SOS: Score-based Oversampling for Tabular Data

BigDyL

3세부

SOS: Score-based Oversampling for Tabular Data

Jayoung Kim[1], Chaejeong Lee[1], Yehjin Shin[1], Sewon Park[2], Minjung Kim[2], Noseong Park[1], Jihoon Cho[2]
Yonsei University[1], Samsung SDS[2]

Motivation

- Class imbalance problems lead to sub-optimal training outcomes.
- Around the class boundary, many samples are placed regardless of their class.
- **It is important to oversample focusing on where classifiers are difficult to classify.**

Contributions:

1. Design **a score-based generative model** for tabular data.
2. Propose **a fine-tuning method**, further enhancing the generation quality.
3. Propose **a style transfer-based oversampling** method to generate samples around the class boundary.

Related Work

Score-based Generative Models (SGMs) [1]

Forward SDE (data → noise)

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\boldsymbol{w}$$

$\mathbf{x}(0)$ ⇄ $\mathbf{x}(T)$

Reverse SDE (noise → data)

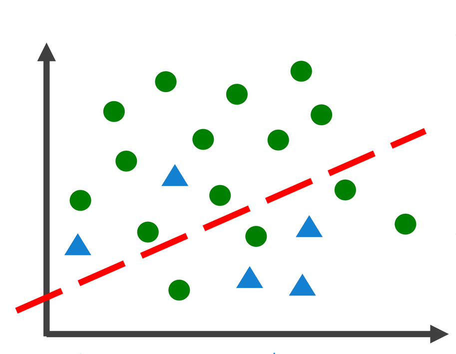
$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x})]dt + g(t)d\bar{\mathbf{w}}$$

1. **SDE-based framework**
   - **Forward SDE** is to add gaussian noises to $\mathbf{x}(0)$.
   - **Reverse SDE** is to remove noises from $\mathbf{x}(T)$.
   - The score network approximates the score function:

$$S_\theta(\mathrm{x}_t, t) \approx \nabla_{\mathrm{x}} \log p_t(\mathrm{x})$$

2. **Denoising score matching loss**
   - Estimate the score of the perturbed data distribution.
   - Collect **the gradient of log transition probability** $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{x}_0)$ during forward SDE.
   - $\theta^* = \arg\min_\theta \mathbb{E}_t \mathbb{E}_{\mathbf{x}_t} \mathbb{E}_{\mathbf{x}_0} \lambda(t) \left[\left\| S_\theta(\mathbf{x}_t, t) - \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t|\mathbf{x}_0) \right\|_2^2 \right]$

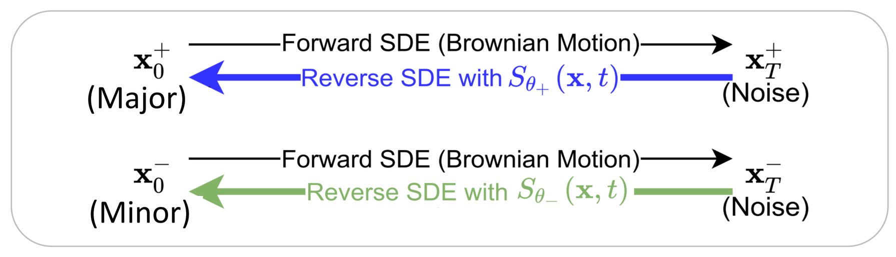
Oversampling



- The samples around the class boundary have **both major and minor characteristics**.
- By generating samples around the class boundary, classifiers can be trained to classify the samples better.

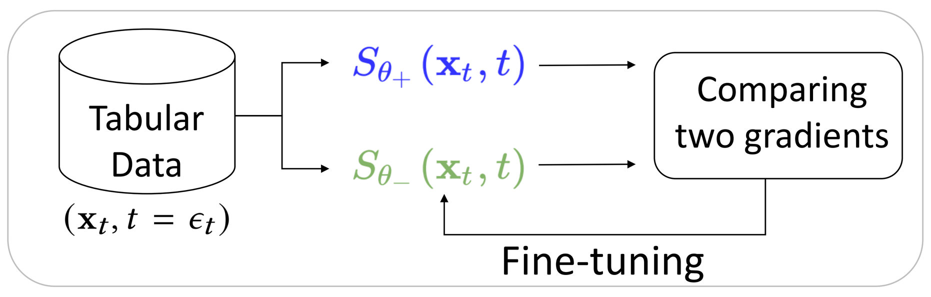
Proposed Method

1. Train a score-based generative model for each class.



- **Separately train two SGMs** for major and minor classes.
- **Smaller steps** are enough to solve the reverse SDE.

2. Fine-tune the minor score network.



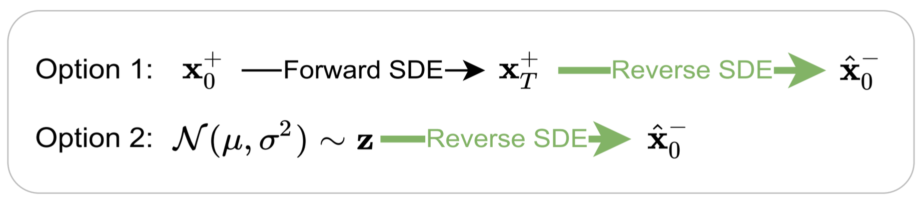
I. **Evaluate scores with each score network** at $(\mathbf{x}_t, t = \epsilon_t)$
   - A record $\mathbf{x}$ is from the entire data regardless of class.
   - A time ϵ_t (a small value close to 0) means the last moment of the reverse SDE.

II. **Calculate an angle** between the gradient of major and minor classes.
   - When the angle is smaller than ξ, their directions are similar.

III. **Decrease the gradient of the minor score network** by a factor of w.

$$L(x, t) = \| S_\theta(\mathbf{x}_t, t) - w g_{x,t} \|_2^2$$

3. Oversample minor class records.

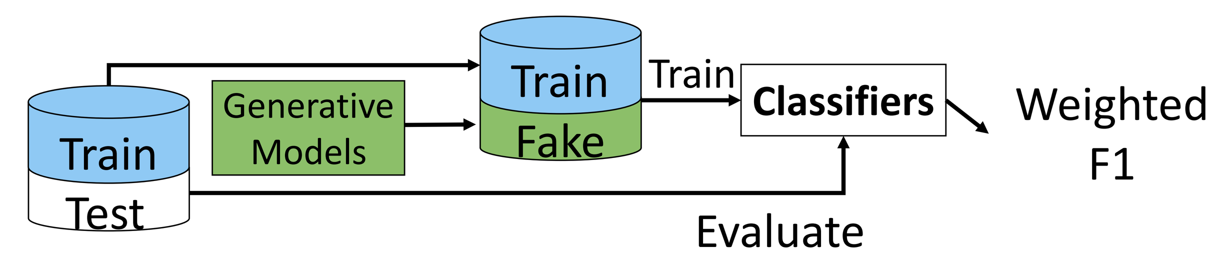


I. **Style transfer-based oversampling**
   - Select the major class record $\mathbf{x}_0^+$.
   - Derive a noisy vector $\mathbf{x}_T^+$.
   - Transfer $\mathbf{x}_T^+$ to $\hat{\mathbf{x}}_0^-$ **using the minor's reverse SDE.**
   - $\mathbf{x}_T^+$ **contains information on its original record.**
   - Generate a sample **around the class boundary.**

II. **Plain score-based oversampling**
   - Sample a noisy vector $\mathbf{z} \sim \mathcal{N}(\mu, \sigma^2)$.
   - Follow the standard use of SGMs.

Experiments

Evaluation Methods



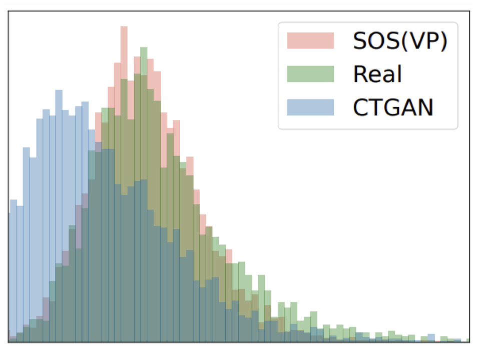
- **Weighted F1** is to give a higher weight to a smaller class.

Experimental Results

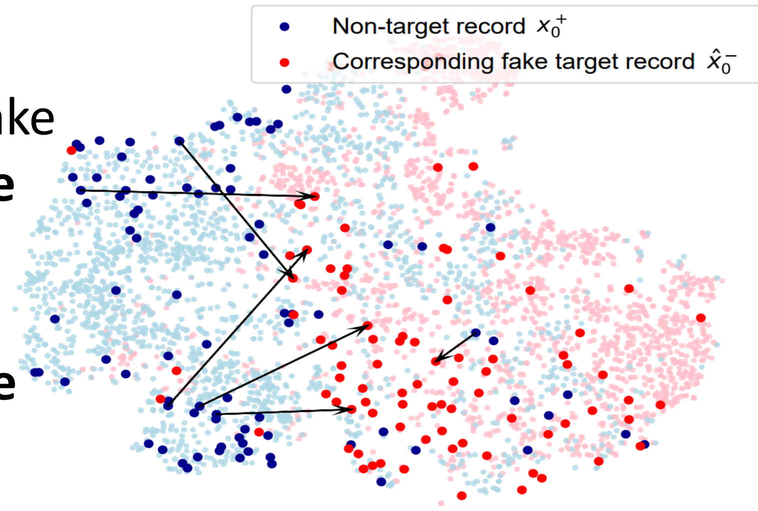
	Methods	Single Minority				Multiple Minority	
		Default	Shoppers	Surgical	WeatherAUS	Buddy	Satimage
	Identity	0.515±0.035	0.601±0.039	0.687±0.004	0.657±0.016	0.603±0.010	0.817±0.004
Baselines	SMOTE	0.561±0.025	0.648±0.004	0.678±0.008	0.674±0.025	0.584±0.005	0.846±0.005
	B-SMOTE	0.561±0.029	0.640±0.042	0.671±0.004	0.663±0.022	0.595±0.003	0.845±0.005
	Adasyn	0.558±0.023	0.630±0.045	0.662±0.007	0.658±0.022	0.608±0.002	0.841±0.008
	MedGAN	0.532±0.028	0.620±0.062	0.686±0.003	0.656±0.022	0.598±0.011	0.835±0.019
	VEEGAN	0.495±0.076	0.607±0.065	0.680±0.117	0.661±0.025	0.555±0.036	0.840±0.031
	TableGAN	0.423±0.115	0.571±0.097	0.704±0.001	0.579±0.066	0.570±0.019	0.813±0.013
	TVAE	0.536±0.035	0.610±0.060	0.681±0.004	0.652±0.018	0.552±0.044	0.846±0.031
	CTGAN	0.545±0.022	0.605±0.059	0.701±0.004	0.659±0.020	0.593±0.009	0.833±0.015
	OCT-GAN	0.531±0.018	0.639±0.029	0.692±0.082	0.656±0.018	0.551±0.015	0.837±0.011
	BAGAN	0.525±0.005	0.610±0.005	0.668±0.004	0.663±0.002	0.555±0.013	0.834±0.011
SOS	VE	0.571±0.003	**0.675±0.004**	0.709±0.003	0.672±0.002	0.607±0.007	0.854±0.002
	VP	0.559±0.006	0.658±0.003	0.712±0.002	0.680±0.002	0.607±0.011	**0.857±0.006**
	Sub-VP	**0.574±0.003**	0.673±0.002	**0.714±0.001**	**0.680±0.003**	**0.608±0.002**	0.855±0.004

- Identity is a minimal requirement for oversampling.
- SOS clearly outperforms all baseline methods and increases the F1 score after oversampling in all cases.

Column-wise histogram & t-SNE plot



- SOS successfully captures the real distribution, but CTGAN fails.



- The scatter plot shows real and fake records with **style transfer-based oversampling.**
- **Solid red dots are around the class boundary.**

Reference

[1] Song et al., Score-Based Generative Modeling through Stochastic Differential Equations. In ICLR, 2021.