Oversampleing for t Mixture Models in Imbalanced Classification

Sunghyun Han

Advisor : Prof. Byungtae Seo

Department of Statistics, Sungkyunkwan University

June 12, 2025

Outline

IMBALANCE DATA

- **Imbalance Data** refers to a state that the raio of data varies severely for each class.
- Notably, data imbalances frequently occur in the problem of credit fraud, medical diagnostics.
- We consider the case where it is more important to predict the minorities class.

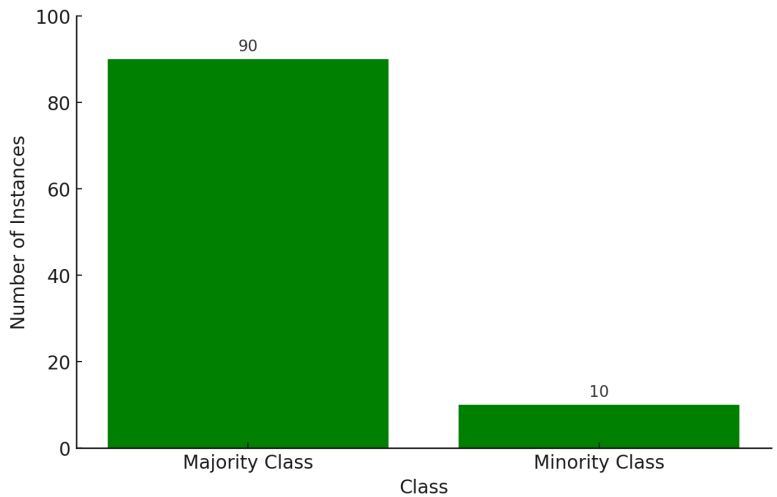
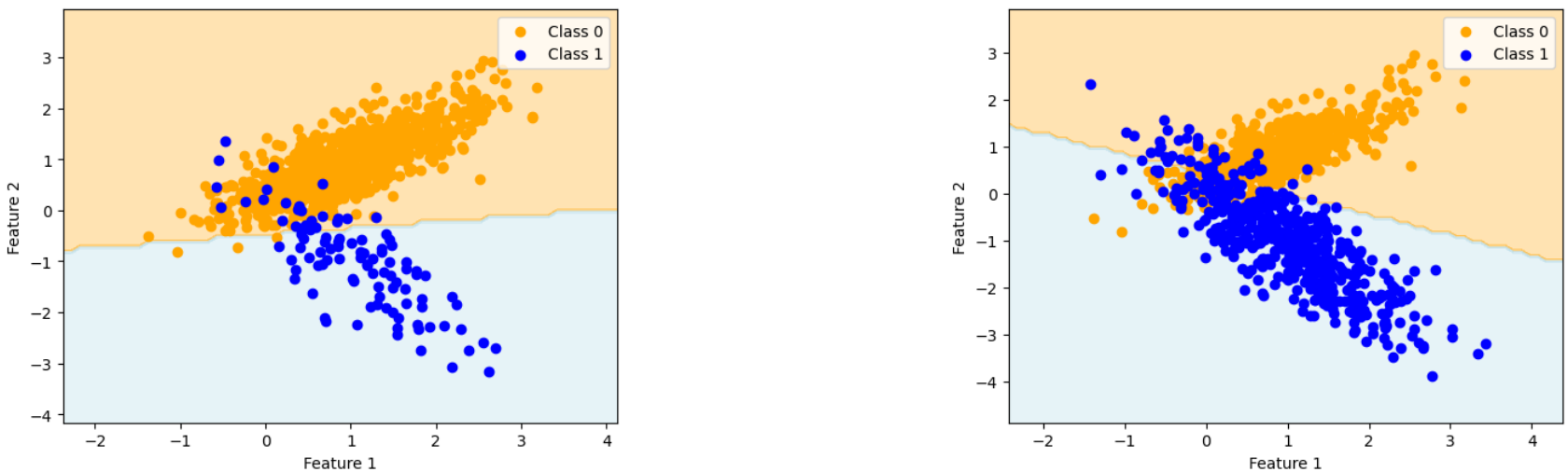


Figure 1: Visualize imbalance Data.

PROBLEMS CAUSED BY DATA IMBALANCES

- In the case of imbalanced data, there is a problem that the classifiers are biased toward the majority class.
- Also, it can lead to an overfitting problem.
- These problems adversely affect the prediction of the minority class.



(a) Imbalance data.

(b) Balance data.

Figure 2: Comparison of Decision Boundaries for Imbalanced and Balanced Data.

PRELIMINARY

- **Small disjuncts** are caused by a rare case. It represents the pattern of actual data, but refers to disjuncts with a small number of data.
- **Noise** is incomplete data and errors that can occur during the step of gathering and preprocessing data.
- **Outlier** is a sample that is noticeably separated from other samples.
- **Class overlap** refers to a situation in which two or more different classes of data coexist in a specific area of the entire data space.

SMALL DISJUNCTS

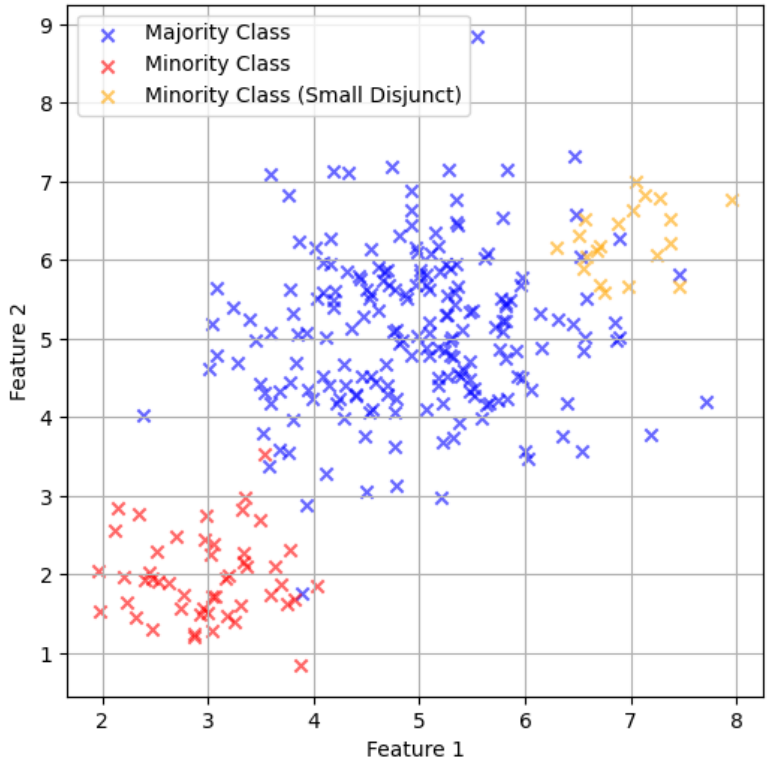


Figure 3: Visualize small disjuncts

- The coverage of a disjuncts is defined as the number of trainging examples if correctly classifies. When the coverage of disjuncts is small, it is called **small disjuncts**.(Holte et al. (1989))

Overlap

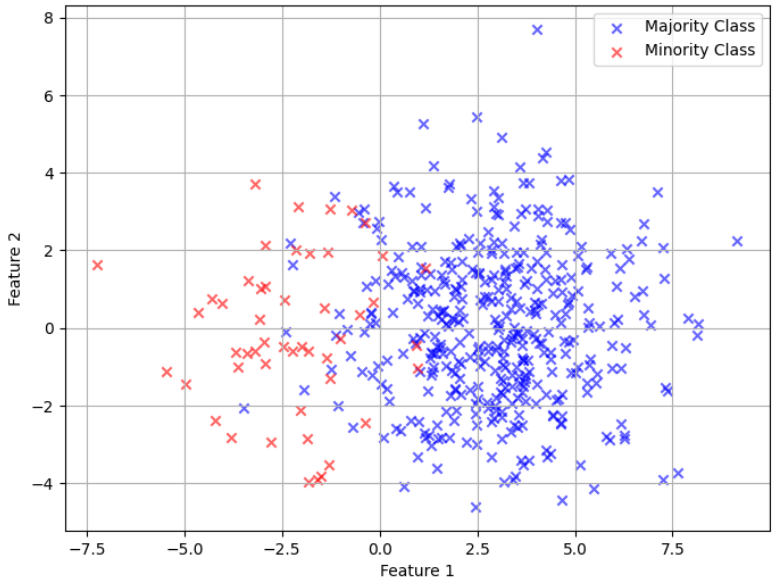


Figure 4: Visualize Overlap

- A global definition of class overlap is based on the existence of the regions populated by examples from different classes.(Santos et al. (2023))

PROBLEM OF SMOTE

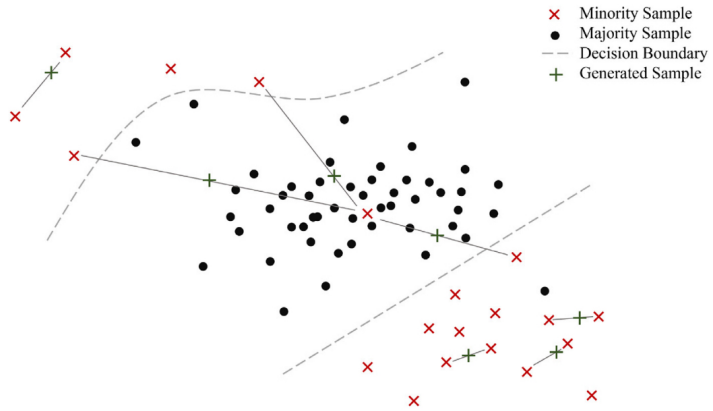


Figure 5: Generate of synthetic example (SMOTE), Douzas et al. (2018)

- SMOTE generates synthetic samples only along the line segments connecting minority samples, limiting its ability to expand the minority distribution.
- It may generate minority samples in majority regions in the presence of noise.
- The issues of within-class imbalance and small disjuncts are ignored.
- Distort the actual data distribution.

WHY APPLY CLUSTERING IN OVERSAMPLING?

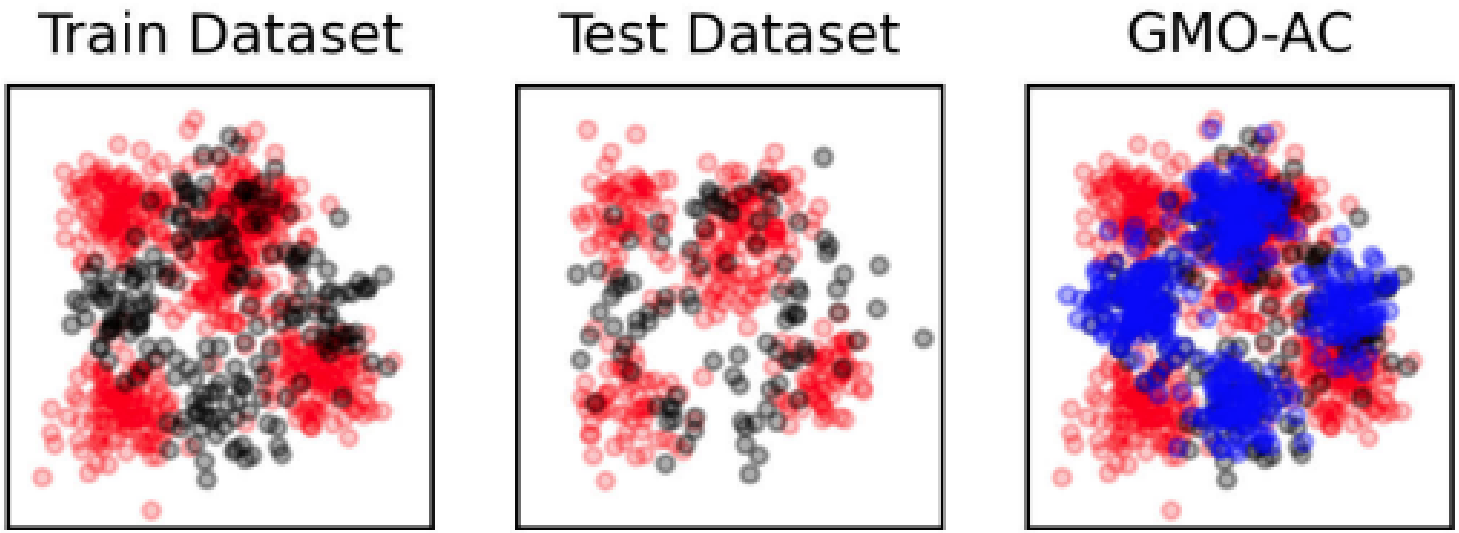
Class imbalances versus small disjuncts - Jo and Japkowicz (2004)

- In small and complex datasets, class imbalances come accompanied with the problem of small disjuncts, which in turn causes a degration in standard classifiers' performance.
- Small disjuncts were approximated using unsupervised learning(e.g, k-means).

GMO-AC : Gaussian-Based Minority Oversampling With Adaptive Outlier Filtering and Class Overlap Weighting- Yang and Cha (2024)

- SMOTE generates synthetic samples in a star or tree-shaped pattern around minority instances, which may not reflect the actual data distribution.
- GMO-AC estimates the minority class distribution using a GMM(Gaussian mixture model) and generates new samples following each sub-component, better capturing the underlying structure.

Illustration of GMO-AC Oversampling



- The left and center plots show the original training and test datasets.
- The right plot shows the synthetic samples (in blue) generated by GMO-AC.
- GMO-AC estimates the structure of the minority class using a Gaussian mixture model after filtering outliers.
- Class overlap is quantified and used to guide sampling, resulting in synthetic samples that reinforce the central structure of each minority cluster.

STEP 1: REMOVING OUTLIERS WITH GMM

1. Fit an initial GMM to the minority class and compute the Mahalanobis distances between each minority sample and its assigned GMM component.

$$d_{i,k} = \sqrt{(\mathbf{x}_{i,k} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x}_{i,k} - \boldsymbol{\mu}_k)}, \quad i = 1, \ldots, N_k$$

2. Sort $d_{i,k}$ in descending order, and iteratively remove the top half of the points, i.e., $x_{(1),k}, \ldots, x_{(N_k/2),k}$. After each removal, compute the determinant of the updated covariance matrix.

$$C_k = \left\{ |\Sigma_{0,k}|, |\Sigma_{-1,k}|, \ldots, |\Sigma_{-\lfloor N_k/2 \rfloor,k}| \right\}$$

3. Detect the break point p via segmented linear regression:

$$|\Sigma_{-i,k}| = \begin{cases} \alpha_1 i + \beta_1, & \text{if } i < p \\ \alpha_2 i + \beta_2, & \text{if } i \geq p \end{cases}$$

   and discard all points before p as outliers.
4. Re-estimate GMM parameters $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}$ using the filtered data.

STEP 2: QUANTIFYING CLASS OVERLAP

1. Filtering Majority Samples Closer than Minority Samples

$$d_{\max,k} = \max_i d_{i,k}$$

$$D_k^{\text{maj}} = \left\{ x_j^{\text{maj}} \,\middle|\, d_{j,k}^{\text{maj}} \leq d_{\max,k} \right\}$$

D_k^{maj} denotes majority samples that are closer to the k-th minority cluster center than minority samples themselves, and are filtered due to potential overlap.

2. Overlap ratio (F-statistic)

$$F_{0,k} = \frac{\sum_{i=1}^{N_k^+} (x_i^{\min} - \mu_k)^\top \Sigma_k^{-1} (x_i^{\min} - \mu_k) / N_k^+}{\sum_{j=1}^{N_k^-} (x_j^{\text{maj}} - \mu_k)^\top \Sigma_k^{-1} (x_j^{\text{maj}} - \mu_k) / N_k^-}$$

where N_k^+ and N_k^- denote the numbers of minority and majority samples associated with component k, respectively.

3. Class overlap probability:

$$\text{Prob}_k = P(F_k < F_{0,k}), \quad F_k \sim F_{N_k^+ \cdot M,\, N_k^- \cdot M}$$

4. Normalized class overlap weight:

$$\delta_k = \frac{\text{Prob}_k}{\sum_{j=1}^{K} \text{Prob}_j}$$

STEP 3: GENERATION OF SYNTHETIC DATA WITH CONSIDERATION OF CLASS OVERLAP

Sampling distribution:

$$f(x_{\text{syn}}; \Psi) = \sum_{k=1}^{K} [\beta\, \delta_k + (1 - \beta)\, \pi_k] \cdot \mathcal{N}(x_{\text{syn}} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \quad \beta \in [0, 1]$$

- π_k: mixture weight of the k-th Gaussian component in the re-estimated GMM.
- δ_k: degree of class overlap (normalized overlap probability for component k).
- β: tuning parameter that adjusts how strongly class-overlap information affects sampling.

Synthetic samples x_{syn} are drawn from this weighted mixture. When $\beta = 0$, sampling uses only GMM weights. When $\beta = 1$, sampling prioritizes components with high overlap δ_k.

Motivation

- **Robust global structure estimation** : The t mixture model (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.

t Mixture Models

- Multivariate t distribution's probability density function

$$t(x;\boldsymbol{\mu},\boldsymbol{\Sigma},\nu) = \frac{\Gamma\left(\frac{\nu+d}{2}\right)|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{(\pi\nu)^{\frac{d}{2}}\Gamma\left(\frac{\nu}{2}\right)\left\{1+\frac{1}{\nu}\delta(x,\boldsymbol{\mu};\boldsymbol{\Sigma})\right\}^{\frac{\nu+d}{2}}}$$

- t Mixture model

$$f(x;\Psi) = \sum_{k=1}^{K}\pi_k t(x|\mu_k,\Sigma_k,v_k),$$

$\Gamma(t) = \int_0^\infty e^{-x}x^{t-1}\,dx, t > 0$
$\delta(x,\boldsymbol{\mu};\boldsymbol{\Sigma}) = (x-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu})$
$\mu_k \in \mathbb{R}^d$: Mean vector of the k-th component
$\Sigma_k \in \mathbb{R}^{d\times d}$: Covariance matrix of k-th
v_k : The degree of freedom of the k-th component component

t Mixture Models

- Latent variable

$$z_{jk} = \begin{cases} 1, & \text{The j-th observation is from the k-th component} \\ 0, & \text{otherwise} \end{cases}$$

- Hierarchical structure for EM algorithms

$$\boldsymbol{X}_j \mid (u_j, z_{jk} = 1) \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k / u_k),$$
$$U_j \mid (z_{jk} = 1) \sim \text{Gamma}\left(\frac{1}{2}\nu_k, \frac{1}{2}\nu_k\right)$$

- Complete-Data Probability Density Function

$$f(x, z, u; \Psi) = \prod_{k=1}^{K} \prod_{j=1}^{n} (\pi_j t(x|\mu_k, \Sigma_k, \nu_k))^{\mathbf{z}_{jk}}$$

t Mixture Models

- Complete log-likelihood

$$L_c(\Psi) = L_{1c}(\pi) + L_{2c}(\nu) + L_{3c}(\epsilon)$$

$$\log L_{1c}(\boldsymbol{\pi}) = \sum_{k=1}^{K}\sum_{j=1}^{n} z_{jk} \log \pi_k,$$

$$\log L_{2c}(\boldsymbol{\nu}) = \sum_{k=1}^{K}\sum_{j=1}^{n} z_{jk}\Bigg\{ - \log\Gamma\left(\frac{1}{2}\nu_k\right) + \frac{1}{2}\nu_k \log\left(\frac{1}{2}\nu_k\right) + \frac{1}{2}\nu_k(\log u_j - u_j) - \log u_j \Bigg\},$$

$$\log L_{3c}(\boldsymbol{\epsilon}) = \sum_{k=1}^{K}\sum_{j=1}^{n} z_{jk}\Bigg\{ - \frac{1}{2}p\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_k| - \frac{1}{2}u_j(\boldsymbol{x}_j - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_k) \Bigg\}$$

E-STEP : LATENT VARIABLE ESTIMATION

$$z_{jk}^{(t)} = \frac{\pi_k^{(t)}\ t\big(\mathbf{x}_j \mid \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}, \nu_k^{(t)}\big)}{\sum_{k=1}^{K} \pi_k^{(t)}\ t\big(\mathbf{x}_j \mid \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)}, \nu_k^{(t)}\big)},$$

$$u_{jk}^{(t)} = \frac{\nu_k^{(t)} + d}{\nu_k^{(t)} + (\mathbf{x}_j - \boldsymbol{\mu}_k^{(t)})^\top \big(\boldsymbol{\Sigma}_k^{(t)}\big)^{-1} (\mathbf{x}_j - \boldsymbol{\mu}_k^{(t)})}$$

$$\mathbb{E}_{\psi^{(t)}}\left[\log u_{jk} \mid \mathbf{x}_j, z_{jk} = 1\right] = \log u_{jk}^{(t)} + \psi\Big(\tfrac{\nu_k^{(t)}+d}{2}\Big) - \log\Big(\tfrac{\nu_k^{(t)}+d}{2}\Big),$$

where $\psi(\alpha) = \dfrac{d}{d\alpha}\log\Gamma(\alpha) = \dfrac{\Gamma'(\alpha)}{\Gamma(\alpha)}$ is the digamma function.

M-STEP: PARAMETER UPDATES

$$\pi_k^{(t+1)} = \frac{1}{n}\sum_{j=1}^{n} z_{jk}^{(t)}, \qquad \boldsymbol{\mu}_k^{(t+1)} = \frac{\sum_{j=1}^{n} z_{jk}^{(t)} u_{jk}^{(t)} \mathbf{x}_j}{\sum_{j=1}^{n} z_{jk}^{(t)} u_{jk}^{(t)}}$$

$$\boldsymbol{\Sigma}_k^{(t+1)} = \frac{\sum_{j=1}^{n} z_{jk}^{(t)} u_{jk}^{(t)} (\mathbf{x}_j - \boldsymbol{\mu}_k^{(t+1)})(\mathbf{x}_j - \boldsymbol{\mu}_k^{(t+1)})^\top}{\sum_{j=1}^{n} z_{jk}^{(t)} u_{jk}^{(t)}}$$

$$\frac{\partial Q_{2j}(\nu_k \mid \Psi^{(k)})}{\partial \nu_k} = -\psi\left(\frac{\nu_k}{2}\right) + \log\left(\frac{\nu_k}{2}\right) + 1 + \frac{1}{n_k^{(t)}}\sum_{j=1}^{n} z_{jk}^{(t)}\left(\log u_{jk}^{(t)} - u_{jk}^{(t)}\right) = 0$$

The last equation is solved numerically to update the degrees of freedom ν_k.

OVERSAMPLING

- Synthetic samples x_{syn} are generated according to the probability density function of the estimated t mixture model:

$$f(x_{syn}; \widehat{\Psi}) = \sum_{k=1}^{K} \widehat{\pi}_k \, t(x_{syn} \mid \widehat{\mu}_k, \widehat{\Sigma}_k, \widehat{v}_k), \quad \widehat{\Psi} = \left\{ \widehat{\pi}_k, \widehat{\mu}_k, \widehat{\Sigma}_k, \widehat{v}_k \right\}_{k=1}^{K}$$

PSEUDOCODE

Algorithm 1 Algorithm for Oversampling via t Mixture Models

Input: Minority data $X^+ \in \mathbb{R}^{n \times m}$; components K; synthetic data size η; tolerance ε
Output: Synthetic data set D_{syn}
for $k \leftarrow 1$ **to** K **do**
    $\pi_k, \mu_k, \Sigma_k, \nu_k \leftarrow$ init values by Kmeans clustering
end for
$D \leftarrow \varnothing$; $\ell_{\text{old}} \leftarrow -\infty$
repeat
    $\ell_{\text{old}} \leftarrow \ell_{\text{new}}$
    for $j \leftarrow 1$ **to** n **do**
        for $k \leftarrow 1$ **to** K **do**
            compute z_{jk}, u_{jk}
        end for
    end for
    for $k \leftarrow 1$ **to** K **do**
        update π_k, μ_k, Σ_k
        update ν_k via Newton-Raphson iteration
    end for
    recompute log-likelihood ℓ_{new}
until $|\ell_{\text{new}} - \ell_{\text{old}}| < \varepsilon$
for $k \leftarrow 1$ **to** K **do**
    $n_k \leftarrow \text{round}(\pi_k \times \eta)$
    $X_{syn} \leftarrow$ draw n_k samples from $t_{\nu_k}(\mu_k, \Sigma_k)$
    $D_{syn} \leftarrow D_{syn} \cup X_{syn}$
end for
return D

Evaluation Metrics for Imbalanced Classification

Accuracy can be misleading when the majority class dominates, so we focus on F1-score, AUROC, and G-mean. Let TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall (Sensitivity)} = \frac{TP}{TP + FN}, \quad \text{Specificity} = \frac{TN}{TN + FP}.$$

F1-score: harmonic mean of precision and recall,

$$\text{F1-score} = 2\,\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

AUROC: area under the ROC curve, which plots sensitivity vs. $(1 - \text{Specificity})$ over all thresholds.
G-mean: geometric mean of sensitivity and specificity,

$$\text{G-mean} = \sqrt{\text{Sensitivity} \times \text{Specificity}} = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}}.$$

CASE1: SUBCOMPONENT SELECTION VIA BIC

Table 1. Experimental data configuration (Case 1)

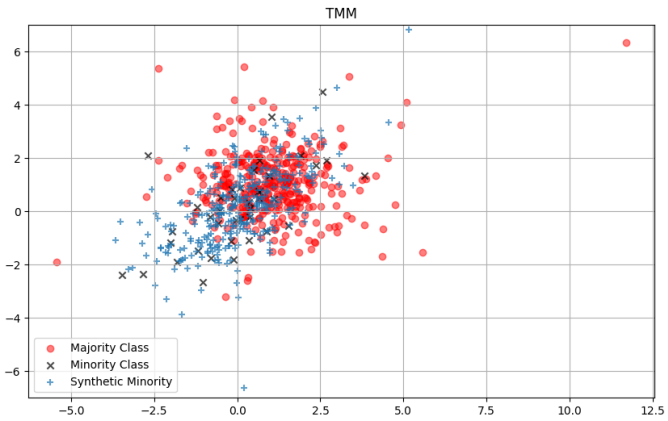
	Mean	Covariance	Degree of freedom	Proportion
Majority	$[1, 1]$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	5	90%
Minority	$[0, 0]$	$\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$	3	10%

- Using the true parameters listed in Table 1, synthetic data were sampled from the corresponding multivariate t-distribution.

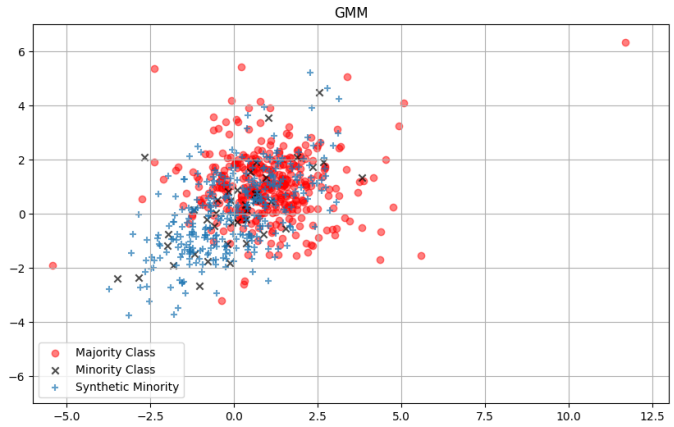
Table 2. BIC-based selection of subcomponents (Case 1)

n	Method	Clusters	
		1	2
n=500	TMM	89%	11%
	GMM	54%	46%
n=1500	TMM	98%	2%
	GMM	19%	81%

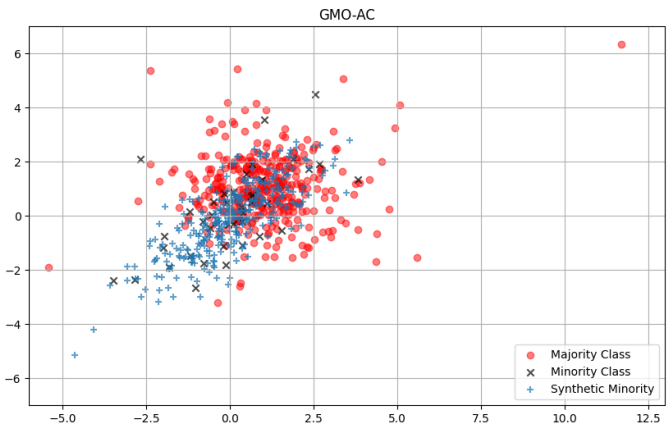
Case 1: Oversampling Visualization



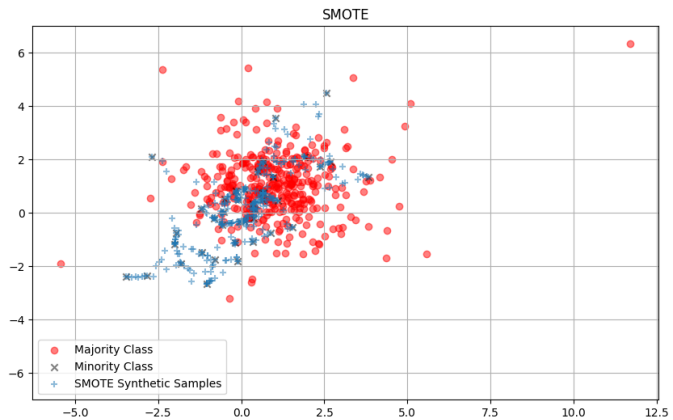
TMM



GMM



GMO-AC



SMOTE

Figure 6: Visual comparison of data distributions after each oversampling method (Case 1).

Case 1: Classification Performance Summary

Table 3. Classification performance metrics for Case 1

		SVM			Random Forest		
n	Oversampling	F1-Score	AUROC	G-mean	F1-Score	AUROC	G-mean
500	TMM	**0.372±0.095**	**0.777±0.091**	**0.707±0.110**	**0.310±0.076**	**0.738±0.095**	**0.674±0.091**
	GMM	0.367±0.107	0.760±0.099	0.691±0.109	0.302±0.078	0.723±0.099	0.662±0.092
	GMO-AC	0.357±0.095	0.758±0.094	0.694±0.113	0.299±0.085	0.719±0.098	0.651±0.104
	SMOTE	0.350±0.090	0.754±0.094	0.691±0.094	0.292±0.098	0.706±0.092	0.604±0.130
	No manipulation	0.065±0.119	0.632±0.122	0.104±0.175	0.249±0.154	0.720±0.091	0.385±0.197
1500	TMM	**0.383±0.051**	**0.772±0.055**	**0.721±0.053**	0.304±0.036	**0.731±0.054**	**0.674±0.046**
	GMM	0.378±0.057	0.767±0.057	0.707±0.058	0.295±0.037	0.722±0.052	0.664±0.046
	GMO-AC	0.370±0.051	0.763±0.054	0.712±0.056	**0.305±0.037**	0.725±0.054	0.674±0.048
	SMOTE	0.367±0.052	0.760±0.052	0.708±0.051	0.300±0.051	0.699±0.061	0.624±0.063
	No manipulation	0.097±0.098	0.618±0.076	0.184±0.154	0.260±0.090	0.714±0.057	0.421±0.086

CASE2: SUBCOMPONENT SELECTION VIA BIC

Table 4. Experimental data configuration (Case 2)

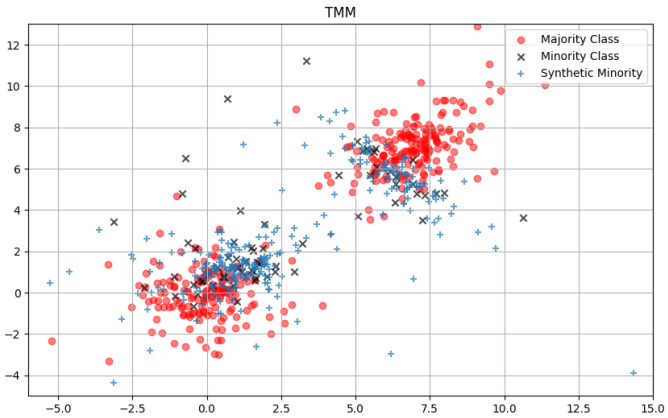
	Mean	Covariance	Degree of freedom	Proportion
Majority	$[0, 0]$, $[7, 7]$	$\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}$, $\begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}$	5, 4	42%, 42%
Minority	$[1, 1]$, $[6, 6]$	$\begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}$, $\begin{bmatrix}1 & -0.5\\-0.5 & 1\end{bmatrix}$	3, 3	10%, 6%

- Using the true parameters listed in Table 4, synthetic data were sampled from the corresponding multivariate t-distributions. where both the majority and minority classes consist of two subclusters.

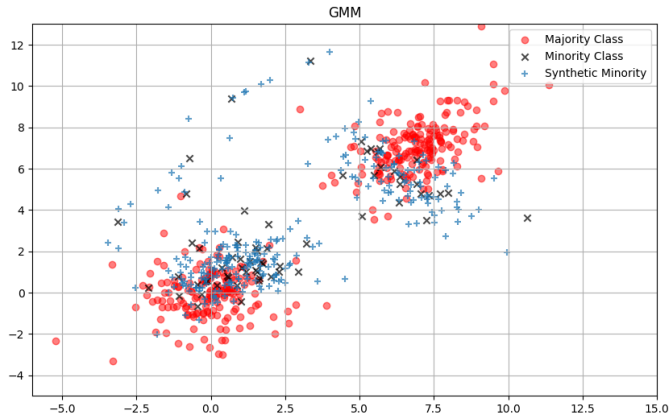
Table 5. BIC-based selection of subcomponents (Case 2)

n	Method	1	2	3	4	5
500	TMM	0%	97%	3%	0%	0%
	GMM	0%	52%	30%	11%	7%
1500	TMM	0%	100%	0%	0%	0%
	GMM	0%	1%	35%	30%	24%

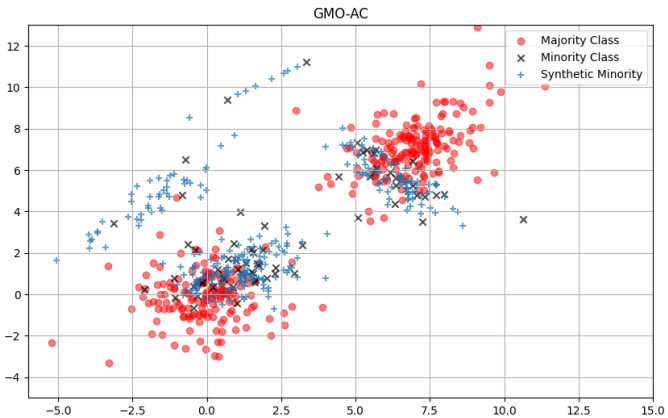
Case 2: Oversampling Visualization



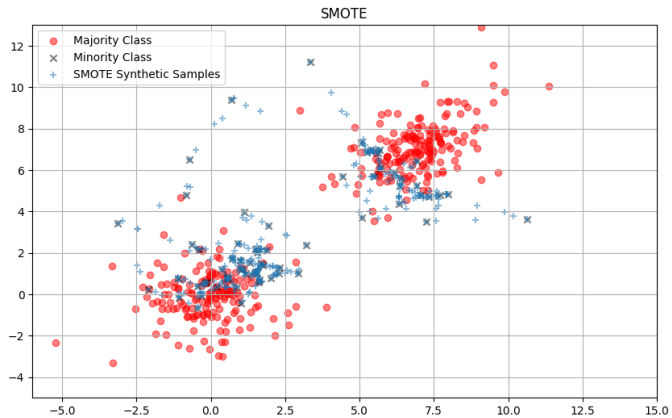
TMM



GMM



GMO-AC



SMOTE

Figure 7: Visual comparison of data distributions after each oversampling method (Case 2).

Case 2: Classification Performance Summary

Table 6. Classification performance metrics for Case 2

		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

Case 3: Subcomponent Selection via BIC

Table 7. Experimental data configuration (Case 3)

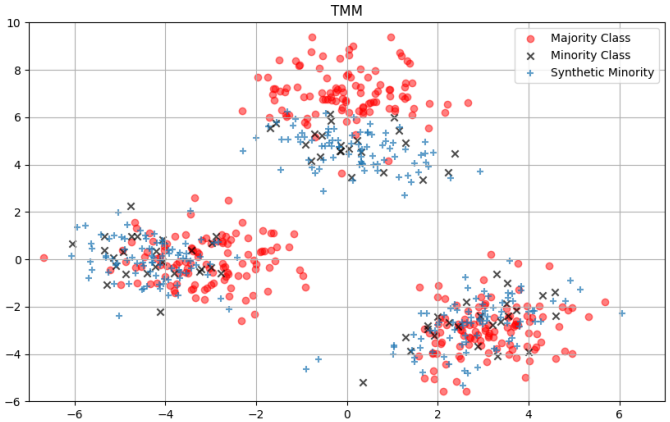
	Mean	Covariance	Proportion
Majority	$[-3, 0]$, $[0, 7]$, $[3, -3]$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	27.3%, 27.3%, 27.3%
Minority	$[-4, 0]$, $[0, 5]$, $[3, -2.5]$	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$, $\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$	6%, 6%, 6%

- Case 3 simulation data were generated from multivariate normal distributions with the true parameters listed in Table 7, where both the majority and minority classes consist of three subclusters.

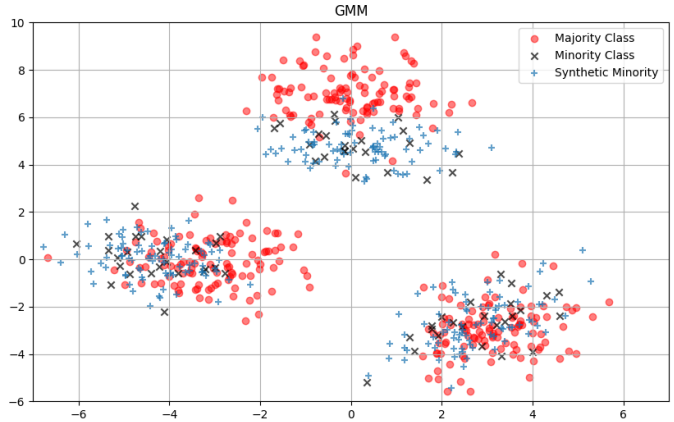
Table 8. BIC-based selection of subcomponents (Case 3)

n	Method	3	4
500	TMM	99%	1%
	GMM	97%	3%
1500	TMM	100%	0%
	GMM	100%	0%

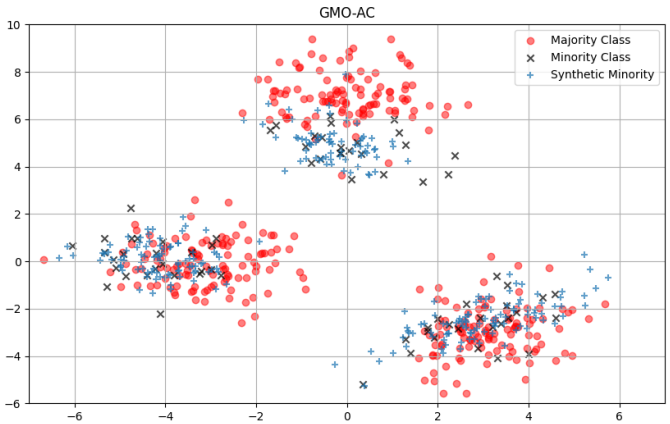
Case 3: Oversampling Visualization



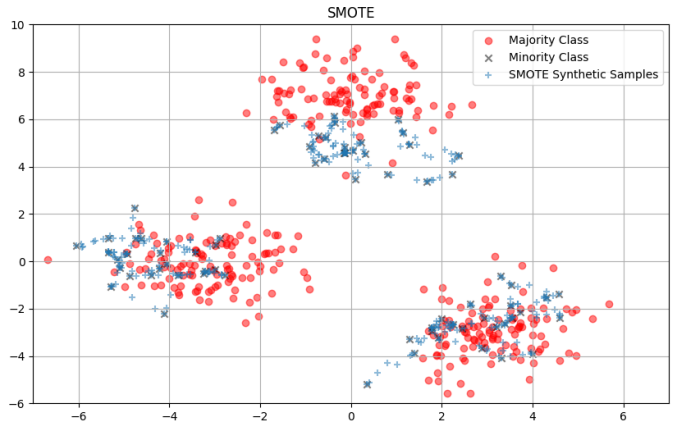
TMM



GMM



GMO-AC



SMOTE

Figure 8: Visual comparison of data distributions after each oversampling method (Case 3).

Case 3: Classification Performance Summary

Table 9. Classification performance metrics for Case 3

		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

CASE4: SUBCOMPONENT SELECTION VIA BIC

Table 10. Experimental data configuration (Case 4)

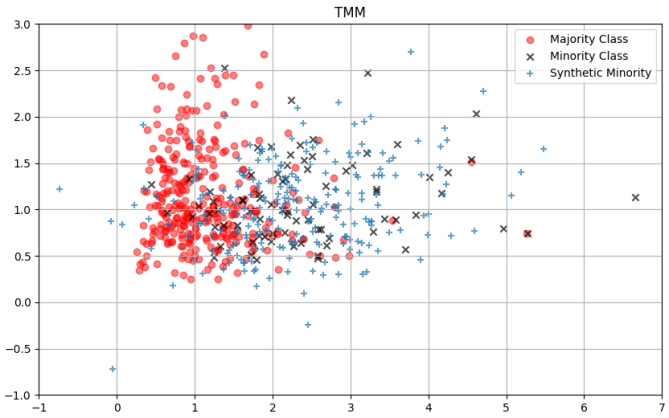
	Mean	Covariance	Proportion
Majority	$[0, 0]$	$\begin{bmatrix} 0.25 & 0 \\ 0 & 0.25 \end{bmatrix}$	80%
Minority	$[0.7, 0]$	$\begin{bmatrix} 0.25 & 0 \\ 0 & 0.25 \end{bmatrix}$	20%

- Case 4 simulation data were generated from log-normal distributions with the true parameters listed in Table 10, and both TMM and GMM models were fitted. The number of subcomponents selected by BIC is summarised in Table 2.

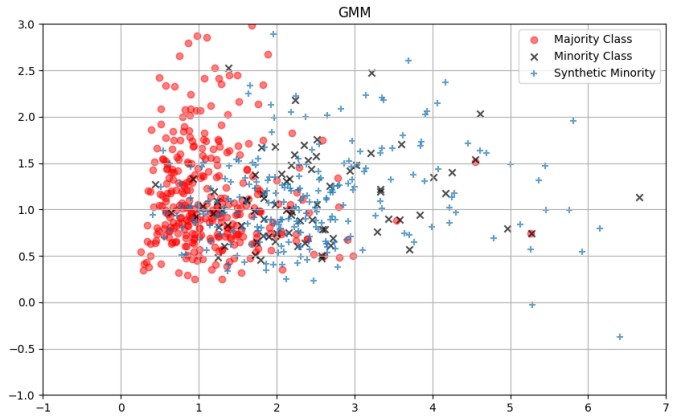
Table 11. BIC-based selection of subcomponents (Case 4)

n	Method	1	2	3	4	5
500	TMM	53%	38%	4%	3%	2%
	GMM	11%	49%	35%	2%	3%
1500	TMM	0%	74%	26%	0%	0%
	GMM	0%	25%	50%	19%	6%

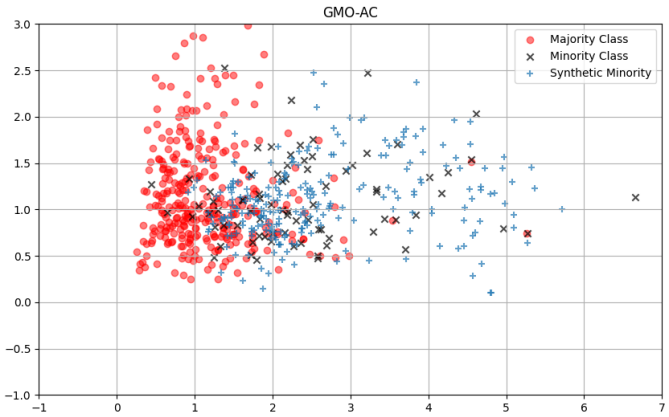
Case 4: Oversampling Visualization



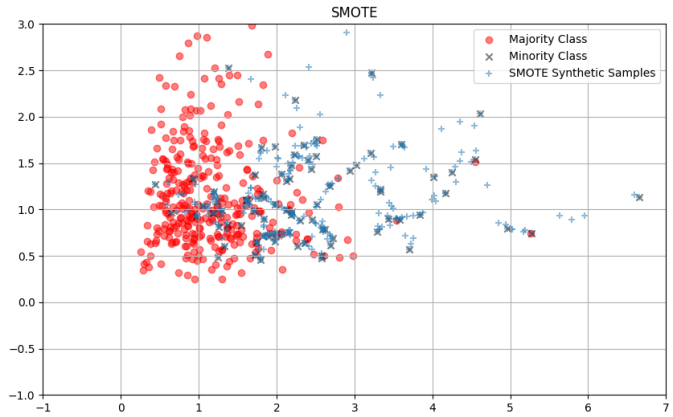
TMM



GMM



GMO-AC



SMOTE

Figure 9: Visual comparison of data distributions after each oversampling method (Case 4).

Case 4: Classification Performance Summary

Table 12. Classification performance metrics for Case 4

		SVM			Random Forest		
n	Oversampling	F1-Score	AUROC	G-mean	F1-Score	AUROC	G-mean
500	TMM	**0.570±0.069**	0.819±0.050	**0.756±0.056**	0.514±0.073	0.783±0.059	0.715±0.061
	GMM	0.560±0.068	**0.820±0.050**	0.751±0.057	0.509±0.072	0.783±0.062	0.713±0.062
	GMO-AC	0.559±0.064	0.816±0.050	0.753±0.053	**0.514±0.071**	**0.785±0.058**	**0.717±0.060**
	SMOTE	0.552±0.061	0.818±0.048	0.749±0.051	0.488±0.082	0.775±0.060	0.682±0.071
	No manipulation	0.424±0.120	0.754±0.071	0.543±0.105	0.458±0.103	0.783±0.058	0.600±0.086
1500	TMM	**0.560±0.041**	0.814±0.032	0.754±0.034	0.513±0.035	0.791±0.031	0.720±0.032
	GMM	0.558±0.040	0.815±0.032	0.754±0.034	0.510±0.032	0.790±0.030	0.718±0.028
	GMO-AC	0.556±0.037	0.813±0.032	**0.754±0.032**	**0.518±0.038**	**0.792±0.033**	**0.724±0.033**
	SMOTE	0.557±0.038	**0.818±0.031**	0.753±0.032	0.494±0.045	0.775±0.035	0.690±0.040
	No manipulation	0.449±0.058	0.748±0.045	0.753±0.032	0.459±0.056	0.782±0.034	0.601±0.047

PIMA INDIANS DIABETES DATASET SUMMARY

Target Variable	**Diabetes** (0 = non-diabetic, 1 = diabetic; 500 vs. 268 samples)
Continuous Variable	Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age
Imbalance Ratio	$\approx 1.87 : 1$

- Collected from health exams of adult women (age 21), originally used to study diabetes risk factors.
- Eight continuous features measuring glucose levels, blood pressure, body composition, insulin, pedigree score, and age.
- Binary classification : predict diabetes diagnosis (0 vs. 1).
- Moderate class imbalance motivates the use of oversampling techniques.

PIMA DATASET: BIC SELECTION & CLASSIFICATION PERFORMANCE

Table 13. BIC-based subcomponent counts (Pima diabetes dataset)

Method	1	2	3	4	5
TMM	100%	0%	0%	0%	0%
GMM	0%	88%	2%	4%	6%

Table 14. Classification performance summary (Pima diabetes dataset)

	SVM			**Random Forest**		
Method	F1-Score	AUROC	G-mean	F1-Score	AUROC	G-mean
TMM	0.633±0.040	0.807±0.032	0.713±0.033	**0.669±0.038**	0.820±0.029	**0.743±0.031**
GMM	0.632±0.039	0.807±0.032	0.712±0.033	0.664±0.037	0.820±0.029	0.738±0.030
GMO-AC	0.635±0.043	0.809±0.033	0.714±0.035	0.657±0.038	0.819±0.030	0.732±0.031
SMOTE	**0.641±0.043**	0.810±0.033	**0.719±0.036**	0.661±0.037	0.818±0.029	0.736±0.031
No manipulation	0.579±0.056	**0.812±0.034**	0.658±0.043	0.628±0.049	**0.821±0.029**	0.704±0.040

Abalone Dataset Summary

Target Variable	**Rings = 9** (majority, 689 samples) vs. **Rings = 18** (minority, 42 samples)
Continuous Variable	Length, Diameter, Height, WholeWeight, ShuckedWeight, VisceraWeight, ShellWeight
Imbalance Ratio	$\approx 16:1$

- Originally a regression dataset to predict abalone age (determined by shell ring count), it is here reformulated as a binary classification task to predict whether an abalone belongs to the **Rings = 9** class or the **Rings = 18** class.
- Seven continuous size/weight features.
- Severe class imbalance (16:1).

Abalone Dataset: BIC Selection & Classification Performance

Table 15. BIC-based subcomponent counts (Abalone dataset)

Method	2	3	4
TMM	100%	0%	0%
GMM	66%	29%	5%

Table 16. Classification performance summary (Abalone dataset)

	SVM			**Random Forest**		
Method	F1	AUROC	G-mean	F1	AUROC	G-mean
TMM	0.491±0.093	**0.935±0.038**	0.839±0.073	0.384±0.091	**0.882±0.058**	**0.751±0.102**
GMM	0.493±0.091	**0.935±0.038**	**0.839±0.072**	0.386±0.095	0.879±0.059	0.747±0.105
GMO-AC	**0.502±0.109**	0.930±0.041	0.835±0.073	**0.387±0.099**	0.862±0.063	0.729±0.100
SMOTE	0.475±0.084	0.931±0.039	0.837±0.066	0.359±0.119	0.843±0.068	0.600±0.131
No manipulation	0.104±0.110	0.890±0.060	0.167±0.175	0.275±0.177	0.833±0.076	0.374±0.199

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 method preserves the global data structure and reinforces minority class centers even under class overlap.
- Across four simulations and two real-world 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.

Burgess-Hull, A. J. (2020). Finite mixture models with student t distributions: an applied example. *Prevention Science*, 21(6):872–883.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Douzas, G., Bacao, F., and Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and smote. *Information sciences*, 465:1–20.

He, H., Bai, Y., Garcia, E. A., and Li, S. (2008). Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. Ieee.

Holte, R. C., Acker, L., Porter, B. W., et al. (1989). Concept learning and the problem of small disjuncts. In *IJCAI*, volume 89, pages 813–818.

Jo, T. and Japkowicz, N. (2004). Class imbalances versus small disjuncts. *ACM Sigkdd Explorations Newsletter*, 6(1):40–49.

McLachlan, G. J. and Krishnan, T. (2007). *The EM algorithm and extensions*. John Wiley & Sons.

McLachlan, G. J., Lee, S. X., and Rathnayake, S. I. (2019). Finite mixture models. *Annual review of statistics and its application*, 6:355–378.

Santos, M. S., Abreu, P. H., Japkowicz, N., Fernández, A., and Santos, J. (2023). A unifying view of class overlap and imbalance: Key concepts, multi-view panorama, and open avenues for research. *Information Fusion*, 89:228–253.

Santos, M. S., Abreu, P. H., Japkowicz, N., Fernández, A., Soares, C., Wilk, S., and Santos, J. (2022). On the joint-effect of class imbalance and overlap: a critical review. *Artificial Intelligence Review*, 55(8):6207–6275.

Vuttipittayamongkol, P., Elyan, E., and Petrovski, A. (2021). On the class overlap problem in imbalanced data classification. *Knowledge-based systems*, 212:106631.

Weiss, G. M. (1995). Learning with rare cases and small disjuncts. In *Machine learning proceedings 1995*, pages 558–565. Elsevier.

Yang, S. J. and Cha, K. (2024). Gmo-ac: Gaussian-based minority oversampling with adaptive outlier filtering and class overlap weighting. *IEEE Access*.

Yang, S. J. and Cha, K. J. (2021). Gmote: Gaussian based minority oversampling technique for imbalanced classification adapting tail probability of outliers. *arXiv preprint arXiv:2105.03855*.