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

Lee Carlin

May 3, 2022

Abstract

Modern tabular data-sets frequently include high cardinality categorical and heterogeneous numerical variables. Such datasets are common in health services (e.g. health and medical records), financial technology services (e.g. personal credit default rates), cyber security, e-commerce and advertising (e.g. user profiling data). High cardinality categorical and heterogenous variables pose significant challenges and difficulties to the analysis and interpretation of data by statistical methodologies.

MISSING.Generating synthetic tabular data characterized by high-cardinality and heterogenous variables . More specifically, I am interested to fit a latent variable model using the learned joint-distribution of such data. To achieve this we explore the highly flexible deep neural architectures of deep generating models. This proposal is composed of the following trajectories: (a) learning the joint-distribution of imbalanced, high-cardinality heterogenous tabular data in order to impute missing values and, (b) to provide a privacy-preserving generative model for data synthesis.

Project Page

1 Plan

Goal

Develop a competitive general variational autoencoder framework for high-cardinality and heterogenous tabular data.

Contribution

1. Use entity embedding to embed categorical variables, and use the ELBO to determine those embeddings.
2. Counter imbalance data (conditioning/ ensemble of consensus and long-tail to rebalance?/...)

Evaluation

1. Evaluate against simulated data (using a predefined joint-distribution)
2. Evaluate ML utility

2 Roadmap

Table 1: Roadmap

Step	Tasks	Due date
MVP		

3 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)				

4 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	

5 Something

dfd

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.

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*.

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.