Algoritmo: αDistance Borderline-ADASYN-SMOTE Mejorado

Algoritmo: αDistance Borderline-ADASYN-SMOTE Mejorado

Referencia base: Feng & Li, 2021.
Mejora incorporada: filtrado adicional de muestras ruidosas basado en umbral de α'_n.

1. **División de vecinos:**
   Para cada muestra minoritaria p_i, obtener sus m vecinos más cercanos y dividirlos en:
   - Vecinos de clase minoritaria: cantidad $pnum$
   - Vecinos de clase mayoritaria: cantidad $nnum$

2. **Cálculo de pesos inversos:**
   Para cada vecino p_j de p_i, calcular el peso basado en la distancia:

$$\alpha_j = \frac{1}{\text{dist}(p_i, p_j)}$$

3. **Suma de pesos por clase:**

$$\alpha'_p = \sum \alpha_j \text{ (de vecinos minoritarios)}, \quad \alpha'_n = \sum \alpha_j \text{ (de vecinos mayoritarios)}$$

4. **Identificación inicial de muestras peligrosas:**
   Si $\alpha'_n > \alpha'_p$, entonces p_i se considera una muestra *potencialmente peligrosa*.

5. **Filtrado de muestras ruidosas (mejora propuesta):**
   Calcular el umbral de ruido θ como el percentil 25 de todos los α'_n del conjunto de muestras minoritarias.
   Si $\alpha'_n < \theta$, entonces p_i se considera una muestra *ruidosa* y se descarta del conjunto peligroso.

   Ejemplo:
   Supongamos que para un subconjunto de muestras minoritarias, los valores de α'_n son:

$$\alpha'_n = [0.10,\ 0.15,\ 0.20,\ 0.25,\ 0.30,\ 0.40,\ 0.50,\ 0.70,\ 0.90]$$

La cantidad total de valores es $n = 9$. Para calcular el percentil 25 (primer cuartil), usamos la fórmula:

$$P_{25} = \frac{25}{100} \cdot (n+1) = 0.25 \cdot 10 = 2.5$$

El percentil 25 se encuentra entre las posiciones 2 y 3 (valores 0.15 y 0.20), por lo tanto:

$$\theta = 0.15 + 0.5 \cdot (0.20 - 0.15) = 0.175$$

Cualquier muestra con $\alpha'_n < 0.175$ será considerada *ruidosa* y no se utilizará para la generación de muestras sintéticas.

Por lo tanto, los valores $\alpha'_n = 0.10$ y 0.15, al ser menores que $\theta = 0.175$, corresponden a muestras *ruidosas*. Estas muestras serán descartadas del conjunto de muestras peligrosas, y no participarán en los siguientes pasos de generación sintética.

6. **Cálculo total de ejemplos sintéticos:**

$$G = (N - n) \cdot \beta$$

Donde:

- N: número de muestras de la clase mayoritaria
- n: número de muestras de la clase minoritaria
- $\beta \in [0, 1]$: proporción deseada de balance.

7. **Distribución proporcional del total:**
Para cada muestra *peligrosa no ruidosa* p_i, calcular:

$$r_i = \frac{\Delta_i}{m}, \quad \hat{r}_i = \frac{r_i}{\sum r_i}, \quad g_i = \hat{r}_i \cdot G$$

Donde:

- Δ_i: número de vecinos mayoritarios de p_i
- m: cantidad total de vecinos
- r_i: proporción de vecinos mayoritarios para p_i
- $\hat{r}_i$: proporción normalizada
- g_i: cantidad de muestras sintéticas a generar para p_i

8. **Generación de muestras sintéticas:**
Para cada p_i, generar g_i muestras sintéticas mediante interpolación:

$$s = p_i + \lambda \cdot (p_z - p_i), \quad \lambda \in [0, 1]$$

Donde p_z es un vecino minoritario aleatorio de p_i.