Deep Correlation-Aware Kernelized Autoencoders for Anomaly Detection in Cybersecurity

December 19, 2022

Table of Contents

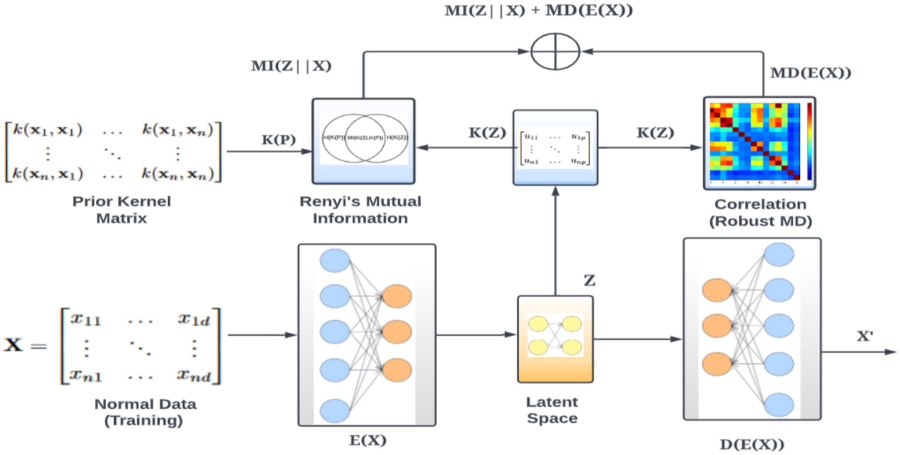
Motivation

- Discriminating anomalies from normal data is difficult in high-dimensional space.
- Empirical evidence shows that retaining valuable properties of input data in latent space helps in the better reconstruction of OOD data.
- Reconstruction error alone fails to consider useful correlation information in the feature space.
- Real-world sensor data is often skewed and non-Gaussian in nature, making mean-based estimators unreliable for skewed data.
- Latent space regularization becomes important in order to preserve the correlation of input space in latent space.
- Anomalies are rare and require expertise to label - Unsupervised representation learning.
- Consider modeling the near and far anomalies separately.

Current Approaches

- Distance Based: Euclidean. Mahalanobis, Minkowski, Nearest Neighbor,
- Clustering Based: Hierarchical clustering, K-means clustering, density-based clustering like DBSCAN.
- Subspace Approach: Combine distance-based and density estimation with some joint learning techniques.

Architecture



Objective function

The robust information theoretic autoencoder enabled with the robust MD metric is trained by optimizing the following loss function

$$\mathcal{L} = \alpha \cdot D_M(Z) + \beta \cdot \mathcal{L}_e(X, D(E(X)) + \max\left(MI_{D_{CS}}(Z||X)\right) \quad (1)$$

Robust Hybrid Error with MD in Latent Space

$$\mathcal{L} = \alpha \cdot D_M(Z) + \beta \cdot \mathcal{L}_e(X, D(E(X))) \tag{2}$$

The robust form of the Mahalanobis distance (D_M) is calculated based on how many standard deviations an encoded sample z_i is from the median encoded data in the latent space. In the encoded space, it is estimated as

$$\hat{D}_M\left(Z\right) = \sqrt{\left(Z - median\right)^T R^{-1} \left(Z - median\right)},$$

where $\hat{R}$ is the estimated feature-based correlation matrix of encoded data in the latent space and the robust correlation co-efficient(ρ) is given by:

$$\rho_{Z_i, Z_j} = \frac{\mathbf{E}[(Z_i - median_i)(Z_j - median_j)]}{MAD_{z_i}, MAD_{z_j}},$$

where the MAD is given by

$$MAD_{z_i} = median|Z_i - median(Z_i)|$$

Matrix-based Renyi's Entropy and Joint Entropy

Let $\mathcal{G}$ be the gram matrix obtained from evaluating a positive definite kernel on all pairs of samples in the original input space. Then the matrix-based analogue of Renyi's α entropy [Yu+21] of order 2 for a normalized positive semi-definite matrix X of size $N \times N$, can be defined as

$$\hat{H}_2(X) = -\log_2(tr(X^2)) = -\log_2\left(\sum_{i=1}^{N} \lambda_i(X)^2\right), \tag{3}$$

where $X_{i,j} = \frac{1}{N}\frac{\mathcal{G}i,j}{\sqrt{\mathcal{G}_{ii}\mathcal{G}_{jj}}}$ and $\lambda_i(X)$ denotes the i^{th} eigenvalue of X. We used this measure to estimate the entropy of the original data space.

Similarly, the matrix-based Renyi's entropy for the latent space can be given as

$$\hat{H}_2(Z) = -\log_2(tr(Z^2)) = -\log_2\left(\sum_{i=1}^{N}\lambda_i(Z)^2\right), \tag{4}$$

From the perspective of information theory, the dependence measure or the total correlation quantifies how much a feature variable $m = \{m^1, m^2, m^3\} \in \mathbf{R}^d$, either latent space or original space, deviates from the statistical independence in each dimension d and is expressed as

$$\sum_{i=1}^{d} H(m^i) - H(m^1, m^2, m^3, ...m^d), \tag{5}$$

where $H(m)$ may be the entropy of the input space or the latent space or the joint entropy between the latent space and the input space expressed as a difference of the joint entropy and the individual entropy of the independent features.

Now, the matrix-based analogue of α order joint entropy between latent space Z and prior space X, defined as

$$\hat{H}_2\left(X, Z\right) = H_2\left(\frac{X \circ Z}{tr\left(X \circ Z\right)}\right), \tag{6}$$

where $\circ$ represents the Hadamard product of matrices.
Based on the above definitions, we calculate the mutual information between latent and prior space with the help of the matrix-based normalized Renyi's entropy of the latent space, input space and the joint entropy between the prior space and latent space data.

Mutual Information in terms of Cauchy Schwarz divergence

The mutual information between input and latent space when expressed in terms of Cauchy-Schwarz divergence is as follows

$$\hat{D}_{CS}(p_x||p_z) = \log_2 \frac{H_2(X)H_2(Z)}{{H_2}^2(X,Z)}, \tag{7}$$

where $H_2(X,Z)$ is the second-order joint entropy between latent and prior space. **Empirical evidence suggests that, the larger the deviation of this subspace from the mutual independence in each dimension, the higher the potential that it is easier to distinguish outliers from normal observations.**

Datasets

We selected two benchmark datasets: CSE-CIC-IDS2018 and NSL-KDD for our experiments.

- **CSE-CIC-IDS2018:** This is a publicly available cybersecurity dataset that is made available by Communications Security Establishment (CSE) and the Canadian Cybersecurity Institute (CIC). We selected 29 continuous features to build the binary classification model.
- **NSL-KDD:** This is also a publicly available benchmark cybersecurity dataset made available by CIC. It has a total of 43 different features of internet traffic flow. We selected 20 most influential features after a correlation analysis of the features.

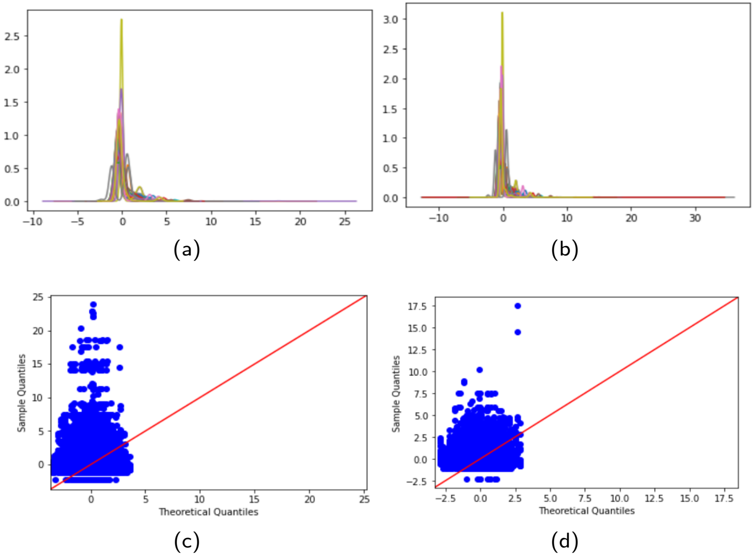


Figure 1: (a) Histogram of normal samples; (b) Histogram of anomaly samples (c) QQ plot of normal

Table 1: The skewed features in CSE-CIC-IDS dataset

Features	Skewness	Kurtosis
Fwd Pkt Len Mean	6.255196	92.777752
Flow Byts/s	20.927526	503.025265
Bwd IAT Min	10.222297	133.542522
Pkt Len Min	9.092836	127.003666
Fwd Seg Size Avg	6.255196	92.777752
Bwd IAT Mean	15.838105	133.542522
Fwd Pkt Len Min	9.047784	123.113359

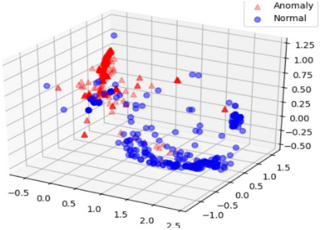
Baseline Models

- **DKAE:** The Deep Kernelized Autoencoder [**kampffmeyer2017deep**] has a kernel alignment loss that is calculated as the normalized Frobenius distance between the latent dimension code matrix and the prior kernel matrix and a reconstruction loss.
- **DAGMM:** The Deep Autoencoding Gaussian Mixture Model [Zon+18] is an unsupervised anomaly detection model that optimizes the parameters of the deep autoencoder and the mixture model simultaneously using an estimation network to facilitate the learning of a Gaussian Mixture Model (GMM).
- **VAE:** VAE leverages a probabilistic encoder-decoder network and the reconstruction probability is used for detecting anomalies. Although it performs latent dimension regularization, it assumes a parametric distribution of the data which is mostly Gaussian.
- **MD-based Autoencoder:** This model [Den+18] also leverages MD in the latent space with mean as the estimator of location and the covariance in the feature space.

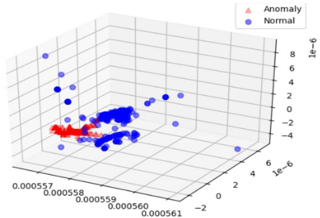
Improve reconstruction error for generative modeling

- The proposed autoencoder shows an improvement of **10%-15%** in **MSE** while reconstructing out-of-distribution data compared to the DKAE model which uses reconstruction error and kernel misalignment error.
- This is mainly attributed to the effectiveness of the robust correlation matrix and the robust position and scale estimator in reconstructing OOD test data.

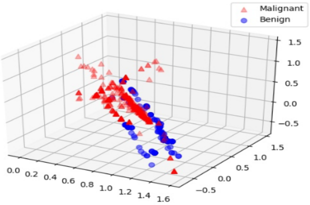
Latent dimension visualization of learned AE on test data with PCA

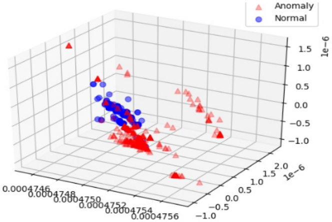


(a)



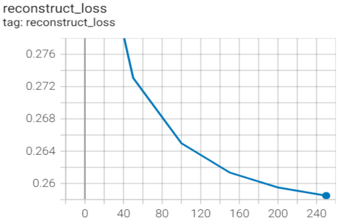
(b)



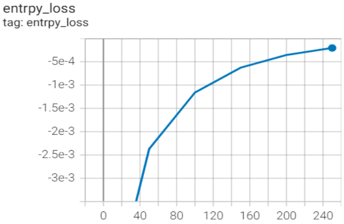


The latent dimension visualization shows that the normal and anomaly samples are well separated with the learned robust MD autoencoder(on a different scale) compared to the Euclidean distance. For e.g, if the calculated Euclidean distance of a normal sample is 1.0 and an anomaly is 0.5, the new robust MD distance of normal sample is 0.1 and anomaly is 0.005, thus making the relative distance between a normal and anomaly sample 20 times with the MD distance(especially for the near anomalies). Therefore, it can detect the near anomalies which lie close to the latent dimension manifold distinctively in the latent space.

Training



(a)



(b)

Figure 3: (a) Robust MD loss; (b) Mutual Information Gain

Results

Model	Type	CSE-CIC-IDS Dataset				NSL-KDD			
		Accuracy	Precision	Recall	AUC	Accuracy	Precision	Recall	AUC
DRMDIT-AE	Near	**71.5**	70.7	**73.4**	**74.5**	**92.4**	**94.6**	**93.1**	**96.0**
	Far	78.9	80.1	**82.4**	**78.9**	**91.2**	**95.2**	**95.4**	**97.3**
DKAE	Near	50	75.2	50.8	71.4	82.0	81.6	92.2	95.6
	Far	79.8	85.2	79.8	74.5	74.5	74.7	74.7	74.5
DAGMM	Near	67.6	68.1	67.9	67.9	95.4	86.9	89.3	88.8
	Far	66.6	67.1	66.8	65.1	94	86.8	89.1	91.9
VAE	Near	61.9	62.1	66.6	65.7	84.2	84.9	88.0	86.4
	Far	75.9	74.5	72.2	73.3	87.0	85.3	81.0	83.6
MD-AE	Near	57.3	57.2	52.2	53.4	82.3	83.9	82.5	82.6
	Far	72.8	70.6	71.6	71.6	81.3	83.6	83.5	79.5

Conclusion and Future Scope

- We propose a correlation-aware deep kernelized autoencoder that leverages the robust MD in latent feature space and the principle of Renyi's mutual information maximization between prior and latent space in order to detect anomalies in cyber-security data.
- The MAD- and median-based MDs and their robust correlation estimators are useful indicators of specific kinds of anomalies especially when the data is non-Gaussian.
- In the future, we would like to explore further in this direction and try to address the important problem of domain generalization in the field of cybersecurity, where a model trained on a number of known kinds of attacks can generalize to unseen attacks, also known as zero-day attacks.

References I

[BCV13] Yoshua Bengio, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives". In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.

[CBK09] Varun Chandola, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey". In: *ACM computing surveys (CSUR)* 41.3 (2009), pp. 1–58.

[Den+18] Taylor Denouden et al. "Improving reconstruction autoencoder out-of-distribution detection with mahalanobis distance". In: *arXiv preprint arXiv:1812.02765* (2018).

[Ham01] Frank R Hampel. "Robust statistics: A brief introduction and overview". In: *Research report/Seminar für Statistik, Eidgenössische Technische Hochschule (ETH)*. Vol. 94. Seminar für Statistik, Eidgenössische Technische Hochschule. 2001.

[Hub11] Peter J Huber. "Robust statistics". In: *International encyclopedia of statistical science*. Springer, 2011, pp. 1248–1251.

References II

[Pri10] Jose C Principe. *Information theoretic learning: Renyi's entropy and kernel perspectives*. Springer Science & Business Media, 2010.

[RC93] Peter J Rousseeuw and Christophe Croux. "Alternatives to the median absolute deviation". In: *Journal of the American Statistical association* 88.424 (1993), pp. 1273–1283.

[Ren+21] Jie Ren et al. "A simple fix to mahalanobis distance for improving near-ood detection". In: *arXiv preprint arXiv:2106.09022* (2021).

[Yu+21] Shujian Yu et al. "Measuring dependence with matrix-based entropy functional". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 12. 2021, pp. 10781–10789.

[Zon+18] Bo Zong et al. "Deep autoencoding gaussian mixture model for unsupervised anomaly detection". In: *International conference on learning representations*. 2018.