Meta-features informed WGAN for tabular data

Roman Netrogolov, Irina Deeva
AI Institute, ITMO University Saint-Petersburg, Russia

Source code

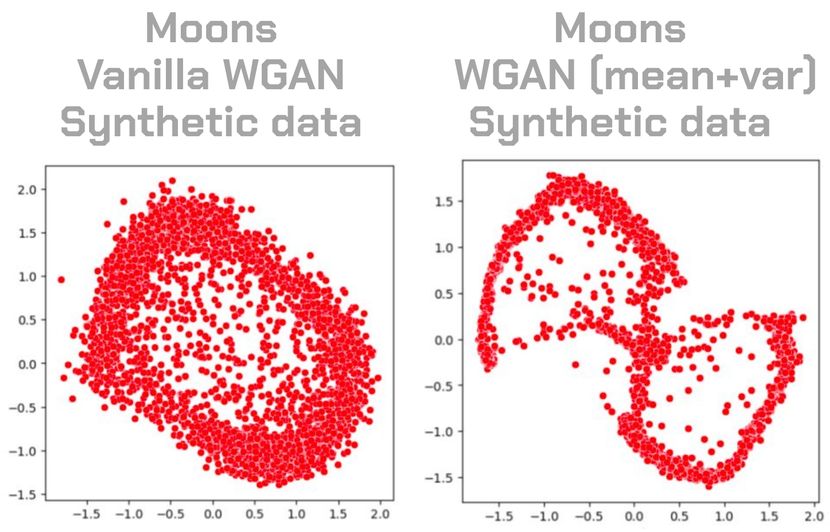
ICDM Workshop 2025

Introduction

Problem: Standard GANs struggle with tabular data generation due to mode collapse, training instability, and poor preservation of complex statistical relationships.

Hypothesis: Can explicit incorporation of high-level statistical characteristics (meta-features) into the generation process improve synthetic tabular data quality?

Contribution: Integrate meta-feature distributions directly into WGAN-GP training using W_1 distance to guide generators toward preserving statistical relationships while maintaining diversity.



Proposed Approach

Meta-features are measurable dataset characteristics capturing structure and complexity: mean, variance, covariance, IQR, eigenvalues, and correlation structures.

Loss function:

$$\mathcal{L}_D = \mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}\left[D(\tilde{x})\right] - \mathbb{E}_{x\sim\mathbb{P}_r}\left[D(x)\right] + \lambda_{gp}\,\mathbb{E}_{\hat{x}\sim\mathbb{P}_{\hat{x}}}\left[\left(\|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1\right)^2\right]$$

$$\mathcal{L}_G = -\mathbb{E}_{\tilde{x}\sim\mathbb{P}_g}\left[D(\tilde{x})\right] + \lambda L_{mfs}(\mathcal{M}_r, \mathcal{M}_g)$$

where λ is the meta-feature loss weight, L_{mfs} measures the W_1 distance between meta-feature distributions M_r (real) and M_g (generated)

Core Innovation: Align meta-feature distributions between real and synthetic data using W_1 distance instead of pointwise matching.

Rationale: Provides geometrically meaningful distribution comparison with stable gradients. The gradient penalty constraints allow alignment of higher-order statistical moments, bridging statistical and geometric aspects.

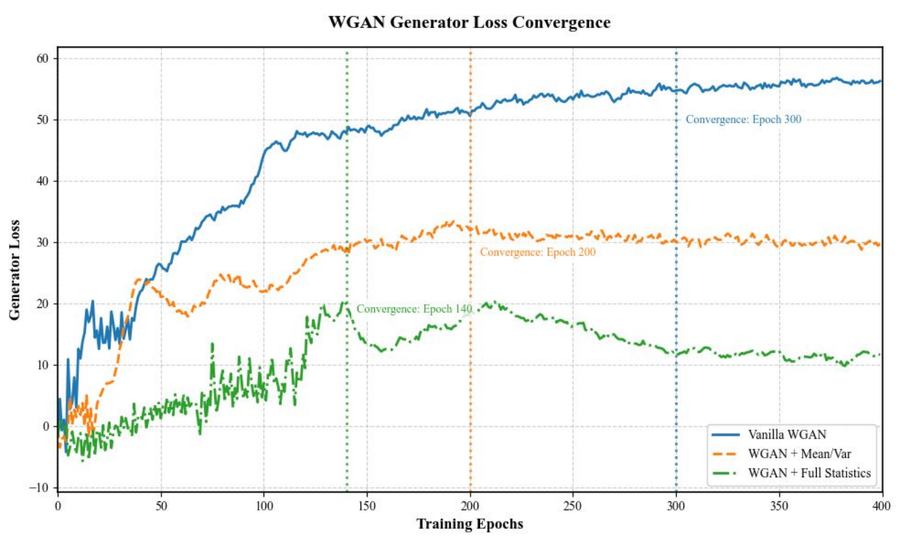
Sample N_{mfs} instances from real and generated data, compute meta-feature vectors φ, minimize W_1 distance between distributions.

Experiments

Performance comparison of synthetic data generators evaluated using XGBoost regressor trained on synthetic data and tested on real data. Corr Dist measures dissimilarity between correlation matrices of real and synthetic data using cosine distance. Mean JS Div quantifies the average distributional difference across all features.

WGAN_mfs (mean,var) uses only mean and variance as meta-features, while WGAN_mfs (full) employs the complete set including mean, variance, covariance, eigenvalues, and IQR.

Training Stability: Even when adversarial loss becomes unstable, meta-feature guidance prevents training collapse and preserves synthetic data quality. Loss trajectories show smooth convergence with the MF component.



Dataset	Model	R2	RMSE	Mean JS div	Corr dist
Regression	Real	0.996	6.33	-	-
	WGAN_mfs (mean,var)	**0.985 ± .008**	**12.07 ± 3.19**	**0.017 ± 0.009**	0.039 ± 005
	WGAN_mfs (full)	0.535 ± 0.299	66.82 ± 22.63	0.117 ± 0.334	0.222 ± 0.046
	CTGAN	0.825 ± 0.047	41.66 ± 5.49	0.029 ± 0.013	**0.005 ± 0.002**
	Vanilla WGAN	0.325 ± 0.039	83.74 ±2.07	0.034 ± 0.019	0.056 ± 0.016
California	Real	0.803	0.479	-	-
	WGAN_mfs (mean,var)	0.249 ± 0.118	0.923 ± 0.069	0.122 ± 0.045	0.056 ± 0.026
	WGAN_mfs (full)	0.389 ± 0.100	0.834 ± 0.068	0.117 ± 0.045	0.014 ± 0.005
	CTGAN	**0.447 ± 0.035**	**0.688 ± 0.013**	**0.052± 0.027**	**0.003 ± 0.001**
	Vanilla WGAN	0.208 ± 0.245	0.942 ± 0.144	0.138 ± 0.038	0.105 ± 0.054
Abalone	Real	0.450	2.369	-	-
	WGAN_mfs (mean,var)	0.185 ± 0.187	2.804 ± 0.317	0.050 ± 0.036	0.003 ± 0.003
	WGAN_mfs (full)	0.157 ± 0.118	2.779 ± 0.161	0.048 ± 0.025	**0.000 ± 0.000**
	CTGAN	**0.263 ± 0.021**	**2.728 ± 0.085**	**0.028 ± 0.011**	0.001 ± 0.000
	Vanilla WGAN	0.003 ± 0.256	3.180 ± 0.508	0.063 ± 0.040	0.006 ± 0.005