo M[2†]

kingdjako@gmail.com[1†] , behou.nguessan@uvci.edu.ci[2†]

Institut National Polytechnique Félix Houphouët-Boigny (INP-HB), EDP, CÔTE D'IVOIE[1†]

Université Virtuelle de Côte d'Ivoire[2†]

Summary

Software vulnerabilities are becoming more and more increasing, their role is to harm the computer systems of companies, governmental organizations and agencies. The main objective of this paper is to propose a method that will cluster future software vulnerabilities that may spread. This method is developed by combining the Multiple Correspondence Analysis (MCA), the Elbow procedure and the Kmeans Algorithm. A simulation was done on a dataset of 15713 observations. This simulation allowed us to identify families of future vulnerabilities. This model was evaluated using the silhouette index.

Key words:

ACM, Unsupervised learning, Vulnerabilities, Cvss, Kmeans

1. Introduction

A vulnerability is defined as a weakness in a computer system that allows an attacker to damage the integrity of that system. There are several types of vulnerabilities including software vulnerabilities that are the focus of our work. A software vulnerability is a weakness, bug or vulnerability that can be exploited to breach privacy, service continuity and data integrity [1]. Software vulnerabilities are among the most commonly used vulnerabilities by hackers to compromise systems. For example, the organization (Cvedetail), counted 18325 in 2020.In general, successful software attacks exploit well-known vulnerabilities [2]. Despite the importance of software vulnerabilities, there is an inability of organizations and businesses to seriously combat software attacks. Indeed, the most used defense tools are very basic and obsolete to fight cyber threats. Therefore, the current trend is to generalize predictive Ethical Hacking as a framework for effectively combating cybercrime. With this in mind, preventive knowledge or detection of security vulnerabilities is more than necessary. However, the requirement for such knowledge will quickly become a hindrance to the practice of penetration testing because the discovery of flaws is an ongoing work that evolves faster than the learning and training capabilities of security specialists. This raises the problem of controlling future vulnerabilities that could lead to systems being compromised. Vulnerabilities are hidden phenomena, so their discovery must be accelerated if we are to be effective. This is to make a kind of Predictive Ethical Hacking in which vulnerabilities are not limited to known vulnerabilities but have been limited to other unknown vulnerabilities detected by activities of simulation and analysis of existing data on vulnerabilities already discovered.

The purpose of this work is to attempt to identify unknown types of vulnerabilities from the exploitation of qualitative data describing existing vulnerabilities. We turn to unsupervised learning techniques. Specifically, we will use a combination of Multiple Correspondence Analysis (ACM), the Elbow procedure and the kmeans algorithm. The data used consists of qualitative descriptions of 15713 software vulnerabilities.

2. Multiple Correspondence Analysis (MCA)

2.1 Definition

We consider $\mathbf{P}$ qualitative variables $\mathbf{P} \geq \mathbf{3}$ denoted $\{\boldsymbol{X_j}\,; \mathbf{j} = \mathbf{1}, \ldots, \mathbf{p}\}$, respectively possessing $\boldsymbol{c_j}$ modalities, with $\mathbf{c} = \sum_{\mathbf{j=1}}^{\mathbf{p}} \mathbf{c_j}$. It is assumed that these variables are observed on n individuals, each assigned the weight $\frac{\mathbf{1}}{\boldsymbol{n}}$.. Let $\mathbf{X} = [\boldsymbol{X_1}| \cdots |\boldsymbol{X_p}]$ be the complete disjunctive table of observations ($\boldsymbol{X}$ **est** $\boldsymbol{n} * \mathbf{c}$). We call Multiple Correspondence Analysis (MCA) of the variables $(\boldsymbol{X_1}, \ldots\ldots, \boldsymbol{X_P})$ relative to the considered sample, the Correspondence Factorial Analysis performed either on the matrix $\mathbf{X}$. We note $\boldsymbol{n_k^j}\, (\mathbf{1} \leq \mathbf{j} \leq \mathbf{p}, \mathbf{1} \leq \mathbf{k} \leq \boldsymbol{c_j})$ the number of the k-th modality of $\boldsymbol{X_j}$, $\boldsymbol{n_k^j}\, (\mathbf{1} \leq \mathbf{j} \leq \mathbf{p}, \mathbf{1} \leq \mathbf{k} \leq \boldsymbol{c_j})$ and $\boldsymbol{D_j} = \frac{\mathbf{1}}{\boldsymbol{n}} \mathrm{diag}\ (\boldsymbol{n_1^j}, \ldots\ldots\ldots, \boldsymbol{n_{c_j}^j}\ )$ et $\Delta = \ \mathbf{diag}\,(\mathbf{D_1} \ldots \mathbf{D_p})$ .We note: Δ is a diagonal matrix of order c and $\boldsymbol{D_j}$ is diagonal matrix of order $\mathbf{c_j}$ $\ \mathbf{1} < \mathbf{j} \leq \mathbf{p}$ [3].

Manuscript received September 5, 2021

Manuscript revised September 20, 2021

https://doi.org/10.22937/IJCSNS.2021.21.9.46

2.2 Complete disjunctive table

Let X be a qualitative variable with c modalities. We call the variable $X_{(k)}$ defined by $X_{(k)}(i) = \begin{cases} \mathbf{1} \;\; \boldsymbol{if}\; \mathbf{X(i)} = x_{(k)} \\ \mathbf{0\ otherwise} \end{cases}$ the indicator variable for the k-th modality of x(k = 1, ..., c) where i is an individual, in our case a vulnerability and $x_{(k)}$ is the k-th modality of **X**. We will note $\boldsymbol{n_{(k)}}$ the number of $\boldsymbol{x_{(k)}}$. We call the matrix of the indicators of the modalities of X and we will note X, the matrix n×c of general term: $x_i^k = X_{(k)}(i)$. Let us now consider P qualitative variables $X^1, \dots\dots, X^P$.

We note c_j the number of modalities of X_j, $c = \sum_{j=1}^{p} c_j$ and X_j the matrix of the indicators. We then call the matrix X, n×c, obtained by concatenation of the matrices X_j, the complete disjunctive array: $\mathbf{X} = [X_1| \cdots |X_p]$ [3].

3. Elbow and Kmeans methods

3.1 ELBOW Method (or Elbow Rule)

In clustering, the Elbow method is a heuristic used to determine the optimal number k of clusters in a data set. The method consists in plotting the explained variation as a function of the number of clusters and choosing the elbow of the curve as the number of clusters to use. It allows to determine this optimal value of k. In clustering, the inertia is the sum of the squares of the distances between each centroid of a cluster and the different observations included in the same cluster. The ELBOW method tries to find a number k of clusters so that the selected clusters minimize the intra-class inertia in the same cluster. The variance of the clusters is calculated as follows:

$$wcss_k = \sum_j \sum_{x_i \to c_j} d(c_{j,}^k x_i)^2$$

c_j^k: The center of the cluster (the centroid)

x_i: The ith observation in the cluster with centroid c_j^k

$d(c_j^k, x_i)^2$: The distance (Euclidean or otherwise) between the cluster center and the point x_i

3.2 K-means

K-means is a vector quantization method. It is an alternate minimization method which, given an integer **k**, will seek to separate a set **X** of observations into **k** clusters.

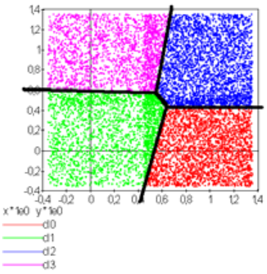


Fig 1:Clustering on a set of 2D points, 4 clusters

3.3 Description

Given a set of $\boldsymbol{n}$ observations, $(\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_n})$, where each observation is a real vector of dimension d, the k -means algorithm aims to partition the n observations into $\mathbf{k}\ (\leq\ \mathbf{n})$ sets $\boldsymbol{S_1}, \boldsymbol{S_2}, \dots, \boldsymbol{S_k}$ so as to minimize the sum of squares intra-cluster (WCSS). Formally, the objective is to find:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \epsilon S_i} \|x - \mu_i\|^2 = \arg\min_{S} \sum_{i=1}^{k} |S_i| Var S_i$$

where $\boldsymbol{S_i} = \boldsymbol{S_1}, \boldsymbol{S_2}, \dots, \boldsymbol{S_k}$

$\boldsymbol{\mu_i}$ is the average of the points of $\boldsymbol{S_i}$

4. Related Work

4.1 Detection of vulnerabilities by unsupervised learning techniques

Computer security researchers have published work on this G. Schaffrah et al [4], have carried out research in the field of flow-based vulnerability detection. This work provides a classification of attack and defense techniques and shows how flow-based techniques can be used to detect scans, worms, botnets, and denial of service (DoS) attacks. Zhengjie et al [5] propose a method based on the combination of the K-means algorithm and the particle swarm optimization algorithm (Kmeans-OEP). The experiments were performed on the KDD CUP 99 database. They have shown the efficiency of the proposed method and also show that the method has a higher detection rate and a lower detection error rate. K. Kumar et al [6] presented an approach to identify vulnerabilities stored in weblogs. They present a new approach based on the K Means algorithm to analyze data using different attributes like protocol, port number, etc. in order to detect vulnerabilities. In this process, they used preprocessing techniques to remove unwanted attributes from weblog data.

Gupta et al [7] propose an approach to the detection of vulnerabilities. Based on the combination of the K-Means algorithm and association rules. The experiments were carried out on KDD CUP 99 data. This approach makes it possible to determine a good detection rate only in the case of a denial of service (DOS) attack but is limited in the case of other types of vulnerabilities.

4.2 Limits of existing works

The work presented above for the identification of vulnerabilities uses the KDD CUP 99 database.
This database only contains quantitative data, which does not make it possible to assess the levels of vulnerabilities on data providing qualitative descriptors.

Then we found that the work of Gupta et al only identifies denial of service attacks. To overcome these shortcomings, we have turned to the use of a database from cevdetail.com. This database describes vulnerabilities mainly using qualitative variables. Another advantage of our approach lies in the fact that little research work has been carried out on this database for the discovery of new vulnerabilities. We propose an Machine learning modeling approach combining Multiple Correspondence Analysis (MCA), the Kmeans method and the Elbow method in order to identify families of potential or not yet discovered software vulnerabilities.

5. Model Construction

5.1 Principle

The study dataset is the database collected from the research website www.cvedetail.com. This database only contains qualitative variables. The model we have developed will help identify potential vulnerabilities and unknown vulnerabilities in applications. To achieve these goals, we wrote an algorithm called **IdSoftVul.** This algorithm follows the following steps:

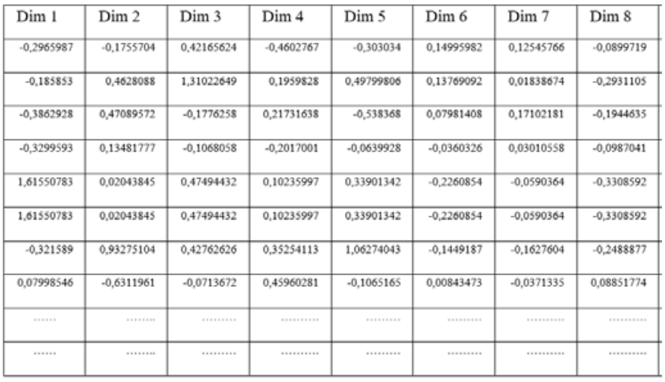
Step 1: Transform our database which contains only qualitative variables using the MCA technique.
Step 2: Apply the Kmeans Algorithm on the transformed database;
Step 3: Apply the Elbow method to determine the optimal number of clusters.
Step 4: Evaluate the model by the silhouette score.

```
ALGORITHM : IdSoftVul
Entry
D : Vulnerability database with qualitative variables
$\Delta_{MCA}$ : Database that has been transformed by an MCA
NCluster, ScoreSilhouette, i: integer

BEGIN
    // Import the qualitative data
    Importer (D)
    $\Delta_{ACM} \leftarrow$ ACM(D)
    Lire ($\Delta_{ACM}$)
   // Build the model
    NCluster ← 1
    For i ← NCluster to 11 do
        $Vul \leftarrow$ KMEANS(NCluster[i])
        $IdVul \longleftarrow Vul.fit(\Delta_{MCA})$
        Apply the Elbow Method on the inertia of Vul
        Display the number of clusters by the Elbow method
    END
   // Evaluation of the model by the Silhouette Score
    For i ← 1 to 11 do
        $ScoreSilhouette \leftarrow SILHOUETTE(IdVul, \Delta_{MCA})$
        Print ($ScoreSilhouette$)
    END
END
```

5.2 Results and discussion

The simulations carried out with IdSoftVul using the python language, allowed us to transform the vulnerability database of cvedetails.com by recoding it into numerical data, as illustrated below:



Dim 1	Dim 2	Dim 3	Dim 4	Dim 5	Dim 6	Dim 7	Dim 8
-0,2965987	-0,1755704	0,42165624	-0,4602767	-0,303034	0,14995982	0,12545766	-0,0899719
-0,185853	0,4628088	1,31022649	0,1959828	0,49799806	0,13769092	0,01838674	-0,2931105
-0,3862928	0,47089572	-0,1776258	0,21731638	-0,538368	0,07981408	0,17102181	-0,1944635
-0,3299593	0,13481777	-0,1068058	-0,2017001	-0,0639928	-0,0360326	0,03010558	-0,0987041
1,61550783	0,02043845	0,47494432	0,10235997	0,33901342	-0,2260854	-0,0590364	-0,3308592
1,61550783	0,02043845	0,47494432	0,10235997	0,33901342	-0,2260854	-0,0590364	-0,3308592
-0,321589	0,93275104	0,42762626	0,35254113	1,06274043	-0,1449187	-0,1627604	-0,2488877
0,07998546	-0,6311961	-0,0713672	0,45960281	-0,1065165	0,00843473	-0,0371335	0,08851774
......	....	....	....	....	....	....	....
......	....	....	....	....	....	....	....

Fig 2:Recode database preview

Then, applying by the ELBOW method, the graph 3 shows that the number k of cluster k can take the values three (3) and four (4) as shown below:

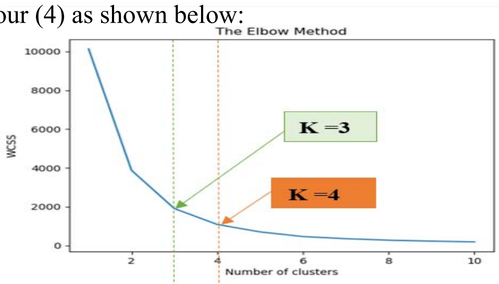


Fig 3:ELBOW method

Finally, IdSoftVul allows us to identify families of unknown software vulnerabilities for which businesses or organizations should be careful.

Case 1: For K = 3, these clusters represent three families of unknown vulnerabilities.
They are assessed by the silhouette index with an average silhouette score of 0.55 represented by the vertical red line.

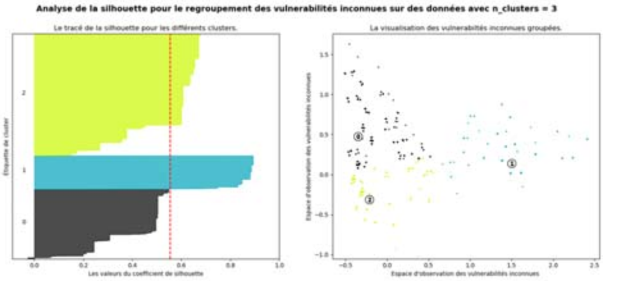


Fig 4:Families of unknown vulnerabilities for k = 3

Case 2: For K = 4, These clusters represent four families of unknown vulnerabilities. They are assessed by the silhouette index with an average silhouette score of 0.54 represented by the vertical red line.

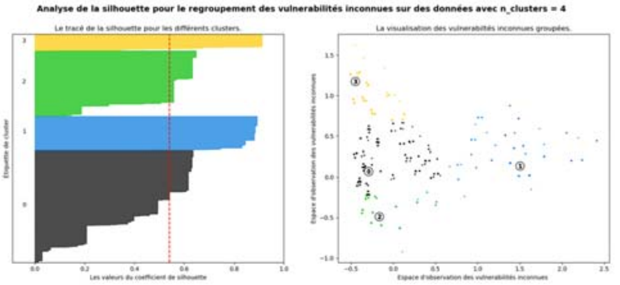


Fig 5: Families of unknown vulnerabilities for k = 4

The average silhouette score in Case 1 and Case 2ϵ [0.51, 1]. According to the Silhouette Scale, this means that the identification of the families of software vulnerabilities discovered are good qualities, as shown in the table below.

Table 1:IdSoftVul performance

CAS	Number of clusters	Average score silhouette	Nature of the structure
Cas 1	3	0.55	High
Cas 2	4	0.54	High

In short, we can say that the IdSoftVul algorithm is a good model for identifying software vulnerabilities.

References

[1] Peter Mell, Karen Scarfone, et Carnegie Mellon, « A Complete Guide to the Common Vulnerability Scoring System Version 2.0 », 2007. https://www.first.org/cvss/v2/guide

[2] Mike Schiffman et Cisco CIAG, « Guide complet du CVSS v1 », 2005.

[3] Wikistat, « Analyse factorielle multiple des correspondances (AFCM) ». 2016.Disponible sur: https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-explo-afcm.pdf

[4] G.Schaffrath et al, « An Overview of IP Flow-Based Intrusion Detection Communications Surveys & Tutorials », *IEEE*, 2010.

[5] Zhengjie Li et al, «Anomaly Intrusion Detection Method Based on K-means Clustering Algorithm with Particle Swarm Optimization,», *International Conference of Information Technology, Computer Engineering and Management Sciences*, 2011.

[6] K. Kumar et al, « «Identifying Network Anomalies Using Clustering Technique in Weblog Data, », *International Journal of Computers & Technology, , vol. 2 , n° %13*, juin 2012.

[7] C. Gupta, A. Sinhal, et R. Kamble, « Intrusion Detection based on K-Means Clustering and Ant Colony Optimization: A Survey », *IJCA*, vol. 79, n° 6, p. 30-35, oct. 2013, doi: 10.5120/13747-1555.

Diako Doffou Jérome received the degrees of Maitrise MIAGE from the University of Nanguy Abrogoua in 2000 and 2005, respectively. He received the Master MIAGE degree in 2012. After working as a high school teacher (from 2012) in a vocational training center, he is a PhD student in computer science since 2017 at Institut National Polytechnique Félix Houphouët-Boigny. His research focuses on cyber defense and artificial intelligence.

Behou Gerard N'Guessan has a doctorate in computer engineering. He holds a master's degree in media engineering from the University of May 08, 2005 in Guelma, Algeria. He obtained his PhD at the University Nangui Abrogoua, Abidjan-Côte d'Ivoire in the Faculty of Applied Basic Sciences. He is a member of the Research Laboratory in Computer Science and Telecommunications of the Institut National Polytechnique Houphouët Boigny (INP-HB), Abidjan, Cote d'Ivoire, member of the Laboratory of Data Engeering and Artificial Intelligence and associate member of the Research Unit and Digital Expertise of the Virtual University of Côte d'Ivoire. His research interests include mathematical modeling, media engineering, traditional medicine, and application inventor. His work focuses on their method of research and training in traditional medicine. He is currently working as a Master Assistant at the Virtual University of Côte d'Ivoire in Abidjan (Côte d'Ivoire).

Achiepo Odilon Yapo Melaine holds a PhD in Mathematics and Information Technology. He holds a master's degree in Computer Science with the option of Knowledge Extraction from Data. He obtained his PhD at the Institut National Polytechnique Houphouet Boigny, Yamoussokro-Côte d'Ivoire. He is a member of the Data Engeering and Artificial Intelligence Laboratory and an associate member of the Research and Digital Expertise Unit of the Université Virtuelle de Côte d'Ivoire. He is also an Engineer Statient Economist (ISE). His research focuses on mathematical modeling using Artificial Intelligence techniques. He is currently working as a lecturer at the Virtual University of Côte d'Ivoire in Abidjan (Côte d'Ivoire).