Algoritmo Propuesto: Sobremuestreo Híbrido αDBASMOTE + AR-ADASYN

Algoritmo: Sobremuestreo Híbrido αDBASMOTE + AR-ADASYN

Fusión: αDBASMOTE (Feng & Li, 2021) + AR-ADASYN (Park & Kim, 2024)

1. **Identificación de muestras peligrosas:**
   Para cada muestra minoritaria p_i, obtener sus m vecinos más cercanos. Calcular pesos inversos:
   $$\alpha_j = \frac{1}{\text{dist}(p_i, p_j)}$$
   Dividir vecinos en mayoritarios/minoritarios y calcular:
   $$\alpha'_p = \sum \alpha_j \text{ (min)}, \quad \alpha'_n = \sum \alpha_j \text{ (may)}$$
   Si $\alpha'_n > \alpha'_p$, se considera muestra *peligrosa*.

2. **Filtrado de ruido (mejora propia):**
   Se observa que algunas muestras minoritarias, aunque se ubiquen cerca de vecinos mayoritarios, tienen a sus vecinos minoritarios demasiado alejados. Esta condición sugiere que dichas muestras no representan adecuadamente a la clase minoritaria en su vecindad y podrían introducir ruido si se usan para generar datos sintéticos.

   Para mitigar este efecto, se propone un filtrado basado en percentiles: se calcula el percentil 25 de todos los valores α'_n, que representan la suma de pesos inversos de los vecinos mayoritarios. Aquel p_i cuya α'_n se encuentre por debajo de este umbral será considerada una muestra *ruidosa* y se descartará del proceso de generación. Esto permite excluir las instancias menos confiables de forma estadísticamente fundamentada.

3. **Cálculo global de sintéticos:**
   $$G = (N - n) \cdot \beta$$
   donde N: mayoritarios, n: minoritarios, $\beta \in [0, 1]$: proporción deseada.

4. **Distribución proporcional del total:**
   Para cada muestra peligrosa 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 es el número de vecinos mayoritarios.

5. **Definición del área segura (AR-ADASYN):**
   Para cada p_i, seleccionar dos vecinos minoritarios x_{nn1}, x_{nn2}, calcular:

$$\theta' = \arccos\left(\frac{(x_{nn1} - p_i) \odot (x_{nn2} - p_i)}{\|x_{nn1} - p_i\| \cdot \|x_{nn2} - p_i\|}\right)$$

$$\theta = \min(\theta', \pi - \theta')$$

$$r = \max(\|x_{nn1} - p_i\|, \|x_{nn2} - p_i\|)$$

6. **Generación de datos sintéticos (AR-ADASYN):**
   Para cada $j = 1, \ldots, g_i$:

   - Elegir ángulo aleatorio $\alpha \in [0, \theta]$
   - Construir matriz de rotación:

$$R(\alpha) = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}$$

   - Calcular vector base $v = x_{nn1} - p_i$
   - Elegir radio aleatorio $\beta \in [0, r]$
   - Generar:

$$x_{\text{syn}}^{(j)} = p_i + \beta \cdot R(\alpha) \cdot \frac{v}{\|v\|}$$