LCT-GAN - Improving the efficiency of tabular data synthesis via latent embeddings

Viktor Velev

Supervisor(s): Zilong Zhao [1], Lydia Chen [1]

[1]*Distributed Systems Group at EEMCS, Delft University of Technology, The Netherlands*

Introduction

While data sharing is crucial for knowledge development, privacy concerns and strict regulation unfortunately limit its full effectiveness. Synthetic tabular data emerges as an alternative to enable data sharing while fulfilling regulatory and privacy constraints.

In this paper, we build upon the state-of-the-art in tabular data synthesis CTAB-GAN [1]. We introduce LCT-GAN, which consists of an embedding solution and a novel conditional GAN operating on latent space.

Background

GANs

Generative Adversarial Networks (GANs) [2] are a dual neural network architecture, which aims to learn a dataset's probability distribution, thus has the ability to generate new samples of data.

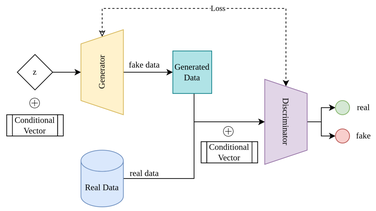

Figure 1. Architecture diagram of a Conditional GAN

Autoencoders

Autoencoder [3] is a neural network architecture, consisting of encoder (down-sampling), bottleneck (low-dimensional space) and decoder (up-sampling). If after training, the output is the same as the input, the network has learned a latent representation of the data in the bottleneck.

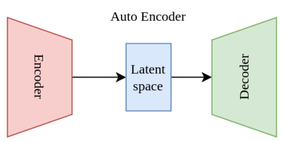

Figure 2. Architecture diagram of an autoencoder

Motivation and Problem Outline

- **Privacy is a key challenge** for sharing data in industry due to strict privacy regulations.
- **Synthetic tabular data is an emerging solution** to enable scientific discoveries while respecting data privacy.
- Encoding tabular columns leads to very high-dimensional data.
- High-dimensional data leads to **a lot of computational overhead**.

Figure 3. Illustration of the dimensionality explosion problem

Research Question

Can we improve the efficiency of high-dimensional tabular data synthesis via latent embeddings?

Methodology

- Introducing **latent embeddings** via autoencoder as dimensionality reduction
- **Novel Conditional GAN operating on latent space**.

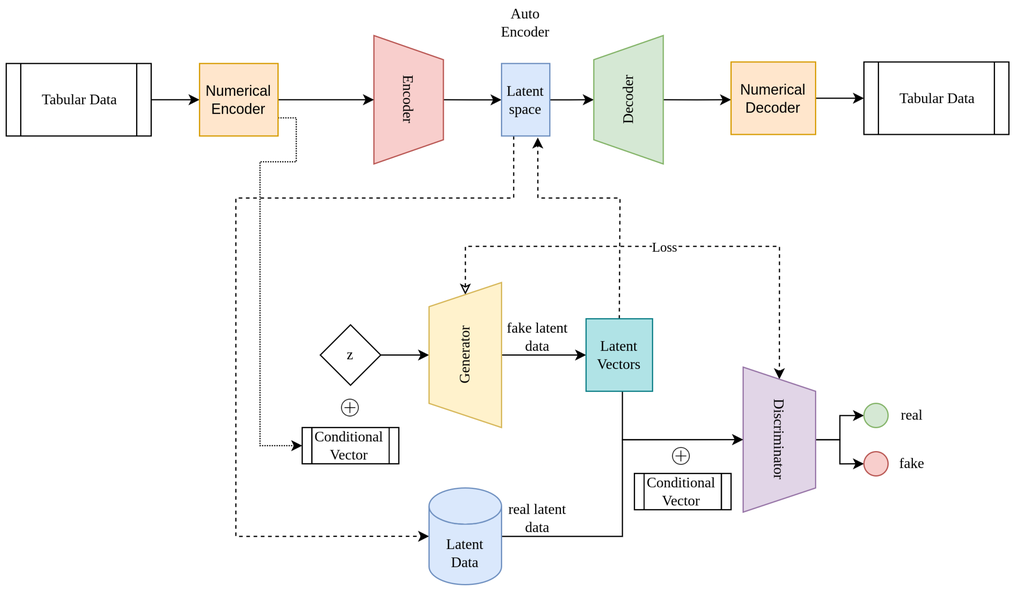

Figure 4. LCT-GAN Architecture overview

Evaluation

- **Statistical similarity**
  - Wasserstein Distance
  - Jenson-Shannon Divergence
  - Difference in pair-wise correlation
- **Machine Learning based analysis**
  - Accuracy, F1 score, AUC
  - Linear Regression, Decision Trees, Random Forest MLP
- Popular datasets: Adult, Credit, Covertype, Loan

Results

In high-dimensional datasets (such as Covertype), we achieved up-to **30% improvement in specific evaluation metrics**, compared to the current state-of-the-art, alongside **up-to 200x faster training per epoch** and **up-to 5x lower memory footprint**.

Approach	Training Time	Accuracy Diff	AUC	F1-Score	Avg. WD	Avg. JSD	Diff Corr	s/epoch
CTAB-GAN	15 minutes	36.83585	0.39111	0.43352	0.02652	0.03939	5.18157	606.00
	30 minutes	37.77141	0.39083	0.42852	0.02393	0.03989	4.93570	612.76
	60 minutes	37.44809	0.38852	0.42228	0.02506	0.04036	4.86093	608.14
LCT-GAN	15 minutes	**29.15499**	**0.24819**	**0.29968**	0.08067	0.05759	5.38522	**2.93**
	30 minutes	**27.52151**	**0.23035**	**0.32369**	0.05780	0.04830	**4.81472**	**2.95**
	60 minutes	**24.45717**	**0.20414**	**0.27737**	0.05125	0.04768	**3.77253**	**2.93**

Figure 5. Autoencoder bottleneck size - 64, trained for 1000 epochs (30 minutes). Dataset - Covertype.

Takeaways

We successfully showed that

- We can **embed tabular data in latent space**
- Conditional **GANs are able to learn complex relationships in latent space** in the context of tabular data
- **Efficiency can be improved** by using LCT-GAN
- Latent embeddings hold a lot of potential and can be further optimized.

References

[1] Z. Zhao, A. Kunar, R. Birke, and L. Y. Chen, "Ctab-gan: Effective table data synthesizing," pp. 97–112, 17–19 Nov 2021. [Online]. Available: https://proceedings.mlr.press/v157/zhao21a.html

[2] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," *CoRR*, vol. abs/1704.00028, 2017. [Online]. Available: http://arxiv.org/abs/1704.00028

[3] G. E. Hinton and R. S. Zemel, "Autoencoders, minimum description length and helmholtz free energy," in *Proceedings of the 6th International Conference on Neural Information Processing Systems*, ser. NIPS'93. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993, p. 3–10.