Uso de las dCF-integrales en sistemas de clasificación basados en reglas difusas para abordar problemas no balanceados

José Antonio Sanz Delgado, Mikel Sesma Sara

Introducción

Los **problemas de clasificación no balanceados binarios** son problemas de dos clases

- Clase mayoritaria (negativa): muchos ejemplos
- Clase minoritaria (positiva): pocos ejemplos

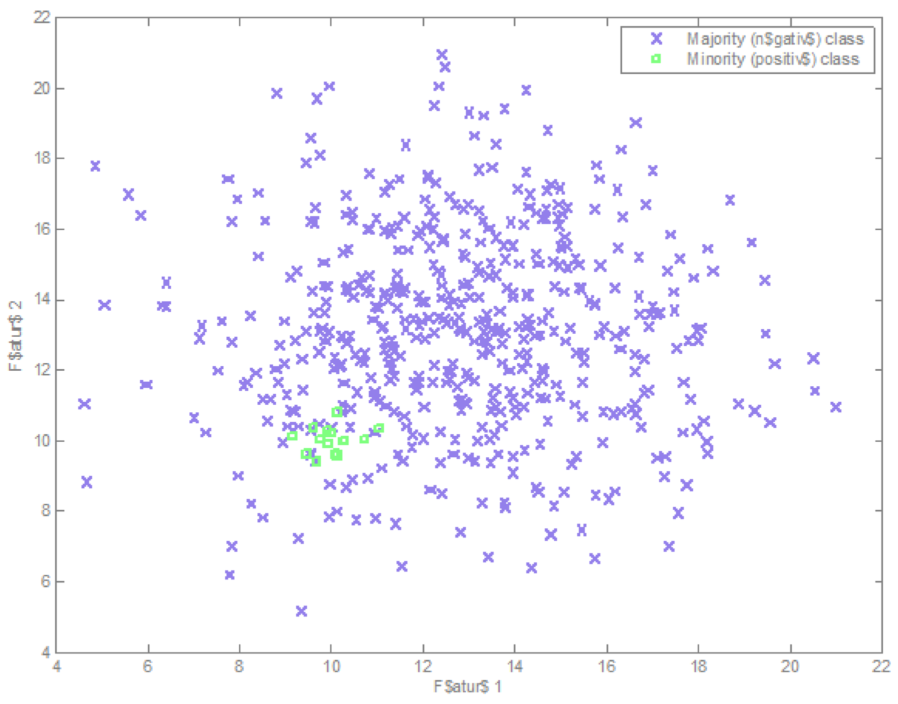

Figura 1: Problema de clasificación no balanceado

Dificultades para el aprendizaje del modelo

- Conjuntos pequeños de instancias (small disjuncts)
- Solapamiento entre las clases
- Tendencia a aprender el concepto de la clase mayoritaria
- Métricas: porcentaje de acierto (accuracy) puede acarrear tomar decisiones erróneas

Metodologías para abordar problemas de clasificación no balanceados

- Soluciones a nivel de datos (muestreo)
- Soluciones a nivel algorítmico
  - Modificación interna del modelo
  - Sensibles al coste
  - Ensembles

Sistemas de Clasificación Basado en Reglas Difusas (SCBRDs)

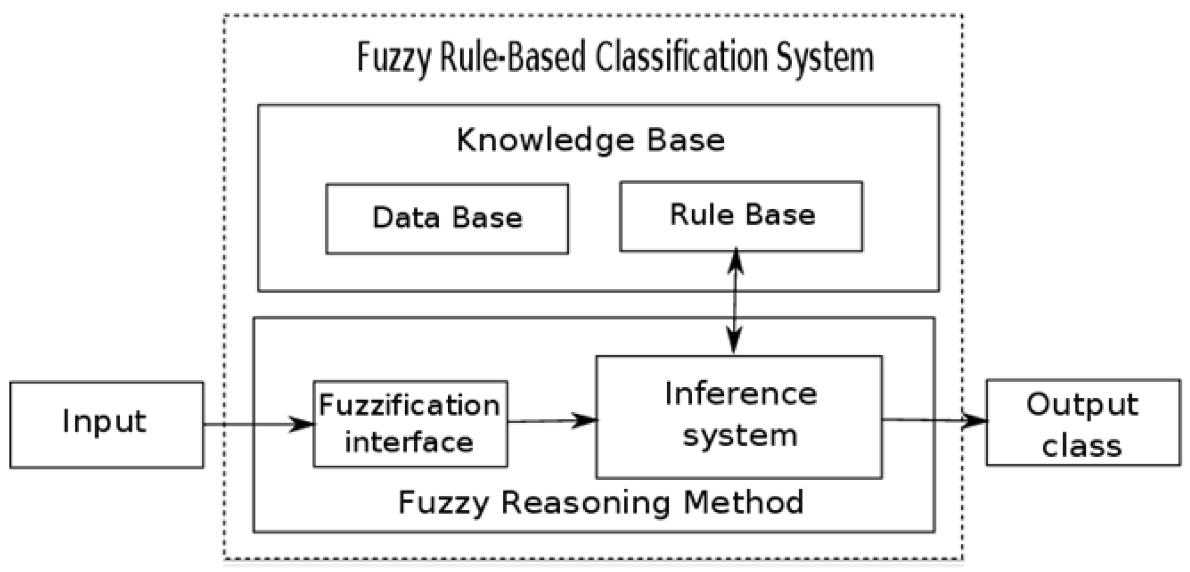

Figura 2: Esquema de un SCBRD

- Estructura de las reglas

Regla R_j : SI x_1 es A_{j1} y . . . y x_n es A_{jn} ENTONCES Clase = C_j con RW_j

- **Normalmente**, tras el aprendizaje del SCBRDs, la **base de reglas está compuesta por**
  - **Clase mayoritaria: muchas reglas y generales**
    - Cortas (pocos antecedentes)
  - **Clase minoritaria: pocas reglas y específicas**
    - Largas (muchos antecedentes)

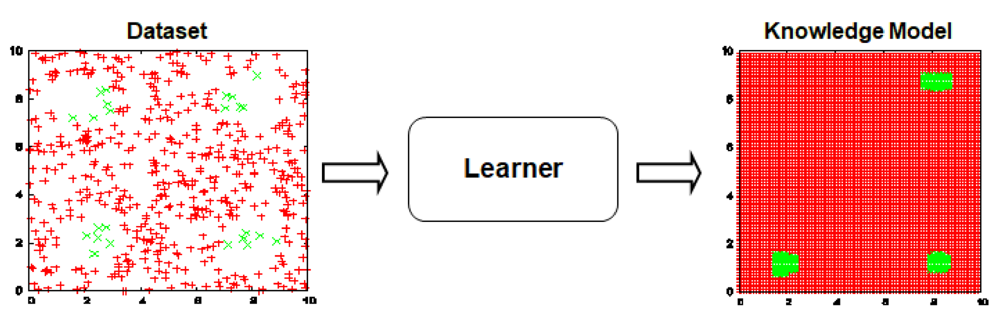

Figura 3: Espacio de entrada cubierto por las reglas de un SCBRD

Motivación

FARCI [1] (Fuzzy Association Rule-based Classifier for Imbalanced classification problems) **es un SCBRDs**, basado en FARC-HD, **que aborda problemas de clasificación no balanceados sin utilizar técnicas de muestreo**. Para ello, modifica los siguientes aspectos de FARC-HD

1. **Aprendizaje de reglas de asociación**: utiliza el **lift**
2. **Selección de reglas**: se **favorece** a las reglas de la **clase positiva**
3. **Proceso evolutivo**: **F1-score** como función de fitness
4. **Grado de emparejamiento**: se utiliza la **media geométrica (MG)** para afrontar el problema de reglas cortas contra reglas largas
   - Ejemplo
     - Regla A con 1 antecedente
       - Grado de Pertenencia (GP): 0.5
     - Regla B con 3 antecedentes
       - GPs: 0.8, 0.8 y 0.75
     - Cálculo del grado de emparejamiento
       - Producto
         - Regla A: 0.5
         - Regla B: 0.48
       - Media Geométrica
         - Regla A: 0.5
         - Regla B: 0.78

Método de razonamiento difuso de FARCI

$$\widehat{y_p} = \underset{k\in\{1,\dots,C\}}{\arg\max} \left(\sum_{\substack{j=1 \\ Class\,(R_j=k)}}^{R} \mu_{A_j}(e_p) \times RW_j \right)$$

$$\mu_{A_j}(e) = MG\left(\mu_{A_{k1}}(e_1), \dots, \mu_{A_{kn}}(e_n)\right)$$

- El **uso de la suma puede acarrear malas decisiones cuando hay muchas reglas disparadas de una clase y pocas de la otra**
  - Ejemplo
    - **Clase minoritaria** con 3 reglas disparadas cuyos valores a sumar son
      - Regla A: 0.2
      - Regla B: 0.25
      - Regla C: 0.3
    - **Clase mayoritaria** con 1 regla disparada cuyo valor a sumar es
      - Regla D: 0.7
    - Suma (grado de asociación por clases)
      - Clase minoritaria: 0.75
      - Clase mayoritaria: 0.7

Propuesta

Sustituir la suma por dCF-integrales [2] para obtener la información global asociada a las clases.

Las dCF-integrales son funciones basadas en la integral Choquet

$$\mathcal{C}_m(x) = \sum_{i=1}^{n} \left(x_{(i)} - x_{(i-1)}\right) \times m\left(A_{(i)}\right)$$

En las que **se sustituye**

- **La resta por una función de disimilitud restringida: δ**
- **El producto por una función bivariada F**

$$\mathcal{C}_{F,m,\delta}(x) = x_{(1)} + \sum_{i=2}^{n} F\left(\delta\left(x_{(i)}, x_{(i-1)}\right), m\left(A_{(i)}\right)\right)$$

Marco Experimental

- 44 datasets (IR>9) aplicando 5fcv
- 19 dCF-integrales [2]
  - $m(A) = \left(\frac{|A|}{n}\right)^q$, con $q > 0$
  - $\delta(x,y) = \sqrt{|x-y|}$
  - 19 funciones F
- F1-score
- Test de Wilcoxon

Resultados

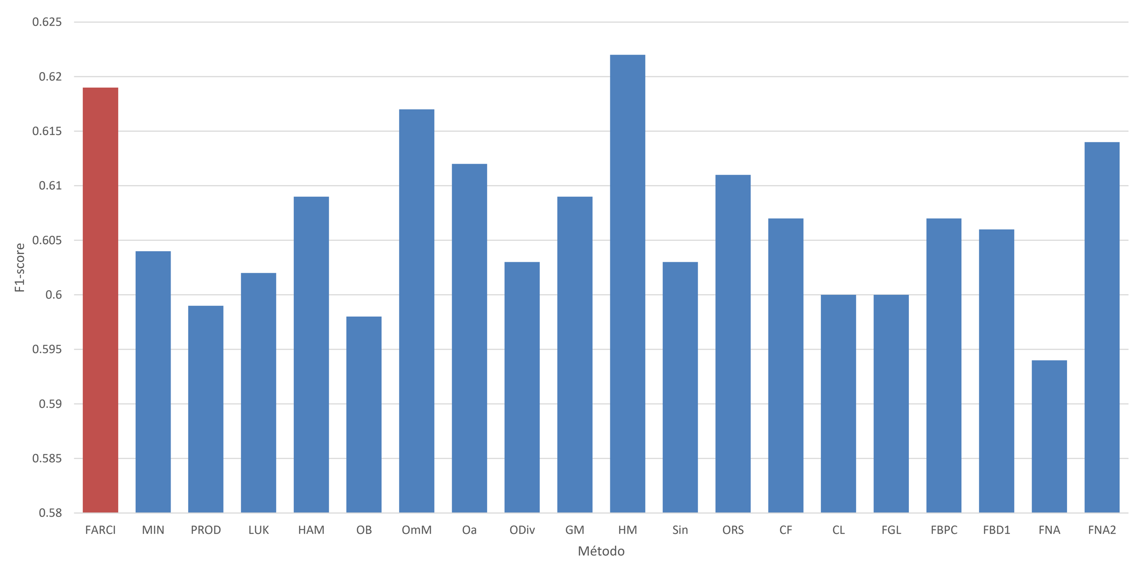

Figura 4: F1-score medio en test de cada método

	FARCI	MIN	PROD	LUK	HAM	O_B	O_{mM}	O_a	O_{Div}	GM	HM	Sin	O_{RS}	C_F	C_L	F_{GL}	F_{BPC}	F_{BD1}	F_{NA}	F_{NA2}
Media	0.619	0.604	0.599	0.602	0.609	0.598	0.617	0.612	0.603	0.609	0.622	0.603	0.611	0.607	0.600	0.600	0.607	0.606	0.594	0.614
Victorias	6	3	4	3	3	2	6	3	3	4	7	4	7	3	3	8	5	6	1	5

Tabla 1: Resultados detallados de todos los métodos

	MIN	PROD	LUK	HAM	O_B	O_{mM}	O_a	O_{Div}	GM	HM	Sin	O_{RS}	C_F	C_L	F_{GL}	F_{BPC}	F_{BD1}	F_{NA}	F_{NA2}
P-valor	<0.01	<0.01	0.04	0.199	0.020	0.784	0.302	0.434	0.149	0.866	0.134	0.636	0.080	0.165	0.241	0.241	0.121	<0.01	0.923
R+	270.5	269.5	320.5	385.0	294.5	471.5	406.5	428.0	371.5	509.5	366.5	454.5	346.5	376.0	394.5	394.5	362.0	271.5	487.0
R-	719.5	720.5	669.5	605.0	695.5	518.5	583.5	562.0	618.5	480.5	623.5	535.5	643.5	614.0	595.5	595.5	628.0	718.5	503.0

Tabla 2: Test de Wilcoxon comparando cada función F (R+) contra FARCI (R-)

Conclusiones y trabajo futuro

- Resultados competitivos
  - Pero menos de lo esperado
  - Algunas funciones F ofrecen potencial
    - HM, F_{NA2}, O_{mM} y O_{RS}
- $\delta(x,y) = \left(\sqrt{x} - \sqrt{y}\right)^2$ obtuvo peores resultados
- Trabajo futuro: D-XC-integrales [3]

Referencias

1. J. Sanz, M. Sesma-Sara, H. Bustince. "A fuzzy association rule-based classifier for imbalanced classification problems", Information Sciences, 577, 265-279, 2021.
2. J. Wieczynski, G. Lucca, G. P. Dimuro, E. N. Borges, J. A. Sanz, T. d. C. Asmus, J. Fernández, and H. Bustince, "dCF -integrals: Generalizing cF-integrals by means of restricted dissimilarity functions," IEEE Transactions on Fuzzy Systems, vol. 31, no. 1, pp. 160–173, 2023.
3. J. Wieczynski, J. Fumanal-Idocin, G. Lucca, E. N. Borges, T. D. C. Asmus, L. R. Emmendorfer, H. Bustince, and G. P. Dimuro, "D-XC Integrals: On the generalization of the expanded form of the choquet integral by restricted dissimilarity functions and their applications," IEEE Transactions on Fuzzy Systems, vol. 30, no. 12, p. 5376 – 5389, 2022.

arin
Artificial Intelligence & Machine Learning Research

José Antonio Sanz – **joseantonio.sanz@unavarra.es**

Departamento de Estadística, Informática y Matemáticas, Universidad Pública de Navarra

upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa