Continuous Diffusion for Mixed-Type Tabular Data

Markus Mueller, Kathrin Gruber, Dennis Fok

Econometric Institute, Erasmus University Rotterdam

NEURAL INFORMATION PROCESSING SYSTEMS

Challenges

The marginal distribution of mixed-type features can vary drastically: any two **continuous features** may be subject to different levels of discretization and bounds (even after applying common pre-processing techniques); and any two **categorical features** may have different associated categories and exhibit different balancing.

Score Matching and Score Interpolation

The denoising **score matching objective** [3] for the **continuous features** is

$$\min_{\boldsymbol{\theta}} \mathbb{E}_t \Big[\lambda(t) \mathbb{E}_{\mathbf{x}_0} \mathbb{E}_{\mathbf{x}_t|\mathbf{x}_0} \big[\| s_{\boldsymbol{\theta}}(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p_{0t}(\mathbf{x}_t|\mathbf{x}_0) \|_2^2 \big] \Big]. \tag{1}$$

Given timestep t and noisy data $\mathbf{x}_t$, the minimizer of the above objective is $\mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_t,t)}[\nabla_{\mathbf{x}_t} \log p_{0t}(\mathbf{x}_t|\mathbf{x}_0)]$, which enables **score interpolation** [2] for the **categorical features**:

$$\mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_t,t)} \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{x}_0) = \frac{\mathbb{E}_{p(\mathbf{x}_0|\mathbf{x}_t,t)}[\mathbf{x}_0] - \mathbf{x}_t}{\sigma_t^2}. \tag{2}$$

We add Gaussian noise to the embeddings $\mathbf{x}_0$ to produce $\mathbf{x}_t$ and train a neural network to predict the ground-truth category via cross-entropy loss. Thus, we plug in $p(\mathbf{x}_0|\mathbf{x}_t, t)$ to Eq. (2) to interpolate the categorical feature scores. The embeddings are trained alongside the rest of the model.

We combine the score matching and the cross-entropy losses into a joint loss function:

$$\mathcal{L}_{\text{joint}}(\boldsymbol{\theta}) = \frac{1}{\alpha K_{\text{cat}} + (1-\alpha) K_{\text{cont}}} \Big[(1-\alpha) \sum_{i=1}^{K_{\text{cont}}} \ell_{\text{cont}}^{(i)}(\boldsymbol{\theta}) + \alpha \sum_{j=1}^{K_{\text{cat}}} \ell_{\text{cat}}^{(j)}(\boldsymbol{\theta}) \Big]. \tag{3}$$

Feature-Specific Adaptive Noise Schedules

We allow each feature (or group of features) to follow separate SDEs (similar to non-uniform diffusion [1]) to explicitly acknowledge feature heterogeneity:

The i-th continuous feature follows a diffusion process given by

$$\mathrm{d}x_{\text{cont}}^{(i)} = f_{\text{cont},i}(x_{\text{cont}}^{(i)}, t)\mathrm{d}t + g_{\text{cont},i}(t)\mathrm{d}w_t^{(i)}, \tag{4}$$

and the evolution of the embedding of the j-th categorical feature is governed by

$$\mathrm{d}\mathbf{x}_{\text{cat}}^{(j)} = \mathbf{f}_{\text{cat},j}(\mathbf{x}_{\text{cat}}^{(j)}, t)\mathrm{d}t + g_{\text{cat},j}(t)\mathrm{d}\mathbf{w}_t^{(j)}, \tag{5}$$

with $\mathbf{x}_{\text{cat}}^{(j)}$ the d-dimensional embedding of $x_{\text{cat}}^{(j)}$ in Euclidean space.

We specify the feature-specific timesteps, $t_{\text{cont},i}(t)$ and $t_{\text{cat},j}(t)$, as a function of the global time t, and set the drift coefficients to zero and the feature-specific diffusion coefficients to $g_{\text{cont},i}(t) = \sqrt{2t_{\text{cont},i}(t)}$ and $g_{\text{cat},j}(t) = \sqrt{2t_{\text{cat},j}(t)}$, respectively.

We use **timewarping** [2] to learn the feature- (or group-) specific noise schedules, by training monotonic piece-wise linear functions F_i to fit the relevant losses. Normalizing and inverting F_i allows us to map from $t \to t_{\text{type}_i,i}$ with $t \sim \mathcal{U}_{[0,1]}$.

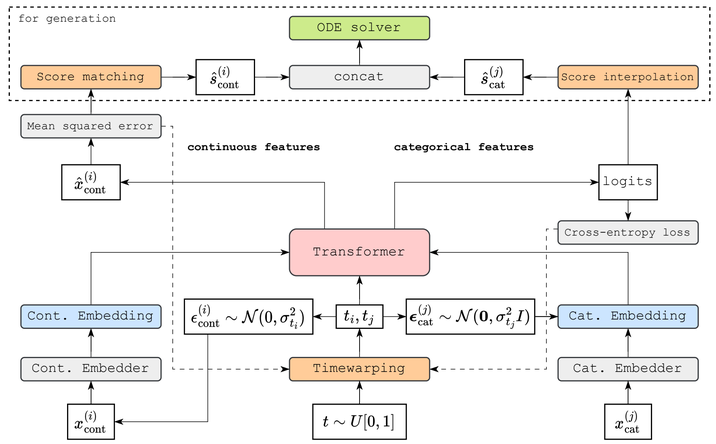
Investigated Noise Schedules

single	single noise schedule common to all features
per type	separate noise schedules per data type
single cont.	single noise schedule for cont. feat., individual noise schedules for cat. feat.
per feature	each feature has its own noise schedule

Abstract

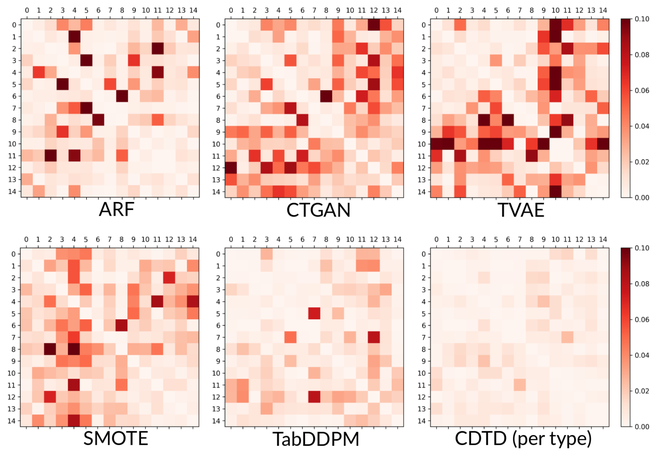
We design a score-based generative model to generate new samples from the joint distribution of mixed-type tabular data. Therefore, we combine **score matching** for continuous features [3] and **score interpolation** for categorical features [2] with **learnable, possibly feature-specific noise schedules**. Our **CDTD** approach emphasizes the increased feature heterogeneity in mixed-type tabular data and captures feature dependencies exceptionally well.

CDTD Model Framework



Absolute Differences in Correlations (Adult Dataset)

We obtain the pairwise correlations between real and synthetic train sets for three different metrics: Pearson correlation coefficient (cont/cont), correlation ratio (cont/cat) and Theil's uncertainty coefficient (cat/cat).



Experimental Results

Machine Learning Efficiency

We compare the test set performance of models trained on the real training data and on the synthetic data for 4 different models: decision tree, random forest, linear model, catboost. We report the absolute difference of the average performances on three benchmark datasets. For the adult and churn datasets, we consider the macro-averaged F1 score; for the nmes dataset, we consider the MSE.

	ARF	CTGAN	TVAE	SMOTE	TabDDPM	CDTD (single)	CDTD (per type)	CDTD (single cont.)	CDTD (per feature)
adult	0.025	0.029	0.025	<u>0.013</u>	0.016	0.010	0.008	**0.007**	0.014
churn	0.091	0.145	0.088	**0.014**	0.390	0.069	<u>0.056</u>	0.064	0.059
nmes	**0.379**	2.387	1.758	0.818	7.371	0.667	0.623	<u>0.587</u>	0.608

bold = best performance; <u>underline</u> = second best performance

Detection Accuracy

We tune and train a catboost model to differentiate between real and fake samples. The performance is evaluated on a test set with equal fake and real proportions.

	ARF	CTGAN	TVAE	SMOTE	TabDDPM	CDTD (single)	CDTD (per type)	CDTD (single cont.)	CDTD (per feature)
adult	0.918	0.988	0.931	0.337	<u>0.600</u>	0.561	0.559	**0.557**	0.583
churn	0.847	0.977	0.915	**0.339**	0.998	0.850	<u>0.832</u>	0.847	0.849
nmes	0.986	0.992	0.990	<u>0.868</u>	0.998	**0.647**	0.653	0.649	0.652

Distance to Closest Record (Privacy)

For each synthetic sample, we determine the minimum Euclidean distance to any real training observation. We report the difference between the average distance of the synthetic samples and the average distance of the real test set samples.

	ARF	CTGAN	TVAE	SMOTE	TabDDPM	CDTD (single)	CDTD (per type)	CDTD (single cont.)	CDTD (per feature)
adult	0.610	1.905	0.438	-0.443	0.061	0.015	0.023	0.032	0.042
churn	0.761	2.260	0.793	-0.123	2.453	0.667	0.599	0.644	0.650
nmes	0.266	0.858	0.007	-0.025	0.911	-0.016	-0.027	-0.022	-0.023

L_2 Distance of Correlation Matrices

	ARF	CTGAN	TVAE	SMOTE	TabDDPM	CDTD (single)	CDTD (per type)	CDTD (single cont.)	CDTD (per feature)
adult	0.585	0.499	0.632	0.503	<u>0.227</u>	0.104	**0.093**	0.098	0.125
churn	0.602	2.678	0.753	**0.283**	4.942	0.475	<u>0.441</u>	0.498	0.514
nmes	0.669	1.390	2.317	<u>0.658</u>	3.305	0.588	0.557	0.544	**0.543**

References

[1] Georgios Batzolis, Jan Stanczuk, Carola-Bibiane Schönlieb, and Christian Etmann. Non-Uniform Diffusion Models. *arXiv preprint arXiv:2207.09786*, 2022.

[2] Sander Dieleman, Laurent Sartran, Arman Roshannai, Nikolay Savinov, Yaroslav Ganin, Pierre H. Richemond, Arnaud Doucet, Robin Strudel, Chris Dyer, Conor Durkan, Curtis Hawthorne, Rémi Leblond, Will Grathwohl, and Jonas Adler. Continuous diffusion for categorical data. *arXiv preprint arXiv:2211.15089*, 2022.

[3] Aapo Hyvärinen. Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

mueller@ese.eur.nl