RIOWA: operadores IOWA restringidos aplicados a imputación de datos faltantes

A. Indurain-Ibero[1,2][0009−0003−8456−5542], B. Bedregal[3][0000−0002−6757−7934], M. Ferrero-Jaurrieta[2][0000−0002−6854−3437], J. Fernández[2][0000−0003−4427−3935], and H. Bustince[2][0000−0002−1279−6195]

[1] Tracasa Instrumental, C. Cabárceno 6, 31621 Sarriguren, España
[2] Universidad Pública de Navarra, 31006 Pamplona, España {alfonso.indurain, mikel.ferrero, fcojavier.fernandez, bustince}@unavarra.es
[3] Fusión de datos Operado, 59072-970 Natal, Brazil bedregal@dimap.ufrn.br

Abstract. El presente documento es un *key work* del trabajo titulado 'RIOWA: Restricted Induced Ordered Weighted Averaging operators, applied to missing data imputation' [1] y que ha sido publicado en la revista *Applied Soft Computing*. En este *key work* presentamos unos nuevos operadores basados en los operadores IOWA, los aplicamos en un problema de imputación de datos faltantes y demostramos la mejora que ofrecen frente a los operadores OWA e IOWA clásicos.

Keywords: Fusión de datos · Operador OWA · Operador IOWA · Imputación de datos

1 Introducción

Los operadores Ordered Weighted Averaging (OWA) y sus extensiones, como los operadores Induced OWA (IOWA), se utilizan ampliamente en la fusión de datos porque permiten asignar diferentes niveles de importancia a cada dato. Esta característica resulta especialmente útil en aplicaciones como la imputación de datos faltantes.

Sin embargo, cuando existen datos duplicados, los operadores clásicos no pueden distinguir si esas repeticiones son reales o consecuencia de problemas de calidad en los datos. Esto puede generar sesgos hacia los valores repetidos y afectar negativamente la representación de la distribución real de los datos.

Para solucionar este problema, en este trabajo se propone una nueva familia de operadores denominada Restricted IOWA (RIOWA), que trabaja únicamente con valores únicos. Además, se estudian sus propiedades fundamentales y se evalúa su aplicación en la imputación de datos faltantes, comparando su rendimiento con los operadores OWA e IOWA tradicionales.

2 Definiciones previas

Dado $n \in \mathbb{N}$, tal que $n \geq 2$, $N_n = \{1, \ldots, n\}$ y $\mathbf{x} = (x_1, \ldots, x_n) \in [0,1]^n$. Un vector $\mathbf{w} = (w_1, \ldots, w_n) \in [0,1]^n$ es un vector de pesos si $\sum_{i=1}^{n} w_i = 1$ y es un vector de subpesos si $\sum_{i=1}^{n} w_i \leq 1$.

Definition 1 ([2]) *Sea* $\mathbf{w} \in [0,1]^n$ *un vector de pesos, el operador Ordered Weighted Averaging (OWA), es la función* $O_{\mathbf{w}} : [0,1]^n \to [0,1]$ *definida para cada* $\mathbf{x} \in [0,1]^n$ *por:* $O_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{n} w_i x_{(i)}$, *donde* $(\cdot)$ *es una permutación decreciente de* $\mathbf{x}$.

Definition 2 ([3]) *Sea* $\mathbf{w} \in [0,1]^n$ *un vector de pesos,* $\mathbf{u} \in \mathbb{R}^n$, *y* $\mathbf{a} \in [0,1]^n$. *El operador IOWA asociado a* $\mathbf{w}$ *es una aplicación* $I_{\mathbf{w}} : (\mathbb{R} \times [0,1])^n \to [0,1]$ *tal que:* $I_{\mathbf{w}}(\langle u_1, a_1\rangle, ..., \langle u_n, a_n\rangle) = \sum_{i=1}^{n} w_i a_{\sigma(i)}$ *donde* σ *es cualquier permutación de* $\{1, ..., n\}$ *tal que* $u_{\sigma(i)} \geq u_{\sigma(j)}$ *cuando* $i < j$.

3 RIOWA: operador OWA restringido inducido

Sea $n \geq 2$. Denotamos $\mathcal{V} = \bigcup_{m=1}^{n} \mathcal{V}_m$, donde $\mathcal{V}_m = \{(x_1, \ldots, x_m) \in [0,1]^m \mid x_i \neq x_j$ para todo $i \neq j \in N_m\}$. Un vector inductor para $\mathbf{x}' \in \mathcal{V}_m$ es un vector $\mathbf{A} = (a_1, \ldots, a_m) \in N_n^m$ tal que $\sum_{i=1}^{m} a_i = n$. Denotamos con $\mathcal{A}_{\mathbf{x}'}$ al conjunto de todos los posibles vectores inductores de $\mathbf{x}' \in \mathcal{V}$.

Para todo $\mathbf{x} = (x_1, \ldots, x_n) \in [0,1]^n$ establecemos lo siguiente:

1. $m = \text{Card}(\{x_1, \ldots, x_n\})$
2. $\mathbf{x}' = (x'_1, \ldots, x'_m)$ con $x'_1 = x_1$ y para todo $j \in N_m$, $x'_j = x_k$, donde $k = \min\{i \in N_n \mid x_i \neq x'_l$ para todo $l \in N_{j-1}\}$
3. $\tau : N_m \to N_n$ es $\tau(i) = \text{Card}(\{j \in N_n \mid x_j = x'_i\})$, para todo $i \in N_m$
4. $\mathbf{A} = (\tau(1), \ldots, \tau(m))$
5. $\sigma : N_m \to N_m$ es una permutación tal que:
   - $a_{\sigma(i)} \geq a_{\sigma(i+1)}$, ó
   - $a_{\sigma(i)} = a_{\sigma(i+1)}$ y $x'_{\sigma(i)} > x'_{\sigma(i+1)}$

   para todo $i \in N_{m-1}$

Nótese que $\mathbf{x}' \in \mathcal{V}$ y $\sum_{i=1}^{m} \tau(i) = n$ y por tanto: (1) $\mathbf{A} = (1, \ldots, 1)$ si y sólo si $m = n$, por tanto, $x_i \neq x_j$ para todo $i \neq j \in N_n$; y (2) $\mathbf{A} = (n)$ si y sólo si $m = 1$, por tanto, $x_i = x_j$ para todo $i, j \in N_n$.

Example 1. Sea $\mathbf{x} = (0.5, 0.7, 0.5, 0.3, 0.7, 0.1, 0.3, 0.2, 0.3)$ entonces: $n = 9$, $m = \text{Card}(\{0.1, 0.2, 0.3, 0.5, 0.7\}) = 5$, $\mathbf{x}' = (0.5, 0.7, 0.3, 0.1, 0.2)$, $\mathbf{A} = (\tau(1), \ldots, \tau(m)) = (2, 2, 3, 1, 1)$, $\sigma = (\sigma(1), \ldots, \sigma(m)) = (3, 2, 1, 5, 4)$, y $\mathbf{x}'_\sigma = (0.3, 0.7, 0.5, 0.2, 0.1)$.

Definition 3 *Sea* $\mathbf{w} = (w_{\mathbf{x}})_{\mathbf{x} \in [0,1]^n}$ *sea una familia de vectores de subpesos tal que* $w_{\mathbf{x}} \in [0,1]^{m_{\mathbf{x}}}$. *Entonces el operador RIOWA (Restricted Induced Ordered Weighted Averaging) basado en* $\mathbf{w}$ *es una aplicación* $R_{\mathbf{w}} : [0,1]^n \to [0,1]$ *tal que para cada* $\mathbf{x} = (x_1, \ldots, x_n) \in [0,1]^n$: $R_{\mathbf{w}}(\mathbf{x}) = \sum_{i=1}^{m_{\mathbf{x}}} w_i x'_{\sigma(i)}$

Como consecuencia de esta definición para el nuevo operador RIOWA, se tiene que todo operador OWA es un operador RIOWA. La demostración está incluida en el trabajo original [1].

4 Aplicación a imputación de datos faltantes

Para comprobar el rendimiento de los nuevos operadores RIOWA, hemos implementado una arquitectura de imputación de datos faltantes introducida en [4]. Esta arquitectura, basada en modelos de *clustering* y funciones de agregación, genera múltiples datos candidatos que posteriormente son agregados utilizando operadores OWA e IOWA.

Utilizando esta misma arquitectura para 6 conjuntos de datos, la hemos modificado para utilizar nuestros operadores RIOWA y hemos comparado su rendimiento frente a los operadores originales.

Dataset	OWA	IOWA	RIOWA
1	0.200 ± 0.003	0.207 ± 0.003	**0.172 ± 0.002**
2	**0.432 ± 0.003**	**0.432 ± 0.003**	0.441 ± 0.004
3	0.284 ± 0.005	0.285 ± 0.005	**0.274 ± 0.006**
4	0.727 ± 0.003	0.729 ± 0.003	**0.724 ± 0.003**
5	0.191 ± 0.004	0.190 ± 0.004	**0.168 ± 0.005**
6	0.245 ± 0.003	0.249 ± 0.003	**0.223 ± 0.004**
Average	0.347 ± 0.188	0.349 ± 0.188	**0.334 ± 0.198**

Table 1. Resultados obtenidos con un 5% de datos faltantes.

Dataset	OWA	IOWA	RIOWA
1	0.364 ± 0.003	0.304 ± 0.009	**0.240 ± 0.003**
2	**0.491 ± 0.002**	0.494 ± 0.002	0.498 ± 0.002
3	0.362 ± 0.003	0.354 ± 0.005	**0.322 ± 0.004**
4	0.749 ± 0.002	0.751 ± 0.003	**0.742 ± 0.003**
5	0.286 ± 0.002	0.255 ± 0.005	**0.210 ± 0.003**
6	0.382 ± 0.002	0.330 ± 0.005	**0.283 ± 0.003**
Media	0.439 ± 0.151	0.415 ± 0.167	**0.382 ± 0.185**

Table 2. Resultados obtenidos con un 20% de datos faltantes.

Como se puede ver en las Tablas 1, 2, los resultados obtenidos por los tres operadores son constantes. En los conjuntos de datos 1, 3, 4, 5 y 6, el operador RIOWA obtiene un error medio inferior al de los otros operadores, mientras que su rendimiento es inferior en el conjunto de datos 2, donde destaca el operador OWA.

Del mismo modo, en la figura 1 se incluye la comparación con varios modelos base de imputación de datos. Como se puede observar, el rendimiento del operador RIOWA utilizado en la arquitectura empleada supera al resto de modelos en la mayoría de los conjuntos de datos.

Finalmente, realizamos un pequeño estudio para comprobar qué modificaciones incluidas en el operador influencian en mayor medida su rendimiento y qué incluimos en la Tabla 3. En esta tabla se puede ver claramente que el rendimiento de los RIOWA depende principalmente de la eliminación de instancias duplicadas en los datos.

Conjunto de datos	RIOWA	RIOWA(manteniendo duplicados)	RIOWA(usando vectores de pesos)
1	0.240 ± 0.002	0.303 ± 0.007	0.240 ± 0.003
2	0.498 ± 0.002	0.494 ± 0.002	0.498 ± 0.002
3	0.324 ± 0.004	0.353 ± 0.005	0.322 ± 0.004
4	0.741 ± 0.003	0.751 ± 0.003	0.742 ± 0.003
5	0.210 ± 0.003	0.255 ± 0.005	0.210 ± 0.003
6	0.283 ± 0.003	0.330 ± 0.005	0.283 ± 0.003

Table 3. Análisis de los RIOWA. Desentrelazado de las contribuciones de la eliminación de datos duplicados y de la relajación de pesos.

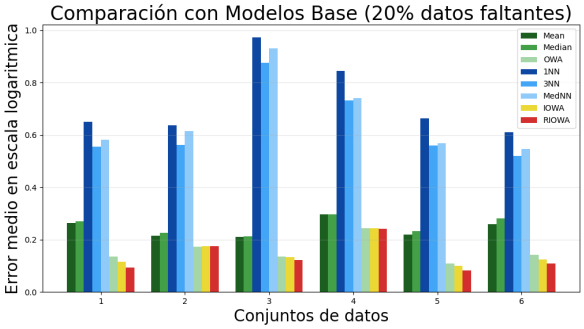


Fig. 1. Comparación frente a otros modelos de imputación de datos faltantes con un 20% de datos faltantes

5 Lineas de Trabajo Futuro

Como continuación de este trabajo, tenemos pensado seguir desarrollando estos operadores, extendiendo su formulación al dominio intervalar y estudiar la posibilidad de prescindir del vector inductor y por tanto, formular los Restricted OWA.

Acknowledgments

Este estudio esta financiado en parte por la Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código 001, N° Proceso 88887.917043 /2023-00, por el proyecto PID2022-136627NB-I00 apoyado por MCIN/AEI/10.13039 /501100011033/FEDER, UE, y por Tracasa Instrumental y el Departamento de Interior, Servicios SCiviles y Justicia del Gobierno de Navarra. Financiación para la publicación en Acceso Abierto dada por la Universidad Pública de Navarra.

References

1. A. Indurain-Ibero, B. Bedregal, M. Ferrero-Jaurrieta, J. Fernandez, H. Bustince, RI-OWA: Restricted induced ordered weighted averaging operators, applied to missing data imputation, Applied Soft Computing, 197 (2026) 115188.
2. R. R. Yager, On ordered weighted averaging aggregation operators in multicriteria decisionmaking, IEEE Transactions on systems, Man, and Cybernetics 18 (1) (1988) 183–190.
3. R. R. Yager, D. P. Filev, Induced ordered weighted averaging operators, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 29 (2) (1999) 141–150.
4. P. Keerin, N. Iam-On, J. J. Liu, T. Boongoen, Q. Shen, Summarising multiple clustering-centric estimates with owa operators for improved knn imputation on microarray data, Fuzzy Sets and Systems 473 (2023) 108718.