UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Network anomaly detection

Digital forensics

Master Degree in Cybersecurity, Master Degree in ICT for Internet and Multimedia

Grimaldi Alberto, Cinthya Celina Tamayo Gonzalez

Network anomaly

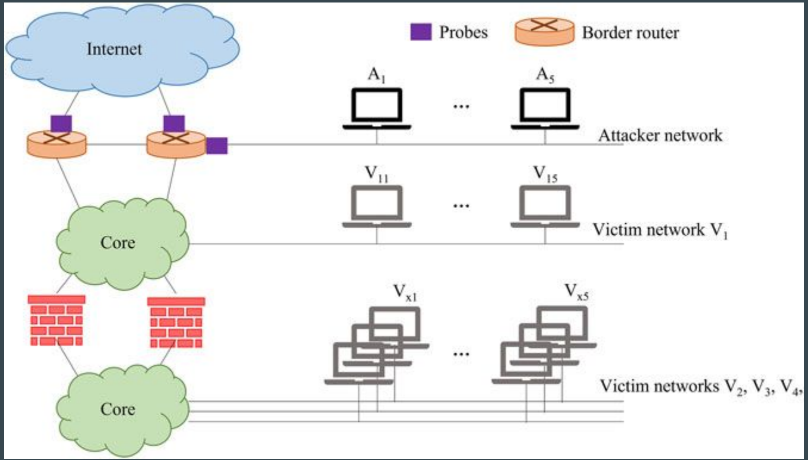
- Given a set of data: ”nominal” samples and “anomalous” samples, assume that anomalies are created by a different generation process with respect with the nominal samples.
- **Well-Defined Anomaly Distribution (WDAD)** assumption: anomalies are drawn from a known distribution

Anomaly can be caused by different factors:

1. *Non-human error*
2. *Human error*
3. *Malicious human activity*
    - Network attack (anomalous traffic)
    - Image tampering or deepfake detection
    - Fraud detection
    - OS attacks

Dataset

- UGR'16 dataset has been selected
  - It's composed by two parts:
    - clean subset: includes real background traffic → training
    - test subset: combination of real background and controlled attack traffic → testing

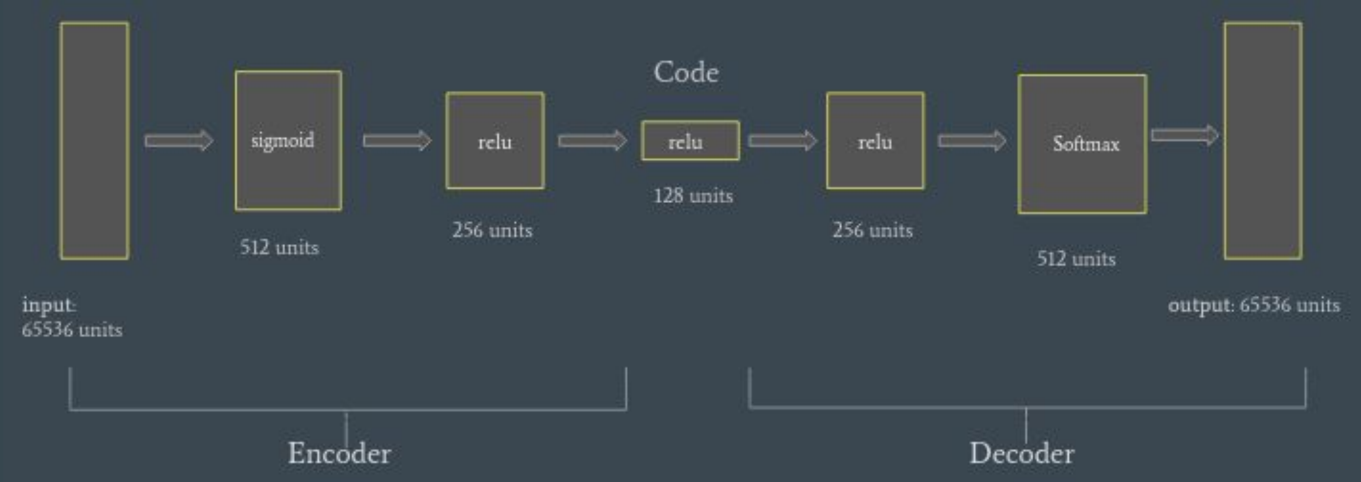


- Types of attack considered:
  - DoS11: one-to-one DoS where attacker A1 attacks the victim V21;
  - DoS53: the five attackers (A1 – A5) attack three victims.
  - Scan11: one-to-one scan attack where attacker A1 scans the victim V41;
  - Scan44: four-to-four scan attack where the attackers A1, A2, A3 and A4 scan the victims V21, V11, V31 and V41, respectively.

Goal and implementation

Develop and design an *anomaly detection system* which is able to detect anomalous conditions in form of attacks:

1. Clean
2. Dos 11
3. Dos 53
4. Scan 11
5. Scan 14



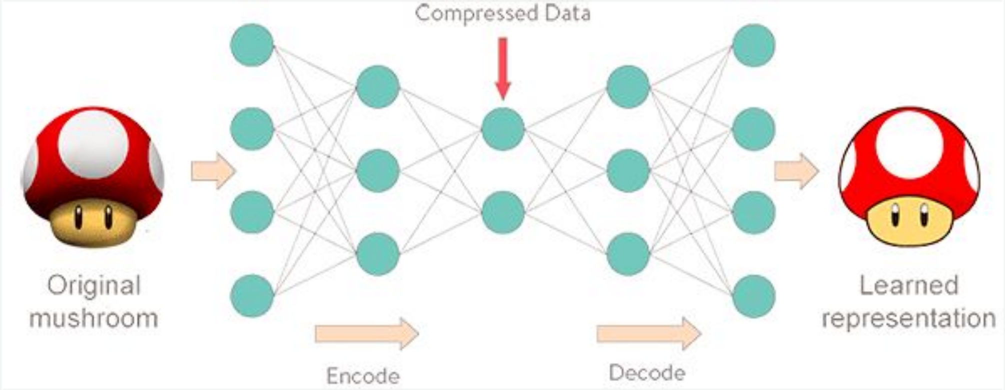
Method used:

- We opted for a deep learning approach to detect anomalies → Autoencoder
- Classification of the type of attack by using unsupervised segmentation
  → K-means clustering
  → Agglomerative Clustering

Grimaldi Alberto
Cinthya Celina Tamayo Gonzalez

Autoencoder

- What is an autoencoder?
  - Is a type of **artificial neural network** used to learn efficient codings of unlabeled data.
  - The autoencoder learns a hidden representation (through encoding) by learning the main features of the data and then reconstructing in based to the hidden representation.



- Basic architecture
  encoder → the input layer
  decoder → the output layer

- Fields of application
  - facial recognition
  - feature detection
  - anomaly detection

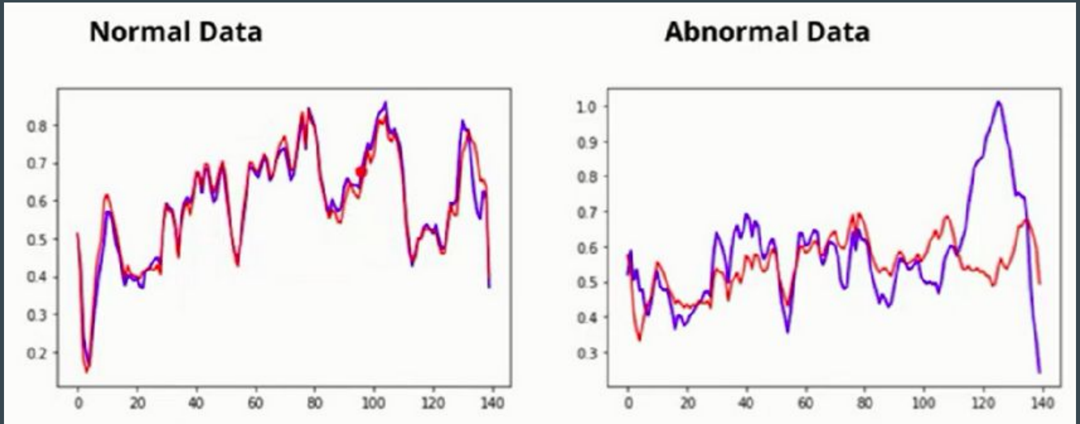
UNIVERSITÀ DEGLI STUDI DI PADOVA

How do we detect an anomaly with autoencoders?

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Autoencoders are naturally lossy due the compression of information. This loss is known as reconstruction error.

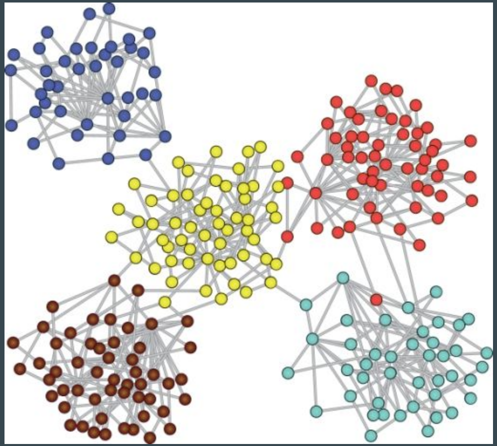
- We train the autoencoder in normal data (without) anomalies and get the nominal error
- In the test data, if the error is greater than a threshold value, then we have anomaly data.



Clustering

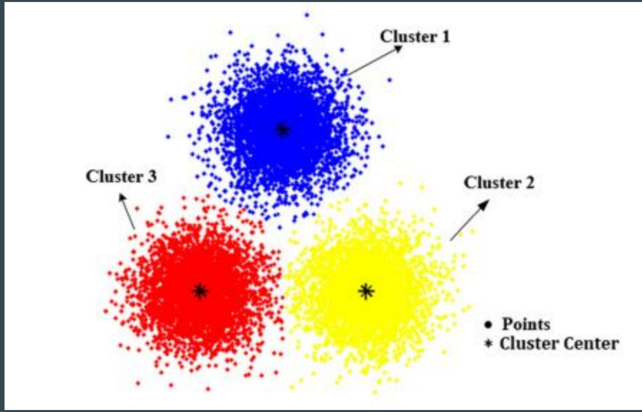
Agglomerative Clustering

Each object is initially a single-element cluster and combine the most similar elements until the target.



K means

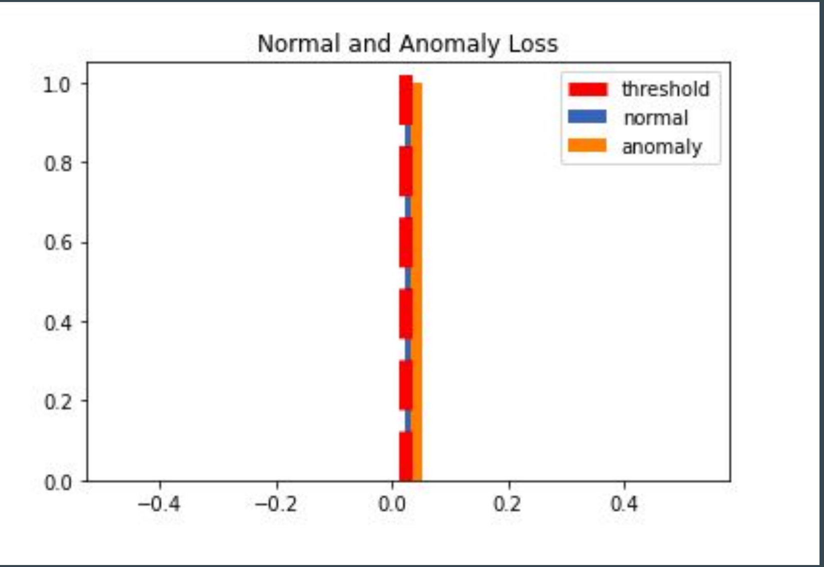
Starts with a randomly selected centroids, which are used as the beginning points for every cluster, and then iteratively optimize the positions of the centroids



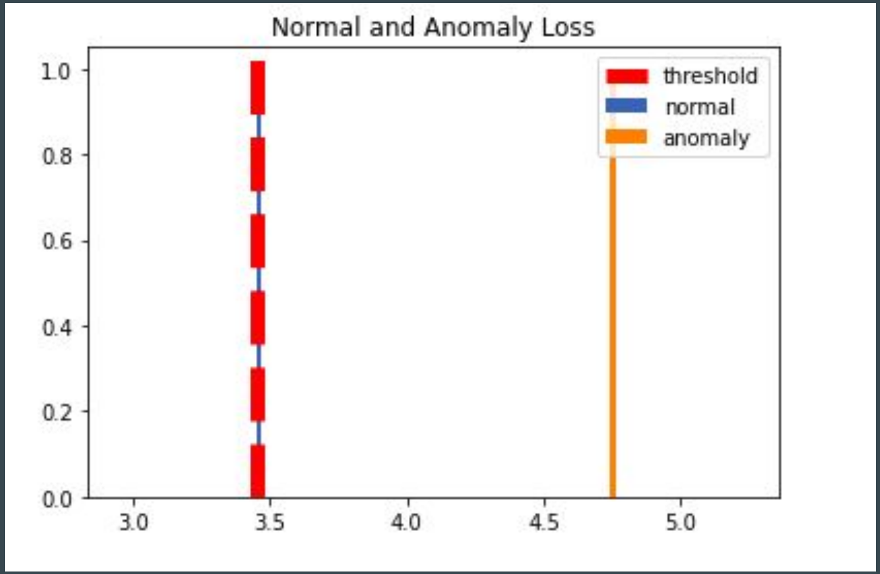
Grimaldi Alberto
Cinthya Celina Tamayo Gonzalez

Experimental results

MSLE

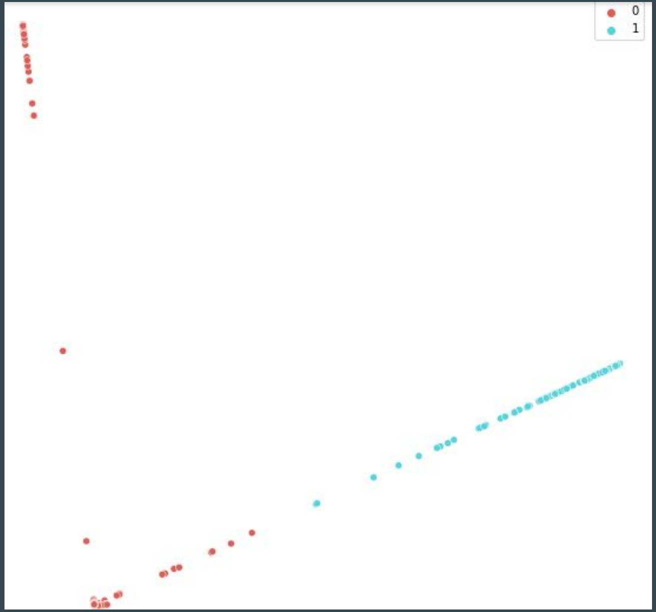


MSE



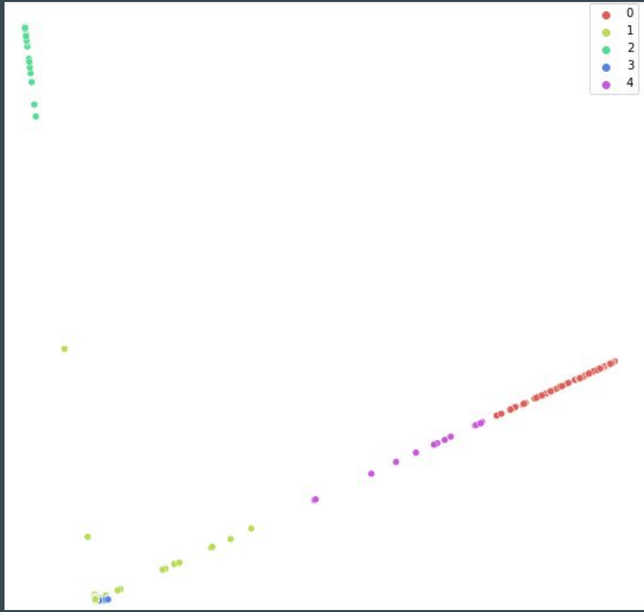
Grimaldi Alberto
Cinthya Celina Tamayo Gonzalez

Experimental results II

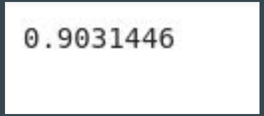


Silhouette Coefficient

PCA
Agglomerative clustering

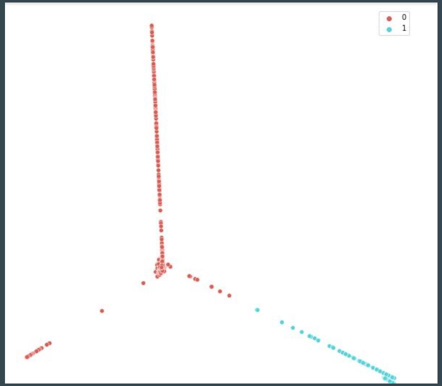


Silhouette Coefficient



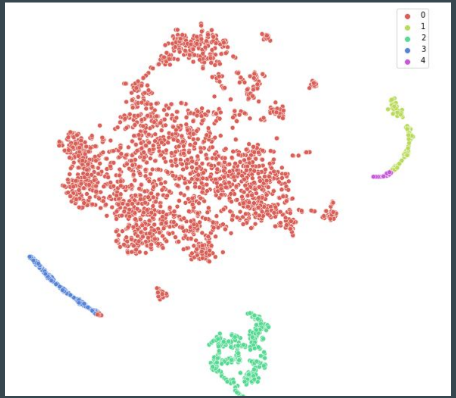
Experimental results II

0.85686904

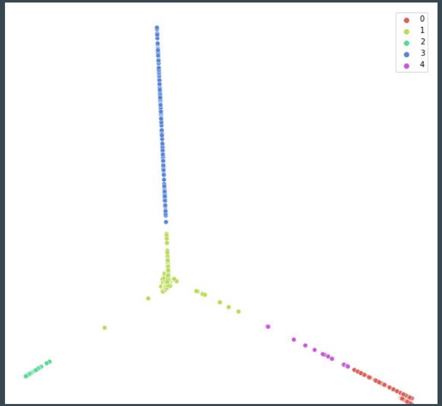


TSNE
K means

Isomap
Agglomerative clustering



0.90360093



Conclusions

- Key point of the project: adequate hidden representations inside the autoencoder.
  - No too simple, no too complex: reaching the best model is an empirical process.
- The greater the distance between anomalies and nominal data, the better.
- In this case no linear dimensionality reduction helped us to get the best model but linear reduction performed good as well.
- With good hidden representations, independently the method of clustering, the result is good enough.
- Resources limitations can limit the improvement.

Grimaldi Alberto
Cinthya Celina Tamayo Gonzalez