Analysis of ML methods and algorithms for phishing and spoofing detection

INNOvative SOC/INNOSOC project

Contents

Introduction 3
Experiments 4
Testing objective 4
Dataset and preprocessing 5
Results: 5
K-Means 5
HDBSCAN 7
Summary 10
Conclusion 10

Introduction

Phishing is a deceptive cyber tactic used by attackers to obtain sensitive information from unsuspecting victims. This method usually involves sending fraudulent e-mails or messages that imitate legitimate sources such as banks, government agencies or well-known companies. The aim is to trick recipients into revealing personal data such as login credentials, credit card numbers or other sensitive data such as a social security number.

Phishing is a typical example of a social engineering attack. Social engineering attacks are various techniques that exploit human psychology for malicious purposes. E-mail phishing is probably the best-known example; it involves attackers sending e-mails that appear to come from trustworthy sources in order to deceive recipients and induce them to provide sensitive information or click on malicious links. These e-mails often imitate the appearance and language of legitimate correspondence from banks, service providers or even friends and colleagues, which makes them difficult to distinguish from genuine communications.

Phishing also extends to text messages, known as SMS phishing or *smishing*. In these attacks, the perpetrator sends text messages that lure the recipient into clicking on a malicious link or providing personal data. These messages may appear to come from a trustworthy organization and often create a sense of urgency to react immediately to the message received. *Vishing*, or voice phishing, is another variant, in which attackers use phone calls to obtain sensitive information. In these situations, the attacker often impersonates a legitimate representative of an organization in order to obtain information or direct action from the victim of the attack.

Impersonation and pretexting are sophisticated tactics used in social engineering attacks, in which the attacker pretends to hold a certain authority or creates a convincing scenario to exploit the trust of the victim of the attack. By assuming the identity of an authoritative person such as a police officer, IT technician or company manager, attackers can manipulate the targets of the attacks into disclosing confidential information or taking actions that compromise security. Creating convincing scenarios requires detailed knowledge of the victim or their organization, which makes these attacks particularly insidious and effective.

Baiting (baiting = luring, enticing) and *tailgating* (tailgate = back door) are other widespread social engineering methods. *Baiting* attacks involve luring the victims of the attack with the promise of goods or services in order to steal their personal data or infect their systems with malicious software. This may take the form of downloading free software or access to exclusive content. *Tailgating*, on the other hand, is a physical security breach tactic in which an unauthorized person gains access to a restricted area by following someone who is authorized. This attack exploits the social norm of

politeness and the natural tendency to help others, which allows attackers to penetrate secured environments.

Elicitation and manipulation are subtle but powerful social engineering techniques. These methods involve obtaining information through a seemingly innocent conversation in which the attacker engages the target victim in a discussion and discreetly steers them towards obtaining the desired information. Attackers often manipulate targets emotionally, exploiting human tendencies such as the desire to be liked or the fear of confrontation. Gaining trust through rapport and empathy is crucial in these interactions, as it lowers the defences of the target of the attack and makes them more susceptible to revealing sensitive information.

Each social engineering attack shows the complexity and diversity of the tactics that attackers use to exploit human psychology. They highlight the need for constant vigilance in both personal and work environments in order to prevent these manipulative techniques. Understanding the nature of these attacks is the first step towards developing effective strategies to deal with them, while emphasizing the importance of awareness and education in the ongoing fight against cyber threats.

To combat phishing, individuals and organizations must be vigilant and sceptical of unsolicited communications. Key preventive measures include not clicking on links or downloading attachments from unknown sources, verifying the authenticity of requests for personal data and using the anti-phishing tools available in most modern web browsers and e-mail services. Education and training in identifying phishing attempts are key to building the first line of defence against these cyberattacks.

Experiments

Testing objective

The objective of the pilot testing was to **verify and compare the suitability of the K-Means and HDBSCAN methods for identifying phishing** in data in the form of **network logs**. The testing was intended to determine whether these methods are able to:

- identify recurring patterns and dense regions of logs
- detect anomalies and outliers that may indicate phishing or suspicious events
- compare the differences in the approach of both algorithms, in particular the ability to detect outliers and the cluster structure
- provide a comprehensible interpretation of the clusters.

Dataset and preprocessing

For the pilot testing, we worked with a set of 57,000 logs obtained from traffic at FRI UNIZA, originally containing approximately 160 variables, of which 41 were identified as relevant for identifying phishing. Several of these variables were categorical, and it was therefore necessary to preprocess them using a combination of one-hot encoding and the hashing trick, which expanded the dataset to 2,009 variables suitable for clustering.

Before clustering, all variables were standardized using StandardScaler in order to remove differences in value ranges. Subsequently, PCA dimensionality reduction to 50 components was applied, which allowed efficient work with the clustering algorithms while preserving the variability of the data.

This step was necessary and time-consuming, and it is of fundamental importance for the preparation of the machine learning model. Logs originating from different systems contain different variables, and therefore, when creating the final model for identifying phishing, it will be a major challenge to develop a robust script capable of processing unexpected or new variables as well.

Results:

K-Means

When applying the K-Means algorithm to our log dataset, we initially assumed 3 clusters. The resulting distribution was very uneven:

Cluster 0: 57,695 points

Cluster 1: 2 points

Cluster 2: 24 points

The K-Means results are visualized in Figures 1 and 2.

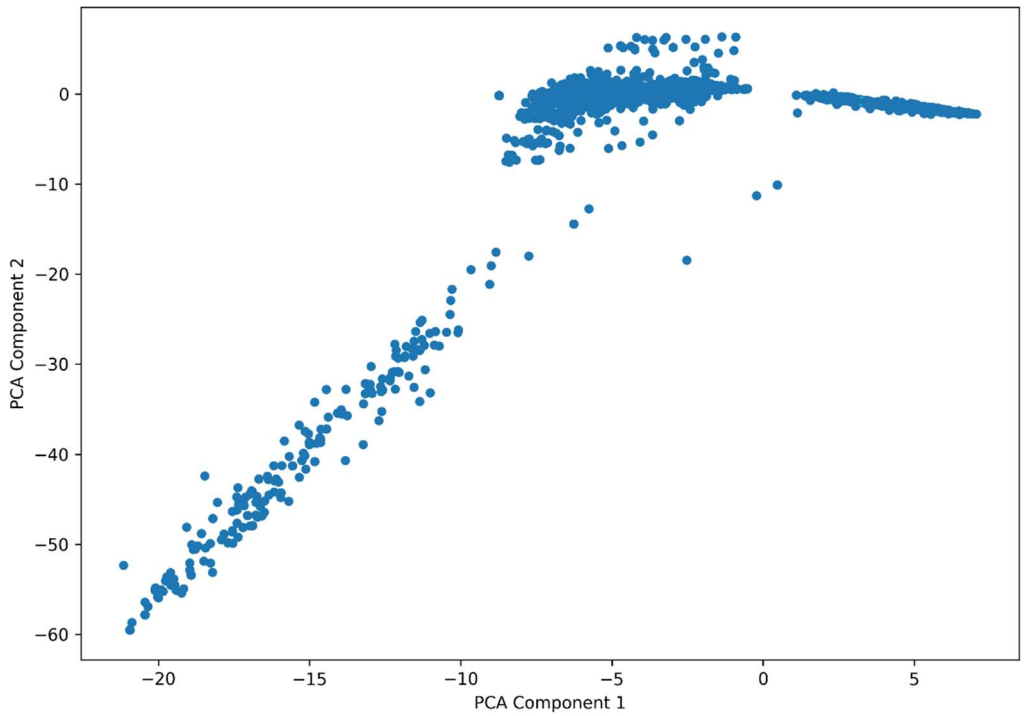


Fig. 1: 2D visualization of K-Means

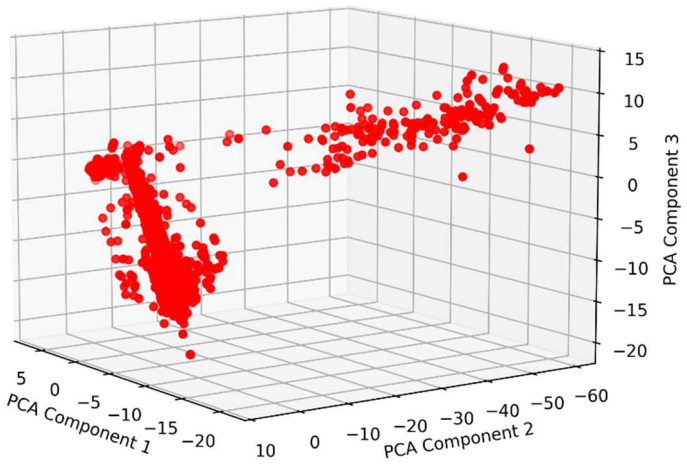


Fig. 2: 3D visualization of K-Means

Changing the assumed number of clusters (e.g. to 4 or 5) did not yield a better distribution of the data. Some new clusters contained only 1–2 points, while the main cluster still captured the majority of the data. This uneven distribution reflects the natural structure of the dataset, in which most logs are of a common type and only a very small number of logs represent suspicious or phishing events.

K-Means assigns every point to some cluster, without the possibility of labelling points as lying outside a cluster (outliers). As a result, extreme or non-standard values are automatically included in one of the clusters, often in small or marginal clusters. This may reduce the ability of K-Means to detect atypical or suspicious logs on its own.

Overall, it appears that K-Means is suitable for a fast and deterministic division of data into a predetermined number of clusters, but its ability to capture finer differences or identify deviations in logs is limited. The uneven distribution in this case corresponds to the natural imbalance of the data, in which the number of phishing logs is very low.

HDBSCAN

The HDBSCAN algorithm was applied to the log dataset with the parameters min_cluster_size=3000 and min_samples=20. These settings made it possible to identify 6 main clusters and at the same time to mark outliers that did not fall into any cluster.

The resulting distribution of the number of logs in the individual clusters was as follows:

Outliers (label = -1): 14,913 points
Cluster 0: 13,493 points
Cluster 1: 6,702 points
Cluster 2: 7,704 points
Cluster 3: 7,891 points
Cluster 4: 3,363 points
Cluster 5: 3,655 points

Interpretation of the results

1) Outliers (-1):
   - the algorithm marked 14,913 logs as outliers that did not belong to any dense cluster
   - these logs probably contain atypical or potentially suspicious events, which is an advantage for detecting phishing or anomalies
2) Main clusters (0–5):
   - the clusters are relatively evenly represented, with the largest clusters containing 13,493 and 7,891 points and the smaller clusters 3,363 and 3,655 points
   - this more even distribution compared with K-Means makes it possible to better capture finer differences between logs and to identify the density structure in the data
3) Significance of the parameters:
   - min_cluster_size=3000 ensured that only significant clusters containing a sufficient number of points are formed
   - min_samples=20 determined the sensitivity to density and contributed to the detection of outliers

The HDBSCAN results are visualized in Figures 3 and 4.

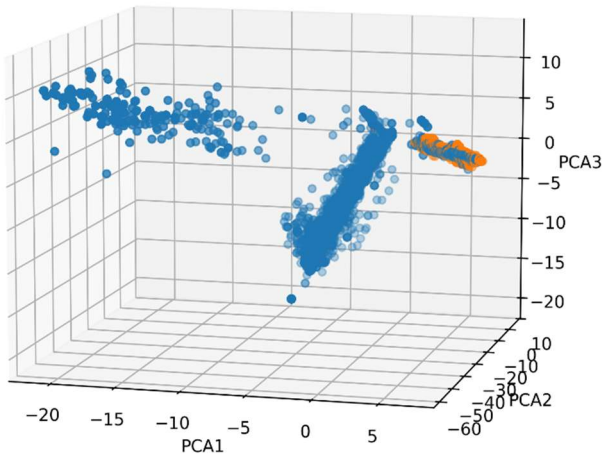


Fig. 3: 3D visualization of HDBSCAN

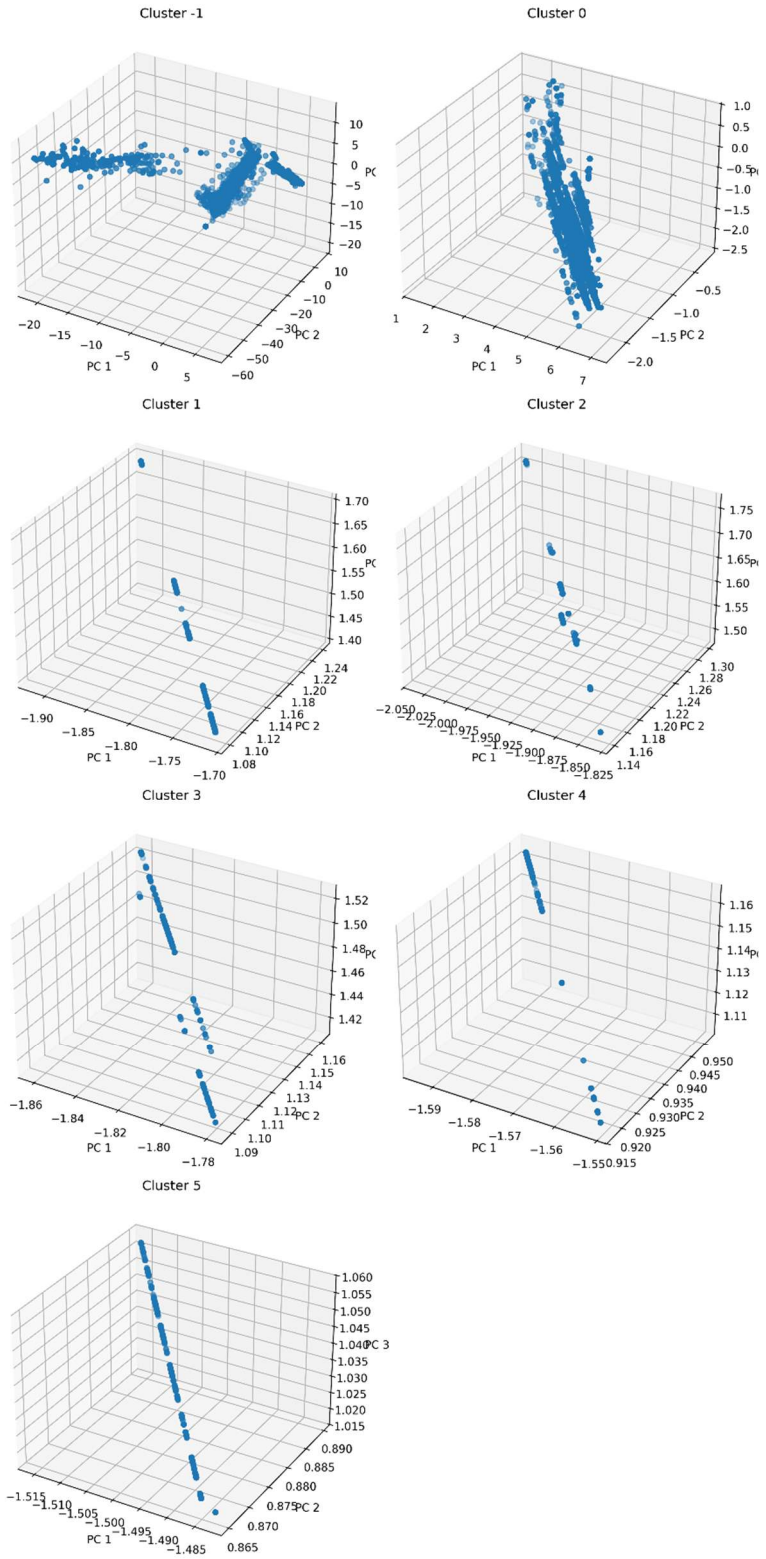


Fig. 4: 3D visualization of the individual clusters with HDBSCAN

In Figure 4 we can see that the logs assigned to the individual clusters are distributed in parallel planes, while the logs marked as outliers are located outside these main planes. This visual effect suggests that HDBSCAN can effectively distinguish typical behaviour

from atypical or suspicious events, unlike K-Means, which assigns every point to some cluster.

In this context, using the distance of a log from the plane of the main cluster as a probabilistic phishing score represents a new and innovative way of interpreting the data. This approach makes it possible to quantify the degree of deviation of individual logs and potentially to identify suspicious or malicious events with finer granularity than the usual binary labelling of outliers.

At the same time, this concept needs to be verified in further research in order to confirm its effectiveness in the automated detection of anomalies and phishing attacks, and to make it possible to develop a robust system that combines the distance score with other log characteristics (e.g. time, IP addresses, event type) for more accurate identification of suspicious activities.

Summary

HDBSCAN has proved to be a suitable tool for analysing logs with an uneven distribution, as it is able to:

- identify dense regions of logs
- mark outliers that may represent phishing or other suspicious events
- maintain a more even distribution of clusters compared with K-Means, which facilitates the interpretation of the results and further analysis
- a new and innovative approach: if a new log is assigned to one of the clusters, it is usually neither an anomaly nor a potential attack. However, if a log is not assigned to any cluster, its degree of severity can be quantified using the distance from the nearest cluster, thereby producing a probabilistic phishing score. This approach allows the risk of suspicious events to be quantified more finely and supports automated anomaly detection.

Overall, HDBSCAN provides a robust and flexible approach to clustering network logs, especially for large and heterogeneous datasets, while the new concept of probabilistic outlier scoring represents a promising area for further research and verification of its effectiveness.

More information on the description of the clustering algorithms mentioned can be found in the supporting document Clustering.pdf

Conclusion

K-Means is suitable when a fast and deterministic decomposition is required; however, it does not detect outliers and distributes points unevenly in data of uneven density.

HDBSCAN is suitable for robust log analysis because it can detect both dense regions and outliers, which is a significant advantage in identifying phishing and other anomalies

in network logs. Unlike algorithms that assign every point to a cluster, HDBSCAN makes it possible to mark atypical or outlying logs (outliers) that do not meet the density criteria required for a cluster to form. In this way, the algorithm effectively filters out noise and focuses on logs that represent repeatable and significant behavioural patterns.

In the pilot testing, HDBSCAN proved more suitable for capturing finer differences between logs that could be indicators of suspicious or non-standard events. The algorithm is able to divide the data into clusters according to the density of points, while preserving the hierarchical structure and making it possible to identify smaller but significant groups of logs that could be “lost” in massive clusters when using classic K-Means.

The main advantage of HDBSCAN over K-Means is therefore its ability to:

- automatically determine the number of clusters according to data density
- robustly identify outliers that may indicate anomalies or phishing activities
- preserve fine differences between logs, which improves the interpretation and analytical value of the results.

Overall, HDBSCAN provides a flexible and robust approach to log analysis that takes into account the natural variability of the data and enables effective detection of suspicious events without the need for a predetermined number of clusters.