**Literature review**

**CTGAN:** Conditional Tabular GAN (Lei Xu et al., 2019) <https://proceedings.neurips.cc/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf> According to authors, existing methods of tabular data generation suffer from imbalanced discrete columns and mode collapse for continuous variables. Solution is conditional tabular GAN, where conditions are mask-vectors, which go along with input vectors. Input vectors transformed in such way - continuous variables are represented with Gaussian Mixture Model as a modes vector and alpha value, discrete – with One Hot Encoder. We can use conditional vectors in our goal of creating samples from specified category.

**TVAE**: Table Variational Autoencoder (Lei Xu et al., 2019) Also, authors tested tabular Autoencoder, which outperforms CTGAN on simulated and real datasets. “To the best of our knowledge, there are no alternative VAE like models that outperform TVAE and have public source code” (Akim Kotelnikov, 2023). Preprocessing the same as in CTGAN. But comparing to CTGAN, TVAE less secure as it works directly with real users data and the possibility of data leakage is high, when GAN training process is divided, and generator does not contact with real data.

Изображение выглядит как текст, снимок экрана, Шрифт, диаграмма

Автоматически созданное описание

*Figure 1. CTGAN architecture*

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

*Figure 2. Tabular Variational AutoEncoder model*

**CTAB-GAN+** (Zilong Zhao, 2022) <https://arxiv.org/abs/2204.00401> Some improvements of CTGAN: i) adding downstream losses to conditional GANs for higher utility synthetic data ii) novel encoders for mixed data – not just Gaussian Mixture model – MinMax for Gaussian, Long tails handler.

**Table GAN** (Aoting Hu, 2018) <https://arxiv.org/pdf/2107.13190.pdf>

Изображение выглядит как текст, диаграмма, снимок экрана, линия

Автоматически созданное описание

Изображение выглядит как текст, линия, снимок экрана, Шрифт

Автоматически созданное описание

Here authors use convolution on table, not linear operations. Also, they use classifier along with discriminator to catch wrongly generated rows. Preprocessing contains just scaling to [-1, 1] by MinMaxScaler. Not so effective as previous models:

Изображение выглядит как текст, снимок экрана, Шрифт, число

Автоматически созданное описание

Performance of models. L – likelihood probability of distributions, clf – F1 metric for classification task, reg – R2 for regression. Table from (Lei Xu et al., 2019).

**corGAN** [**https://paperswithcode.com/paper/cor-gan-correlation-capturing-convolutional**](https://paperswithcode.com/paper/cor-gan-correlation-capturing-convolutional)(Amirsina Torf, 2020)state-of-the-art on paperswithcode! <https://github.com/astorfi/cor-gan> They use 1D convolutional autoencoders, and typical GAN adversarial blocks. Where elements are discrete – they transform them from continuous to discrete. No specific preprocessing described in article. Also, in this article author introduces some security enhancement techniques.

**Sparse Autoencoder** <https://web.stanford.edu/class/cs294a/sparseAutoencoder.pdf> (Alireza Makhzani et al, 2013) – we train autoencoder in such way that only few elements hidden layers are activated by particular sample. I do not find it in articles, but I think, that we can use that particular activations/latent-vector elements to produce samples from particular category. <https://arxiv.org/abs/1312.5663>

**Attentive Transformers**: Attentive Interpretable Tabular Learning (Arik S.O. et al, 2019) <https://arxiv.org/pdf/1908.07442.pdf> Good for classifications and regression, but generation is not highlighted in article.

**Diffusion:** TabDDPM: Diffusion model for tabular generation (Akim Kotelnikov, 2023). <https://proceedings.mlr.press/v202/kotelnikov23a/kotelnikov23a.pdf> Authors use diffusion process, like in text-to-image models with consequent denoising. Don’t know how to use conditions for modeling in this diffusion model yet, but method looks good and this is state-of-the-art ]: <https://github.com/yandex-research/tab-ddpm>

**SMOTE**: SMOTE: Synthetic Minority Over-sampling Technique (N. V. Chawla, 2011) <https://arxiv.org/pdf/1106.1813.pdf> - generate by appending a random k-nearest neighbor of random sample with random coefficient. This approach do not use networks at all. So, we can choose specific samples which satisfy condition and oversample them by random neighbor appending. But the method is about increasing of size of minor class, not about generating new samples from distribution. So samples can’t be so diverse as in neural-network generative models.

**Conditional autoencoder pre-training and optimization algorithms for personalized care of hemophiliac patients (**Cédric Buche, 2023) <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9905812/> They predict value from past values, not rows in table, but use condition – dose of the medicine. The architecture is hard and not similar to previous articles, but contains autoencoders.

**MASON** - MultiAgent Simulation Environment <https://www.researchgate.net/publication/220165043_MASON_A_Multiagent_Simulation_Environment> used for modeling of transactions, maybe will be useful in recommendation simulation

**CODE**

GITHUB with implemented GANs for private data generation <https://github.com/BorealisAI/private-data-generation>

TabDDPM <https://github.com/yandex-research/tab-ddpm>

CTGAN <https://github.com/sdv-dev/CTGAN>

**DATASETS**

TABULAR DATASETS (NO RECCOMENDATIONS):

**Adult Census** : The dataset comprises of census attributes like age, gender, native country etc and the goal is to predict whether a person earns more than $ 50k a year or not. <https://archive.ics.uci.edu/ml/datasets/adult>

**NHANES Diabetes** : National Health and Nutrition Examination Survey (NHANES) questionnaire is used to predict the onset of type II diabetes. <https://github.com/semerj/NHANES-diabetes/tree/master/data>

**Give Me Some Credit** : Historical data are provided on 250,000 borrowers and task is to help in credit scoring, by predicting the probability that somebody will experience financial distress in the next two years. <https://www.kaggle.com/c/GiveMeSomeCredit/data>

**Home Credit Default Risk** : Home Credit makes use of a variety of alternative data including telco and transactional information along with the client's past financial record to predict their clients' repayment abilities. <https://www.kaggle.com/c/home-credit-default-risk/data>

**Adult Categorical** : This dataset is the same as the Adult Census dataset, but the feature values for continuous attributes are put in buckets. We evaluate Private-PGM's performance on this dataset. <https://github.com/ryan112358/private-pgm/tree/master/data>

TABULAR DATASETS (RECOMMENDATIONS):

**Restaurants&Consumer** <https://archive.ics.uci.edu/dataset/232/restaurant+consumer+data> user profiles and ratings for restaurants. 113 users

Datasets for recommendation systems (пока не нашел с юзер инфой): <https://cseweb.ucsd.edu/~jmcauley/datasets.html>

Есть еще идеи взять банковские транзакции с пользовательскими данными, где отображается ID юзера и ID айтема в транзакциях

Еще есть идея найди данные соцсетей, где можно взять отдельно нужные данные пользователей, и их взаимодействия с постами. Но это вряд ли законно, а найти в общем доступе датасет не получается.

*other articles:*

Banksim: A bank payments simulator for fraud detection research Inproceedings

1. Wasserstain GAN – CTGAN trained GAN with Wasserstein distance https://arxiv.org/abs/1701.07875

2. PacGAC - allows cope with mode collapse - when data is generated with little diversity https://arxiv.org/abs/1712.04086

Mode collapse happens when the generator can only produce a single type of output or a small set of outputs. This may happen due to problems in training, such as the generator finding a type of data that is easily able to fool the discriminator and thus keeps generating that one type

1. Autoencoder on Gaussian Mixture model https://www.researchgate.net/publication/348947965\_Collaborative\_Filtering\_Based\_on\_a\_Variational\_Gaussian\_Mixture\_Model/download?\_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6Il9kaXJlY3QiLCJwYWdlIjoiX2RpcmVjdCJ9fQ

2. Variational AutoEncdoer https://jaan.io/what-is-variational-autoencoder-vae-tutorial/

3. Kullback-Leibler divergence <https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence>

1. Jensenn-Hannon divergence https://towardsdatascience.com/how-to-understand-and-use-jensen-shannon-divergence-b10e11b03fd6?source=read\_next\_recirc-----f3ddc8dff254----0---------------------258b436c\_ea89\_4e3e\_8cc1\_f2729dfb1ae2-------