Uso de las $d\mathrm{C}F$-integrales en sistemas de clasificación basados en reglas difusas para abordar problemas no balanceados

José Antonio Sanz y Mikel Sesma-Sara
Departamento de Estadística, Informática y Matemáticas
Universidad Pública de Navarra
Pamplona, España
joseantonio.sanz, mikel.sesma@unavarra.es

***Resumen*—Los problemas de clasificación no balanceados son muy habituales en el mundo real y suponen un reto para los sistemas de aprendizaje puesto que tienden a ignorar la clasificación de los ejemplos de la clase minoritaria, que generalmente es la de mayor interés. Existen numerosas formas de abordalos, pero en esta contribución utilizamos un sistema de clasificación basado en reglas difusas, llamado FARCI, que afronta estos problemas sin necesidad de aplicar técnicas de muestreo de datos. Nuestra propuesta es modificar la fase de agregación de la información dada por las reglas disparadas en el proceso de inferencia. En concreto, proponemos utilizar las $d\mathrm{C}F$-integrales, que son una generalización de la integral Choquet, que además de reemplazar el producto por una función F, también cambia la diferencia por una función de disimilitud restringida, d.**

***Index Terms*—Clasificación no balanceada, sistemas de clasificación basados en reglas difusas, funciones de agregación, generalizaciones de la integral Choquet.**

I. Introducción

Los problemas de clasificación no balanceados representan un gran desafío en aprendizaje automático, ya que la distribución de clases en los datos no es homogénea. Esta situación aparece con frecuencia en contextos críticos para la sociedad, como el diagnóstico médico, la detección de fraudes o la predicción de fallos en sistemas. En tales casos, la clase minoritaria suele ser la más relevante, por lo que es esencial diseñar métodos de clasificación que manejen adecuadamente este desequilibrio para no comprometer la toma de decisiones.

Los sistemas de clasificación basados en reglas difusas (SCBRD) han demostrado ser herramientas eficaces para resolver problemas de clasificación gracias a su capacidad para representar el conocimiento de forma interpretable y gestionar la incertidumbre presente en los datos. Estos sistemas permiten generar reglas lingüísticas que facilitan la comprensión del modelo por parte de expertos humanos, a la vez que ofrecen resultados competitivos en términos de precisión.

Una propuesta reciente dentro de los SCBRD para abordar problemas de clasificación no balanceados es el método llamado FARCI [1] (Fuzzy Association Rule-based Classification for Imbalanced datasets), una extensión del clasificador FARC-HD [2] específicamente adaptada para manejar el desbalanceo de clases. En FARCI se modifican todas las fases del proceso de aprendizaje de FARC-HD para mejorar su rendimiento en contextos no balanceados, logrando resultados competitivos frente a otros métodos del estado del arte.

Un componente fundamental de FARCI es su método de inferencia difusa, llamado de combinación aditiva, en el que usa la suma normalizada como función de agregación para fusionar la información de todas las reglas disparadas durante la clasificación de un ejemplo. Es decir, se suman los grados de asociación de todas las reglas disparadas por cada clase, y se selecciona aquella con el valor agregado más alto. Esta estrategia resulta sencilla e intuitiva, pero puede generar sesgos cuando existe un número significativamente mayor de reglas disparadas de una clase respecto a la otra, lo cual es frecuente en escenarios de clasificación no balanceada.

Dado este posible sesgo, surge la hipótesis de que el uso de funciones de agregación alternativas podría mejorar el comportamiento del sistema. En este sentido, se han propuesto numerosas generalizaciones de la integral Choquet [3], que permiten una fusión más flexible y controlada de la información. Estas generalizaciones incluyen tanto funciones de agregación como de pre-agregación y han sido ampliamente estudiadas en la literatura reciente para abordar problemas de clasificación estándar. Recientemente, se ha desarrollado una nueva generalización de la integral Choquet, denominada $d\mathrm{C}F$-integral [4], basada en funciones de disimilitud restringidas. En concreto, se sustituye la diferencia clásica en la integral de Choquet por una función de disimilitud restringida.

En este trabajo proponemos el uso de las $d\mathrm{C}F$-integrales para reemplazar la suma normalizada utilizada actualmente en el método de inferencia de FARCI. Concretamente, utilizamos las dos mejores funciones de disimilitud y 19 de las 21 funciones de agregación presentadas en [4]. Evaluamos el rendimiento de la nueva propuesta en el mismo marco experimental que el utilizado en [1], pero centrado únicamente en los 44 conjuntos de datos con un elevado ratio de no balanceo ($IR \geq 9$).

Partially funded by MCIN/AEI/10.13039/501-00011033/ FEDER,UE (PID2022-136627NB-I00).

II. Propuesta

En este trabajo utilizamos FARCI [1] como SCBRD, que está basado en FARC-HD [2] que es un clasificador diseñado para abordar problemas de clasificación estándar. Para abordar problemas de clasificación no balanceados, en FARCI se propuso el uso de funciones de agregación promedio para modelar la intersección de los antecedentes de las reglas y además se modificaron las 3 etapas de aprendizaje de FARC-HD:

- En el algoritmo A-priori reemplaza la confianza como medida de calidad de las reglas por el $lift$.
- Se modifica la filosofía del uso del esquema de ponderación de ejemplos, aplicada para seleccionar las reglas más interesantes, para potenciar la selección de reglas de la clase minoritaria.
- El algoritmo genético utiliza el F-score en la función de $fitness$, puesto que es una métrica apropiada para este tipo de problemas.

En la fase de inferencia, FARCI utiliza la combinación aditiva como método de razonamiento difuso. Es decir, para agregar la información local de las reglas disparadas y obtener la información global (por clases), se aplica como función de agregación la suma normalizada. Posteriormente, se predice la clase asociada al valor agregado más alto.

Sin embargo, en la práctica se disparan más reglas de la clase positiva; por ejemplo, en la mejor configuración de FARCI el número de reglas generadas de la clase positiva casi duplica al de la negativa. Este hecho puede provocar que el uso de la suma no sea equitativo para ambas clases, ya que el valor agregado de la clase positiva tiende a ser mayor incluso en situaciones en las que los valores, de manera individual, sean inferiores a los de la clase negativa por el hecho de tener más reglas disparadas de dicha clase.

Por ello, en este trabajo proponemos el uso de las $d\mathbf{C}F$-integrales en vez de la suma normalizada para obtener la información global en el proceso de inferencia ya que su carácter no promedio puede mejorar el comportamiento del sistema en estas situaciones.

III. Estudio Experimental

Para evaluar el rendimiento de la nueva propuesta utilizamos la mejor medida difusa (cardinalidad exponencial), las dos mejores funciones de disimilitud restringidas (δ_2 y δ_5) y probaremos 19 de las 21 funciones de agregación presentadas en la Tabla II de [4] (no utilizamos el producto drástico ni el mínimo nilpotente), numeradas correlativamente. Utilizamos el mismo estudio experimental que el utilizado en [1] pero nos centramos en los 44 datasets con un ratio de no balanceo alto ($IR \geq 9$).

La Tabla I resume los resultados. Cada fila corresponde a una de las 19 funciones de agregación (F) y las columnas muestran los valores obtenidos para $d = \delta_2$, ya que con $d = \delta_5$ el rendimiento fue inferior. Para cada F se muestran la media en test (Med. Tst) del F-score calculada sobre los 44 conjuntos de datos[1], junto con el número de ellos en los que logra el mejor resultado; y el p-valor (p-val) del test de Wilcoxon que compara FARCI (rango positivo, R+) con la función F (rango negativo, R-).

Tabla I
Resultados del estudio experimental utilizando $d = \delta_2$. En **negrita** está resaltado el mejor resultado.

F	Med. Tst (Vict.)	p-val (R+;R-)
1	0.604 (3)	$<$ 0,01 (270.5;719.5)
2	0.599 (4)	$<$ 0,01 (269.5;720.5)
3	0.602 (3)	0.04 (320.5;669.5)
4	0.609 (3)	0.199 (385.0;605.0)
5	0.598 (2)	0.02 (294.5;695.5)
6	0.617 (6)	0.784 (471.5;518.5)
7	0.612 (3)	0.302 (406.5;583.5)
8	0.603 (3)	0.434 (428.0;562.0)
9	0.609 (4)	0.149 (371.5;618.5)
10	**0.622** (7)	0.866 (509.5;480.5)
11	0.603 (4)	0.134 (366.5;623.5)
12	0.611 (7)	0.636 (454.5;535.5)
13	0.607 (3)	0.08 (346.5;643.5)
14	0.600 (3)	0.165 (376.0;614.0)
15	0.600 (8)	0.241 (394.5;595.5)
16	0.607 (5)	0.241 (394.5;595.5)
17	0.606 (6)	0.121 (362.0;628.0)
18	0.594 (1)	$<$ 0,01 (271.5;718.5)
19	0.614 (5)	0.923 (487.0;503.0)
FARCI	0.619 (6)	-

IV. Conclusiones

Los resultados experimentales muestran que el uso de HM (10), F_{NA2} (19), O_{mM} (6) y de O_{RS} (12) como funciones F en las $d\mathbf{C}F$-integrales, al usar $d = \delta_2$, tienen potencial para ser utilizadas en el proceso de inferencia de FARCI. Al utilizar $d = \delta_5$, los resultados, en general, han sido inferiores aunque algunas funciones F también pueden presentar cierto potencial. Por ello, el uso de las $d\mathbf{C}F$-integrales podrá ser la base de futuros estudios en los que se trate de mejorar el rendimiento de este clasificador utilizando generalizaciones de la integral Choquet.

Referencias

[1] J. Sanz, M. Sesma-Sara, H. Bustince, A fuzzy association rule-based classifier for imbalanced classification problems, Information Sciences 577 (2021) 265–279.

[2] J. Alcala-Fdez, R. Alcala, F. Herrera, A fuzzy association rule-based classification model for high-dimensional problems with genetic rule selection and lateral tuning, IEEE Transactions on Fuzzy Systems 19 (5) (2011) 857–872.

[3] G. P. Dimuro, J. Fernández, B. Bedregal, R. Mesiar, J. A. Sanz, G. Lucca, H. Bustince, The state-of-art of the generalizations of the choquet integral: From aggregation and pre-aggregation to ordered directionally monotone functions, Information Fusion 57 (2020) 27–43.

[4] J. Wieczynski, G. Lucca, G. P. Dimuro, E. N. Borges, J. A. Sanz, T. d. C. Asmus, J. Fernández, H. Bustince, dc_F-integrals: Generalizing c_F-integrals by means of restricted dissimilarity functions, IEEE Transactions on Fuzzy Systems 31 (1) (2023) 160–173.

[1] Los resultados detallados se pueden ver en https://github.com/JoseanSanz/FARCI_Choquet_integral_generalizations