ARIA

APPLIED RESEARCH IN ACTION

TabPFGen – Tabular Data Generation with TabPFN

TabPFGen creates an energy-based generative model for tabular data using the discriminative, in-context learning capabilities of pre-trained TabPFN transformer

Apoorv Dankar

Gennady Pekhimenko
ACADEMIC SUPERVISOR

Guangwei Yu
INDUSTRY SUPERVISOR

Table 1: Average AUC over OpenML-CC18 test sets, with error bars over 3 runs. Top 4 rows: synthetic data as augmentation. Bottom 4 rows: synthetic data as replacement.

Model	Original	SMOTE	CTGAN	TVAE	NF	RTVAE	TabDDPM	TabPFGen
XGB	$0.924 \pm 3e{-4}$	$0.926 \pm 3e{-4}$	$0.912 \pm 2e{-4}$	$0.914 \pm 7e{-4}$	$0.912 \pm 4e{-4}$	$0.917 \pm 3e{-4}$	$0.927 \pm 3e{-4}$	$\mathbf{0.936} \pm 4e{-4}$
RF	$0.906 \pm 3e{-4}$	$0.906 \pm 2e{-3}$	$0.898 \pm 1e{-3}$	$0.904 \pm 1e{-3}$	$0.894 \pm 3e{-4}$	$0.907 \pm 2e{-3}$	$0.911 \pm 7e{-4}$	$\mathbf{0.918} \pm 3e{-4}$
LR	$0.920 \pm 7e{-4}$	$0.914 \pm 3e{-3}$	$0.904 \pm 3e{-3}$	$0.909 \pm 6e{-3}$	$0.901 \pm 9e{-4}$	$0.906 \pm 8e{-3}$	$0.885 \pm 3e{-4}$	$\mathbf{0.921} \pm 7e{-4}$
TabPFN	$0.934 \pm 2e{-3}$	$0.927 \pm 1e{-3}$	$0.930 \pm 1e{-3}$	$0.931 \pm 1e{-3}$	$0.928 \pm 3e{-4}$	$0.932 \pm 1e{-3}$	$0.929 \pm 5e{-4}$	$\mathbf{0.937} \pm 2e{-4}$
XGB	N/A	$0.907 \pm 4e{-4}$	$0.842 \pm 8e{-4}$	$0.858 \pm 2e{-3}$	$0.700 \pm 6e{-4}$	$0.795 \pm 9e{-4}$	$0.812 \pm 3e{-4}$	$\mathbf{0.913} \pm 6e{-4}$
RF	N/A	$0.894 \pm 1e{-3}$	$0.837 \pm 6e{-4}$	$0.844 \pm 5e{-4}$	$0.676 \pm 2e{-3}$	$0.774 \pm 3e{-4}$	$0.814 \pm 9e{-4}$	$\mathbf{0.897} \pm 7e{-4}$
LR	N/A	$0.893 \pm 2e{-3}$	$0.843 \pm 6e{-4}$	$0.873 \pm 1e{-3}$	$0.722 \pm 3e{-3}$	$0.854 \pm 7e{-4}$	$0.876 \pm 3e{-4}$	$\mathbf{0.916} \pm 1e{-3}$
TabPFN	N/A	$0.920 \pm 8e{-4}$	$0.888 \pm 4e{-4}$	$0.887 \pm 3e{-4}$	$0.705 \pm 2e{-3}$	$0.862 \pm 1e{-3}$	$0.894 \pm 7e{-4}$	$\mathbf{0.925} \pm 3e{-4}$

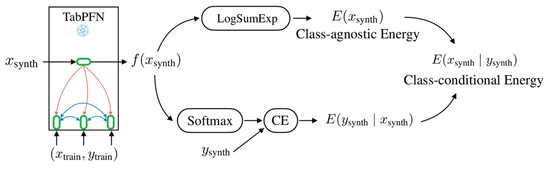
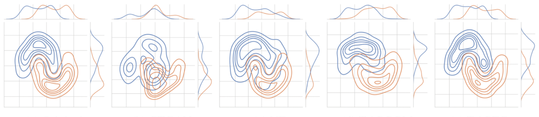


Figure 1: TabPFGen Overview. We backpropagate from the class-conditional energy to x_{synth} for gradient generation. CE denotes cross entropy; blue and red arrows represent attention.



(a) Original (b) CTGAN (c) NF (d) TabDDPM (e) TabPFGen

Figure 2: Contour and marginal density plots of: (a) original two-moons dataset; (b)-(d) synthetic data generated using baseline methods; (e) synthetic data generated by TabPFGen

PROJECT SUMMARY

Advances in deep generative modelling have not translated well to tabular data. We argue that this is caused by a mismatch in structure between popular generative models and discriminative models of tabular data. We thus devise a technique to turn TabPFN -- a highly performant transformer initially designed for in-context discriminative tabular tasks -- into an energy-based generative model, which we dub TabPFGen. This novel framework leverages the pre-trained TabPFN as part of the energy function and does not require any additional training or hyperparameter tuning, thus inheriting TabPFN's in-context learning capability. We can sample from TabPFGen analogously to other energy-based models. We demonstrate strong results on standard generative modelling tasks, including data augmentation, class-balancing, and imputation, unlocking a new frontier of tabular data generation.

layer 6
AI at TD

Computer Science
UNIVERSITY OF TORONTO

Master of Science in Applied Computing