Improvements to CTGAN for Tabular Data Synthesis

Jimmy Woo
MScAC
University of Toronto
jimmywoo@cs.toronto.edu

Sumant Bagri
MScAC
University of Toronto
sbagri@cs.toronto.edu

Vignesh Edithal
MScAC
University of Toronto
edithal@cs.toronto.edu

Abstract

Synthesizing tabular data involves learning distributions for both categorical and continuous variables. Training datasets for such models may include under-represented categorical features as well as multi-modal continuous variables and devising a singular learning procedure for such a problem is a non-trivial task. A particular generative model, CTGAN, has been shown to outperform several Bayesian network baselines by using a conditional generative adversarial network. In this project, we propose extensions to the original implementation of CTGAN. We broadly classify our work into three research directions: 1) Reducing sparsity in input data representations, 2) Adding constraints to data generation and 3) Learning better estimates of column distributions. We will evaluate the performance of our extensions using statistical properties of the synthetically generated data as well as machine learning efficacy as described in the original work.

1 Introduction

Generative modeling in machine learning focuses on learning domain representations with the objective of synthetically generating data that closely resembles the reality. The main idea behind generative modeling is to assume that real world observations come from a distribution $x \sim p(x)$. At the same time, the generative model can be viewed as an estimating distribution $q_\theta(x)$ described by a set of parameters θ. Then, we can formulate the objective of generative modeling as trying to ensure that $q_\theta(x)$ resembles $p(x)$ as closely as possible (using some statistical divergence: $D(p||q) : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$). Data synthesis has been one of the most actively researched domains in machine learning in the recent past resulting in the creation of various state-of-the-art techniques such as GANs (Generative-Adversarial Netwokrs), VAE (Variational AutoEncoders) and LDMs (Latent Diffusion Models). Each of these techniques have been used to further create domain-efficient models for image-synthesis, text-synthesis as well as tabular data synthesis.

Tabular data synthesis involves generating a sequence of fake, columnar data that is statistically similar to the original table. In this project, we explore one such technique for tabular data synthesis called CTGAN. We will work on extending the original implementation by Xu u. a. (2019) such that model is able to learn more precise representations of the columnar data.

2 Related Works

Xu und Veeramachaneni (2018) propose Tabular GAN (TGAN) which represents continuous variables and discrete variables using Gaussian Mixture Model (GMM) and One Hot Encoding (OHE) using soft max, respectively. To the GAN Generator loss, they add KL divergence between the original and learned distribution of the variable representation. Their model is evaluated based on ML performance and a statistical criteria such as Nearest Neighbor Mutual Information (NNMI). Three data sets from the UCI repository namely CoverType, KDD 99 and Census Income are used for evaluation. Xu u. a. (2019) propose a conditional generator (which they term as CTGAN) and training by sampling technique to overcome the issue of class imbalance. The former uses a conditional vector mask representing a given discrete variable and the latter is used to sample an appropriate data item from the training data. Zhao u. a. (2021) (CTAB-GAN) introduces a classifier to CTGAN architecture to account for semantic integrity with respect to the conditional vector.

Kingma und Welling (2014) introduce Variation Auto Encoder (VAE) which make it possible to generate synthetic data using the latent space shared by the encoder decoder architecture. This is achieved by regularizing the learned latent space in order to allow the decoder to generalize. Wu und Goodman (2018) use a product-of-experts model and sub-sampling training paradigm for ELBO terms to learn a joint distribution across modalities which makes VAE robust to missing data and missing modalities. A VAE-GAN model was proposed by Larsen u. a. (2015) which combines high quality generative models like GAN with methods that produce an encoded representation. They share the parameters of the VAE Decoder and GAN Generator and then use the GAN Discriminator as a basis of VAE reconstruction error.

3 Proposed Methods

In this project, we propose a study of three potential avenues for extensions to CTGAN Xu u. a. (2019) that may improve the tabular data synthesis process. All extensions will be evaluated based on statistical properties of the simulated data as well as machine learning efficacy as detailed in Xu u. a. (2019).

Compact representation of input data CTGAN relies on one-hot encoded vectors of discrete variables, and one-hot representation of continuous variables from mode-specific normalization Xu u. a. (2019), resulting in a sparse input representation to the generator. It has been shown that the quality of the generated synthetic data is sensitive to column permutations due to this sparsity, and that an latent vector representation generated from an autoencoder can mitigate this issue Zhu u. a. (2022). We aim to explore various techniques that can be applied to generating this latent representation such as entity embeddings for categorical variables Guo und Berkhahn (2016), denoising autoencoders Vincent u. a. (2008), variational autoencoders Kingma und Welling (2014), VIMEYoon u. a. (2020), etc.

Constrained data generation There are no explicit constraints on the allowed values of numerical variables for the synthetically generated data from CTGAN. Currently, invalid synthetic data is simply rejected after generation. By adding explicit constraints, the generative model may learn a distribution that resembles the training dataset more closely. We aim to explore adding constraints in two different ways: range-constraints (i.e. allowed range of values) and feature-constraints (i.e. imposing relationships between features). One way to impose these constraints is by adding constraint terms in the loss functionNobari u. a. (2021).

Estimation of column distributions CTGAN uses variational gaussian mixture models (VGM) in the mode-specific normalization process to estimate each continuous feature using multiple gaussians. Inspired by Larsen u. a. (2015), we aim to explore if variational autoencoders can replace the VGMs and be used in conjunction with the remaining pipline of CTGAN.

References

[Guo und Berkhahn 2016] GUO, Cheng ; BERKHAHN, Felix: Entity Embeddings of Categorical Variables. In: *CoRR* abs/1604.06737 (2016). – URL `http://arxiv.org/abs/1604.06737`

[Kingma und Welling 2014] KINGMA, Diederik P. ; WELLING, Max: Auto-Encoding Variational Bayes. In: BENGIO, Yoshua (Hrsg.) ; LECUN, Yann (Hrsg.): *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, URL `http://arxiv.org/abs/1312.6114`, 2014

[Larsen u. a. 2015] LARSEN, Anders Boesen L. ; SØNDERBY, Søren K. ; WINTHER, Ole: Autoencoding beyond pixels using a learned similarity metric. In: *CoRR* abs/1512.09300 (2015). – URL `http://arxiv.org/abs/1512.09300`

[Nobari u. a. 2021] NOBARI, Amin H. ; CHEN, Wei ; AHMED, Faez: *Range-GAN: Range-Constrained Generative Adversarial Network for Conditioned Design Synthesis*. 2021. – URL `https://arxiv.org/abs/2103.06230`

[Vincent u. a. 2008] VINCENT, P. ; LAROCHELLE, H. ; BENGIO, Y. ; MANZAGOL, P.-A.: Extracting and composing robust features with denoising autoencoders. In: *International Conference on Machine Learning proceedings*. 2008

[Wu und Goodman 2018] WU, Mike ; GOODMAN, Noah D.: Multimodal Generative Models for Scalable Weakly-Supervised Learning. In: *CoRR* abs/1802.05335 (2018). – URL `http://arxiv.org/abs/1802.05335`

[Xu u. a. 2019] XU, Lei ; SKOULARIDOU, Maria ; CUESTA-INFANTE, Alfredo ; VEERAMACHANENI, Kalyan: Modeling Tabular data using Conditional GAN. In: WALLACH, H. (Hrsg.) ; LAROCHELLE, H. (Hrsg.) ; BEYGELZIMER, A. (Hrsg.) ; ALCHÉ-BUC, F. d'(Hrsg.) ; FOX, E. (Hrsg.) ; GARNETT, R. (Hrsg.): *Advances in Neural Information Processing Systems* Bd. 32, Curran Associates, Inc., 2019. – URL `https://proceedings.neurips.cc/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf`

[Xu und Veeramachaneni 2018] XU, Lei ; VEERAMACHANENI, Kalyan: Synthesizing Tabular Data using Generative Adversarial Networks. In: *CoRR* abs/1811.11264 (2018). – URL `http://arxiv.org/abs/1811.11264`

[Yoon u. a. 2020] YOON, Jinsung ; ZHANG, Yao ; JORDON, James ; SCHAAR, Mihaela van der: VIME: Extending the Success of Self- and Semi-supervised Learning to Tabular Domain. In: LAROCHELLE, H. (Hrsg.) ; RANZATO, M. (Hrsg.) ; HADSELL, R. (Hrsg.) ; BALCAN, M.F. (Hrsg.) ; LIN, H. (Hrsg.): *Advances in Neural Information Processing Systems* Bd. 33, Curran Associates, Inc., 2020, S. 11033–11043. – URL `https://proceedings.neurips.cc/paper/2020/file/7d97667a3e056acab9aaf653807b4a03-Paper.pdf`

[Zhao u. a. 2021] ZHAO, Zilong ; KUNAR, Aditya ; SCHEER, Hiek V. der ; BIRKE, Robert ; CHEN, Lydia Y.: CTAB-GAN: Effective Table Data Synthesizing. In: *CoRR* abs/2102.08369 (2021). – URL `https://arxiv.org/abs/2102.08369`

[Zhu u. a. 2022] ZHU, Yujin ; ZHAO, Zilong ; BIRKE, Robert ; CHEN, Lydia Y.: *Permutation-Invariant Tabular Data Synthesis*. 2022. – URL `https://arxiv.org/abs/2211.09286`