Oversampling for t Mixture Models in Imbalanced Classification

Sunghyun Han [1] , Byungtae Seo [1]

[1] Department of Statistics, Sungkyunkwan University, Korea

Abstract

Class imbalance is a common issue in real-world datasets, often leading to biased classification models that underperform on the minority class. This problem becomes more challenging when compounded by class overlap, small disjuncts, outliers, and noise. To address these complexities, we propose an oversampling method based on the *t* mixture models (TMM) instead of commonly used Gaussian mixture models. Our approach clusters the minority class using TMM to capture its underlying structure and generates synthetic samples based on each cluster's distribution. The heavy tails of the *t* distribution make the proposed method more robust to outliers, reducing spurious clusters and enabling more reliable modeling of the data structure. We evaluate our method on various simulated datasets using SVM and Random Forest classifiers and standard imbalanced classification metrics, including F1-score, G-mean, and AUROC.

Motivation

Limitation of SMOTE in expanding minority regions: SMOTE creates synthetic points by linearly interpolating between nearby minority samples, disregarding the data-generating distribution. SMOTE does not truly enlarge the minority class support, which can lead to overfitting on a narrow region. To mitigate this, we perform oversampling grounded in the estimated statistical distribution using the *t* mixture model to broaden the minority region in a principled manner.

Robust global structure estimation : The *t* mixture models (TMM) estimates the underlying cluster configuration without the distortions that Gaussian models suffer in the presence of heavy-tailed data, yielding a faithful representation of the overall data geometry.

Enhanced minority representation under overlap and outliers : Because the *t* distribution has heavier tails, the TMM can robustly estimate the true centers of minority clusters even in the presence of outliers, while also capturing the extended tail regions of the distribution.

Data Complexity Factors

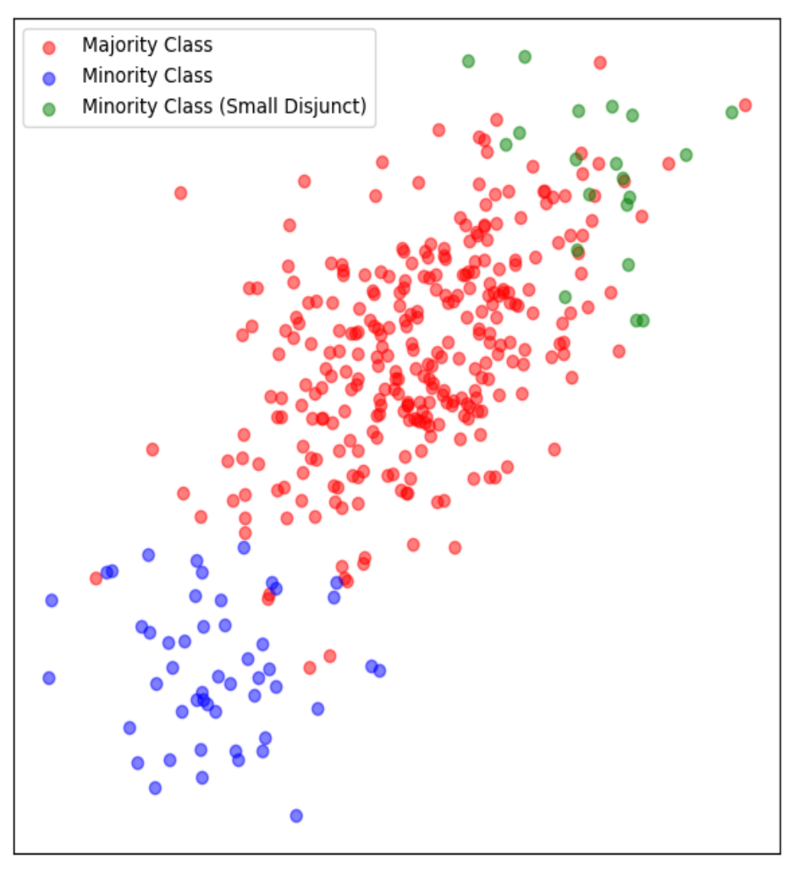

Fig 1. Small disjuncts

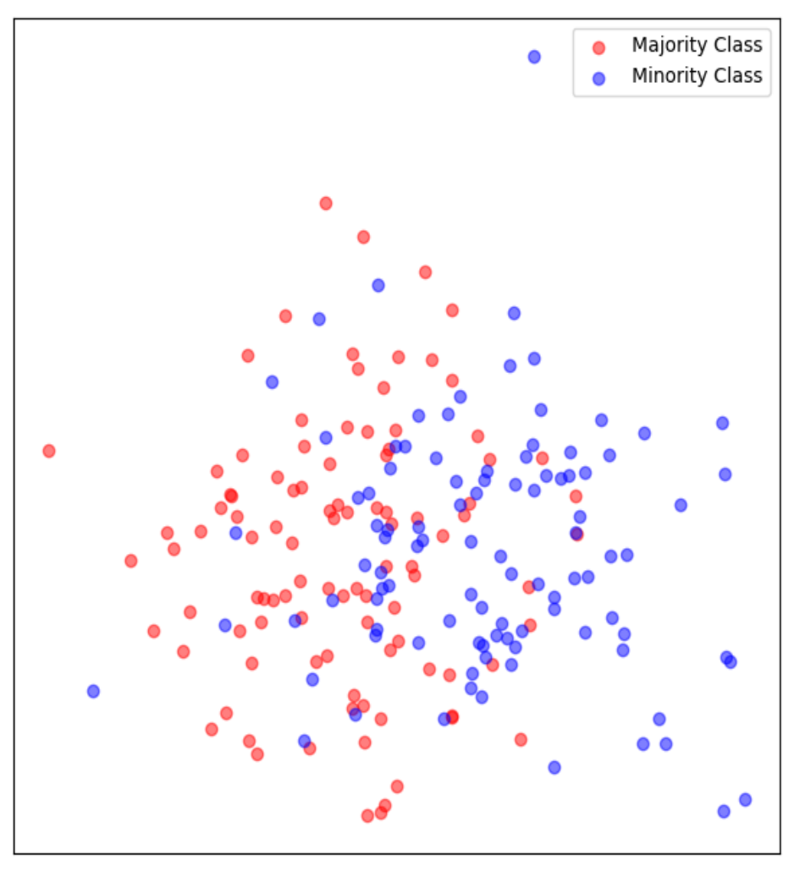

Fig 2. Class overlap

Small Disjuncts

- *A disjunct* is the decision region and its accompanying rule that a classifier carves out to explain a subset of the data. A single class can be represented by several disjuncts scattered across the feature space, each containing a different number of correctly classified training samples.
- Disjuncts with low coverage are referred to as *small disjuncts*. These disjuncts can be *identified* through a clustering approach.
- they generalize poorly and are easily derailed by attribute noise, missing values, or label errors, making them a disproportionate source of misclassification.

Noise

- Random noise amplifies complexity without contributing useful information. It refers to arbitrary fluctuations unrelated to the true data distribution, often caused by sensor errors, input mistakes, or complex environmental factors.

Outlier

- Outlier refers to an observation that significantly deviates from the rest of the data distribution. It can distort the representativeness of the dataset and degrade the model's generalization performance.
- However, some minority-class samples may appear as outliers due to data sparsity, even though they potentially contain important patterns.

Class Overlap

- Class overlap refers to regions of the data space where samples from two or more distinct classes coexist. While there is still no universally accepted formula for quantifying the degree of overlap, numerous metrics have been proposed. Research on this topic generally falls into four complementary perspectives: feature overlap, structural overlap, multiresolution overlap, and instance overlap.

t Mixture Models based Oversampling

Probability density function

$$f(x;\Psi) = \Sigma_{k=1}^{K} \pi_k \cdot t(x;\mu_k,\Sigma_k,\nu_k)$$

· $t(x;\mu_k,\Sigma_k,\nu_k) = \Gamma\left(\frac{\nu_k+d}{2}\right)|\Sigma_k|^{\frac{-1}{2}} / (\pi\nu_k)^{\frac{d}{2}}\,\Gamma\left(\frac{\nu_k}{2}\right)\left[1+\frac{1}{\nu_k}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\right]^{\frac{\nu_k+d}{2}}$,
· Observation x ∈ $\mathbb{R}^{n\text{x}d}$,
· Mean vector $\mu_k \in \mathbb{R}^d$,
· Covariance matrix $\Sigma_k \in \mathbb{R}^{d\text{x}d}$,
· Degree of freedom ν_k, Mixing proportion π_k.

Estimate *t* mixture model

- The *t* mixture models' parameters are iteratively estimated via the Expectation–Maximization (EM) algorithm, using the hierarchical structure between the auxiliary latent variable *u* and the observation *x*.

Generate synthetic sample

1. We estimate the *t* mixture models (TMM) using the original minority class samples.
2. The probability density function for generating synthetic data is precisely this estimated TMM.
3. Random sampling from estimated TMM.

Simulation

Case 1 : *t* mixture model based simulation dataset

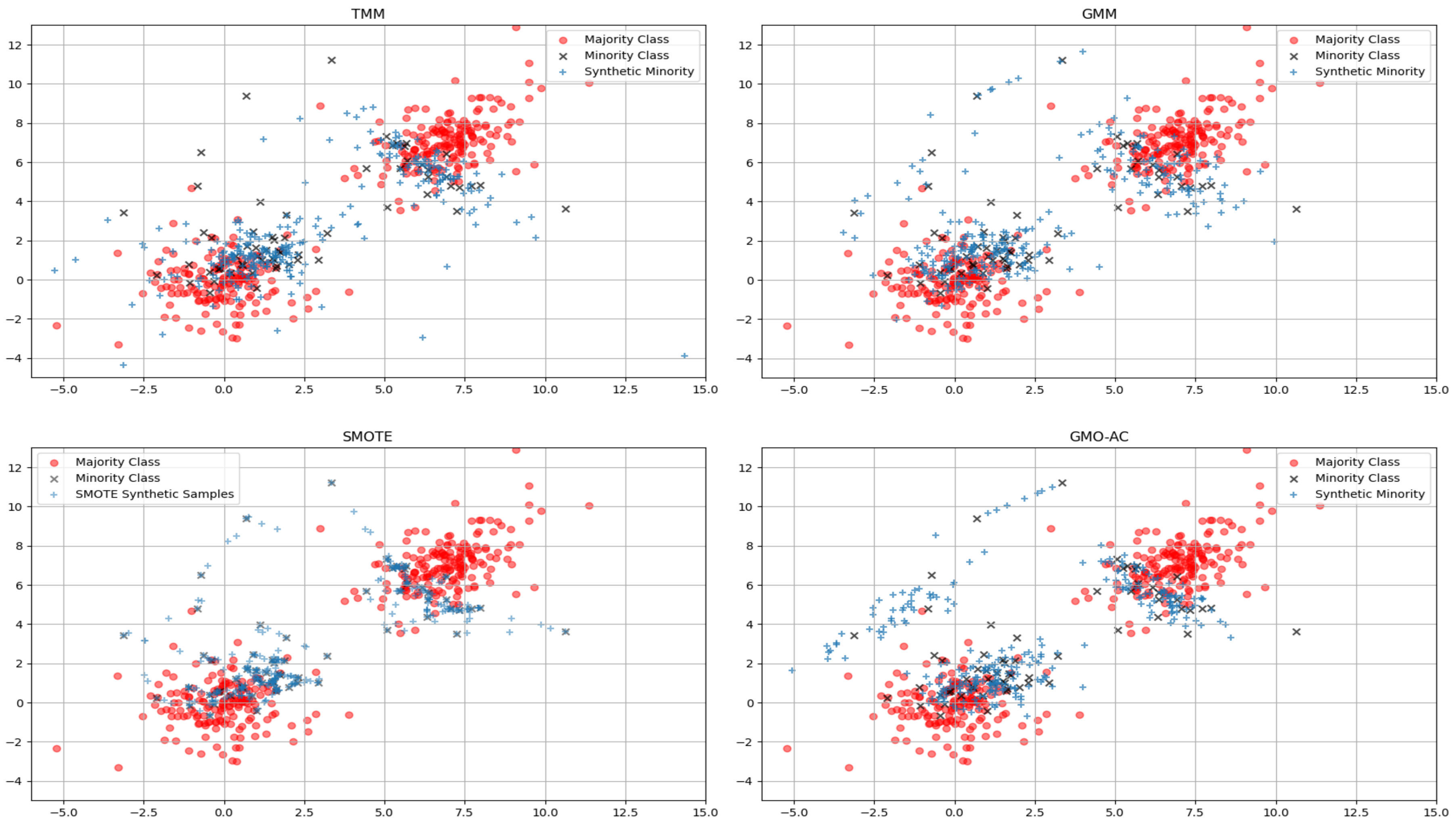

Fig 3. Visual comparison of data distributions after each oversampling Method (case1)

		SVM			Random Forest		
n	Oversampling	F1-Score	AUROC	G-mean	F1-Score	AUROC	G-mean
500	TMM	**0.463±0.072**	0.791±0.059	0.720±0.064	**0.427±0.069**	**0.767±0.061**	**0.692±0.066**
	GMM	0.459±0.063	**0.792±0.054**	0.713±0.062	0.410±0.076	0.752±0.067	0.676±0.074
	GMO-AC	0.457±0.059	0.784±0.058	**0.723±0.056**	0.420±0.075	0.757±0.060	0.681±0.073
	SMOTE	0.451±0.067	0.786±0.060	0.713±0.065	0.389±0.083	0.738±0.059	0.631±0.083
	No manipulation	0.004±0.025	0.738±0.064	0.008±0.048	0.333±0.125	0.736±0.067	0.489±0.124
1500	TMM	0.473±0.038	0.801±0.034	**0.739±0.034**	**0.428±0.037**	**0.767±0.038**	**0.699±0.036**
	GMM	**0.476±0.037**	**0.801±0.033**	0.738±0.033	0.420±0.044	0.762±0.038	0.691±0.041
	GMO-AC	0.465±0.041	0.794±0.036	0.731±0.037	0.427±0.044	0.760±0.039	0.694±0.041
	SMOTE	0.467±0.040	0.799±0.035	0.736±0.037	0.416±0.050	0.746±0.040	0.658±0.047
	No manipulation	0.001±0.007	0.749±0.054	0.005±0.028	0.343±0.073	0.759±0.039	0.506±0.067

Table 1. Classification performance metrics (case1)

Case 2 : Gaussian mixture model based simulation dataset

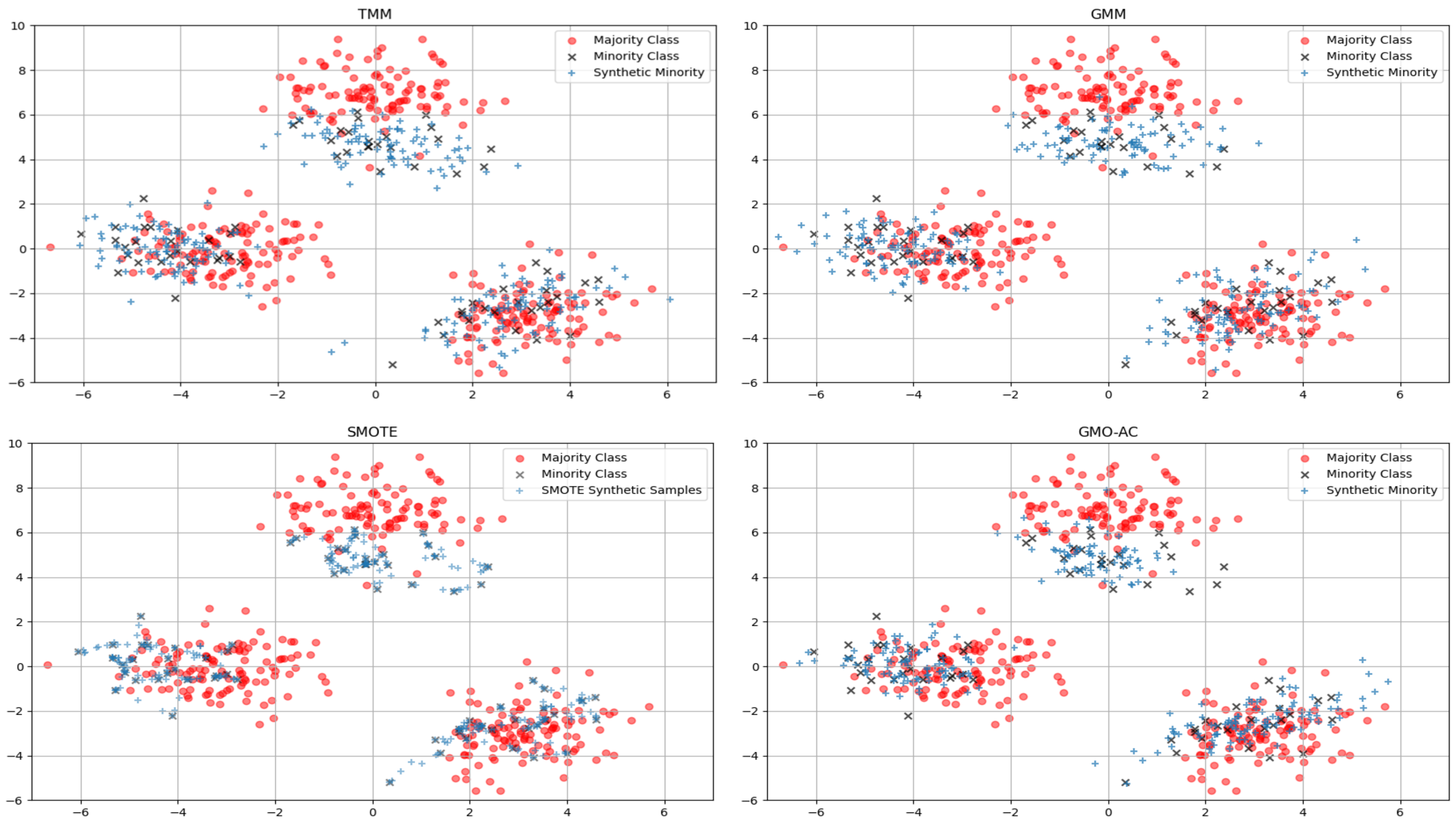

Fig 4. Visual comparison of data distributions after each oversampling Method (case2)

		SVM			Random Forest		
n	Oversampling	F1-Score	AUROC	G-mean	F1-Score	AUROC	G-mean
500	TMM	0.450±0.079	0.769±0.064	0.684±0.075	**0.439±0.073**	**0.752±0.071**	**0.679±0.068**
	GMM	0.453±0.078	**0.771±0.064**	0.686±0.075	0.423±0.075	0.745±0.069	0.664±0.071
	GMO-AC	0.443±0.068	0.763±0.068	0.683±0.071	0.435±0.062	0.747±0.063	0.673±0.058
	SMOTE	**0.454±0.074**	0.769±0.066	**0.689±0.072**	0.427±0.080	0.741±0.065	0.647±0.072
	No manipulation	0.037±0.089	0.751±0.061	0.061±0.134	0.372±0.115	0.733±0.065	0.522±0.105
1500	TMM	0.462±0.036	**0.791±0.034**	0.701±0.034	0.437±0.044	0.764±0.038	0.682±0.042
	GMM	**0.463±0.039**	0.791±0.035	0.702±0.036	0.433±0.041	0.763±0.038	0.678±0.039
	GMO-AC	0.452±0.033	0.786±0.035	0.700±0.032	**0.440±0.044**	**0.764±0.036**	**0.684±0.041**
	SMOTE	0.459±0.031	0.789±0.033	**0.702±0.031**	0.433±0.049	0.754±0.035	0.654±0.044
	No manipulation	0.219±0.101	0.760±0.039	0.339±0.119	0.387±0.063	0.754±0.035	0.535±0.053

Table 2. Classification performance metrics (case2)

Conclusion

▶ Class imbalance hampers minority-class prediction when rare examples are overwhelmed by the majority.

▶ We propose *t* mixture model based oversampling to robustly capture heavy tails and outliers, avoiding GMM's tendency to overestimate components or form spurious groups.

▶ Unlike SMOTE, our statistically grounded method preserves global structure and broadens minority coverage based on the data distribution.

▶ Across two simulations datasets, oversampling with the *t* mixture model showed superior or comparable performance in F1-score, AUROC, and G-mean relative to SMOTE and GMM-based methods.