ISSN 2395-8618

Tratamiento y evaluación de datos incompletos mediante el clasificador Gamma Pydra y Missing Values Accuracy (MVA)

Juan Ignacio Vilchis García

***Resumen*—Los datos incompletos son un fenómeno real, actual y en algunas situaciones impredecible que suele presentarse en diversas áreas del conocimiento humano. En los estudios enfocados al análisis de datos, la ausencia de un correcto tratamiento para valores perdidos suele presentar una serie de inconvenientes importantes en el resultado final del fenómeno estudiado. El presente artículo propone, en primer lugar, un método que posee ocho variantes para el tratamiento de valores faltantes con base en el clasificador asociativo Gamma; y, en segundo lugar, una medida de desempeño como base para la mejor comprensión de la evaluación en la clasificación con valores perdidos. Las propuestas de este estudio son comparadas contra dos métodos existentes en el estado del arte para tratar valores perdidos pertenecientes al enfoque de imputación. En la fase experimental se utilizaron diez bancos de datos sin valores perdidos. Adicionalmente se implementó un módulo para generar valores perdidos en los bancos de datos completos para poder controlar el número de valores perdidos inducidos y de esta forma poder comparar los resultados con diferentes porcentajes de pérdida. Vale la pena mencionar que las propuestas presentan desempeños competitivos frente al estado del arte con base en la medida propuesta, la cual permite analizar el comportamiento de la clasificación centrándose en los valores perdidos.**

***Palabras clave*—Datos incompletos, clasificadores asociativos, valores faltantes, imputación de media, medidas de rendimiento, clasificación supervisada, técnicas de tolerancia.**

Incomplete Data Treatment and Evaluation based on the Gamma Pydra Classifier and Measure Missing Values Accuracy (MVA)

***Abstract*—Incomplete data is a real, current, and in some situations, unpredictable phenomenon that often occurs in diverse areas of human knowledge. In studies focused on data analysis, the lack of an adequate treatment for missing values tends to give rise to important deficiencies in the final results of the studied phenomenon. This work presents on one hand, a method based on the associative gamma classifier (introducing eight variants) to tackle the problem of missing data in pattern classification tasks; and on the other hand, a performance measure aimed at better understanding the evaluation of classification under a missing values scenario. The proposals were compared against two existent state-of-the-art methods to treat missing values belonging to the imputation approach. Ten complete data sets were used for the present study. Additionally, a module to generate missing values in order to control the quantity of induced missing values and thus be able to do a comparison between distinct data missing percentages was implemented. It is noteworthy that the proposals are competitive against the state of the art based on the proposed measure which allows the analysis focused on missing values classification.**

***Index Terms*—Incomplete data, associative classifiers, missing values, mean imputation, performance measures, supervised classification, tolerance techniques.**

I. Introduction

La materia prima para toda investigación son los datos; y el análisis, síntesis y tratamiento que se les da a éstos son cruciales para obtener resultados adecuados. Al trabajar con datos, siempre existe la posibilidad de encontrarse con valores perdidos, los cuales también conocidos son como datos faltantes, y dan lugar a la aparición de datos incompletos. Los valores perdidos o faltantes en toda investigación, son un problema que presentan ciertas características: son reales, actuales, y suelen presentarse de forma impredecible [1, 2]. Cuando se realiza una investigación, por lo general se asume el escenario de contar con los datos íntegramente. Sin embargo, pocas veces se plantea la posibilidad de tener que hacer frente a datos faltantes, lo cual se refleja en situaciones singulares, como el hecho de que la bibliografía donde el tema central son los datos, contiene poca o nula información sobre cómo tratar con un valor perdido.

De este modo, es cada vez más difícil ignorar el hecho de que los bancos de datos del mundo real contienen valores perdidos [1, 3] y se pueden encontrar en: minería de datos [1], estudios epidemiológicos [4], datos educativos [5], estudios de intervención [6], datos proteómicos [7], por mencionar algunos ámbitos.

La obtención de resultados de calidad a partir de datos incompletos suele tornarse una tarea difícil [8], porque las tareas de clasificación y análisis con datos incompletos puede presentar sesgos, análisis infundados, inconsistencias del fenómeno estudiado, entre otras adversidades; adicionalmente, si se disminuye la muestra de datos completos, también disminuirá el efecto de los tests estadísticos que sean aplicados [1, 4, 5, 9].

Dado que es imposible garantizar que siempre se contará con datos completos, la motivación para desarrollar el presente trabajo es atender la necesidad de mejorar los tratamientos para

Manuscript received 15/01/2022, accepted for publication on 15/02/2024.

Juan Ignacio Vilchis García (autor correspondiente) estudia en el Instituto Politécnico Nacional (IPN), en el Centro de Investigación en Computación (CIC), CDMX, México, (correo: jvilchisg2018@cic.ipn.mx).

ISSN 2395-8618

valores perdidos mencionada en una considerable cantidad de literatura, con respecto al tema [9].

Un punto de vital importancia es la evaluación de los métodos para tratar valores perdidos, de manera que reflejen el desempeño precisamente en datos faltantes, aislándolos de los datos completos, para enfocar la atención en el fenómeno de interés. De manera natural, si el resultado es satisfactorio en datos incompletos, esto ayudará a mejorar el desempeño al evaluar bancos de datos que presenten conjuntamente los casos completo e incompleto de los datos.

II. Tratamientos para valores perdidos

A. Generalidades sobre los valores perdidos

Los datos faltantes fueron definidos en una taxonomía por Donald Bruce Rubin, convirtiéndose en el estándar para cualquier discusión sobre este tema [10]. Esta clasificación, también conocida como los tres mecanismos de datos faltantes, es:

- Datos perdidos al azar (MAR: Missing at random).
- Datos completamente perdidos al azar (MCAR: Missing completely at random).
- Datos no perdidos al azar (NMAR: Not missing at random).

B. Enfoque de supresión

El método más común para tratar datos faltantes es excluirlos. El uso de este enfoque simplemente se limita a eliminar del experimento todos los patrones (instancias, casos, sujetos) con algún valor perdido. De esta forma, el experimento se lleva a cabo en los datos que no fueron excluidos. Este método es una opción apropiada en algunas situaciones, debido a que es muy fácil de implementar, además de que es la opción por defecto de algunos sistemas que se utilizan para análisis de datos [1,5,6,9,11]. Se recomienda su uso cuando sólo una cantidad pequeña de los datos contiene valores perdidos.

Sin embargo, también hay documentación [12–15] donde se afirma que usar eliminación por lista tiene serios inconvenientes. En ocasiones, cuando se tiene un gran número de patrones o atributos, puede ser ineficiente, así como inaceptable, el descartar una gran cantidad de sujetos. Este enfoque a menudo resulta en un decremento sustancial del tamaño de la muestra para el análisis.

C. Enfoque de imputación

El principio fundamental de este enfoque es calcular el posible valor faltante con base en los datos que sí se tienen completos; algunos métodos de este enfoque son los siguientes:

1) Imputación media incondicional

Una de las maneras más fáciles de imputar es reemplazar cada valor faltante con la media aritmética de los valores no faltantes de la variable en cuestión. A este enfoque se le conoce como Imputación media incondicional. También se pueden utilizar la mediana y la moda para reemplazar los valores perdidos [8,12]. Este enfoque tiene la ventaja de utilizar todos los casos, pero también tiene algunas desventajas ya que, aunque se han añadido casos, no se ha añadido nueva información y los cambios en el banco de datos se pueden considerar falsos [12].

2) Imputación media condicional

La imputación media condicional o también conocida como *class mean imputation* se realiza reemplazando cada valor perdido de la variable en cuestión, por la media aritmética de la clase a la que pertenece, si los valores son numéricos; para los valores categóricos se utiliza la moda. Dicho de otra manera, esta técnica divide el banco de datos por clases y se aplica la imputación media independiente por clase [5].

3) Imputación por regresión

Este método consiste en la regresión de la variable que tiene datos perdidos, en las otras variables independientes, produciendo así un modelo para estimar el valor de un dato perdido. La ventaja sobre la imputación media incondicional es que el dato imputado está de alguna manera condicionado con información que ya se tenía, mientras que la desventaja de este enfoque se debe a que aumenta artificialmente las correlaciones [12].

4) Imputación hot deck y cold deck

Una forma diferente para imputar es a través de un tipo pareo. Es decir, para cada unidad con un dato perdido X, se debe encontrar una unidad con valores similares de Y, para posteriormente tomar su valor X para reemplazarlo en el primero. Presenta ventajas cuando sólo una pequeña cantidad de datos tienen valores perdidos [16].

En contraste con el enfoque anterior donde las imputaciones provienen de una fuente interna, el método de imputación cold deck usa datos externos al banco de datos que se desea imputar. Los datos externos por lo general suelen ser una versión previa del mismo banco de datos.

5) Imputación deductiva

En este tipo de imputación, los datos faltantes se deducen de la información o de las normas existentes, utilizando:

- Información de observaciones relacionadas [17],
- Imputación basada en reglas lógicas [18],
- Último valor recolectado [19].

6) Estimación de máxima verosimilitud

Al aplicar este método, se asume que los datos completos son generados por un modelo descrito por una función de probabilidad, la cual capta la relación entre el banco de datos y los parámetros del modelo de datos y con ello se pueden realizar estimaciones verosímiles de los parámetros. Este enfoque se inicia estimando los parámetros del modelo con los datos completos con la función de máxima verosimilitud. Estos parámetros se utilizan para predecir los valores faltantes que se presenten. Este proceso será repetitivo hasta que se logre llegar a la máxima convergencia y así obtener la máxima probabilidad [20].

7) Imputación Múltiple

Imputación múltiple es un método que promedia las estimaciones obtenidas de n conjuntos de datos imputados que

ISSN 2395-8618

se obtienen a través de un proceso repetido varias veces, con los promedios se genera una estimación global. Este enfoque se basa en fundamentos bayesianos y consiste en tres procesos: el proceso de imputación, el proceso de análisis y el proceso de combinación.

8) Imputación mediante inteligencia artificial

Para una variable con valores perdidos, los valores de reemplazo son estimados tratando la variable con datos perdidos como un objetivo y usando las variables restantes como entrada a un algoritmo de inteligencia artificial, machine learning o computación inteligente. Entre las diferentes alternativas utilizadas con éxito y reportadas en la literatura científica se puede mencionar las siguientes:

- Árboles de decisión [21]. Para propósitos de eficiencia, la construcción de los árboles puede ser perezosa, donde se construye solo la ruta necesaria.
- Vecino más cercano (KNN) [22–24]. Este método implica la definición de una medida de distancia adecuada donde la distancia está en función de las variables auxiliares. Una ventaja de este método es que los valores observados son usados para la imputación. Por otro lado, algunos valores podrían ser usados varias veces para la imputación, mientras que otros podrían no ser usados.
- Agrupamiento (*clustering*). Se hace uso de "*k-means clustering*", donde la disimilitud intra-cluster (que se busca minimizar) es medida por la suma de las distancias desde el centroide, el cual representa el valor medio de los objetos en un grupo; para lograr esto se pueden usar diferentes funciones de distancia, como la distancia euclidiana [25].
- Redes neuronales artificiales (RNAs). Son arquitecturas computacionales que combinan unidades simples en un arreglo que puede mostrar un comportamiento complejo [26]. El tipo más común de RNAs usadas para imputación de datos perdidos son conocidas como *Feedforward*, donde las terminales de entrada reciben valores de las variables completas X, mientras que la salida proporciona la variable imputada Y [20].
- Lógica difusa, cuyos sistemas son diseñados específicamente para tratar con incertidumbre matemática y proporcionar herramientas formales para hacer frente a la imprecisión intrínseca de muchos problemas [27]. Para tratar datos perdidos, usualmente se incorporan otras técnicas [28]. Algunas aplicaciones son: *Fuzzy cluster imputation* [29], *Fuzzy majority imputation* [30], *o Fuzzy preference relation* [31].
- Conjuntos rugosos (*rough sets*), cuya teoría fue propuesta como una herramienta matemática para tratar información vaga o imprecisa [32]. Hoy en día varios enfoques basados en conjuntos rugosos han sido exitosamente aplicados en el campo del descubrimiento del conocimiento y, en especial, el tratamiento de valores perdidos es una de sus aplicaciones importantes; Por ejemplo, usando conjuntos rugosos en conjunción con árboles de decisión.

D. Enfoque de manipulación directa de los datos

Este enfoque agrupa algoritmos que son considerados tolerantes o que pueden tratar a los valores perdidos sin hacer uso de los dos enfoques antes mencionados; es decir, ni suprimen ni calculan los valores perdidos. Dentro de este enfoque se pueden encontrar los siguientes:

1) Algoritmo CART

El algoritmo CART (Classification and Regression Trees) fue propuesto en 1984 [33]. Al encontrarse con casos que poseen valores faltantes, CART realiza una especie de división o separación de las ramas para asignar los casos con valores perdidos a ramas del árbol, donde el impacto de tener valores perdidos puede ser menor utilizando enfoques probabilísticos.

2) Algoritmo C4.5

En 1993 Ross Quinlan propuso un método alternativo al algoritmo CART y a su vez una mejora del algoritmo ID3 [34]. Este método calcula la ganancia de información esperada asumiendo que el valor faltante está distribuido de acuerdo con los valores observados en el subconjunto de los datos en ese nodo del árbol; de esta forma puede trabajar con valores perdidos [35]. El algoritmo C4.5 parece ser una de las mejores técnicas para tratar los valores perdidos [36].

3) Algoritmo KNN

Anteriormente se mencionó que el algoritmo de los k vecinos más cercanos es útil para realizar imputación y con esto sustituir el valor perdido con uno calculado. Por otro lado, si el algoritmo se combina con otras técnicas emergentes como lo es el MI (algoritmo de información mutua) [37], o bien, usando otras métricas de distancia, es posible realizar una clasificación aun cuando los patrones a clasificar presentan valores perdidos.

4) Algoritmos ALVOT

Los algoritmos de votación surgieron aproximadamente en 1965. Son modelos de clasificación supervisada desarrollados dentro del enfoque lógico-combinatorio para reconocimiento de patrones [38]. Estos modelos se basan en precedencias parciales que permiten trabajar con descripciones de objetos en términos de variables numéricas y no numéricas e incluso admiten valores perdidos. La desventaja de usar los algoritmos ALVOT es que la complejidad es de tipo exponencial, lo que significa un alto costo computacional.

5) Algoritmo Gamma DMI

Es una extensión al clasificador asociativo Gamma, que permite manejar datos mezclados e incompletos en el enfoque asociativo [39]. Los resultados de los experimentos que tuvieron lugar utilizando este enfoque, fueron competitivos comparándose contra otros clasificadores que permiten el manejo de datos mezclados e incompletos. Además, cabe mencionar que una parte de la experimentación tuvo una etapa de pre procesamiento, utilizando selección de rasgos.

6) Naïve Associative Classifier (NAC)

Es un novedoso modelo de aprendizaje supervisado, cuyas principales cualidades radican en su simplicidad, transparencia, transportabilidad y precisión. Tanto la creación, el diseño, la implementación y la aplicación del NAC se basan en un operador de similitud original, cuyo nombre es operador de similitud de datos mixtos e incompletos (MIDSO). El modelo

ISSN 2395-8618

TABLA I
DEFINICIÓN DE LOS OPERADORES ALFA Y BETA

$\alpha : A \times A \to B$			$\beta : B \times A \to A$		
x	y	$\alpha(x, y)$	x	y	$\beta(x, y)$
0	0	1	0	0	0
0	1	0	0	1	0
1	0	2	1	0	0
1	1	1	1	1	1
			2	0	1
			2	1	1

fue probado en el área financiera, superando a varios clasificadores de patrones de última generación [40].

III. MATERIALES Y MÉTODOS

En la presente sección se presentan los materiales necesarios para el desarrollo de la propuesta presentada en la sección 4 y el esquema experimental descrito en la sección 5. Este apartado está formado por tres secciones.

En primer lugar, se presenta el clasificador asociativo Gamma, el cual forma parte de los modelos asociativos Alfa-Beta y que es la base de la presente propuesta. Además de lo anterior se presentan algunos conceptos clave para entender el clasificador mencionado.

En segundo lugar, se proporciona la información correspondiente al método de validación utilizado en la experimentación: Stratified k-fold cross-validation.

La tercera parte de la sección tiene como función primordial presentar una descripción de los bancos de datos que se utilizaron en la fase experimental. De manera que en esta sección se pueden encontrar bancos de datos que poseen ciertas características necesarias para la adecuada evaluación de los experimentos, a saber: bancos de datos completos con valores numéricos positivos y sin desbalance de clases.

A. Clasificador asociativo Gamma

El clasificador asociativo Gamma [41] es parte de los modelos asociativos Alfa-Beta [40, 42–48]. Dicho clasificador ha mostrado desempeños competitivos en los campos donde ha sido posible aplicarlo. Para poder comprender este clasificador es necesario mencionar de manera sucinta los operadores Alfa y Beta, el operador Gamma y el código Johnson-Möbius modificado.

1) Operador Alfa y operador Beta

Son los operadores fundamentales de los modelos asociativos Alfa-Beta. El operador alfa se utiliza en la fase de aprendizaje del clasificador. Por otro lado, el operador beta se desempeña en la fase de recuperación. Estos operadores fueron definidos de manera tabular y sus propiedades demostradas en [49]. A continuación, se muestra la tabla 1 donde se presentan tales operadores, dados los conjuntos A = {0,1} y B = {0,1,2}. En [49–52] se utilizan estos operadores y se demuestran sus propiedades.

2) Operador Gamma de similitud generalizado

Este operador es de suma importancia para el clasificador Gamma y para el presente trabajo de tesis y se define como un operador de similitud; es decir, el resultado de aplicarlo indica si dos vectores o elementos son parecidos, dado un grado de disimilitud θ, de lo contrario se indica que no son parecidos. De manera que dos vectores diferentes se pueden considerar similares dependiendo del grado de disimilitud θ. La siguiente definición del operador Gamma de similitud generalizado fue tomada de [53].

Definición: Sean el conjunto $A = \{0,1\}$, dos números $n, m \in \mathbb{Z}+$, $n \leq m$, $\boldsymbol{x} \in An$ y $\boldsymbol{y} \in Am$ dos vectores binarios, $n{-}dimensional$ y $m{-}dimensional$ respectivamente, con la $i{-}ésima$ componente representada por xi y yi, respectivamente; y θ un número entero no negativo. Se define el operador Gamma de similitud generalizado $\gamma g\,(x, y, \theta)$ de la siguiente manera: $\gamma g\,(x, y, \theta)$ tiene como argumentos de entrada 2 vectores binarios $\boldsymbol{x}$ y $\boldsymbol{y}$, y un número entero no negativo θ, y la salida es un número binario que se calcula así:

$$\gamma_g(x, y, \theta) = \begin{cases} 1 \; si \; m - u\beta[\alpha(\boldsymbol{x}, \boldsymbol{y}) mod\; 2] \leq \theta \\ 0 \qquad\qquad en\; otro\; caso \end{cases}.$$

Si $m > n$, truncar los $m - n$ bits izquierdos de $\boldsymbol{y}$ antes de realizar el cálculo.

3) Johnson-Möbius modificado

Los vectores con los que trabaja el clasificador Gamma necesitan codificarse con el código Johnson-Möbius modificado [52] cuyo algoritmo se muestra a continuación:

Algoritmo 1. Código Johnson-Möbius modificado

1. Sea un conjunto de números reales $\{r1, r2, \ldots ri, \ldots rn\}$ donde n es un número entero positivo fijo.
2. Si uno de los números del conjunto (por ejemplo, ri) es negativo, crear un nuevo conjunto transformado a través de la operación "restar ri a cada uno de los n números" $\{t1, t2, \ldots ti, \ldots tn\}$ donde $tj = rj - ri \;\forall\; j \in \{1,2, \ldots n\}$ y particularmente $ti = 0$. Nota: si hay más de un negativo, se trabaja con el menor.
3. Escoger un número fijo d de decimales y truncar cada uno de los números del conjunto transformado (los cuales son no negativos) precisamente a d decimales.
4. Realizar un escalamiento de $10d$ en el conjunto del paso 3, para obtener un conjunto de n enteros no negativos $\{e1, e2, \ldots ei, \ldots em, \ldots en\}$ donde em es el mayor.
5. El código Johnson-Möbius modificado para cada $j = 1, 2, \ldots n$ se obtiene al generar $(em - ej)$ ceros concatenados por la derecha con ej unos.

4) Algoritmo del clasificador Gamma

A continuación, se presenta el algoritmo del clasificador Gamma, tal y como se propuso originalmente en [53,54] , donde se presenta y discute a mayor detalle:

Sean los números $k, m, n, p \in \mathbb{Z}^+, \{x^\mu \mid \mu = 1,2,\ldots,p\}$ el conjunto fundamental de patrones de cardinalidad p, donde $\forall\, \mu\; x^\mu \in \mathbb{R}^n$ y $\boldsymbol{y} \in \mathbb{R}^n$ es un vector real $n - dimensional$ a ser clasificado. Se asume que el conjunto fundamental está particionado en m clases diferentes, donde cada clase tiene

ISSN 2395-8618

cardinalidad k_i , $i = 1,2,\ldots,m$, por lo que $\sum_{i=1}^{m} k_i = p$. Para clasificar el patrón $\boldsymbol{y}$, se realiza lo siguiente:

Algoritmo 2. Clasificador Gamma

1. Codificar cada componente de cada patrón del conjunto fundamental con el código Johnson-Möbius modificado, obteniéndose un valor $e_m = \bigvee_{i=1}^{p} x_j^i$
2. Calcular el parámetro de paro $p = \bigwedge_{j=1}^{n} e_m(j)$.
3. Codificar cada componente del patrón a clasificar con el código Johnson-Möbius modificado, utilizando las mismas condiciones que se utilizaron para codificar las componentes de los patrones fundamentales. En caso de que alguna componente del patrón a clasificar sea mayor al e_m correspondiente $\left(y_\xi > e_m(\xi)\right)$, igualar esa componente a $e_m(\xi)$ y guardar su valor anterior en la variable $nmax_\xi$.
4. Realizar una transformación de índices en todos los patrones del conjunto fundamental, de manera que el índice único que tenía un patrón originalmente en el conjunto fundamental, por ejemplo x^μ, se convierta en dos índices: uno para la clase (por ejemplo, la clase i) y otro para el orden que le corresponde a ese patrón dentro de esa clase (por ejemplo w). Así, la notación para el patrón x^μ será ahora x^{iw}.
5. Inicializar θ a 0.
6. Realizar la operación $\gamma_g\left(x_j^{iw}, y_j, \theta\right)$ para cada clase y para cada componente de cada uno de los patrones fundamentales que corresponden a esa clase, y del patrón a clasificar, considerándose $nmax_\xi$ como la dimensión del patrón binario y_ξ .
7. Calcular la suma ponderada c_i de los resultados obtenidos en el paso 6, para cada clase $i = 1,2,\ldots,m$:

$$c_i = \frac{\sum_{w=1}^{k_i} \cdot \sum_{j=1}^{n} \cdot \gamma_g\left(x_j^{iw}, y_j, \theta\right)}{k_i}.$$

8. Si existe más de un máximo entre las sumas ponderadas por clase, incrementar θ en 1 y repetir los pasos 6 y 7 hasta que exista un máximo único, o se cumpla con la condición de paro $\theta \geq p$.
9. Si existe un máximo único, asignar al patrón a clasificar la clase correspondiente a ese máximo:

$$C_y = C_j \quad \text{tal que} \quad \bigvee_{i=1}^{m} c_i = c_j$$

10. En caso contrario: si λ es el índice más pequeño de clase que corresponde a uno de los máximos, asignar al patrón a clasificar la clase C_λ.

B. Stratified k-fold cross-validation

Al aplicar un clasificador, es necesario poder evaluar la capacidad de clasificar correctamente muestras o patrones que aún no conoce, mismos que no estuvieron presentes en su etapa de construcción o entrenamiento. Para este fin tenemos métodos de validación cruzada que consisten en hacer particiones al conjunto fundamental de datos. El fin es obtener dos subconjuntos denominados conjunto de entrenamiento y conjunto de prueba, y con ellos evaluar el comportamiento del clasificador ante datos que no conoce [55].

La validación cruzada estratificada con k-folds consiste en particionar aleatoriamente el conjunto original de datos en K subconjuntos disjuntos, aproximadamente de igual tamaño denominados folds, pero de forma que cada clase se distribuya proporcionalmente en cada fold; de esta manera, la distribución de clases en cada fold es representativa del banco de datos original. Cada uno de estos K subconjuntos es utilizado en turno como el conjunto de prueba, mientras que los folds restantes (K-1) son utilizados como el conjunto de entrenamiento, proceso que se realiza K veces [56].

C. Bancos de datos

En esta sección se describen de forma breve los 10 bancos de datos utilizados en los experimentos descritos en la sección 5. Dichos bancos de datos fueron seleccionados con base en los siguientes criterios:

- Primero, el banco de datos no debe contener valores faltantes. Es decir, se trabajó con bancos de datos con valores completos. Esto tiene como objetivo el poder ingresar valores faltantes de forma aleatoria y controlada, permitiendo manejar la cantidad deseada de datos con valores perdidos.
- Segundo, los valores de los atributos correspondientes a cada patrón deben ser de tipo numérico.
- Tercero, los valores numéricos deben ser positivos.

Cuarto, el número de patrones que pertenecen a cada clase, debe ser balanceado, tomando un *imbalance ratio* (IR)< 1.5, el cual se calcula a partir de (1) [57–59].

$$mbalance\ ratio = \frac{cardinalidad\ de\ la\ clase\ mayor}{cardinalidad\ de\ la\ clase\ menor}.$$

Los 10 bancos de datos recopilados se describen a continuación.

1) Iris plant data set

Tomado de [60], el banco de datos iris plant fue creado por R.A Fisher en 1936, y sigue siendo hasta nuestros días un banco de datos clásico en el ámbito del reconocimiento de patrones. Este banco de datos contiene tres clases, con 50 instancias cada una, dando un total de 150, donde cada clase es un tipo de la planta iris (iris setosa, iris versicolor e iris virginica). Cada patrón está compuesto por cuatro atributos, los cuales hacen referencia al largo y ancho del sépalo y al largo y ancho del pétalo. Cabe mencionar que estas medidas están dadas en centímetros, y se presentan como un tipo de datos real, es decir con punto decimal y positivos. Este banco de datos no presenta desbalance de clases.

2) Breast Cancer Coimbra data set

Tomado de [60], este banco de datos posee información de 64 pacientes con diagnóstico de cáncer de pecho y 52 controles sanos, de forma que se tienen 116 patrones. Los patrones

ISSN 2395-8618

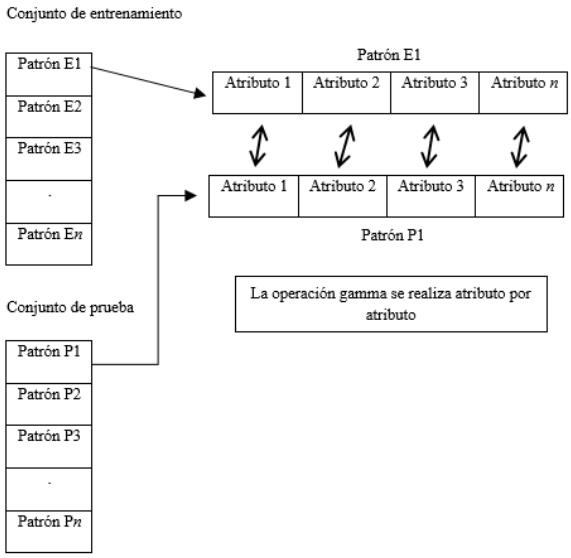

Fig. 1. Operación del operador gamma

poseen nueve atributos de tipo cuantitativo, positivos. El banco de datos no contiene valores perdidos y tiene un IR = 1.23.

3) Seeds data set

Recopilado de [60], este banco de datos está formado por medidas geométricas que pertenecen a tres diferentes tipos de grano de trigo. En su creación, se seleccionaron al azar 70 muestras de tres variedades de trigo, a saber: kama, rosa y Canadian, de forma tal que el banco de datos cuenta con tres clases con 70 patrones por clase, con un IR = 1. Cada patrón consta de siete atributos de tipo real, positivos, y no presenta valores perdidos.

4) Wine data set

Obtenido de [60, 61], este banco de datos está relacionado a la tarea de determinar el origen de tres vinos. El banco tiene 178 patrones de tipo numérico, positivos, agrupados en tres clases, con un IR= 1.47. En las referencias mencionadas existen otros dos bancos de datos relacionados, los cuales sí presentan desbalance de clases.

5) The Monk's Problem data set

Este banco de datos tiene íntima relación con una tarea simple que se creó con el fin de comparar algoritmos de aprendizaje. La tarea fue bautizada "the three MONK's problems". De los tres bancos de datos disponibles, para este se ha seleccionado el monk-2, el cual fue tomado de [60,61]. El banco de datos monk-2 consta de 432 patrones, divididos en dos clases, donde cada patrón tiene seis atributos de tipo entero; no hay valores perdidos y el IR= 1.1.

6) Ultrasonic flowmeter diagnostics data set

Este banco de datos se recopiló de [60], y consiste en el diagnóstico de fallas de cuatro flujómetros ultrasónicos líquidos. Cada flujómetro generó un banco de datos identificados con las letras A, B, C y D. El banco utilizado en la fase experimental fue el banco A. Este banco es biclase y a diferencia de los otros tres posee clases balanceadas; tiene 87 patrones con 36 atributos positivos.

7) Sonar, mines vs rocks data set

Asociado a la clasificación de señales de sonar, este banco de datos tomado de [60], consiste en 208 patrones divididos en 97 para la clase mina y 111 para la clase roca. Cada uno de los atributos toma valores desde 0.0 hasta 1.0, el desbalance de clase es de 1.14 y no hay datos perdidos.

8) Vehicle silhouettes data set

Este banco de datos tomado de [60,61], tiene como objetivo clasificar las siluetas en uno de cuatro tipos de vehículos. El banco posee 846 patrones formados por 18 atributos de tipo real, positivos. Posee un IR = 1.09 y no hay valores perdidos.

9) Parkinson speech with multiple types of sound recordings data set

Recopilado de [60], este banco de datos trata la clasificación de pacientes que poseen la enfermedad de Parkinson contra los controles sanos. Dicho banco tiene 1040 patrones con 26 atributos, y es preciso hacer notar que el banco de datos original tiene 28 atributos, de los cuales se eliminaron dos: la etiqueta que indica el id del paciente y la variable UPDRS que está directamente relacionada con la enfermedad del Parkinson. El banco de datos no tiene valores perdidos, las clases están igualmente distribuidas y los atributos son de tipo real, positivos.

10) Libras movement data set

Este banco de datos se obtuvo de [61], cuenta con un total de 360 patrones, los cuales se dividen en 15 clases y cada una de estas indica un tipo de movimiento de la mano en LIBRAS (Lenguaje brasileño de señales). Las clases se presentan de manera balanceada con 24 instancias por clase, contando además con un total de 90 atributos, los cuales representan las coordenadas de cada movimiento. El tipo de datos de todo el conjunto es real, sin valores perdidos.

IV. Propuestas

En la presente sección se describen las propuestas de solución a los problemas que dan origen y motivación a este trabajo. Para llevar a cabo una clasificación supervisada es necesario realizar los experimentos en dos conjuntos mutuamente excluyentes denominados conjunto de entrenamiento y conjunto de prueba, los cuales a su vez son subconjuntos del conjunto de datos original.

De forma que al llevar a cabo el proceso de clasificación utilizando el clasificador Gamma, se comparan uno a uno los patrones del conjunto de prueba con los patrones del conjunto de entrenamiento, aplicándoles la operación gamma, a su vez, atributo por atributo como se muestra en la figura 1.

Dada esta forma de operación, cuando hay valores perdidos en cualquiera de los dos conjuntos o en ambos, podemos encontrarnos con cuatro casos diferentes que se mencionan a continuación y para los cuales se propone la nomenclatura C (complete) M (missing):

– Valores completos en ambos patrones (CC),

ISSN 2395-8618

TABLA II

OPERACIÓN GAMMA ORIGINAL; DONDE: X = PATRÓN DE ENTRENAMIENTO, Y = PATRÓN DE PRUEBA, C = VALORES COMPLETOS, ? = PATRÓN CON AL MENOS UN VALOR PERDIDO, γ (X, Y, θ) = DEPENDE DE LA OPERACIÓN GAMMA

X	**Y**	**Salida**
C	C	$\gamma(X, Y, \theta)$
?	C	No es possible operar
C	?	
?	?	

TABLA III

OPERACIÓN GAMMA ORIGINAL A NIVEL ATRIBUTO; DONDE: X_A = ATRIBUTO DEL PATRÓN DE ENTRENAMIENTO, Y_A = ATRIBUTO DEL PATRÓN DE PRUEBA, C = VALOR COMPLETO, ? = VALOR PERDIDO

X_a	Y_a	**Problema**	**Solución Propuesta**
C	C	Casi ideal: no hay problema	
?	C	No es possible operar	Extensión al operador gamma de similitud generalizada
C	?		
?	?		

Variante 0

X_a	Y_a	**Salida**
C	C	$\gamma(X_a, Y_a, \theta)$ = {0 | 1}
?	C	0
C	?	0
?	?	0

Variante 1

X_a	Y_a	**Salida**
C	C	$\gamma(X_a, Y_a, \theta)$ = {0 | 1}
?	C	1
C	?	0
?	?	0

Variante 2

X_a	Y_a	**Salida**
C	C	$\gamma(X_a, Y_a, \theta)$ = {0 | 1}
?	C	0
C	?	1
?	?	0

Variante 3

X_a	Y_a	**Salida**
C	C	$\gamma(X_a, Y_a, \theta)$ = {0 | 1}
?	C	1
C	?	1
?	?	0

Variante 4

X_a	Y_a	**Salida**
C	C	$\gamma(X_a, Y_a, \theta)$ = {0 | 1}
?	C	0
C	?	0
?	?	1

Variante 5

X_a	Y_a	**Salida**
C	C	$\gamma(X_a, Y_a, \theta)$ = {0 | 1}
?	C	1
C	?	0
?	?	1

Variante 6

X_a	Y_a	**Salida**
C	C	$\gamma(X_a, Y_a, \theta)$ = {0 | 1}
?	C	0
C	?	1
?	?	1

Variante 7

X_a	Y_a	**Salida**
C	C	$\gamma(X_a, Y_a, \theta)$ = {0 | 1}
?	C	1
C	?	1
?	?	1

- Valores perdidos en el patrón de entrenamiento (MC),
- Valores perdidos en el patrón de prueba (CM),
- Valores perdidos en ambos patrones (MM).

El primer caso no implica problema para el operador gamma. Sin embargo, los casos restantes no pueden ser resueltos de forma satisfactoria, como se muestra en la tabla 2.

A. El método Pydra

El método Pydra que se presenta a continuación deriva su nombre del término Pyro-Hydra, que en la mitología japonesa era un dragón con ocho cabezas por las cuales lanzaba fuego. Este método consta de ocho variantes (haciendo referencia a las 8 cabezas mencionadas) y se basa en la extensión del operador gamma de similitud generalizado, específicamente para darle la capacidad de operar ante la presencia de valores perdidos.

La propuesta comienza por retomar los escenarios mostrados en la tabla 2 donde pueden aparecer los valores perdidos al llevar a cabo una clasificación con el clasificador Gamma. Para los tres escenarios presentados donde el clasificador no puede operar, se identifica en primer lugar que el problema está en el patrón en cuestión; y dado que un patrón está compuesto por atributos, la raíz del problema se encuentra en los atributos, los cuales pueden contener valores perdidos, de forma que el problema se replantea en la tabla 3.

Posteriormente, tomando en cuenta los dos posibles resultados de la operación gamma original (1: los atributos son similares, 0: no son similares), tenemos un total de ocho variantes posibles. Las ocho variantes propuestas se presentan a continuación, siendo estas identificadas por un índice (0-7) y su implementación en el clasificador Gamma se denomina Gamma Pydra; para todas las variantes: Xa = atributo del patrón de entrenamiento, Ya = atributo del patrón de prueba, C = valor completo, ? = valor perdido, □ = operación gamma, 0 = atributos diferentes, 1 = atributos similares.

Cabe aclarar que los resultados definidos en cada una de estas variantes son a nivel de atributo, resolviendo así el problema de valor perdido en el atributo y propagando el resultado a nivel patrón y éste a su vez a nivel banco de datos, lo que ofrece una solución íntegra en el proceso de clasificación.

B. Missing Values Accuracy (MVA)

Cuando se trabaja con un ambiente controlado de valores perdidos, lo cual significa que es posible conocer el valor original que posee un valor perdido generado artificialmente tal como se trabajó en la parte experimental de este trabajo, se pone en evidencia la necesidad de evaluar el desempeño obtenido en los patrones inducidos con valores perdidos.

Por lo anterior se propone la medida *missing values accuracy*, la cual representa el porcentaje de patrones con valores perdidos que fueron clasificados de forma correcta, de acuerdo con la expresión (2):

$$\underset{accuracy}{missing\ values} = \frac{\underset{clasificados\ corectamente}{patrones\ con\ valores\ perdidos}}{patrones\ con\ valores\ perdidos}, \quad (2)$$

Mediante esta expresión es posible determinar el desempeño que obtuvo el clasificador en los patrones con valores perdidos independientemente de otras medidas de desempeño.

V. EXPERIMENTOS

Esta sección muestra los elementos que son parte fundamental de los experimentos realizados. En primer lugar, se presenta un módulo para introducir valores perdidos en bancos de datos con valores completos; después, se muestran los bancos de datos que se utilizaron en el módulo y posteriormente el clasificador con diferentes implementaciones

ISSN 2395-8618

TABLA IV
BANCOS DE DATOS INDUCIDOS CON VALORES PERDIDOS

Banco de datos	C	P	A	UD	IR
Iris	3	150	4	600	1.00
B. Cancer Coimbra	2	116	9	1,044	1.23
Seeds	3	210	7	1,470	1.00
Wine	3	178	13	2,314	1.47
Monk 2	2	432	6	2,592	1.11
U. F. Diagnostics A	2	87	36	3,132	1.48
Sonar vs Rocks	2	208	60	12,480	1.14
V. Silhouettes	4	846	18	15,228	1.09
Parkinson Speech	2	1040	26	27,040	1.00
Libras Movement	15	360	90	32,400	1.00

TABLA
CLASIFICADORES IMPLEMENTADOS; UMI REPRESENTA UNCONDITIONAL MEAN IMPUTATION, CMI REPRESENTA CONDITIONAL MEAN IMPUTATION

Nombre del clasificador Gamma	Enfoque	Tratamiento	Implementación
UMI	Imputación	Imputación media incondicional	Como pre procesamiento de datos
CMI		Imputación media condicional	
Pydra 0 (PY0)	Manipulación directa de los datos	Método pydra	Directamente en el clasificador
Pydra 1 (PY1)		Método pydra1	
Pydra 2 (PY2)		Método pydra2	
Pydra 3 (PY3)		Método pydra3	
Pydra 4 (PY4)		Método pydra4	
Pydra 5 (PY5)		Método pydra5	
Pydra 6 (PY6)		Método pydra6	

de métodos para tratar valores perdidos. Finalmente se muestra el procedimiento abordado para la experimentación.

A. *Módulo para introducir valores perdidos*

Con el propósito de poner a prueba y evaluar el desempeño de los métodos para tratar valores perdidos, fue necesario utilizar bancos de datos con valores perdidos generados artificialmente, de manera que se creó un módulo que recibe un banco de datos completo y de acuerdo con un parámetro que representa el porcentaje del total de datos que se desea sean producidos con valores perdidos, regresa un banco de datos con valores perdidos. Cabe mencionar que el tipo de valores perdidos inducidos es del tipo MCAR, utilizando el siguiente algoritmo:

Algoritmo 3

1. Se obtiene el universo de datos multiplicando el total de patrones por el total de atributos.
2. Cada elemento de este conjunto universal es designado por un índice.
3. De acuerdo con el porcentaje de pérdida del universo de datos, se obtiene el número de elementos a perder.
4. Se obtiene al azar un índice correspondiente al universo de datos.
5. a) Si el elemento no tiene valor perdido se le asigna "?" y se resta 1 al número de elementos a perder. b) Si el elemento ya tiene asignado "?" se regresa al paso 4.
6. a) Si el número de elementos a perder es mayor a 0 se regresa al paso 4. b) Si el número de elementos a perder es igual a 0 el algoritmo termina.

B. *Bancos de datos con valores perdidos*

Los bancos de datos presentados en la sección de materiales y métodos subsección C fueron artificialmente alterados con valores perdidos mediante el módulo introducido en la sección de propuestas subsección A.

Los porcentajes de pérdida utilizados fueron hasta 50% con intervalos de 5%. Los porcentajes mencionados son considerados aditivos, lo cual significa que se inicia con un 5% de valores perdidos y se realizan los experimentos; posteriormente, se agrega 5% para generar 10% y se realizan los experimentos; consecuentemente, se agrega 5% de valores perdidos para generar el 15% y se llevan a cabo los experimentos correspondientes; y así posteriormente, hasta alcanzar el 50% de valores perdidos.

La tabla 4 resume la información de los bancos de datos utilizados, los cuales se ordenan en forma ascendente de acuerdo con el universo de datos. La nomenclatura utilizada es la siguiente: C, número de clases; P, número de patrones; A, número de atributos; UD, universo de datos; IR: imbalance ratio.

C. *Clasificadores implementados*

Para poder comparar los resultados del método propuesto contra métodos del estado de arte, se implementaron en el clasificador Gamma 2 métodos clásicos de imputación.

En este apartado no se consideró el enfoque de supresión, debido a que la medida de desempeño a evaluar trabaja sólo con valores perdidos y dicho enfoque los elimina por completo. Los clasificadores implementados se muestran en la tabla 5.

D. *Procedimiento experimental*

El objetivo de la etapa experimental en este trabajo es evaluar el impacto de la clasificación en patrones con valores perdidos, mediante un ambiente controlado de valores perdidos y clasificadores que implementan tratamientos para datos incompletos.

Por otro lado, los bancos de datos utilizados son los que se mencionan en la tabla 4, los cuales son alterados con valores faltantes en incrementos aditivos de 5% hasta llegar a 50%; mientras que se utilizaron los clasificadores de la tabla 5.

ISSN 2395-8618

TABLA VI
RESULTADOS GLOBALES EN MVA OBTENIDOS EN EL BANCO DE DATOS IRIS

Clasificador	Porcentaje de valores perdidos									
	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
UMI	84.24	82.49	78.59	76.12	69.55	67.07	64.01	60.77	56.99	54.10
CM I	97.93	98.05	98.90	99.21	99.31	98.94	98.95	99.01	99.42	99.43
PY0	86.45	84.96	84.25	83.53	83.05	81.75	79.84	77.39	75.80	72.39
PY1	79.39	76.83	72.86	68.65	65.49	63.40	61.82	59.47	52.80	53.40
PY2	82.94	79.59	75.79	73.40	69.77	67.96	63.88	60.37	55.64	50.79
PY3	76.94	72.84	67.43	65.22	60.81	58.63	58.10	54.48	49.46	46.93
PY4	85.40	82.97	80.12	75.23	69.98	65.26	62.92	60.44	56.91	53.03
PY5	78.61	74.74	70.27	66.54	64.74	62.66	56.69	55.32	48.18	47.87
PY6	86.45	84.96	84.25	83.53	83.05	81.75	79.84	77.39	75.80	72.39
PY7	79.39	76.83	72.86	68.65	65.49	63.40	61.82	59.47	52.80	53.40

TABLA VII
RESULTADOS GLOBALES EN MVA OBTENIDOS EN EL BANCO DE DATOS WINE

Clasificador	Porcentaje de valores perdidos									
	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
UMI	80.43	77.00	77.94	73.83	70.88	66.70	63.08	58.75	56.92	52.36
CM I	89.21	91.15	93.31	95.1	96.2	97.96	98.82	99.38	99.83	99.89
PY0	80.31	79.04	77.65	77.36	77.52	76.17	75.53	73.66	72.5	71.20
PY1	81.65	73.81	69.61	66.77	63.45	60.74	60.47	56.67	54.95	52.47
PY2	80.55	79.04	77.72	74.61	70.71	68.75	65.73	63.14	57.81	53.14
PY3	82.00	74.11	69.98	66.6	62.58	58.87	56.88	53.74	50.00	48.65
PY4	80.32	78.35	77.01	72.75	71.69	68.34	64.71	59.60	55.73	51.29
PY5	82.23	73.61	67.35	61.14	59.70	56.49	56.24	52.34	48.92	47.52
PY6	80.31	79.04	77.65	77.36	77.52	76.17	75.53	73.66	72.5	71.20
PY7	81.65	73.81	69.61	66.77	63.45	60.74	60.47	56.67	54.95	52.47

Asimismo, el método de validación utilizado fue stratified 10 *fold cross validation*, repitiendo cada experimento diez veces y promediando los resultados. A fin de mantener todos los experimentos bajo las mismas condiciones, los folds creados son compartidos por todos los clasificadores. Adicionalmente, para medir el desempeño de la clasificación sólo en los valores perdidos se utilizó MVA, la medida propuesta en la sección de propuestas subsección B.

El algoritmo utilizado para cada uno de los diez bancos de datos y que se considera un experimento, fue el siguiente:

Algoritmo 4

1. Se agrega al banco de datos un 5% de valores perdidos. Si el porcentaje de valores perdidos es superior a 50% el algoritmo termina.
2. a) Si el porcentaje de valores perdidos es igual a 5% se definen los diez folds y se distribuye el banco de datos en ellos. b) Si el porcentaje de valores perdidos es mayor a 5% se distribuye el banco de datos con valores perdidos en los 10 folds, previamente definidos.
3. Se procede con la clasificación, enfatizando que todos los clasificadores comparten los folds.
4. Se regresa al paso 1.

VI. RESULTADOS Y DISCUSIÓN

En esta sección se muestran los resultados de la experimentación.

A. Resultados en el banco de datos iris

La tabla 6 muestra los resultados obtenidos en el banco de datos *iris plant.*

En primer lugar, se muestra el clasificador UMI, que implementa unconditional mean imputation, y que sigue un comportamiento descendente. Cuando se tiene un 5% de valores perdidos, se clasifican correctamente 84.24% de esos datos; mientras que a mayor porcentaje de valores perdidos la medida MVA tiene a disminuir hasta llegar al 50% de valores perdidos, donde se clasifica el 54.10% de valores perdidos; es decir, aproximadamente la mitad.

En segundo lugar, se puede apreciar que el método que clasifica mejor los valores perdidos es el clasificador CMI (en color negro); es decir, el que utiliza conditional mean imputation. Cuando se presenta un 5% de valores perdidos, CMI es capaz de clasificar correctamente un 97.93% de los mismos. Posteriormente, cuando el banco de datos presenta un 10% de valores perdidos, se clasifica correctamente el 98.05% de dicho porcentaje. Sucesivamente mientras el porcentaje de valores perdidos aumenta, la medida MVA tiende a aumentar, como se observa en el porcentaje correspondiente al 50% de valores perdidos, donde se obtiene un MVA igual al 99.43%.

En cuanto a Pydra los métodos 0 y 6 (en colores gris) muestran resultados superiores e idéntico comportamiento, los cuales clasifican correctamente el 86.45% de los valores perdidos, cuando representan un 5%. El comportamiento de los métodos Pydra es descendente cuando aumenta el porcentaje de valores perdidos, hasta llegar al porcentaje de 50% donde se obtiene un MVA de 72.39%. Las restantes variantes del método propuesto obtienen resultados que rondan los desempeños del clasificador UMI.

B. Resultados en el banco de datos wine

La tabla 7 muestra los resultados obtenidos en el banco de datos wine, donde se observa que los resultados generales son similares al banco de datos iris. El clasificador UMI muestra un

ISSN 2395-8618

TABLA VIII
RESULTADOS GLOBALES

Banco de datos	Clasificadores									
	UMI	CMI	PY0	PY1	PY2	PY3	PY4	PY5	PY6	PY7
Iris	4	1	2	7	6	10	5	9	3	8
B. Cancer Coimbra	8	1	2	5	4	6	9	10	3	7
Seeds	6	1	2	4	7	10	8	9	3	5
Wine	5	1	2	7	4	9	6	10	3	8
Monk 2	9	1	2	4	10	7	8	6	3	5
U. F. Diagnostics A	4	1	2	7	6	10	5	9	3	8
Sonar vs Rocks	2	1	3	5	8	9	6	10	4	7
V. Silhouettes	6	1	2	7	4	9	5	10	3	8
Parkinson Speech	4	1	2	7	6	8	5	10	3	9
Libras Movement	10	1	2	4	6	9	7	8	3	5

comportamiento descendente a mayor porcentaje de valores perdidos. Cuando se presenta un 5% de valores perdidos, este método clasifica correctamente el 84.24% de dichos valores. Cuando se presenta un 50% de valores perdidos, se clasifica correctamente el 54.10% de tales datos.

El clasificador CMI (en color negro) nuevamente es el que mejor clasifica los valores perdidos, mostrando un comportamiento ascendente en cuanto a la medida MVA. Muestra de lo anterior se aprecia en los diversos porcentajes de pérdida. Por ejemplo, cuando se tiene un 5% de valores perdidos, se clasifican correctamente el 97.93% y cuando se llega al 50% de valores perdidos el MVA es igual a 99.43%.

Respecto a los métodos Pydra (en colores gris) es notorio que las variantes 0 y 6 son las que obtienen un desempeño superior a las variantes restantes; se ratifica, además, que su comportamiento es igual entre las dos variantes, las cuales rondan los resultados del clasificador UMI.

C. Resultados globales y finales

Debido a limitaciones prácticas, este documento no puede proporcionar una revisión exhaustiva de cada uno de los resultados obtenidos en los diez bancos de datos utilizados. Dicho lo anterior y después de haber presentado los resultados en 2 bancos de datos, se procede a mostrar en la tabla 8, los resultados globales de los clasificadores mediante un número que representa el lugar que obtienen los mismos, indicando con el número 1 al clasificador que obtuvo la mejor clasificación y con el número 10 al que obtuvo la peor.

Con base en la tabla 8 se muestran los resultados finales a manera de lista en la tabla 9. De esta última se desprende de forma evidente que clasificador CMI es el que clasifica mejor los patrones con valores perdidos, siendo en todos los bancos de datos el primer lugar. Aunado a esto exhibe un comportamiento interesante, ya que por lo general a mayor porcentaje de valores perdidos la medida MVA debería tender a bajar, fenómeno que no sucede en este clasificador y hecho que da pauta a un posible análisis detallado del mismo. Lo anterior también responde al hecho de que en ocasiones se logra clasificar el 100% de patrones con valores perdidos, lo cual de principio parece algo ideal; sin embargo, es posible que el origen de este hecho sea la información artificial que se agrega al momento de la imputación.

Con respecto a los métodos Pydra propuestos, las variantes 0, 6 y 1 son las que obtienen mejores desempeños después de CMI. Cabe mencionar que las 2 primeras mencionadas tienen resultados idénticos en todo momento. De manera que se pueden considerar a las variantes Pydra 0 y 6 como una buena opción para el tratamiento de valores perdidos en los bancos de datos utilizados. Como se puede verificar, las variantes

restantes 2, 4, 7, 3 y 5 obtienen resultados no tan favorables, estando además por debajo del clasificador UMI. El comportamiento de las propuestas Pydra con respecto a la medida MVA es de forma descendente cuando hay mayor porcentaje de valores perdidos.

Finalmente, el clasificador UMI se ubica en la posición 5, por lo que bien podría considerarse una opción promedio. De igual manera que en las variantes Pydra, su comportamiento es descendente en MVA cuando los valores perdidos aumentan.

D. Test estadístico

Debido al número de datos presentados (producto del número de las variantes propuestas y los diferentes porcentajes de valores perdidos), los resultados de los experimentos tienden a ser un poco confusos. Por tal motivo, se utilizaron las tablas 6 y 7 donde se muestran los resultados mediante un ordenamiento del mejor al peor desempeño. Dichas tablas fueron creadas con base en los resultados obtenidos en los experimentos sobre los 10 bancos de datos. Por otro lado, se determinó realizar un análisis de los resultados mediante un test estadístico a fin de que los resultados puedan evidenciar si realmente las diferencias entre los desempeños son significativas.

El test utilizado es el de rangos con signo de Wilcoxon debido a que el problema es de comparación. Los datos presentados cumplen con las características necesarias para llevar a cabo dicho test, y el objetivo es conocer si los resultados se deben al azar o existe diferencia significativa entre ellos. Se tienen en total diez tablas con los resultados de los diez porcentajes de pérdida; sin embargo, como se mencionó previamente, un exhaustivo abordaje de cada una está más allá del alcance de este estudio. Por tal motivo se muestra en la tabla 10 los resultados de los tests estadísticos para el 35% de valores perdidos, siendo éstos los que mejor reflejan el comportamiento general de los resultados. Los resultados de los tests estadísticos se interpretan de la siguiente manera:

- No hay diferencia significativa
- Fila mejor que columna ●
- Fila peor que columna ○

ISSN 2395-8618

TABLA IX.
RESULTADOS FINALES

Clasificador	Lugar
CMI	1
PY0	2
PY6	3
PY1	4
UMI	5
PY2	6
PY4	7
PY7	8
PY3	9
PY5	10

TABLA X
RESULTADOS TESTS DE WILCOXON PARA 35% DE VALORES PERDIDOS

	U	C	0	1	2	3	4	5	6	7
U	■	○	○						○	
C	●	■	●	●	●	●	●	●	●	●
0	●	○	■	●	●	●	●	●		●
1		○	○	■		●		●	○	
2		○	○		■	●			○	
3		○	○	○	○	■	○		○	○
4		○	○			●	■	●	○	
5		○	○	○			○	■	○	○
6	●	○		●	●	●	●	●	■	●
7		○	○			●		●	○	■

Los resultados de la tabla 10 muestran la superioridad del clasificador CMI, el cual muestra diferencia significativa contra todos los clasificadores comparados, y esta tendencia se mantiene a lo largo de los diferentes porcentajes.

En segundo lugar, se aprecia que las variantes Pydra 0 y 6 de manera regular obtienen el segundo puesto, mostrando diferencia significativa contra todos los clasificadores excepto contra CMI. De nuevo la tendencia se mantiene mientras varían los porcentajes. De manera que ya es posible afirmar que las variantes 0 y 6 son una buena opción para tratar valores perdidos.

Finalmente, las variantes 1, 2, 4, y 7 suelen aparecer de manera esporádica en tercer puesto sin un claro dominio por parte de alguna propuesta, tendencia que se mantiene de forma general. Dicho lo anterior sería necesario un análisis para saber en qué bancos de datos y por qué funcionan mejor esas variantes.

VII. CONCLUSIONES

En este estudio se realizó una fuerte experimentación con valores perdidos generados artificialmente. El trabajo ha cubierto la propuesta de un método con ocho variantes para tratar valores perdidos, además de una medida de desempeño propuesta para evaluar la clasificación sólo en valores perdidos.

Se concluyó que, por lo general, la medida propuesta MVA termina impactando de forma directa en el desempeño alcanzado por las otras medidas de desempeño; es decir, un mejor MVA implicará una mejor evaluación en otras medidas. Además de lo anterior, la medida propuesta permite resolver la incertidumbre del comportamiento aislado de un método para tratar valores perdidos evaluándolo sólo con valores perdidos.

En cuanto al método Pydra, las variantes 0 y 6 de la propuesta han mostrado diferencias significativas contra las demás propuestas y contra los métodos existentes en el estado del arte, lo cual convierte al método Pydra en una opción competitiva para tratar valores perdidos cuando se desea aprovechar la ventaja de no tener que hacer un pre procesamiento de los datos. Lo anterior coloca a la propuesta dentro del enfoque de técnicas tolerantes para tratar datos incompletos. En cuanto a la comparación contra los métodos del estado del arte, se puede mencionar que el método unconditional mean imputation presentó resultados mejores que las variantes no sobresalientes de Pydra y peores contra las propuestas prometedoras. Por otro lado, conditional mean imputation fue el que obtuvo la mejor medida en MVA, curiosamente llegando en ocasiones a clasificar el 100% de los patrones con valores perdidos, hallazgo que no ha pasado desapercibido y que es digno de un análisis como trabajo futuro, ya que existe la sospecha de que el método está siendo beneficiado por los valores artificiales agregados al banco de datos. De igual manera, como trabajo futuro se pretende tomar en cuenta bancos de datos con diferentes tipos de complejidad, experimentar con una cantidad mayor de bancos de datos y comparar los resultados contra otros métodos para tratar valores perdidos a fin de mejorar el método propuesto y extenderlo para mitigar diferentes complejidades de datos.

REFERENCIAS

[1] L.O. Silva and L.E. Zárate, "A brief review of the main approaches for treatment of missing data," *Intell. Data Anal.*, vol. 18, pp. 1177–1198, 2014. DOI: DOI 10.3233/IDA-140690.

[2] P.D. Allison, "Missing Data," 2001.

[3] T. Tressy and E. Rajabi, "A systematic review of machine learning-based missing value imputation techniques," *Data Technol. Appl.*, vol. 55, no. 4, pp. 558–585, 2021. DOI: 10.1108/DTA-12-2020-0298.

[4] F. Barzi and M. Woodward, "Imputations of Missing Values in Practice: Results from Imputations of Serum Cholesterol in 28 Cohort Studies," *Am. J. Epidemiol.*, vol. 160, no. 1, pp. 34–45, 2004. DOI: 10.1093/aje/kwh175.

[5] N.M. Gibson and S. Olejnik, "Treatment Of Missing Data At The Second Level Of Hierarchical Linear Models," *Educ. Psychol. Meas.*, vol. 63, no. 2, pp. 204–238, 2003. DOI: 10.1177/ 0013164 402250987.

[6] C. Rioux and T.D. Little, "Missing data treatments in intervention studies: What was, what is, and what should be," *Int. J. Behav. Dev.*, vol. 45, no. 1, pp. 51–58, 2021. DOI: 10.1177/01650 25419880609.

[7] Liang Jin, Yingtao Bi, Chenqi Hu, Jun Qu, Shichen Shen, Xue Wang, and Yu Tian, "A comparative study of evaluating missing value imputation methods in label-free proteomics," *Sci. Rep.*, vol. 11, no. 1, 2021. DOI: 10.1038/s41598-021-81279-4.

[8] C. Zhang, X. Zhu, J. Zhang, Y. Qin, and S. Zhang, "GBKII: An imputation method for missing values," *Lect. Notes Comput. Sci.*, vol. 4426, LNAI, pp. 1080–1087, 2007. DOI: 10.1007/978-3-540-71701-0_122.

[9] K.M. Lang and T.D. Little, "Principled missing data treatments," *Prev. Sci.*, vol. 19, no. 3, pp. 284–294, 2018. DOI: 10.1007/s11121-016-0644-5.

[10] D.B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976. DOI: 10.1093/biomet/63.3.581.

[11] J.D. Kromrey and C.V. Hines, "Nonrandomly missing data in multiple regression: An empirical comparison of common

ISSN 2395-8618

missing-data treatments," *Educ. Psychol. Meas.*, vol. 54, no. 3, pp. 573–593, 1994. DOI: 10.1177/0013164494054003001.

[12] J.L. Schafer and J.W. Graham, "Missing data: Our view of the state of the art.," *Psychol. Methods*, 2002.

[13] R. Little and D. Rubin, "Statistical analysis with missing data," Third Edition, 2019.

[14] D.C. Howell, "The analysis of missing data," *Handb. Soc. Sci. Methodol.*, pp. 1–44, 2008.

[15] L. Ben-Othman and S. Ben-Yahia, "Yet another approach for completing missing values," *Lect. Notes Comput. Sci.*, vol. 4923 LNAI, pp. 155–169, 2008. DOI: 10.1007/978-3-540-78921-5_10.

[16] P. Thionet, M.H. Hansen, W.N. Hurwitz, and W.G. Madow, "Sample Survey Methods and Theory," *Econometrica*, vol. 23, no. 1, pp. 111, 1955. DOI: 10.2307/1905593.

[17] B. Wilmots, Y. Shen, E. Hermans, and D. Ruan, "Missing data treatment: Overview of possible solutions," Steunpunt Mobil. Openb. Werken, 2011. http://www.steunpuntmowverkeersveiligheid.be/sites/default/files/RA-MOW-2011-002.pdf%5Cn.

[18] M.K. Meyers and I. Garfinkel, "Social Indicators and the Study of Inequality," *SSRN Electron. J.*, 2011. DOI: 10.2139/ssrn.1018735.

[19] F.J. Molnar, B. Hutton, and D. Fergusson, "Does analysis using 'last observation carried forward' introduce bias in dementia research?," *CMAJ*, vol. 179, no. 8, pp. 751–753, 2008. DOI: 10.1503/cmaj.080820.

[20] L.A. Galarza-Guerrero, "Comparación mediante simulación de los métodos em e imputación múltiple para datos faltantes," pp. 83, 2013.

[21] J.H. Friedman, R. Kohavi, and Y. Yun, "Lazy decision trees," *Proc. Natl. Conf. Artif. Intell.*, vol. 1, pp. 717–724, 1996.

[22] G. Kalton, "Compensating for missing survey data," Res. Rep. Ser. / Inst. Soc. Res., pp. 157, 1983.

[23] P.S. Kott, J.T. Lessler, and W.D. Kalsbeek, "Nonsampling Error in Surveys.," *J. Am. Stat. Assoc.*, vol. 88, no. 424, pp. 1470, 1993. DOI: 10.2307/2291300.

[24] S. Zhang, "Nearest neighbor selection for iteratively kNN imputation," *J. Syst. Softw.*, vol. 85, no. 11, pp. 2541–2552, 2012. DOI: 10.1016/j.jss.2012.05.073.

[25] S. Zhang, J. Zhang, X. Zhu, Y. Qin, and C. Zhang, "Missing Value Imputation Based on Data Clustering," *Trans. Comput. Sci. I*, pp. 128–138, 2008. DOI: 10.1007/978-3-540-79299-4_7.

[26] F. Rosenblatt, "The perceptron: A probabilistic model for information storage and organization in the brain," *Psychol. Rev.*, vol. 65, no. 6, pp. 386–408, 1958. DOI: 10.1037/h0042519.

[27] J.M.C. Sousa and U. Kaymak, "Fuzzy decision making in modeling and control," *World Sci. Ser. Robot. Intell. Syst.*, vol. 27, no. 27, pp. 335, 2002.

[28] S. Azim and S. Aggarwal, "Hybrid model for data imputation: Using fuzzy c means and multi layer perceptron," *in: Souvenir 2014 IEEE Int. Adv. Comput. Conf. IACC´14*, pp. 1281–1285, 2014. DOI: 10.1109/IAdCC.2014.6779512.

[29] D. Li, J. Deogun, W. Spaulding, and B. Shuart, "Towards missing data imputation: A study of fuzzy K-means clustering method," *Lect. Notes Artif. Intell. (Subseries Lect. Notes Comput. Sci.)*, vol. 3066, pp. 573–579, 2004, doi: 10.1007/978-3-540-25929-9_70.

[30] J.I. Peláez, J.M. Doña, and J.A. Gómez-Ruiz, "Analysis of OWA operators in decision making for modelling the majority concept," *Appl. Math. Comput.*, vol. 186, no. 2, pp. 1263–1275, 2007. DOI: 10.1016/j.amc.2006.07.161.

[31] E. Herrera, F. Chiclana, F. Herrera, and S. Alanso, "Group decision-making model with incomplete fuzzy preference relations based on additive consistency," *IEEE Trans. Syst. Man, Cybern. B Cybern.*, vol. 37, no. 1, pp. 176–189, 2007.

[32] Z. Pawlak, "Rough sets*," Int. J. Comput. Inf. Sci.*, vol. 11, no. 5, pp. 341–356, 1982.

[33] L. Breiman, J. Friedman, C. Stone, and R. Olshen, "Classification and Regression Trees," *Taylor & Francis*, 1984.

[34] J.R. Quinlan, "C4.5 - Programs for Machine Learning," 1993.

[35] Q. Song and M. Shepperd, "A new imputation method for small software project data sets," *J. Syst. Softw.*, vol. 80, no. 1, pp. 51–62, 2007. DOI: 10.1016/j.jss.2006.05.003.

[36] J.W. Grzymala-busse, M. Hu, and N. York, "A Comparison of Several Approaches to Missing Attribute Values in Data Mining," *Rough Sets Curr. Trends Comput.*, pp. 378–385, 2000

[37] J. Larrey-Ruiz, J. Morales-Sánchez, J.L. Sancho-Gómez, R. Verdú-Monedero, and P.J. García-Laencina, "Algoritmo KNN basado en información mutua para clasificación de patrones con valores perdidos," 2008.

[38] J. Ruiz-Shulcloper, "Pattern recognition with mixed and incomplete data," *Pattern Recognit. Image Anal.*, vol. 18, no. 4, pp. 563–576, 2008. DOI: 10.1134/S1054661808040044.

[39] L. Vilahomat, "Extensión al clasificador asociativo Gamma para el manejo de datos mezclados e incompletos," *Universidad Máximo Gómez Báez*, 2015.

[40] Y. Villuendas-Rey, C.F. Rey-Benguría, Á. Ferreira-Santiago, O. Camacho-Nieto, and C. Yáñez-Márquez, "The Naïve Associative Classifier (NAC): A novel, simple, transparent, and accurate classification model evaluated on financial data," *Neurocomputing*, vol. 265, pp. 105–115, 2017. DOI: 10.1016/j.neucom.2017.03.085.

[41] I. López, A. Argüelles, O. Camacho, and C. Yáñez, "Pollutants time series prediction using the Gamma classifier," *Int. J. Comput. Intell. Syst.*, vol. 4, pp. 680–711, 2011

[42] C. Yáñez-Márquez, I. López-Yáñez, M. Aldape-Pérez, O. Camacho-Nieto, A.J. Argüelles-Cruz, and Y. Villuendas-Rey, "Theoretical Foundations for the Alpha-Beta Associative Memories: 10 Years of Derived Extensions, Models, and Applications," *Neural Process. Lett.*, vol. 48, no. 2, pp. 811–847, 2018. DOI: 10.1007/s11063-017-9768-2.

[43] Y. Villuendas-Rey, J.A. Hernández-Castaño, O. Camacho-Nieto, C. Yáñez-Márquez, and I. López-Yañez, "NACOD: A naïve associative classifier for online data," *IEEE Access*, vol. 7, pp. 117761–117767, 2019. DOI: 10.1109/ACCESS.2019.2936366.

[44] A. Rangel-Diaz, Y. Villuendas-Rey, C. Yañez-Marquez, and O. Camacho-Nieto, "The Naïve Associative Classifier with Epsilon Disambiguation," *IEEE Access*, vol. 8, pp. 51862–51870, 2020. DOI: 10.1109/ACCESS.2020.2979054.

[45] M. Aldape-Pérez, A. Alarcón-Paredes, C. Yáñez-Márquez, I. López-Yáñez, and O. Camacho-Nieto, "An associative memory approach to healthcare monitoring and decision making," Sensors (Switzerland), vol. 18, no. 8, 2018. DOI: 10.3390/s18082690.

[46] R. Ramírez-Rubio, M. Aldape-Pérez, C. Yáñez-Márquez, I. López-Yáñez, and O. Camacho-Nieto, "Pattern classification using smallest normalized difference associative memory," *Pattern Recognit. Lett.*, vol. 93, pp. 104–112, 2017. DOI: 10.1016/j.patrec.2017.02.013.

[47] A.V. Uriarte-Arcia, I. López-Yáñez, C. Yáñez-Márquez, J. Gama, and O. Camacho-Nieto, "Data stream classification based on the gamma classifier," *Math. Probl. Eng.*, 2015. DOI: 10.1155/2015/939175.

[48] A.V. Uriarte-Arcia, I. López-Yáñez, and C. Yáñez-Márquez, "One-hot vector hybrid associative classifier for medical data classification," *PLoS One*, vol. 9, no. 4, 2014. DOI: 10.1371/journal.pone.0095715.

[49] C. Yáñez, *"Memorias Asociativas Basadas en Relaciones de Orden y Operadores Binarios*," Instituto Politécnico Nacional, 2002.

ISSN 2395-8618

[50] M.E. Acevedo, "*Memorias Asociativas Bidireccionales Alfa-Beta*," PhD thesis, Instituto Politécnico Nacional, 2006.

[51] M.E. Acevedo-Mosqueda, C. Yáñez-Márquez, and I. López-Yáñez, "Complexity of Alpha-Beta bidirectional associative memories," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 4293, pp. 357–366, 2006. DOI: 10.1007/11925231_34.

[52] R. Flores, "Memorias asociativas Alfa-Beta basadas en el código Johnson-Möbius modificado," *Instituto Politecnico Nacional*, 2006.

[53] I. López, "Teoría y aplicaciones del Clasificador asociativo gamma," *Instituto Politecnico Nacional*, 2011

[54] I. López, "Clasificador Automatico de alto desempeño," *Instituto Politecnico Nacional*, 2007.

[55] J. Han and M. Kamber, "Data mining: Concepts and techniques", *Morgan Kaufmann Publishers*, vol. 49, no. 6, 2000.

[56] R. Kohavi, "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," *Int. Jt. Conf. Artif. Intell.*, vol. 2, pp. 1137–1143, 1995.

[57] L. Victoria, F. Alberto, G. Salvador, P. Vasile, and H. Francisco, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci. (Ny).*, vol. 250, pp. 113–141, 2013.

[58] V. García, J.S. Sánchez, and R.A. Mollineda, "On the effectiveness of preprocessing methods when dealing with different levels of class imbalance," *Knowledge-Based Syst.*, vol. 25, no. 1, pp. 13–21, 2012. DOI: 10.1016/j.knosys.2011.06.013.

[59] A. Orriols and E. Bernadó, "Evolutionary rule-based systems for imbalanced datasets," *Soft Comput.*, vol. 13, pp. 213–225, 2009

[60] A. Frank and A. Asuncion, "UCI Machine Learning Repository," 2010. http://archive.ics.uci.edu/ml.

[61] J. Alcalá-Fernández, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *J. Mult. Log. Soft Comput.*, 2011.