High-VAE: High-Cardinality & Heterogenous Tabular Variational Autoencoder

Lee Carlin

May 8, 2022

Abstract

Modern tabular datasets frequently include high cardinality categorical and heterogeneous numerical variables that pose significant challenges for generative models. Such datasets are common in health services, financial technology services, cyber security, e-commerce and advertising. Learning the joint-distribution of these type of datasets and providing an accurate synthetic data generating process is often required due to privacy concerns. However, current generative models fail to model the intricacies of such data. In this work, we are interested in developing a general variational autoencoder model to accurately fit imbalanced high-cardinality and heterogenous tabular data. To achieve that we propose the following: (a) entity embed the categorical variables using the generative model's loss, (b) handling imbalance variables by adapting the network's loss, and (c) benchmarking our model against state-of-the-art models using suitable non-trivial datasets.

Project Page

1 Introduction

Real-life tabular datasets, particularly the ones often generated or characterized by humans' data, are very heterogenous. These include electronic medical records (EMR), medical charges, personal credit default, e-commerce, and social networks' datasets. Handling and modeling these datasets is challenging because of the relatively large number of mixed categorical and numerical variables. This is in contrast to homogenous-type datasets such as of audio and images, where the data belongs to a single type with very similar statistical properties. Attributes in mixed-type data have widely different statistical properties and often exhibit complex correlation structures with no trivial way to exploit without losing on the quality of the learned model (Xu et al., 2019). The cardinality of the categorical attributes is often high: the number of unique values can be extremely large. In addition, continuous attributes often follow a non-gaussian distribution. Consequently, the joint distribution between variables can be highly complex with non-trivial relationships. Finally, heterogenous tabular datasets are often imbalanced and sparse, adding an additional layer of complexity.

Modern synthetic data generation is usually accomplished by generative models. Particularly, in this proposal I am interested in generative models for heterogenous tabular data. The goal in generative modeling is to derive the joint probability distribution of the observed data. Once the joint probability is learnt, generative models can be used to impute missing values and generate new data instances. However, existing state-of-the-art generative models fail to properly address the challenges introduced by heterogenous tabular data Nazabal et al. (2020). Both classical generative models such as Gaussian Mixture Models (GMMs) and Hidden Markov Chains (HMMs) in addition to Deep Generative Models (DGMs) such as Generative Adversarial Nets (GANs) and Variational Autoencoders (VAEs) are enfeebled by the complex distributions and interactions intrinsic to mixed type data. However, DNNs usually fair better in learning the non-linear relationships between variables over classical methods (Harshvardhan et al., 2020) .

In general, the use of DGMs have many advantages over classical methods. Deep neural networks utilize a constrained number of parameters relative to the amount of data itself to efficiently learn the data distribution. They are flexible and extremely scalable. For example, Yoon et al. (2018) is a missing data imputation method based on GANs which implicitly learns the joint distribution of the data and imputes the missing values using the generator, while the discriminator tries to guess which components were actually missing.

In this research proposal, we focus on variational autoencoders (Kingma and Welling, 2013), a popular and widely used deep generating probabilistic models for homogenous fixed type data. This deep neural network architecture has many advantages in estimating explicitly the joint probability distribution of such data; VAEs offer flexibility of design and automated feature generation, they support parallel and distributed learning by design, and are very scalable. Most importantly, VAE's network's architecture offer the ability to explicitly learn complex relationships and interactions.

However, VAEs are typically applied when the data is homogeneous and complete; the input variables belong to the same type with similar statistical properties such as images, natural language and audio. Yet, many large scale real world datasets contain mixed data types with various statistical properties, which vanilla VAEs fail to model correctly. VAEs are usually trained by maximizing the lower bounds of the likelihood data and these bounds assume that the data is complete and fully observed.

Improving and adapting tabular VAEs to work with high-cardinality heterogenous data can have a drastic effect on their synthetic data generation abilities. My research proposal concentrates on exactly that: how current tabular VAEs are lacking in their abilities to learn the generative model of such data, and what can we do to improve on these deficiencies by proposing a better VAE architecture to handle high cardinality categorical and heterogenous variables. To address these challenges we seek a powerful yet simple variational autoencoder framework; one that can efficiently handle heterogenous mixed variable types and imbalanced high cardinality categorical variables for fitting tabular data for data generation and synthesis.

2 Plan

Goal

Develop a competitive general variational autoencoder framework for high-cardinality and heterogenous tabular data using entity embeddings and imbalance data correction.

Contribution

1. Use entity embedding to embed categorical variables, and use the ELBO to determine those embeddings.
2. Provide a methodology to counter imbalance data during training
3. Provide a benchmarking framework using non-trivial datasets with various suitable metrics for model evaluation.

Evaluation

1. Evaluate synthetic generation using simulated data with predefined joint-distribution and correlation structure.
2. Evaluate the ML utility using various ML tasks.

Development

- Theory
  - Defining the entity embeddings
  - handling imbalance data
- Network's Architecture
  - Mixed type handler
  - Entity embedding layers
  - Suitable loss function

- Benchmarking
    - Collect datasets
    - Generate synthetic data using competitive models
    - Decide on evaluation metrics and test them

3 High-Level Roadmap

Table 1: Roadmap

Step	Tasks	Due date
MVP		

4 Competing Approaches

	Ref.	Architecture	Use-cases	Repo	Datasets
VAEs					
HI-VAE	Nazabal et al. (2020)	Hierarchical Decoder	Imputation	github	1-5
VAEM	Ma et al. (2020)	2-stage: ind. & dep. VAEs	Imputation	github	9-13
VSAE	Gong et al. (2021)	mask & data gen. models	Imputation	n/a	14
RVAE	Akrami et al. (2020, 2022)	Beta-divergence	Outlier robust	github	1,3,4-8
GANs					
medGAN	Choi et al. (2017)		Medical data		
table-GAN	Park et al. (2018)				
TGAN	Xu and Veeramachaneni (2018)				
CTGAN	Xu et al. (2019)				
CTAB-GAN	Zhao et al. (2021)				

5 Datasets:

	Name	Ref.	Summary
1	Adult	UCI-MLR	
2	Breast	UCI-MLR	
3	Credit Default	UCI-MLR	
4	Spam	UCI-MLR	
5	Wine	UCI-MLR	
6	KDDCup 99	KDD	
7	NSL-KDD	CiC	
8	UNSW-NB15	UNSW	
9	Bank	UCI-MLR	
10	Boston	UCI-MLR	
11	Avocado		
12	Energy	UCI-MLR	
13	MIMIC	MIT	de-identified health-related data (~40k)
14	Heart	UCI-MLR	

References

Akrami, H., Aydore, S., Leahy, R. M., and Joshi, A. A. (2020). Robust variational autoencoder for tabular data with beta divergence. *arXiv preprint arXiv:2006.08204*.

Akrami, H., Joshi, A. A., Li, J., Aydöre, S., and Leahy, R. M. (2022). A robust variational autoencoder using beta divergence. *Knowledge-Based Systems*, 238:107886.

Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. F., and Sun, J. (2017). Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*, pages 286–305. PMLR.

Gong, Y., Hajimirsadeghi, H., He, J., Durand, T., and Mori, G. (2021). Variational selective autoencoder: Learning from partially-observed heterogeneous data. In *International Conference on Artificial Intelligence and Statistics*, pages 2377–2385. PMLR.

Harshvardhan, G., Gourisaria, M. K., Pandey, M., and Rautaray, S. S. (2020). A comprehensive survey and analysis of generative models in machine learning. *Computer Science Review*, 38:100285.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Ma, C., Tschiatschek, S., Turner, R., Hernández-Lobato, J. M., and Zhang, C. (2020). Vaem: a deep generative model for heterogeneous mixed type data. *Advances in Neural Information Processing Systems*, 33:11237–11247.

Nazabal, A., Olmos, P. M., Ghahramani, Z., and Valera, I. (2020). Handling incomplete heterogeneous data using vaes. *Pattern Recognition*, 107:107501.

Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., and Kim, Y. (2018). Data synthesis based on generative adversarial networks. *arXiv preprint arXiv:1806.03384*.

Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. (2019). Modeling tabular data using conditional gan. *Advances in Neural Information Processing Systems*, 32.

Xu, L. and Veeramachaneni, K. (2018). Synthesizing tabular data using generative adversarial networks. *arXiv preprint arXiv:1811.11264*.

Yoon, J., Jordon, J., and Schaar, M. (2018). Gain: Missing data imputation using generative adversarial nets. In *International conference on machine learning*, pages 5689–5698. PMLR.

Zhao, Z., Kunar, A., Birke, R., and Chen, L. Y. (2021). Ctab-gan: Effective table data synthesizing. In *Asian Conference on Machine Learning*, pages 97–112. PMLR.