matemática: Una publicación del ICM – ESPOL
2011, Vol. 9, No. 1

TÉCNICAS DE MINERÍA DE DATOS APLICADA A BASES DE DATOS IMPUTADAS. UN CASO DE ESTUDIO

[1]Vallejos Oscar, [2]Valesani Maria, [3]Rigonatto Enzo

Resumen. *El presente trabajo tiene por objeto presentar un caso de estudio sobre la comparación de bases reales y bases de datos imputadas aplicando técnicas de minerías de datos a los efectos de poder concluir si la obtención de información resultante en cada una de ellas son similares o presentan un grado de error aceptable, observando la variabilidad de los patrones de comportamiento en los valores de las variables e interpretando y evaluando los datos una vez aplicado el modelo de minería de datos. El trabajo se estructura de la siguiente manera: Introducción a la teoría de imputación de datos y minería de datos, un detalle no exhaustivo de los operadores de agregación, algoritmo de kmeans, clustering, para luego describir pormenorizadamente el experimento y finalmente presentar las conclusiones y líneas futuras. Al final se presenta una bibliografía abundante sobre el trabajo.*

Palabras Claves: imputación de datos, minería de datos, operador OWA, k-means, clustering.

Abstrat. *This paper aims to present a case study on the comparison of actual bases and databases imputed using data mining techniques for the purpose of being able to conclude whether the resulting information obtained in each are similar or have a degree of error, noting the variability of patterns in the values of the variables and interpreting and evaluating data after applying the data mining model. The paper is structured as follows: Introduction to the theory of imputation of data and data mining, a non-exhaustive detail the aggregation operators, kmeans algorithm, clustering, and then describes in detail the experiment and finally present the conclusions and future lines. In the end he presents an abundant literature on the job.*

Key words. *Data imputation, data mining, OWA operator, k-means, clustering.*

Recibido: Noviembre, 2010
Aceptado: Febrero, 2011

1. INTRODUCCIÓN A LA TEORÍA DE IMPUTACIÓN DE DATOS

La imputación de datos puede ser considerada como la etapa final de un proceso de depuración de datos, ya sea por datos faltantes o valores cuyas reglas de edición han sido fallidas y serán reemplazados por valores aceptables conocidos. Se la puede definir simplemente como promedios o selecciones provenientes de una distribución de predicción de los valores faltantes que se basa en los valores observados. [13] [18] [11].

La razón principal para realizar imputaciones es obtener un conjunto de datos completos y consistentes al cual se le pueda aplicar las técnicas de estadística clásica, de la lógica difusa e incluso minería de datos [12][9][3][13]. Encontrar el mejor método de imputación, o el más eficiente, es una tarea importante ya que se puede cometer errores en las imputaciones de datos individuales, e inclusive, pueden aparecer aumentados al realizar estadísticas agregadas [20][19]. Por lo tanto se puede entender que es razonable estudiar métodos de imputación que conserven características de la variable como pueden ser: preservación de la distribución real del contenido de la variable, su relación con el resto de variables en estudio, etc [5].

Operador OWA

Las bases de datos utilizadas en el experimento fueron aquellas a las que se imputaron los datos faltantes utilizando el operador OWA. (*Ordered Weighted Averaging*). Son muy utilizados en los procesos de toma de decisión y existen en la actualidad numerosos trabajos de investigación en distintas áreas, como para ser utilizado en la imputación de datos faltantes. [18] Esta nueva técnica de agregación basada en un promedio de pesos ordenados (OWA), fue introducido por Yager y a poco tiempo de su aparición, y posteriores adaptaciones, se han convertido en una de las familias de operadores más usada en la actualidad.

Un operador OWA de dimensión n es una aplicación $F : \Re^n \to \Re$, que tiene un vector de ponderaciones asociado $W = \left[w_1, w_2, ..., w_n\right]^T$ tal que

i) $w_i \in [0,1]$, $1 \le i \le n$,

[1] *Vallejos Oscar, Ing., Departamento de Informática, Universidad Nacional del Nordeste; Corrientes. Argentina (e_mail: ovallejos@exa.unne.edu.ar)*
[2] *Valesani María, Ing., Departamento de Informática, Universidad Nacional del Nordeste; Corrientes. Argentina (e_mail: mevalesani@exa.unne.edu.ar)*
[3] *Rigonatto Enzo, Licenciatura en Sistemas de Información, Universidad Nacional del Nordeste; Corrientes. Argentina (e_mail: enzo.rigonatto08@gmail.com)*

ii) $\sum_{i=1}^{n} w_i = 1,$

donde $F(x_1,x_2,...,x_n) = \sum_{ki=1}^{n} w_k x_{jk}$ siendo x_{jk} el k-ésimo elemento más grande de la colección $x_1, x_2, ..., x_n$ [16].

2. MINERÍA DE DATOS

La minería de datos (DM, Data Mining) consiste en la extracción no trivial de información que reside de manera implícita en los datos. Dicha información era previamente desconocida y podrá resultar útil para algún proceso. En otras palabras, la DM, sondea y explora los datos para sacar la información oculta en ellos [4][6][7]. Bajo el nombre de minería de datos se engloba todo un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las bases de datos. Está fuertemente ligado con la supervisión de procesos industriales ya que resulta muy útil para aprovechar los datos almacenados en las bases de datos [10][1]. Las bases de la minería de datos se encuentran en la inteligencia artificial y en el análisis estadístico. Mediante los modelos extraídos utilizando técnicas de minería de datos se aborda la solución a problemas de predicción, clasificación y segmentación [14].

Las técnicas más representativas, son entre otras, Redes Neuronales, Regresión Lineal, Árboles de decisión, modelos estadísticos y Agrupamiento o Clustering. Se describen a continuación las dos utilizadas en el caso de estudio.

Modelos estadísticos.

Es una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales y en la regresión para indicar los diferentes factores que modifican la variable de respuesta.

Agrupamiento o Clustering.

Es un procedimiento de agrupación de una serie de vectores según criterios habitualmente de distancia; se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes. (*K-means; K-medoids*).

Se basa en intentar responder como es que ciertos objetos (casos) pertenecen o "caen" naturalmente en cierto número de clases o grupos, de tal manera que estos objetos comparten ciertas características. Esta definición asume que los objetos pueden dividirse, razonablemente, en grupos que contienen objetos similares. Si tal división existe, ésta puede estar oculta y debe ser descubierta. Este es el objetivo principal de las técnicas o estudios de clustering

• Medidas de Disimilaridad (Similaridad)

Asocian un número (d_{ij}) a un par de Objetos/Datos (i,j), donde: (Sea S el subespacio de objetos a clasificar)

$d_{ij} \geq 0$ para todo $i, j \in S$

$d_{ij} = 0$ para todo $i = j \in S$

$d_{ij} = d_{ji}$ para todo $i, j \in S$

$d_{ij} \leq d_{iz} + d_{zj}$

• Distancias

$d_{ij} = \sum_{k=1}^{P} W_k \left| x_{ik} - x_{jk} \right|$ City-Block

$d_{ij} = \sqrt{\sum_{k=1}^{P} W_k \left(x_{ik} - x_{jk} \right)^2}$ Euclídea

$d_{ij} = \sqrt[\lambda]{\sum_{k=1}^{P} W_k \left(x_{ik} - x_{jk} \right)^{\lambda}}$ $\lambda > 0$ Minkowski

Tipos de Clustering

• Jerárquico (Hierarchical) : dendrogramas, grafos (arboles)

• De partición: División en grupos (SOM, LVQ, etc.)

El algoritmo de las *K-medias* es un algoritmo de partición. Básicamente este algoritmo busca formar clusters (grupos) los cuales serán representados por K objetos. Cada uno de estos K objetos es el valor medio de los objetos que pertenecen a dicho grupo:

1. Inicialmente se seleccionan K objetos del conjunto de entrada. Estos K Objetos serán los centroides iniciales de los *K-grupos*.
2. Se calculan las **distancias** de los objetos (datos) a cada uno de los centroides. Los datos (objetos) se asignan a aquellos grupos cuya distancia es mínima con respecto a todos los centroides.
3. Se actualizan los centroides como el valor medio de todos los objetos asignados a ese grupo

$$c_j = \frac{1}{\left| C_j \right|} \sum_{\forall x \in C_j} z$$

Donde z representa un elemento del conjunto de datos que pertenece al cluster C_j; c_j es un centroide y $\left| C_j \right|$ corresponde al número de elementos en el cluster C_j

4. Se repite el paso 2 y 3 hasta que se satisface algún criterio de convergencia.

Existen una serie matrices que constituyen el fundamento para la implementación de este tipo de algoritmo, entre ellas: 1.- Matriz de datos; 2.- Matriz de distancias; 3.- Matriz de centroides;
4.- Matriz de pertenencias.
Sus diferentes variantes se basan fundamentalmente en la forma de medir distancias entre los datos y los grupos, el criterio para definir la pertenencia de los datos a cada grupo y la forma de actualizar dichos grupos.

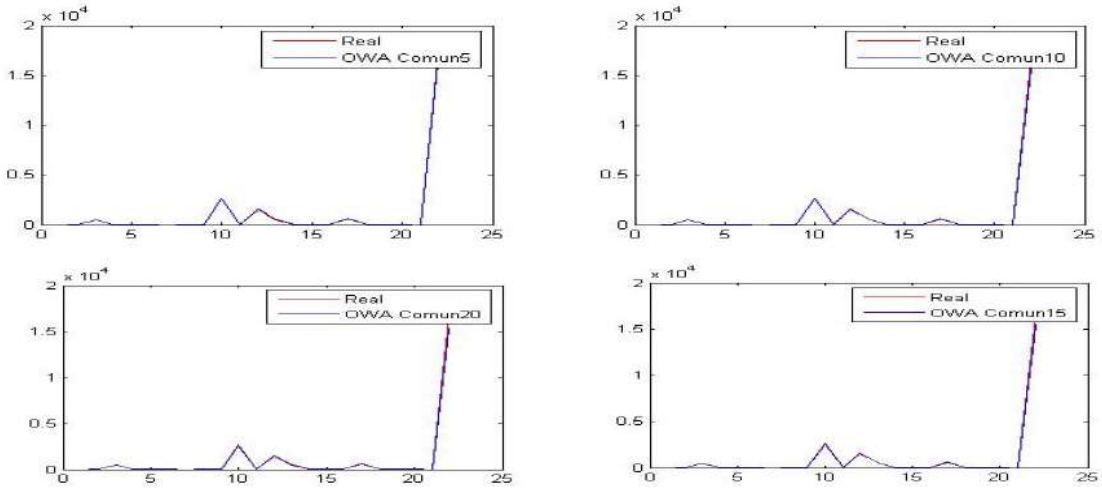
3. DESCRIPCIÓN DE EXPERIMENTO Y ANÁLISIS DE RESULTADOS

A partir de las bases de datos imputadas (sobre distintos porcentuales de datos faltantes) con el operador OWA, ya normalizadas, se realizo la aplicación de dos técnicas de minería de datos.
Se procedió a dividir en dos partes el experimento:
1) Se realizo la aplicación de un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las bases de datos (modelos estadísticos y clustering).
Posteriormente se compararon los resultados obtenidos bajo dichos modelos considerando ciertas medidas de disimilaridad.
2) Se extrajo información implícita de la base real e imputada a los efectos de poder compararlas a partir de la obtención de los resultados [17].

Primera Parte

La primera realizando un análisis estadístico clásico (media, desvío estándar, varianza, diferencias y errores) y luego se aplico el algoritmo de *K-means* Clustering [15].
Para la realización de los experimentos, se utilizo un software científico en el cual se codificó las distintas técnicas para posteriormente ser aplicadas a la base real e imputada.
Se parte de una base de datos completa, la cual esta compuesta por 1964 instancias y 22 tuplas, simulando perdida de datos en distintos porcentuales.

FIGURA 1
Técnicas de minería de datos aplicada a bases de datos imputadas. Un caso de estudio
Desvió Standard de bases real e imputadas por ítems para diversos porcentajes de datos faltantes



En la Figura 1 se aprecia en el grafico donde se trabajo con un 5% de datos faltantes, que la distancia entre datos reales e imputados es minima. Conforme se va aumentado el porcentaje de datos faltantes las distancias con los reales van creciendo.

Luego de realizar las imputaciones de las bases, se procede al análisis estadístico nombrado anteriormente.
Los valores obtenidos en otras mediciones estadísticas son presentados en la Tabla I para su comparación y posterior análisis.

TABLA I
Técnicas de minería de datos aplicada a bases de datos imputadas. Un caso de estudio
Representación de valores obtenidos luego de aplicar el análisis estadístico

	Matriz Real	Matriz Imputada con 15% de ausentismo	Matriz Imputada con 20% de ausentismo
Media	3400	3353.9	3344.8
Varianza	2.7479e+008	2.4097e+008	2.4097e+008
Error total Desvío Estándar	-	1053.7	1272.1
Error total Varianza	-	3,3824e+007	4,0557e+007

Si observamos la tabla I y II, denota claramente una mínima diferencia de los resultados en la BD real y la imputada. Esto implica que el comportamiento de los valores de las bases imputadas posee solo pequeñas diferencias comparadas con la base real.
Luego de analizar los diversos resultados obtenidos con esta primera técnica de minería, se completo este segmento del experimento aplicando el algoritmo *k-means* a las bases.
En las Figuras 2, 3 y 4 se muestran diversos gráficos representando los resultados obtenidos con la aplicación del algoritmo.
Luego de la aplicación del algoritmo *K-means,* se puede observar en los gráficos los diversos comportamientos de los valores de las bases. Como se puede ver las diferencias entre la base real y las bases imputadas son ínfimas, tal como paso ya con la primera técnica de minería aplicada.

FIGURA 2
Técnicas de minería de datos aplicada a bases de datos imputadas. Un caso de estudio
K-means aplicado a la Base real

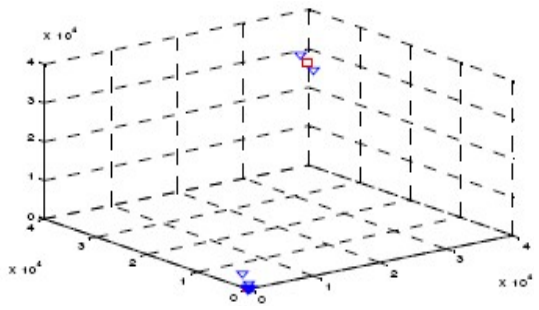


FIGURA 3
Técnicas de minería de datos aplicada a bases de datos imputadas. Un caso de estudio
K-means aplicado a la base imputada con 10% de datos faltantes

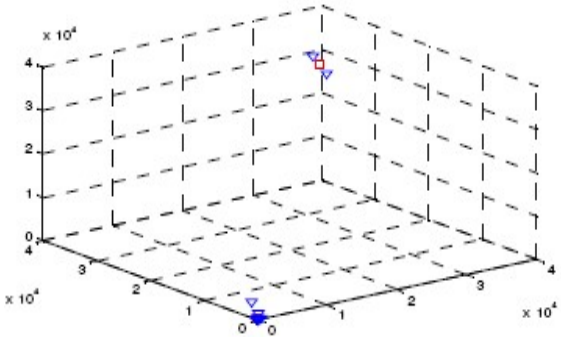
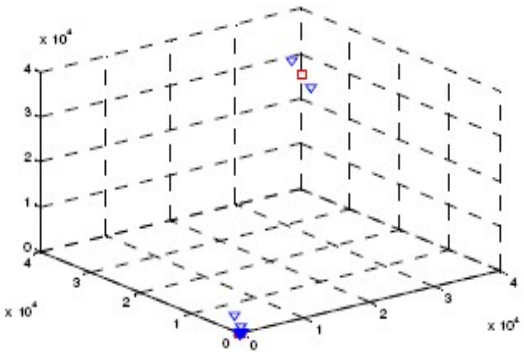


FIGURA 4
Técnicas de minería de datos aplicada a bases de datos imputadas. Un caso de estudio
K-means aplicado a la base imputada con 20 % de datos faltantes



Segunda Parte
Se extrajeron diferentes tipos de información de la totalidad de las tuplas que conforman la base de dato real como aplicación propia de la minería de datos. Se aplicaron estos algoritmos también a la

base de datos imputada a los efectos de poder comparar con las obtenidas a partir de la base real. A manera de ejemplo se presenta dos casos concretos y los resultados obtenidos en ambos, siendo estos una representatividad exacta de los obtenidos en cada una de las tuplas.

Ejemplo A

Se obtuvo la cantidad de instancias distribuidas por año de adquisición del bien en la base real y en la base imputada. Los resultados obtenidos se reflejan en la Tabla II. En ella se aprecia que la diferencia porcentual de los datos obtenidos en ambas bases es mínimo.

Ejemplo B

Se realizo un ordenamiento por el campo de rubro del bien a los efectos de determinar la distribución de las instancias tanto en la base real como en la imputada. Luego se procedió a calcular su porcentual sobre el total en cada una de ellas.

TABLA II

Cuadro comparativo de la información extraída de la base Real y la Imputada (tupla: años)

Años	Real	%	Imputada	%	Diferencia
1989	54	2,75	46	2,34	0,41
1990	126	6,42	111	5,66	0,76
1991	1438	73,29	1492	76,04	-2,75
1992	293	14,93	265	13,51	1,43
1993	24	1,22	24	1,22	0,00
1996	10	0,51	9	0,46	0,05
1997	1	0,05	1	0,05	0,00
2008	16	0,82	14	0,71	0,10
Total	**1962**	**100,00**	**1962**	**100,00**	

En la tabla se aprecia que la diferencia de los valores absolutos y porcentuales de la base real e imputada es mínima.

4. CONCLUSIONES

A partir de las bases imputadas se ha podido extraer información que residía implícitamente en los datos para su posterior utilización en otro proceso, con lo cual podemos afirmar, en este caso, que en estas bases imputadas en un rango de porcentaje entre 5 % y 25 % de aleatoriedad, se puede preparar, sondear y explorar los datos para obtener información oculta. La extracción de conocimiento procesable, implícito en las bases de datos no ha variado de aquel que se extrajo de las bases de datos imputadas. Los patrones de comportamiento observados en los valores de las variables del problema o relaciones entre dichas variables fueron los mismos en las bases de datos reales e imputadas. La interpretación y evaluación de datos con los modelos de la estadística clásica y clustering fue satisfactorio.

REFERENCIAS BIBLIOGRÁFICAS Y ELECTRÓNICAS

[1]. FRAWLEY, W., PIATETSKY-SHAPIRO G. Y MATHEUS, C. (1992): "*Knowledge Discovery in Databases: An Overview*". AI Magazine, pp 213-228.

[2]. BASSEVILLE, M., AND NIKIFOROV, I. V. (1993). "*Detection of Abrupt Changes: Theory and Application*". Englewood Cliffs, N.J.: Prentice Hall.

[3]. YAGER R. (1993). "*Families of OWA operators. Fuzzy Sets and Systems*". 59:125-148.

[4]. AGRAWAL, R., & PSAILA, G. (1995). "*Active Data Mining. In Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95)*", 3–8. Menlo Park, Calif.: American Association for Artificial Intelligence.

[5]. TRENA M. EZZATI-RJCE, MEENA KHARE , DONALD B. RUBIN , RODERICK J. A. LITTLE, JOSEPH L. SCHAFER. (1995) "*A comparison of imputation techniques in the third national health and nutrition examination Surrey*". National Center for Health Statistics, Harvard University, University of Michigan, Pennsylvania State University, 6525 Belcrest Road, Hyattsville, MD 20782.

[6]. AGRAWAL, R.; MANNILA, H.; SRIKANT, R.; TOIVONEN, H.; AND VERKAMO, I. (1996). "*Fast Discovery of Association Rules*". In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 307–328. Menlo Park, Calif.: AAAI Press.

[7]. APTE, C., AND HONG, S. J. (1996). "*Predicting Equity Returns from Securities Data with Minimal Rule Generation*". In Advances in Knowledge Discovery and Data Mining, eds. U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, 514–560. Menlo Park, Calif.: AAAI Press.

[8]. YAGER R. (1996). "*Quantifier Guided Aggregation Using OWA Operators*". International Journal of Intelligent Systems. 11, 49-73.

[9]. YAGER R. (1998b). "*New Modes of OWA Information Fusion*". International Journal of Intelligence Systems. 13, 661-681.

[10]. GOEBEL, M. Y GRUENWALD, L. (1999) "*A survey of data mining and knowledge discovery software Tools*". SIGKDD Explorations, vol. 1, nº 1, pp. 20-33.

[11]. DR. JANN-HUEI JINN. (2000). "*The Effect of Different Imputation Methods on Analytical Statistics of Simple Linear Regression*". Department of Mathematics and Statistics. Grand Valley State University Allendale, Michigan 49401.

[12]. YANG C. YUAN. (2000). "*Multiple Imputation for Missing Data: Concepts and New Development., SAS Institute Inc.*", Rockville, MD. P267-25.

[13]. GARCÍA PÉREZ, A. (2001). "*Métodos avanzados de estadística aplicada*". Madrid. Universidad Nacional de Educación a Distancia.

[14]. LAST, M., KLEIN, A. Y KANDEL, A. (2001). "*Knowledge Discovery in Time Series Databases*". IEEE Transactions on Systems, Man and Cybernetics, vol. 31, Part B, nº 1, pp. 160-169.

[15]. LITTLE, R. & RUBIN, D. B. (2002). "*Statistical Analysis with Missing Data, 2 edn, Jonh Wiley &Sons*".

[16]. PELAEZ J.I., DOÑA J.M. (2003a). "*Majority Additive-Ordered Weighting Averaging: A New Neat Ordered Weighting Averaging Operators Based on the Majority Process*", International Journal of Intelligent Systems 18, 469-481.

[17]. GÓMEZ GARCÍA J., PALAREA ALBALADEJO J., (2006). *"Métodos de inferencia estadística con datos faltantes"*. Estudio de simulación sobre los efectos en las estimaciones. Departamento de Métodos Cuantitativos para la Economía. Universidad de Murcia, Departamento de Informática de Sistemas Universidad Católica. San Antonio, Departament de Informática y Matemática Aplicada. Universitat de Girona. -

ESTADÍSTICA ESPAÑOLA Vol. 48, Núm. 162, págs. 241 a 270.

[18]. GRAJALESA L., LÓPEZ L. (2006). "*Imputación de datos en diseños switchback usando un modelo mixto con errores correlacionados Data Imputation in Switchback Designs Using a Mixed Model with Correlated Errors*". Universidad Nacional de Colombia, Departamento de Estadıstica, Bogota Revista Colombiana de Estadıstica Volumen 29 No 2. pp. 221 a 238. Diciembre.

[19]. DOÑA J.M., QUINTANA O.P., VALESANI M.E., VALLEJOS O.A. (2008). "*Analysis of Agregation Methods in Incomplete Database Systems. Information Processing and Management of Uncertainty in Knowledge-Based System (IPMU 2008)*". ISBN: 9978-84-612-3061-7.

[20]. QUINTANA O.P., VALESANI M.E., VALLEJOS O.A. (2008). "*Imputación de datos desaparecidos utilizando operadores de agregación MA-OWA*". (WICC 2008 – X Workshop de investigadores de Ciencias de la Computación. General Pico, La Pampa, Argentina 5 y 6 de Mayo de).