DSF-GAN: DOWNSTREAM FEEDBACK GENERATIVE ADVERSARIAL NETWORK

Oriel Perets
Software and Information Systems Engineering
Ben Gurion University

Nadav Rappoport
Software and Information Systems Engineering
Ben Gurion University

ABSTRACT

Utility and privacy are two crucial measurements of the quality of synthetic tabular data. While significant advancements have been made in privacy measures, generating synthetic samples with high utility remains challenging. To enhance the utility of synthetic samples, we propose a novel architecture called the DownStream Feedback Generative Adversarial Network (DSF-GAN). This approach incorporates feedback from a downstream prediction model during training to augment the generator's loss function with valuable information. Thus, DSF-GAN utilizes a downstream prediction task to enhance the utility of synthetic samples. To evaluate our method, we tested it using two popular datasets. Our experiments demonstrate improved model performance when training on synthetic samples generated by DSF-GAN, compared to those generated by the same GAN architecture without feedback. The evaluation was conducted on the same validation set comprising real samples. All code and datasets used in this research will be made openly available for ease of reproduction.

1 INTRODUCTION & RELATED WORK

1.1 INTRODUCTION

The use of synthetic tabular data is a topic of high recent interest. Synthetic data serves various purposes, such as preserving privacy by replacing the original samples (Yoon et al., 2020), supplementing scarce datasets to develop models requiring more data points and boosting performance (Che et al., 2017), and balancing datasets to eliminate bias (Perets & Rappoport, 2023). However, the utility of the generated synthetic data is a critical factor. Data utility measures the effectiveness of synthetic data for a specific task when compared to real samples. One common method for assessing synthetic data utility is by evaluating machine learning efficacy, i.e., comparing the performance of a model trained on synthetic samples with one trained using the real samples, evaluated on the same test set (Chin-Cheong et al., 2019). Despite the success of GAN-based approaches in generating realistic and privacy-preserving synthetic samples, the utility of synthetic tabular data lags behind that of real samples in state-of-the-art results (Xu et al., 2019; Rajabi & Garibay, 2022).

1.2 RELATED WORK

A feedback mechanism was first introduced as a way to enhance synthetic samples' similarity. CTAB-GAN (Zhao et al., 2021) and its predecessor CTAB-GAN+ (Zhao et al., 2022) both propose a conditional GAN with a form of feedback that verifies the semantic integrity of synthetic samples' feature values (e.g., a sample whose `sex=female`, and `prostate cancer=true` is not semantically correct) (Zhao et al., 2022). Unlike these approaches, our approach uses the classification or regression performance of a model trained for the specific underlying task the data is aimed at solving. Feedback GAN (FBGAN) (Huh et al., 2019) uses a separate predictor denoted as a function analyzer to optimize the generated gene sequences for desired properties, samples above a threshold are fed back into the discriminator. Similarly, the Feedback Antiviral Peptides GAN (Hasegawa et al., 2022) uses AVPs sampled from the generator mid-training to further train the discriminator. While all of the above approaches propose some sort of feedback mechanism, to the best

of our knowledge, no documented literature proposed using feedback from the actual classifier or regressor trained for the actual downstream task the synthetic data is generated for.

2 METHODOLOGY

The original GAN loss function is given by:

$$\min_G \max_D \mathbb{E}_{x \sim P_{\text{data}}(x)}[\log D(x;\theta_d)] + \mathbb{E}_{z \sim P_z(z)}[1 - \log D(G(z;\theta_g);\theta_d)]$$

Where $P_{data}(x)$ is the real data's distribution. $D(x;\theta_d)$ is the prediction produced by the discriminator D, when fed with a mini-batch of real samples x, w.r.t the discriminator's parameters θ_d. P_z is the noise distribution from which the generator G is sampling to generate the synthetic samples, by employing the function $G(z)$ w.r.t the generator's parameters θ_g. $D(G(z;\theta_g);\theta_d)$ represents the prediction produced by the discriminator D for the a mini-batch of generated sample w.r.t to θ_d. The training procedure consists of two loops optimizing G and D iteratively (Goodfellow et al., 2020).

We demonstrate our architecture using the Conditional Tabular GAN (CTGAN) (Xu et al., 2019) where the loss function is defined as:

$$\min_G \max_{w \in W} V(G,f) = \mathbb{E}_{x \sim P_{data}(x)}[f(x;w)] + \mathbb{E}_{z \sim P_z}[f(G(z;\theta_g);w)] + H$$

Where the discriminator D is replaced with a critic f as proposed by (Arjovsky et al., 2017) and H is the binary cross entropy loss between the conditional vector composed on the real samples and the one composed on the generated samples. Let $loss_G$ denote generator loss, and $loss_D$ denote discriminator loss, our proposed approach (DSF-GAN) facilitates feedback from a downstream task mid-training, adding a loss term of the downstream classifier of regressor to the $loss_G$ scaled by $\lambda \in \mathbb{R}^+$ to control the magnitude of the feedback term.

Case 1 – Classification: let $h(x,\theta_h)$ denote a logistic regression model, with loss function given as log-loss:

$$L = -\left(y \log(\hat{y}) + (1-y)\log(1-\hat{y})\right)$$

Case 2 – Regression: let $h(x,\theta_h)$ denote a linear regression model, with loss function given as RMSE:

$$L = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(\hat{y}_i - y_i)^2}$$

Hence, with the regression or classification loss denoted as L_f, the new $Loss_G$ with feedback is:

$$L_G = \mathbb{E}_{z \sim P_z}[f(G(z;\theta_g);w)] + H + (\lambda * L_f)$$

3 EXPERIMENTS

We used the base GAN architecture presented in (Xu et al., 2019) for this experiment. For empirically evaluating the increase in synthetic data utility contributed by the downstream feedback, we used two distinct datasets. We trained a DSF-GAN for $\frac{N}{2}$ epochs using the original loss function. For each iteration of training in the remaining epochs ($\frac{N}{2}+1$,..,N), we sampled synthetic samples $\hat{x}$ from G and used it to train a logistic regression or linear regression model. We then added the respective loss term to $loss_G$. Post-training, we sampled n samples from the trained model, and used it as a training set for a regression or classification model, we evaluated the model performance using a set-aside validation set comprised of real samples which were excluded from the GAN training. Our results show an increase in utility for the classification and regression tasks and are presented in Tables 2 and 3. Epochs, batch sizes, datasets, and more details are presented in Appendix A.

4 CONCLUSIONS

In this work, we propose a novel GAN architecture, with a feedback mechanism from a downstream task (i.e., the original task the dataset was used to solve). This was done by utilizing the loss

function of the downstream task trained on the synthetic samples mid-training, and evaluated on the real samples. The empirical experiments show increased synthetic data utility, hence proving the potential of this architecture. Many directions for future work are possible. For example, using other forms of feedback or feedback models. This research is another stepping stone in enabling synthetic data's safe and efficient use in machine-learning tasks.

URM STATEMENT

We acknowledge that all authors in this paper meet the URM criteria of ICLR 2024 Tiny Papers Track.

REFERENCES

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN, January 2017. URL https://arxiv.org/abs/1701.07875v3.

Zhengping Che, Yu Cheng, Shuangfei Zhai, Zhaonan Sun, and Yan Liu. Boosting Deep Learning Risk Prediction with Generative Adversarial Networks for Electronic Health Records. In *2017 IEEE International Conference on Data Mining (ICDM)*, pp. 787–792, November 2017. doi: 10.1109/ICDM.2017.93. ISSN: 2374-8486.

Kieran Chin-Cheong, Thomas Sutter, and Julia E Vogt. Generation of heterogeneous synthetic electronic health records using gans. In *workshop on machine learning for health (ML4H) at the 33rd conference on neural information processing systems (NeurIPS 2019)*. ETH Zurich, Institute for Machine Learning, 2019.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11): 139–144, October 2020. ISSN 0001-0782, 1557-7317. doi: 10.1145/3422622. URL https://dl.acm.org/doi/10.1145/3422622.

Kano Hasegawa, Yoshitaka Moriwaki, Tohru Terada, Cao Wei, and Kentaro Shimizu. Feedback-AVPGAN: Feedback-guided generative adversarial network for generating antiviral peptides. *J. Bioinform. Comput. Biol.*, 20(06):2250026, December 2022. ISSN 0219-7200. doi: 10.1142/S0219720022500263. URL https://www.worldscientific.com/doi/abs/10.1142/S0219720022500263. Publisher: World Scientific Publishing Co.

Minyoung Huh, Shao-Hua Sun, and Ning Zhang. Feedback adversarial learning: Spatial feedback for improving generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1476–1485, 2019.

Oriel Perets and Nadav Rappoport. Ensemble synthetic ehr generation for increasing subpopulation model's performance. *arXiv preprint arXiv:2305.16363*, 2023.

Amirarsalan Rajabi and Ozlem Ozmen Garibay. TabFairGAN: Fair Tabular Data Generation with Generative Adversarial Networks. *Machine Learning and Knowledge Extraction*, 4(2):488–501, June 2022. ISSN 2504-4990. doi: 10.3390/make4020022. URL https://www.mdpi.com/2504-4990/4/2/22. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling Tabular data using Conditional GAN, October 2019. URL http://arxiv.org/abs/1907.00503. arXiv:1907.00503 [cs, stat].

Jinsung Yoon, Lydia N. Drumright, and Mihaela van der Schaar. Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics*, 24(8):2378–2388, August 2020. ISSN 2168-2208. doi: 10.1109/JBHI.2020.2980262. Conference Name: IEEE Journal of Biomedical and Health Informatics.

Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y. Chen. CTAB-GAN: Effective Table Data Synthesizing. In *Proceedings of The 13th Asian Conference on Machine Learning*, pp. 97–112. PMLR, November 2021. URL https://proceedings.mlr.press/v157/zhao21a.html. ISSN: 2640-3498.

Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y. Chen. CTAB-GAN+: Enhancing Tabular Data Synthesis, April 2022. URL http://arxiv.org/abs/2204.00401. arXiv:2204.00401 [cs].

A APPENDIX

A.1 FIGURES

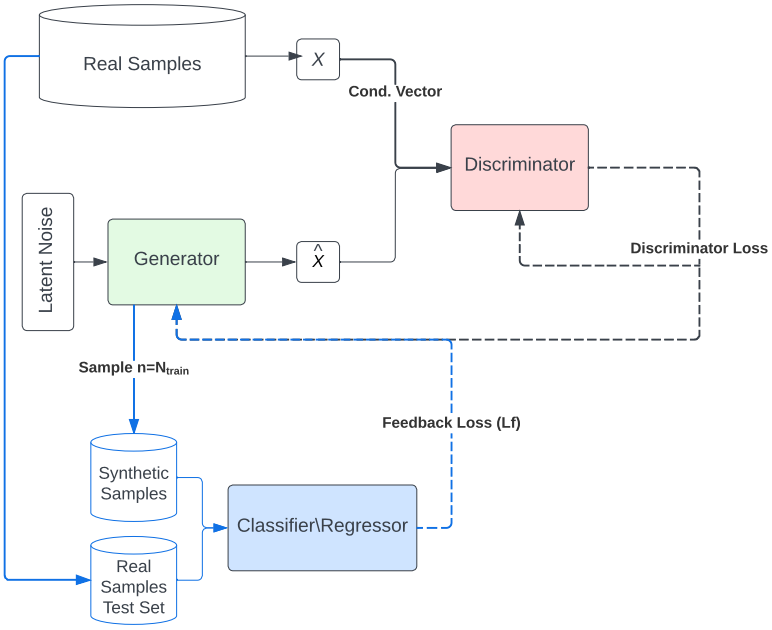

Figure 1: Downstream Feedback Generative Adversarial Network Schematic Flow

A.2 EXPERIMENTAL DETAILS

Before training the GAN model, a train, validation split is performed. The training samples are used to train the GAN models, while the validation set, comprised of real samples, is set aside for evaluating the synthetic data which is ultimately generated by the trained models. Both the GAN models (base architecture and with the feedback mechanism) are trained separately, using the exact same training set, the same number of epochs, batch size, and other parameters e.g., embedding dimensions and learning rate.
To further ensure the stability of the results, we used five-fold Cross-validation and computed the confidence interval, both in the initial split phase before GAN training and in the synthetic data sampling where we sample from the trained model to train the downstream classifier or regressor to evaluate the synthetic data utility.

A.2.1 DATASETS

1. House Price - https://www.kaggle.com/datasets/shree1992/housedata
2. Adult Census Income - https://archive.ics.uci.edu/dataset/2/adult

A.2.2 MODEL TRAINING

Data set	n	Epochs	Batch size	λ
Adult	32,560	100	500	1
House	4,551	500	200	1

Table 1: Training specifics for DSF-GAN

A.3 RESULTS

	No Feedback		**Feedback**	
Data set	Precision	Recall	Precision	Recall
Adult	0.575 ± 0.003	0.441 ± 0.007	0.598 ± 0.003	0.485 ± 0.006

Table 2: Results for classification feedback model, precision and recall for a model evaluated on synthetic generated by the base GAN model, and precision, recall for a model trained on synthetic data generated by the DSF-GAN with the feedback mechanism. All models are evaluated using a validation set comprised of real samples

	No Feedback		**Feedback**	
Data set	RMSE	R^2	RMSE	R^2
House	$0.0118 \pm 1.7e^{-4}$	0.3607 ± 0.018	$0.0115 \pm 5.8e^{-5}$	0.3903 ± 0.006

Table 3: Results for regression feedback model, RMSE and R^2 for a model evaluated on synthetic generated by the base GAN model, and RMSE, R^2 for a model trained on synthetic data generated by the DSF-GAN with the feedback mechanism. All models are evaluated using a validation set comprised of real samples