MAHALANOBIS DISTANCE BASED ADVERSARIAL NETWORK FOR ANOMALY DETECTION

Yubo Hou[1], *Zhenghua Chen*[1]*, *Min Wu*[1], *Chuan-Sheng Foo*[1], *Xiaoli Li*[1], *Raed M. Shubair*[2,3]

[1] Institute for Infocomm Research, A*STAR, 1 Fusionopolis Way #21-01 Connexis, Singapore 138632
[2] Research Laboratory of Electronics (RLE), Massachusetts Institute of Technology (MIT), USA
[3] Department of Electrical and Computer Engineering, New York University (NYU) Abu Dhabi, UAE

ABSTRACT

Anomaly detection techniques are very crucial in multiple business applications, such as cyber security, manufacturing and finance. However, developing anomaly detection methods for high-dimensional data with high speed and good performance is still a challenge. Generative Adversarial Networks (GANs) are able to model the complex high-dimensional data, but they still require large computation in inference stage. This paper proposes an efficient method, known as Mahalanobis Distance-based Adversarial Network (MDAN), for anomaly detection. The proposed MDAN models the data using generative adversarial network (GAN) and detects anomalies by using the Mahalanobis distance. The proposed MDAN outperforms conventional GAN-based methods considerably and has a higher inference speed, when applied to several tabular and image datasets.

***Index Terms*—** Anomaly Detection, Mahalanobis Distance, Generative Adversarial Network

1. INTRODUCTION

Anomaly detection refers to the identification of rare observations which have significant difference from majority data. Due to its practicality, anomaly detection techniques are important in various applications, such as cyber-intrusion detection [1], industrial damage detection [2], fraud detection [3], unusual urban traffic flow detection [4], medical anomaly detection, and video surveillance [5]. All these applications are required to determine whether a new sample belongs to the same existing distribution of normal data, or should be considered an anomaly. This problem cannot be solved by traditional supervised classification methods, because the amount of anomalies is insufficient to be effectively modeled and new emerging anomalies will appear. Therefore, the anomaly detection is usually considered to be an unsupervised one-class classification problem, which is solved by training the model based solely on the normal samples.

The existing anomaly detection models can be categorized into three categories according to anomaly score calculation method: 1) boundary-based methods; 2) density-based methods; and 3) reconstruction-based methods. Boundary-based methods attempt to learn a classification boundary around the normal data, such as one class support vector machine (OC-SVM) [6]. Density-based methods utilize the relation between a sample and its neighbors. A sample is identified as an anomaly if its density is relatively lower than that of its neighbors or it needs more steps to separate from group. The third class of methods utilizes reconstruction error to determine whether a sample is anomalous. For example, auto-encoder (AE) [7] trains a network to reconstruct normal data and identifies anomalies if samples can not be reconstructed well. Deep Autoencoding Gaussian Mixture Model (DAGMM) [8] is another representative method.

Recently, the generative adversarial network (GAN), which is a well-studied deep learning framework, has been used for anomaly detection [5]. For example, [9] proposed a standard GAN-based method for anomaly detection on eye images. However, this method requires a computationally-expensive inference procedure to recover latent space features for each testing sample. The authors in [10] proposed an adversarially-learned anomaly detection (ALAD) method based on bi-directional GANs [11, 12]. ALAD trains an encoder network to recover latent space features, and thus its inference procedure is much more efficient than [9].

In this work, we propose a Mahalanobis Distance-based Adversarial Network (MDAN) for anomaly detection. Similar to ALAD [10], we also use an encoder to learn the latent low-dimensional features for the input data. We further simplify ALAD's bi-directional GAN structure and design new loss functions during the training process. In the inference phase, we directly measure the Mahalanobis distance of learned low-dimensional features to the normal distribution as anomaly score. As such, MDAN can boost the performance for anomaly detection, and also improve the efficiency in both training and inference stages. Experimental results on high-dimensional tabular and image datasets also demonstrate that our proposed MDAN approach is efficient and effective.

This work is supported by the A*STAR IIoT Research Program under the RIE2020 IAF-PP Grant A1788a0023. (* indicates corresponding author.)

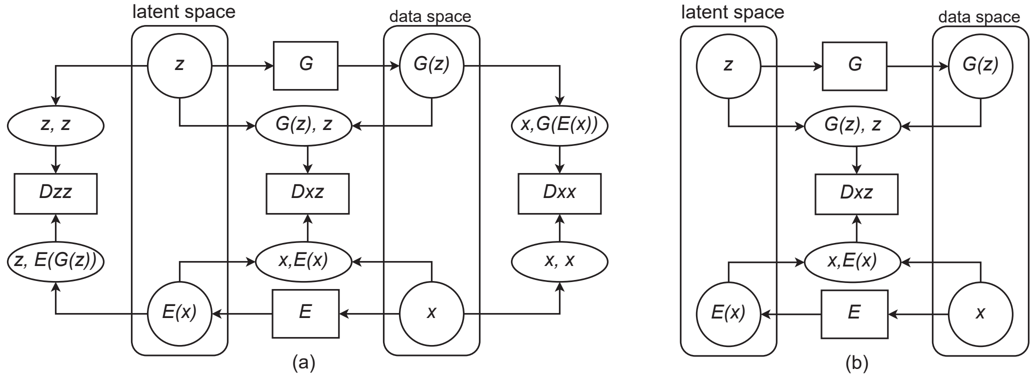

Fig. 1. (a) Structure of ALAD and (b) Training structure of MDAN.

2. METHOD

In this section, we first briefly elaborate the overall idea of GAN-based anomaly detection. We then present our proposed MDAN method. After that, we introduce the GAN architecture in MDAN for training in Section 2.2 and the anomaly inference in Section 2.3.

2.1. GAN and BiGAN

The standard GAN consists of two models, namely generator and discriminator. The generator aims to model the data distribution by mapping a random sample from latent space to the input data space. The discriminator aims to distinguish the real input data and fake data generated by the generator. Let's denote the generator and discriminator as G and D, the distributions for latent space and input data space as p_z and p_{data}. The adversarial learning process will learn both the generator G and discriminator D simultaneously through a min-max game below.

$$\min_{G}\max_{D}\ \mathbb{E}_{x\sim p_{data}}[logD(x)] + \mathbb{E}_{z\sim p_z}[log(1-D(G(z))] \tag{1}$$

BiGAN [11] and AliGAN [12] provide an additional encoder to invert data space to latent space. Hence, we can leverage such an architecture to extract feature and reduce dimensionality for the input data. We denote the encoder as E. Then, the generator G, discriminator D and encoder E can be similarly learned by optimizing the objective function below.

$$\min_{G,E}\max_{D}\ \mathbb{E}_{x\sim p_{data}}[logD(x,E(x))] + \mathbb{E}_{z\sim p_z}[log(1-D(G(z),z)] \tag{2}$$

Recently, GANs have been employed for anomaly detection and achieved notable results [13]. For example, AnoGAN [9] uses the standard GAN for anomaly detection, while ALAD [10] detects anomalies based on BiGAN.

2.2. GAN architecture in MDAN

In this paper, we also leverage BiGAN for anomaly detection. Figure 1(a) shows the structure of the existing ALAD method, while Figure 1(b) shows the structure of our proposed MDAN. As shown in Figure 1(a), ALAD utilizes two additional discriminators D_{xx} and D_{zz}. In particular, the adversarial loss of generator from the discriminators D_{xx} and D_{zz} is considered as the cycle consistency loss in ALAD to stabilize the training.

In MDAN, we removed these two discriminators D_{xx} and D_{zz} to speed up the training process. Correspondingly, we also need to update the the cycle consistency loss to stabilize the training. The generator G should invert encoder E theoretically (i.e., $G(E(x)) \approx x$). However, the real situation is that the model often does not coverage to the saddle point when training in practice. To address this issue, we use the reconstruction errors for both x and z in MDAN as the the cycle consistency loss, which are calculated using mean square error in Equations 3 and 4.

$$loss_{rec_x} = mean((x - G(E(x)))^2) \tag{3}$$

$$loss_{rec_z} = mean((z - E(G(z)))^2) \tag{4}$$

As such, our generator loss is then updated in Equation 5 by including the above reconstruction losses.

$$loss_{gen} = loss_{adv} + loss_{rec_x} + loss_{rec_z}, \tag{5}$$

where the adversarial loss $loss_{adv}$ of generator from the discriminator D_{xz} is calculated by using cross entropy in Equation 6.

$$loss_{adv} = \frac{1}{N}\sum_{n=1}^{N} -log(D(G(z_n), z_n)). \tag{6}$$

The above encoder E, generator G and the discriminator D can be trained using *Adam* in two stages. First, we train the discriminator to distinguish the true samples from the generated fake samples. Second, we train the encoder and the generator so as to fool the discriminator with its generated samples. As the learned encoder E can invert the generator G and thus $E(x)$ can serve as a useful feature representation for input sample x. Next, we will show the anomaly inference in MDAN based on $E(x)$.

2.3. Anomaly detection in MDAN

ALAD utilizes the discriminator D_{xx} to calculate the anomaly score, which also requires encoder and generator to help. We found that the discriminator D_{zz} in ALAD can also be used to calculate the anomaly score. The input of D_{zz} is either random variable from standard normal distribution or $E(G(z))$ during training. It makes the output of encoder E obey normal distribution approximately.

To further speed up the inference procedure, we propose to leverage only the encoder for anomaly detection. In particular, we calculate the Mahalanobis distance of $E(x)$ to the multivariate normal distribution as the anomaly score as shown in Figure 2.

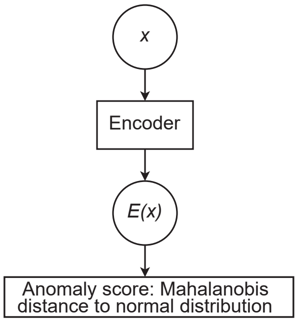

Fig. 2. Inference structure of MDAN.

The Mahalanobis distance is a measure of the distance between a point and a distribution. It can be formulated as:

$$D(z) = \sqrt{(z-\mu)^T S^{-1}(z-\mu)} \tag{7}$$

where S is the covariance matrix and μ is mean of the multivariate normal distribution. In this case, we use multivariable standard normal distribution $N(0, 1)$ to generate random sample z when training the model. Here, S will become an identity matrix and μ will be a zero matrix. So the Mahalanobis distance becomes:

$$D(z) = \sqrt{z^T z} \tag{8}$$

Concretely, having trained a model on the normal data to provide E, G and D, we define an anomaly score function based on the Mahalanobis distance and standard normal distribution:

$$D(z) = \sqrt{E(x)^T E(x)} \tag{9}$$

The $E(x)$ of normal data should obey the normal distribution and its Mahalanobis distance is supposed to be small. If $E(x)$ does not obey the normal distribution, its Mahalanobis distance should be larger than the normal ones, and thus it can be detected as an anomaly.

3. EXPERIMENTS

3.1. Experimental setup

We evaluate the proposed MDAN on public tabular and image datasets. For the tabular data, we adopt the widely used KDD-Cup99 10% dataset [14] which is a network intrusion dataset. For image datasets, we leverage the SVHN dataset [15] containing numbers from 0 to 9, and the CIFAR-10 dataset [16] that have 6 kinds of animals and 4 kinds of transportation. Figure 3 and Figure 4 show some samples of these two image datasets. To quantify the performance of the model, we use the criteria of Precision, Recall, and F1 score for the KDD99 dataset and the criterion of area under ROC (AUROC) for the image datasets. This setting is the same as [10] for fair comparison. Experiments are conducted with Tensorflow running on GPU. The machine for experiments is a workstation with Intel Core i9-7920X CPU, 128G memory, 4TB SSD and Nvidia GeForce GTX 1080Ti.

Fig. 3. Sample data in SVHN.

Table 1. Statistics of the public benchmark datasets

Dataset	Features	Total Instance
KDD99	121	494021
SVHN	3072	99289
CIFAR-10	3072	60000

To verify the performance of the proposed approach, we have compared the performance of the proposed method with the following reference algorithms: One Class Support Vector Machine (OC-SVM) [6], Isolation Forest (IF) [17], Deep Structured Energy Based Model (DSEBM) [18], Deep Autoencoding Gaussian Mixture Model (DAGMM) [8],

Fig. 4. Sample data in CIFAR-10.

AnoGAN [9], and Adversarially Learned Anomaly Detection (ALAD) [10].

3.2. Result and Discussion

We report results in Table 2 on the tabular data and Table 3 on the image data. It can be found that our method is competitive with baseline methods. Results for other methods except MDAN on KDD99, SVHN and CIFAR-10 were obtained from [10]. Results for MDAN on KDD99 are averaged over 10 runs and results for MDAN on SVHN and CIFAR-10 are averages over all tasks over 3 runs, which is the same as [10].

Table 2. Performance on the tabular dataset

Dataset	Model	Precision	Recall	F1 score
KDD99	OC-SVM	0.7457	0.8523	0.7954
	IF	0.9216	0.9373	0.9294
	DSEBM-r	0.8521	0.6472	0.7328
	DSEBM-e	0.8619	0.6446	0.7399
	DAGMM	0.9297	0.9442	0.9369
	AnoGAN	0.8786	0.8297	0.8865
	ALAD	0.9427	0.9577	0.9501
	MDAN	**0.9472**	**0.9623**	**0.9547**

As show in Table 2 and Table 3, the proposed MDAN outperforms all other methods on the KDDCup99 10% and SVHN datasets. And MDAN is quite competitive on the CIFAR-10 dataset. Specifically, on the SVHN dataset, GAN-based methods including AnoGAN, ALAD and MDAN perform better than other methods, which indicates the effectiveness of GAN-based solutions for anomaly detection. Moreover, MDAN significantly improves on AUROC by around 7.51% compared with the second best method, i.e., ALAD. On the CIFAR-10 dataset, the performance of MDAN is close to ALAD and better than other methods. Besides, the variance of MDAN is the lowest, which means MDAN is more stable.

Table 3. Performance on the image datasets

Dataset	Model	AUROC
SVHN	OC-SVM	0.5027 ± 0.0132
	IF	0.5163 ± 0.0120
	DSEBM-r	0.5290 ± 0.0129
	DSEBM-e	0.5240 ± 0.0067
	AnoGAN	0.5410 ± 0.0193
	ALAD	0.5753 ± 0.0268
	MDAN	$\mathbf{0.6185 \pm 0.0516}$
CIFAR-10	OC-SVM	0.5843 ± 0.0956
	IF	0.6025 ± 0.1040
	DSEBM-r	0.6071 ± 0.1007
	DSEBM-e	0.5956 ± 0.1151
	AnoGAN	0.5949 ± 0.1076
	ALAD	$\mathbf{0.6072 \pm 0.1201}$
	MDAN	0.6035 ± 0.0512

Table 4. Training and inference time comparison

Dataset	Model	Training (second per epoch)	Inference (second)
KDD99	ALAD	64.74	7.52
	MDAN	**29.12**	**2.47**
SVHN	ALAD	10.27	6.63
	MDAN	**6.11**	**2.23**
CIFAR-10	ALAD	13.64	3.30
	MDAN	**5.34**	**1.16**

The performance of ALAD is quite competitive and similar to MDAN. So we also compare the speed of these two methods. Table 4 presents the training time and inference time of MDAN and ALAD on different datasets over 10 runs. Training time is measured per epoch on the entire training data and inference time is counted on the entire test data. We can find that MDAN trains model 2 times faster than ALAD, because ALAD has three discriminators and MDAN only has one. During testing, MDAN maps high-dimensional data to low-dimensional features using encoder and detects anomalies by the simplified Mahalanobis distance. While, ALAD uses a discriminator to extract features from the output of encoder and generator and calculates reconstruction error in feature space. Compared to ALAD, MDAN reduces the computation of the inference procedure and speeds up around 3 times.

4. CONCLUSION

In this paper, we proposed a Mahalanobis distance-based adversarial network method MDAN for anomaly detection. It utilizes BiGAN to train the encoder for feature extraction and dimension reduction. Then, it detects anomalies by calculating the Mahalanobis distance of the encoder's output to the normal distribution. The experiment on both tabular data and image data shows that MDAN is both efficient and effective, when compared to conventional GAN-based methods.

5. REFERENCES

[1] M. A. Siddiqui, J. W. Stokes, C. Seifert, E. Argyle, R. McCann, J. Neil, and J. Carroll, "Detecting cyber attacks using anomaly detection with explanations and expert feedback," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2872–2876.

[2] M. Riazi, O. Zaiane, T. Takeuchi, A. Maltais, J. Günther, and M. Lipsett, "Detecting the onset of machine failure using anomaly detection methods," in *International Conference on Big Data Analytics and Knowledge Discovery*. Springer, 2019, pp. 3–12.

[3] M. Ahmed, A. N. Mahmood, and M. R. Islam, "A survey of anomaly detection techniques in financial domain," *Future Generation Computer Systems*, vol. 55, pp. 278–288, 2016.

[4] Y. Djenouri, A. Zimek, and M. Chiarandini, "Outlier detection in urban traffic flow distributions," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 935–940.

[5] R. Chalapathy and S. Chawla, "Deep learning for anomaly detection: A survey," *arXiv preprint arXiv:1901.03407*, 2019.

[6] B. Schölkopf, R. C. Williamson, A. J. Smola, J. Shawe-Taylor, and J. C. Platt, "Support vector method for novelty detection," in *Advances in neural information processing systems*, 2000, pp. 582–588.

[7] J. Chen, S. Sathe, C. Aggarwal, and D. Turaga, "Outlier detection with autoencoder ensembles," in *Proceedings of the 2017 SIAM International Conference on Data Mining*. SIAM, 2017, pp. 90–98.

[8] B. Zong, Q. Song, M. R. Min, W. Cheng, C. Lumezanu, D. Cho, and H. Chen, "Deep autoencoding gaussian mixture model for unsupervised anomaly detection," *International Conference on Learning Representations*, 2018.

[9] T. Schlegl, P. Seeböck, S. M. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised anomaly detection with generative adversarial networks to guide marker discovery," in *International Conference on Information Processing in Medical Imaging*. Springer, 2017, pp. 146–157.

[10] H. Zenati, M. Romain, C.-S. Foo, B. Lecouat, and V. Chandrasekhar, "Adversarially learned anomaly detection," in *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 727–736.

[11] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," *International Conference on Learning Representations*, 2017.

[12] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville, "Adversarially learned inference," *International Conference on Learning Representations*, 2017.

[13] F. Di Mattia, P. Galeone, M. De Simoni, and E. Ghelfi, "A survey on gans for anomaly detection," *arXiv preprint arXiv:1906.11632*, 2019.

[14] D. Dua and C. Graff, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml

[15] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," 2011.

[16] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," Citeseer, Tech. Rep., 2009.

[17] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008, pp. 413–422.

[18] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang, "Deep structured energy based models for anomaly detection," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ser. ICML'16. JMLR.org, 2016, pp. 1100–1109. [Online]. Available: http://dl.acm.org/citation.cfm?id=3045390.3045507