Algoritmo híbrido basado en aprendizaje computacional para el manejo de datos faltantes en aplicaciones OLAP

Hybrid algorithm based machine learning for missing data management applications OLAP

Claudia Liliana Hernández García^{1*} Jorge Enrique Rodríguez Rodríguez¹

Recibido 24 de noviembre de 2014, aceptado 22 de diciembre de 2015

Received: November 24, 2014 Accepted: December 22, 2015

RESUMEN

El artículo presenta el desarrollo y uso de un algoritmo híbrido de aprendizaje computacional para la tarea de relleno de valores faltantes realizada durante la fase de preparación de datos. En primer lugar, se aborda el problema a resolver, el que está orientado al estudio y análisis de diferentes técnicas para el relleno de valores faltantes, con el fin de proponer una técnica híbrida como producto de esta investigación para dicha tarea y asociarla con la tecnología OLAP (Procesamiento Analítico en Línea). Luego, se justifica la metodología de investigación (científica descriptiva-exploratoria con enfoque experimental) aplicada en este proyecto. Se realizó la revisión de técnicas utilizadas en el relleno de valores faltantes; con base en la verificación de las técnicas y los casos de estudio, se seleccionaron métodos basados en vecindad y redes neuronales artificiales, y se propuso una técnica híbrida (KMediasSom) aplicada a un conjunto de datos sintético y a uno real, provenientes de una aplicación OLAP. En seguida, se plantean las pruebas de análisis y resultados con el fin de precisar su aplicabilidad en cuanto a efectividad y complejidad algorítmica se refiere. Finalmente, se presentan las conclusiones, donde se demostró que la técnica híbrida genera mejores resultados que las técnicas usadas por separado.

Palabras clave: Algoritmos híbridos, métodos basados en vecindad, redes neuronales artificiales, valores faltantes, preprocesamiento de datos, aprendizaje computacional, complejidad algorítmica.

ABSTRACT

This paper shows the development and use of a hybrid algorithm of machine learning for the missing values predict task, with this task that is carried out while the data preprocessing phase. Firstly, we go through the problem to solve, which is pointed to the study and analysis of different techniques for the missing values predict in order to suggest a hybrid technique as a product of this research for the task and link it to the OLAP (On-Line Analytical Processing) technology. Then, justifying the research methodology (scientific descriptive - probing) applied in this project. The techniques review was carried out filling out the missing values; based on the techniques proof and the study cases, k-Nearest Neighbors algorithms and artificial neural networks were selected and a hybrid technique (KMediaSom) was suggested, applied to a synthetic data set and a real one, coming from a OLAP; for the algorithms implementation was used Matlab. Right away, the analysis and results are set out in order to specify its applicability about efficacy and time complexity. Results are suitable as for the synthetic data set as for the real one; according to the test signs achieved. Finally, the conclusions, where it is proved that the technique or hybrid algorithms generate better results than the techniques used by separately.

Keywords: Hybrid algorithms, k-Nearest Neighbours algorithms, artificial neural networks, missing data, data pre-processing, machine learning, time complexity.

¹ Universidad Distrital "Francisco José de Caldas". Calle 68 D Bis A sur N° 49 F 70. Bogotá, Colombia.
E-mail: clhernandez@gmail.com; jerodriguezr@udistrital.edu.co

* Autor de correspondencia

INTRODUCCIÓN

El volumen de datos que manejan las organizaciones cada vez es más grande, no sólo porque sus sistemas guardan transacciones más detalladas sino porque actualmente la información histórica también está siendo utilizada como soporte en la toma de decisiones. Para realizar tareas de proyección, las empresas se están valiendo, cada día con mayor frecuencia, de tecnologías de análisis de datos como OLAP o minería de datos. Dentro del tratamiento de datos lo que se denomina limpieza de datos, agrupa tareas como el relleno de valores faltantes que tiene como uno de sus propósitos ayudar en el mejoramiento de la calidad de los datos. Existen diversas técnicas utilizadas para el relleno de valores faltantes, las que se usan dependiendo de algunos factores como la naturaleza de los datos. Asimismo, al verificarse el funcionamiento de las técnicas podría existir la posibilidad de generar una técnica híbrida que produjera mejoras a alguna de ellas y que al realizar la evaluación de indicadores se pudiera establecer la pertinencia de su aplicabilidad, la viabilidad de su implementación y la opción de adaptarla a aplicaciones de tecnología OLAP.

Con este artículo se plantea una técnica híbrida con base en la comparación de técnicas de aprendizaje computacional que suelen ser utilizadas en el relleno de valores faltantes y que puedan enfocarse a aplicaciones OLAP.

Para lograr los planteamientos establecidos se tuvo en cuenta las etapas de investigación [1] ya que el método aplicado en este proyecto fue la investigación científica con enfoque experimental. Además, se utilizaron los casos de estudio como una herramienta que permitió establecer ejemplos reales en los que se emplearon las técnicas, de tal forma que proporcionaron información que ya han sido evaluadas durante investigaciones preliminares y aportaron los resultados como base y argumento para la comparación de las técnicas.

El presente artículo se encuentra organizado en cinco secciones, que recopilan el proceso que se llevó a cabo con el proyecto de investigación. En primer lugar, se hace una descripción del problema a abordar. Posteriormente, las técnicas analizadas junto con la técnica híbrida propuesta. A continuación se aborda la metodología de investigación que se

utilizó. Finalmente, se presenta el análisis de pruebas y resultados, y las conclusiones obtenidas de esta investigación.

DEFINICIÓN DEL PROBLEMA

En el registro de los datos es posible que se presenten diversos errores que pueden llevar a que los datos sean incompletos, con ruido o inconsistentes, características comunes teniendo en cuenta que se pretenden analizar grandes volúmenes de datos almacenados de forma estructurada. En resumen, los datos del mundo real tienden a ser incompletos, tener ruido y ser inconsistentes, la utilización de técnicas de preprocesamiento mejora la calidad de los datos y mejoran la precisión y eficiencia de los posteriores procesos de análisis. La fase de preprocesamiento de datos permite detectar anomalías en los datos y corregirlas a tiempo o reducir el conjunto de datos para el análisis, lo que hace más eficiente la toma de decisiones [2].

Entre los beneficios de utilizar técnicas de preprocesamiento, se encuentran [3]:

- La generación de un conjunto de datos más pequeño que el original, lo que puede mejorar la eficiencia del proceso de análisis o minería de datos.
- La generación de datos de calidad, los que pueden conducir a patrones o reglas de calidad.

En general, el preprocesamiento de los datos engloba todas aquellas técnicas de análisis de datos que permiten mejorar la calidad de un conjunto de datos, de modo que el proceso de análisis o minería de datos se pueda valer de ellos con toda confianza. La etapa de preprocesamiento de datos también enmarca una serie de tareas, dentro de las que se encuentra el relleno de valores faltantes que puede aplicarse al conjunto de datos por medio del borrado de los datos, el reemplazo por la media de los datos o relleno de valores mediante un modelo [4].

La calidad de datos es vista desde diversas dimensiones básicas (exactitud, oportunidad, relevancia, completitud, inteligibilidad y confiabilidad), de las que en este proyecto de investigación se abordó principalmente la completitud de los datos. Los valores de datos faltantes son causa de errores, un

dato sin valor genera ambigüedad ya que puede ser correcto o errado. Su importancia radica en que la eficiencia de la toma de decisiones depende de la calidad de los datos, de tal manera que pequeñas mejoras en las dimensiones de los datos puede conducir a mejoras sustanciales en la información para la toma de decisiones [5]. De ahí que para las organizaciones sea de gran utilidad contar con estudios comprobados de la selección y evaluación de características de técnicas de aprendizaje computacional y utilizar técnicas híbridas que mejoren los resultados obtenidos. En resumen, cualquier esfuerzo realizado para mejorar la calidad de los datos representa beneficios importantes para una organización [5].

En la tecnología OLAP, se considera que los datos deben ser integrados en un *DataWarehouse* (DW) o en un *Datamarts* como prerequisite para análisis eficiente de los datos. De esta forma, el proceso de integración se ejecuta una sola vez y todos los pasos que continúan son ejecutados en el DW o en el *Datamarts* [6]. Así, todo lo que ha dado origen a OLAP, hace que se refiera a una nueva forma para manejar la información de importancia especialmente para la toma de decisiones [7].

Los DW y OLAP necesitan datos históricos, sumariados y consolidados. Los datos están recopilados desde ciertas bases de datos operacionales y otras fuentes externas. Esto requiere capacidad de almacenamiento y una arquitectura típica como la que se muestra en la Figura 1 [7].

Algunos tópicos importantes en el área que involucra tecnología OLAP incluyen limpieza de datos, selección de índices, particionamiento de datos, vistas materializadas y administración del DW. El análisis de datos requiere extracción de datos relevantes del DW, agregando datos y analizando los resultados [8].

Una forma de integrar OLAP con minería de datos es propuesta en [9] especialmente enfocado a la discretización que es un proceso que generaliza un atributo a un intervalo de datos y que reduce y simplifica los datos originales; sin embargo, las tareas asociadas a la discretización aún no están incorporadas en las herramientas utilizadas en aplicaciones OLAP. La minería OLAP es un mecanismo que integra OLAP con minería de datos [10]. De esta forma, se considera una exploración sana de cómo puede ser integrada la visualización

Figura 1. Arquitectura de un sistema típico de DW [8].

de los datos con otras técnicas como aprendizaje inductivo y *clustering* jerárquico [7, 11].

De acuerdo con la arquitectura de un sistema DW, ver Figura 1, lo expuesto en este artículo centra su atención en la limpieza de datos a nivel de relleno de valores faltantes; tarea previa al almacenamiento de los datos en DW y que forma parte de la Extracción, Transformación y Carga (ETL) en dicha arquitectura.

Teniendo en cuenta que se plantea asociar las técnicas de relleno de valores faltantes a aplicaciones de tecnología OLAP, fue importante precisar las siguientes características para las pruebas y experimentación, que se convirtieron en las fronteras del problema:

- Aplica solo para las tablas de hechos y los datos faltantes asociados a las medidas de los cubos OLAP.
- Las medidas en los cubos OLAP siempre son datos de tipo numérico.
- De acuerdo con las condiciones de los datos de un cubo OLAP se habla de datos no supervisados, debido a que el cubo contiene datos consolidados asociados a las diferentes dimensiones. Por lo tanto, no es identificable una clase que pueda ser referenciada para orientar un aprendizaje supervisado; razón por la que se consolidan en el cubo, datos de problemas no supervisados.
- El análisis de datos faltantes, en este caso, fue aplicado solo a un *Datamarts* en Multidimensional Online Analytical Processing (MOLAP) [2].
- Debido a la naturaleza de los datos recolectados del problema, el análisis de valores faltantes se realizó únicamente a datos simples; es decir, no se contempla datos compuestos, como por ejemplo fecha, la que está conformada por los datos simples día, mes y año; esto último se contemplará en trabajos futuros.
- El porcentaje de valores faltantes es cercano al 20%, ya que según [12] se considera como promedio para la evaluación de la exactitud en el proceso de agrupamiento y relleno de valores faltantes. Según [13] al aumentarse el porcentaje de valores faltantes se incrementa el error global debido a que se pierde más información a medida que hay más valores faltantes en los datos.

Basado en lo anterior, se formula la siguiente hipótesis: Si se hace un análisis de diferentes

técnicas para el relleno de valores faltantes; se podría proponer una técnica híbrida para dicha tarea como producto del análisis y emplearla en aplicaciones OLAP.

METODOLOGIA

El método aplicado en este proyecto fue la investigación científica descriptiva-exploratoria con enfoque experimental. De acuerdo con el proceso formal de investigación fue utilizado un método hipotético-deductivo en el que se formuló una hipótesis, que mediante un razonamiento deductivo se validó de manera empírica. Se buscó establecer, con base en la experimentación, un mecanismo de ponderación de los indicadores de evaluación de las técnicas, de forma tal que fuera posible evaluar dicho mecanismo al cambiar el conjunto de datos.

Se definieron las siguientes etapas que permitieron obtener resultados que propendieran por encontrar las técnicas adecuadas para la propuesta de técnica híbrida en el manejo de datos faltantes:

- Análisis y selección de técnicas de aprendizaje computacional utilizadas en el relleno de valores faltantes.
- Definición de dominio y contexto en el que se realizaron las pruebas y comparación de las técnicas (híbrida e individual). Se definió un conjunto de datos sintético con el fin de realizar las pruebas de las técnicas híbrida e individual.
- Selección de herramientas para el preprocesamiento de datos con el fin de verificar las técnicas.
- Definición y aplicación de pruebas para la comparación de la efectividad de las técnicas en el relleno de valores faltantes. Con base en las pruebas realizadas con los datos sintéticos, se realizó la imputación y la evaluación de los indicadores de imputación para aplicar la técnica híbrida al dominio de datos.
- Revisión de resultados de las pruebas y de la complejidad algorítmica. Se realizó el análisis de los datos obtenidos de la imputación con la técnica híbrida y con base en el algoritmo de esta técnica propuesta se definió y revisó su complejidad algorítmica. Asimismo, se revisaron las complejidades algorítmicas de las técnicas individuales.

Los instrumentos con los que se desarrolló esta investigación son principalmente: casos de estudio, pruebas con datos sintéticos y pruebas con datos de casos reales.

ANÁLISIS Y SELECCIÓN DE TÉCNICAS DE APRENDIZAJE COMPUTACIONAL

Los algoritmos o técnicas al igual que en el caso de los lenguajes de programación se eligen de acuerdo con una serie de criterios dentro de los que se pueden destacar [14]: área de aplicación general, complejidad algorítmica y entorno en el cual se ejecutará.

Las Tablas 1 y 2 resumen el resultado de la revisión realizada en la que se establecieron ventajas y desventajas de la utilización de las técnicas, con base en criterios como el área de aplicación, naturaleza de los datos al cual se enfoca o tipo de aprendizaje,

simplicidad y efectividad del algoritmo. Se consultaron otros algoritmos como: algoritmo EM (Expectation Maximization) y el algoritmo GTM (Generative Topographic Mapping) cuyas ventajas y desventajas pueden inferirse en [15-18], respectivamente.

De acuerdo con los resultados de investigación en [24], para dar solución al problema de los datos faltantes son utilizadas técnicas simples y tradicionales como la media y la mediana, pero también existen técnicas especializadas que se denominan *Hot Deck* (*método usado para el relleno de valores faltantes*) dentro de las que se pueden incluir las técnicas K-Medias y SOM. Estas técnicas tienen como rasgo principal que permiten identificar características comunes entre los registros completos y los incompletos. Con base en estas características comunes se realizan agrupaciones y se seleccionan los datos representativos de acuerdo con algún método [25].

Tabla 1. Ventajas técnicas de aprendizaje computacional utilizadas en el relleno de valores faltantes.

Técnica	Ventajas
K-Medias	Simple en su implementación [19]. Algoritmo de aprendizaje no supervisado más simple [20]. Rapidez en encontrar la clasificación adecuada [20]. Más eficiente que los algoritmos jerárquicos [21]. Genera grupos con buena calidad estructural [22].
Mapa Auto - organizativo de Kohonen (SOM)	La red SOM aprende rápidamente [23]. Es capaz de llevar a cabo la clasificación incluso sin la información suficiente [23]. Preserva la topología de los datos de entrada [23].
Imputación de datos con métodos simples y tradicionales	Imputación por la media [24]: la más usada, bajo costo computacional, se usa en variables numéricas, método simple, recomendado solo cuando el porcentaje de valores faltantes no exceda de 5 a 10%. Imputación por la mediana [24]: aplica a variables cuantitativas, se puede aplicar porque funciona bien como una medida de tendencia central.

Tabla 2. Desventajas técnicas de aprendizaje computacional utilizadas en el relleno de valores faltantes.

Técnica	Desventajas
K-Medias	Requiere diversas iteraciones, lo que aumenta el costo computacional [19]. El agrupamiento final depende de los centroides iniciales [19, 21]. La convergencia en el óptimo global no está garantizada iniciales [19].
Mapa Auto - organizativo de Kohonen (SOM)	Al utilizar aprendizaje no supervisado, necesita repetir todo el proceso de aprendizaje para clasificar un nuevo dato [23].
Imputación de datos con métodos simples y tradicionales	Imputación por la media [24]: afecta la correlación con otras variables, al utilizarla para muchos valores hace que la distribución de frecuencias no sea apropiada, no se recomienda su uso en porcentajes de valores faltantes superiores al 10%. Imputación por la mediana [24]: se aplica separadamente a cada característica que contiene valores faltantes.

En cuanto a las ventajas de las técnicas HotDeck se tienen [24]:

- Resulta mejor para volúmenes grandes de datos.
- Conservación de la distribución de los datos.
- Los procedimientos conducen a agrupaciones sencillas.
- No presentan problemas al requerir insertar conjuntos de datos.
- Hay métodos para aprendizaje supervisado y no supervisado.

Para la comparación de las técnicas de aprendizaje, se utilizó el test estadístico t-test (prueba en la que el estadístico utilizado tiene una distribución *t-student*) que se utiliza para certificar si las medias de dos grupos son estadísticamente diferentes. Es decir, que si el valor del estadístico para un grado de confianza y un grado de libertad es mayor que el crítico se dice que hay diferencia significativa entre las medias [26]. En la Tabla 3 se muestran los resultados generados de t-test, con un umbral para la distribución estadística *t-student* de 2,3263, precisión de 0,01 e infinitos grados de libertad, para la comparación de las técnicas de aprendizaje que se aplicaron en la imputación a los datos sintéticos que serán explicados más adelante. En estos datos se puede observar que existe diferencia significativa entre las técnicas analizadas.

Teniendo en cuenta que la prueba t-test generó como resultado que existe diferencia entre las técnicas de aprendizaje y de acuerdo con lo planteado en [24], que un buen método de imputación debe preservar las estimaciones de los parámetros que definen la forma de la distribución de los datos, se propuso evaluar la media aritmética, la desviación estándar, la asimetría y la curtosis para comparar las técnicas de imputación. Entonces, se definió un experimento con base en los datos sintéticos que consistió en calcular los estadísticos en el caso en que los datos sintéticos están completos, cuando

los datos faltantes son imputados con la media, la mediana, K-Medias y SOM.

De esta forma fue posible ver el comportamiento de los estadísticos para las técnicas de imputación en estudio y obtener las conclusiones. Los resultados de las comparaciones de las técnicas se resumen en la Tabla 4.

Con base en los resultados obtenidos para los estadísticos se puede determinar que:

- La media no presenta mayor variación al aplicar las técnicas de imputación; sin embargo, la técnica que menos variación le imprime es la técnica SOM, a pesar que podría decirse que la imputación con media y mediana debieran presentar mejores resultados por ser medidas de tendencia central.
- La desviación estándar presentó menor variación al utilizar la técnica K-Medias. En cuanto a la técnica SOM la variación no fue notoria

Tabla 3. Resultados de aplicar t-test con datos sintéticos.

T-TEST	t obtenido
Media-Mediana	-12,39185779
Media-K-medias	0,958720473
Media-SOM	0,154251763
Mediana-Media	12,39185779
Mediana-K-medias	4,887672558
Mediana-SOM	4,545839709
K-medias-Media	-0,958720473
K-medias-Mediana	-4,887672558
K-medias-SOM	-0,912346724
SOM-Media	-0,154251763
SOM-Mediana	-4,545839709
SOM-K-medias	0,912346724

Tabla 4. Resultados de evaluar estadísticos con datos sintéticos.

	Media	Desviación Estándar	Asimetría	Curtosis
Datos completos	65,85	38,51	-0,68	-1,07
Datos faltantes media	66,12	34,49	-0,79	-0,58
Datos faltantes mediana	69,00	34,98	-0,99	-0,42
Datos faltantes k-medias	65,49	38,35	-0,64	-1,11
Datos faltantes SOM	66,03	37,77	-0,68	-1,04

comparada con las otras técnicas que presentaron diferencias significativas con respecto a las dos técnicas Hot Deck.

- En el caso del coeficiente de asimetría, al utilizar la técnica SOM no se presentó variación, lo que indica que no se alteró la distribución de los datos con respecto al punto central. Las otras técnicas presentaron una diferencia relativamente baja, destacándose K-Medias por ser la diferencia más baja.
- La curtosis se vio afectada al utilizar las técnicas K-Medias y SOM en baja proporción y en igual medida. Las técnicas de imputación por media y mediana presentaron diferencias de mayor medida y significativa con respecto a la diferencia de las técnicas Hot Deck.

En general, se puede concluir que las técnicas K-Medias y SOM son las que menos variación imprimen a la distribución de los datos. Este resultado es un primer argumento para pensar en la selección de estas dos técnicas como la base de la técnica híbrida.

Tal y como se planteó en la definición del problema, las condiciones de los datos de un cubo OLAP permiten hablar de datos no supervisados; entonces, la naturaleza de los datos hizo que se restringiera el espectro de técnicas para abordar el problema. De acuerdo con la literatura consultada [12-13, 27], se observó el uso común de las redes neuronales y las técnicas basadas en vecindad para la imputación de datos, que se tomó como marco dentro del que estuvieron las técnicas que fueron seleccionadas como base para el modelamiento y definición de la técnica híbrida. Para afrontar el relleno de valores faltantes cuando la naturaleza de los datos no es supervisada se analizaron las técnicas por su sencillez y bajo costo computacional; además de considerar los resultados mostrados en la Tabla 4 y su respectivo análisis. Dentro de las redes neuronales hay diferentes topologías de las cuales algunas se usan para datos supervisados, como es el caso del perceptrón multicapa, y otras se usan para datos no supervisados como los mapas autoorganizativos de Kohonen (SOM). En cuanto a las técnicas basadas en vecindad, una de las que se orientan al aprendizaje no supervisado es la denominada K-Medias. De esta forma, SOM y K-Medias fueron las técnicas empleadas en el proceso de definición de la técnica híbrida; haciendo claridad que los algoritmos de

agrupación han sido ampliamente utilizados en la imputación y uno de los algoritmos de agrupamiento no supervisado más conocido y usado es el K-Medias [13]. De acuerdo con [27] K-Medias es uno de los diez algoritmos de minería de datos identificados en el año 2006 como los principales en esta área. Estos diez algoritmos están entre los más influyentes de minería de datos en la comunidad investigativa. La técnica híbrida propuesta se utilizó para la imputación de valores faltantes y no solo para el agrupamiento de los datos.

DESCRIPCIÓN DE LAS TÉCNICAS PARA EL CASO DE ESTUDIO

Algoritmo K-Medias

El algoritmo K-Medias (K-Means) es un método de agrupamiento por vecindad en el que se parte de un número determinado de prototipos y de un conjunto de ejemplos a agrupar. La constante K se refiere al hecho que el algoritmo funciona para un número fijo de grupos (*cluster*), los que son definidos en términos de la proximidad entre los puntos de datos [28-29]. El procedimiento de este algoritmo se describe en [30].

Mediante este algoritmo el espacio de ejemplos de entrada se divide en k clases o regiones, y el prototipo de cada clase estará en el centro de la misma. Dichos centros se determinan con el objetivo de minimizar las distancias euclidianas cuadráticas entre los patrones de entrada y el centro más cercano; es decir, minimizando el valor J (ecuación (1)):

$$J = \sum_{i=1}^k \sum_{n=1}^m M_{in} d_{EUC}^2(X_n - A_i) \quad (1)$$

Donde m es el conjunto de patrones, d_{EUC} es la distancia euclidiana, X_n es el ejemplo de entrada n, A_i es el prototipo de la clase i, y M_{in} es la función de pertenencia del ejemplo n a la región i de forma que vale 1 si el prototipo A_i es el más cercano al ejemplo X_n y 0 en caso contrario.

El modelo de Kohonen y las redes autoorganizativas (Self Organizing Map - SOM)

El modelo Kohonen es una red neuronal artificial que cuenta la capacidad de formar mapas de características de manera similar a como ocurre en el cerebro humano. El objetivo de este paradigma es demostrar que un estímulo externo (información

de entrada) es suficiente para forzar la formación de los mapas. Esta red utiliza un aprendizaje no supervisado de tipo competitivo. Las neuronas de la capa de salida compiten por activarse, pero solo una de ellas permanece activa ante una determinada información de entrada a la red. Los pesos se ajustan en función de la neurona que haya resultado vencedora. En el caso de existir mayor número de entradas que neuronas de salida, más de una deberá asociarse con la misma neurona de salida; es decir, pertenecerán a la misma categoría. En tal caso, los pesos se obtienen como un promedio de dichas entradas [31-32].

El proceso se repite para mejorar el entrenamiento y el mapa topológico de salida, de tal forma que cuantas más veces se presenten los datos, más se reducirán las zonas de neuronas que deben activarse ante entradas parecidas, y así la red realiza una clasificación más selectiva.

En el modelo SOM, el número de neuronas se toma tan grande como se pueda, considerando como mínimo cubrir las características de las entradas, ya que a diferencia de otros modelos de redes neuronales artificiales que tienden al sobreajuste; para SOM, tener más neuronas que entradas no afecta el resultado final. En realidad, al aumentar

el número de neuronas lo que realmente se afecta es el costo computacional [33].

Técnica híbrida

Finalmente, se presenta el algoritmo técnica híbrida propuesta (Tabla 5), basado en las técnicas K-Medias y SOM.

Para la técnica propuesta se implementaron algunas rutinas de las que se muestran los diagramas de actividad. Para la rutina que permite quitar los centroides de los grupos que no tienen elementos se creó una operación denominada QuitarCentroidesCero (ver Figura 2) y la rutina para el cálculo de los centroides de los datos faltantes se denominó CentroidesDatosFaltantes (ver Figura 3).

ANÁLISIS DE PRUEBAS Y RESULTADOS

Descripción del conjunto de datos

Los datos con los que se realizaron las pruebas finales de la técnica híbrida propuesta, corresponden a los datos de una empresa transportadora de un sector de la localidad Ciudad Bolívar, ubicada en Bogotá. La empresa guarda información relacionada con los ingresos y egresos para establecer comparación mes a mes y año tras año de las operaciones.

Tabla 5. Algoritmo técnica híbrida propuesta.

(1)	<i>ALGORITMO KMediasSOM (Información de entrada)</i>
(2)	<i>Fase SOM:</i>
(3)	<i>Inicializar pesos aleatoriamente;</i>
(4)	<i>Repetir</i>
(5)	<i>Presentar el mapa de información de entrada a la red neuronal;</i>
(6)	<i>Determinar la neurona vencedora de la capa de salida;</i>
(7)	<i>Actualizar los pesos;</i>
(8)	<i>Fin Repetir (hasta cuando no haya cambios en los pesos)</i>
(9)	<i>Generar mapa SOM;</i>
(10)	<i>Extraer centroides centr;</i>
(11)	<i>Quitar centroides de grupos que no tienen elementos;</i>
(12)	<i>Fase K-Medias:</i>
(13)	<i>Leer información de entrada;</i>
(14)	<i>Agrupar los datos completos de acuerdo con los centroides centr;</i>
(15)	<i>Repetir</i>
(16)	<i>Calcular los centroides (distancia euclidiana);</i>
(17)	<i>Agrupar los datos;</i>
(18)	<i>Fin Repetir (hasta cuando los centroides no varían)</i>
(19)	<i>Calcular los centroides para los datos faltantes;</i>
(20)	<i>Fase Imputación:</i>
(21)	<i>Rellenar los datos faltantes con base en los centroides;</i>
(22)	<i>FIN ALGORITMO</i>

Figura 2. Diagrama de Actividad de Quitar Centroides Cero.

Teniendo en cuenta, que un *Datamarts* es un subconjunto de datos enfocados a un tema seleccionado, entonces, la información referente a los recaudos por pasajes de usuarios es la información más importante y se estableció el *Datamarts* de recaudos para ser utilizado para las pruebas. Los cubos de datos son considerados una representación para el almacenamiento de los datos. Estos cubos pueden ser representados de diferentes formas por medio de esquemas gráficos que se pueden asociar a un *DataWarehouse* o a un *Datamarts*. Para los *Datamarts*, los esquemas de estrella y de copo de nieve son comúnmente usados; sin embargo, el esquema estrella es el más popular y eficiente [2]. El *Datamarts* de recaudos se construyó con base en los datos asociados a los conductores que son asignados a una ruta y que realizan un recaudo y prestan el servicio a una cantidad de usuarios. El cubo construido para el *Datamarts* de los recaudos considera una tabla de hechos (recaudos), tres dimensiones asociadas (conductores, rutas y

Figura 3. Diagrama de Actividad de Centroides Datos Faltantes.

tiempo con granularidad de mes) y dos medidas a registrar (valor del recaudo y número de usuarios atendidos). La información que será manejada es la correspondiente a los años 2008, 2009, 2010 y 2011.

El esquema estrella del cubo asociado al *Datamarts* de recaudos se muestra en la Figura 4. La implementación de los algoritmos en referencia se realizó usando la herramienta MATLAB, utilizando la documentación de *som_toolbox* [34].

Imputación y análisis de resultados utilizando el conjunto de datos

Se evaluó el resultado del agrupamiento luego de aplicar alguna de las técnicas, para verificar que los grupos tengan la mínima variabilidad dentro de los grupos y la máxima variabilidad entre los grupos. De acuerdo con esto, los índices que se aplicaron para evaluar el desempeño de los algoritmos de agrupamiento son los denominados homogeneidad y separación en los resultados. Otros criterios usados para evaluar el resultado del agrupamiento y de la imputación fueron los indicadores denominados Efectividad (%) y Error Cuadrático Medio promediando la suma de las diferencias al cuadrado entre el valor imputado y el valor real [35]. Para la interpretación de las expresiones asociadas a los indicadores es necesario tener en cuenta la función distancia, que en este caso corresponde a la distancia euclidiana cuya expresión es la ecuación (2) [36].

$$D_{ij} = \sqrt{\sum_{k=1}^M (X_{ij} - X_{jk})^2} \quad (2)$$

Figura 4. Esquema estrella *Datamarts* recaudos empresa transportadora.

Para el análisis de los resultados se realizó la normalización por el método MIN-MAX (ecuación 3) de los indicadores al intervalo [0,1] por medio del que se realiza una transformación lineal de los datos. Con base en los indicadores normalizados fue posible establecer que la técnica híbrida propuesta presenta los mejores resultados en términos de la imputación y se determinó la mejor configuración, la que fue utilizada en la aplicación de *KmediasSom* con los datos reales.

$$v' = \frac{v - \min_A}{\max_A - \min_A} \quad (3)$$

$$(new_max_A - new_min_A) + new_min_A$$

A continuación se relacionan los indicadores para la evaluación de la imputación:

Homogeneidad

$$Homogeneidad(H_{ave}) = \frac{1}{N} \sum_i D(g_i, C(g_i)) \quad (4)$$

Variabilidad de los datos dentro de un grupo. Es decir, refleja que tan compacto es un grupo. En la expresión (ecuación (4)), la homogeneidad es calculada como la distancia promedio entre cada elemento del centro del grupo (*cluster*) al cual pertenece, donde g_i es el i -ésimo elemento y $C(g_i)$ es el centro del grupo al cual pertenece g_i , N es el número total de elementos y D es la función distancia.

Separación

$$Separacion(S_{ave}) = \frac{1}{\sum_{i \neq j} N_{Ci} N_{Cj}} \sum_{i \neq j} N_{Ci} N_{Cj} D(C_i, C_j) \quad (5)$$

Variabilidad entre los grupos. Es decir, distancia general entre los grupos. En la expresión (ecuación (5)), la separación es calculada como la distancia media ponderada entre los centros de los grupos, donde C_i y C_j son los centros del i -ésimo y el j -ésimo grupo y N_{Ci} y N_{Cj} son el número de elementos de los i -ésimo y j -ésimo grupo.

Error cuadrático medio (ECM)

$$ECM = \frac{\sum (X_1 - X_2)^2}{N} \quad (6)$$

Medida del posible error cometido por la imputación, por medio del promedio la suma de las diferencias al cuadrado entre el valor imputado y el valor real (ecuación (6)).

$$\%Efectividad = \frac{\#casos\ hallados}{\#casos\ esperados} * 100 \quad (7)$$

Cercanía de los datos imputados con los reales (ecuación (7)).

Los indicadores hallados tienen un significado particular en el proceso de agrupamiento e imputación de los datos. El análisis de los indicadores en conjunto, permitió definir la configuración más adecuada para la imputación y se concluyó, entre otras cosas, que la calidad de los datos es un factor fundamental en la confiabilidad de un sistema de análisis de datos. Los indicadores que permitieron establecer la configuración mencionada fueron la efectividad cuando sea el mayor valor, el ECM pequeño y considerar que entre más pequeño sea el valor de la homogeneidad y mayor el valor de la separación, mejores son los resultados del agrupamiento y, por ende, de la imputación.

Con base en lo anterior la ponderación de los indicadores puede ser modelada mediante el análisis multiobjetivo o multicriterio, el que trata de identificar la mejor solución considerando simultáneamente múltiples objetivos en conflicto. Generalmente, en los problemas con múltiples objetivos no es posible obtener una solución factible que sea óptima para todos los objetivos, entonces,

como puede haber varias soluciones eficientes (no dominadas). La técnica de combinación lineal de los pesos, es una de las técnicas pertenecientes a este grupo y en esta técnica la función objetivo no es más que la suma de las funciones del problema original multiplicadas por un peso [37]. En este caso particular, se utiliza la combinación lineal de los pesos para calcular los valores de la ponderación asociados a cada una de las configuraciones definidas. La importancia de la ponderación, radica en determinar de forma cuantitativa la relevancia de cada elemento o función objetivo. En general, se asignan pesos aunque su especificación es una cuestión para la cual se han propuesto diversos métodos, la entropía es uno usualmente utilizado cuyo principal interés reside en su objetividad con respecto al decisor, siendo los propios datos del problema los que determinan la importancia relativa de los criterios [38]. Aplicando este método se pueden obtener los pesos asociados a cada una de las funciones objetivo de la evaluación multiobjetivo, que en el caso particular de este proyecto corresponden a los indicadores.

La Tabla 6 muestra los resultados obtenidos para cada indicador y en la Tabla 7 se presentan los valores logrados a partir de la ponderación. Los valores de los indicadores de evaluación y la ponderación conseguidos con los datos de la empresa transportadora permitieron confirmar lo que ya había sido evaluado para la técnica KMediasSom, con los datos sintéticos y es que la técnica KMediasSom en promedio, presenta mejores resultados que los generados por la técnica individual K-Medias.

Tabla 6. Indicadores de imputación KmediasSom y K-Medias con datos reales.

	KMediasSom Configuración Datos <i>Datamarts</i> recaudos	K-Medias - K = 465 - Datos <i>Datamarts</i> recaudos
Efectividad	0,77236	0,73387
ECM (Error Cuadrático medio)	0,35	0,2
Separación	11,40150467	11,40143091
Homogeneidad máxima	1,728546608	1,240864329
Homogeneidad mínima	0,053741722	0

Tabla 7. Resultados de ponderación de indicadores.

	KMediasSom - Configuración - Datos <i>Datamarts</i> recaudos	K-Medias - K = 465 - Datos <i>Datamarts</i> recaudos
Ponderación	2,21617345	2,14361792

Análisis de la complejidad algorítmica del algoritmo KMediasSOM (Tabla 8)

De acuerdo con lo anterior, la complejidad se puede expresar de la siguiente manera:

$$T(n) = C_1 + C_2N_1 + C_3N_1 + C_4N_1 + C_5 + C_6 + C_7 + C_8 + C_9 + C_{10}N_2 + C_{11}N_2 + C_{12} + C_{13} \quad (8)$$

$$T(n) = C_1(C_2 + C_3 + C_4)N_1 + C_5 + C_6 + C_7 + C_8 + C_9 + (C_{10} + C_{11})N_2 + C_{12} + C_{13}$$

Con N_1 (ecuación (8)) como el número de veces que se ejecuta el primer ciclo (Repetir) y N_2 el número de veces que se repite el segundo ciclo. Donde el orden de la complejidad es $O(N_1N_2)$. La técnica híbrida mejora el resultado obtenido para los valores de los indicadores de evaluación de las técnicas como se muestra en la Tabla 6 y conserva el orden lineal de la complejidad de las técnicas iniciales, según [34, 39].

CONCLUSIONES

El resultado de esta investigación demostró que el área de aplicación general y en particular la naturaleza

de los datos es uno de los factores que restringe el espectro de técnicas adecuadas, ya que si los datos no son supervisados entonces las técnicas deben orientarse a este tipo de aprendizaje.

En el contexto de las técnicas de aprendizaje computacional basadas en vecindad y las redes neuronales para el relleno de valores faltantes, sobresalieron las técnicas K-Medias y redes SOM. Con sus características, similitudes, debilidades y fortalezas, permitieron definir una técnica híbrida (KMediasSom) basada en los resultados obtenidos del análisis de las técnicas individuales y de los casos de estudio que confirmaron la posibilidad de proponer una técnica híbrida de este estilo.

Los indicadores Efectividad, ECM (Error Cuadrático Medio), Separación y Homogeneidad permiten evaluar los algoritmos para datos no supervisados. Al aplicar dichos indicadores a los resultados de la técnica híbrida sobre los datos sintéticos, permitieron establecer que la técnica híbrida suministra mejores resultados que la técnica individual (K-Medias). Al aplicar la técnica híbrida KMediasSom a los datos reales de la empresa transportadora y de acuerdo

Tabla 8. Complejidad computacional.

Algoritmo	Costo	Número de veces
(1) ALGORITMO KMediasSOM (Información de entrada)		
(2) Fase SOM:		
(3) Inicializar pesos aleatoriamente;	C_1	I
(4) Repetir		
(5) Presentar el mapa de información de entrada a la red neuronal;	C_2	N_1
(6) Determinar la neurona vencedora de la capa de salida;	C_3	N_1
(7) Actualizar los pesos;	C_4	N_1
(8) Fin Repetir (hasta cuando no haya cambios en los pesos)		
(9) Generar mapa SOM;	C_5	I
(10) Extraer centroides centr;	C_6	I
(11) Quitar centroides que no tienen elementos;	C_7	I
(12) Fase K-Medias:		
(13) Leer información de entrada;	C_8	I
(14) Agrupar los datos completos de acuerdo con los centroides centr;	C_9	I
(15) Repetir		
(16) Calcular los centroides (distancia euclidiana);	C_{10}	N_2
(17) Agrupar los datos;	C_{11}	N_2
(18) Fin Repetir (hasta cuando los centroides no varían)		
(19) Calcular los centroides para los datos faltantes;	C_{12}	I
(20) Fase Imputación:		
(21) Rellenar los datos faltantes con base en los centroides;	C_{13}	I
(22) FIN ALGORITMO		

con los valores obtenidos para los indicadores de evaluación del agrupamiento y de la imputación, se obtuvo que la técnica híbrida proporciona una mayor efectividad para la imputación.

Terminada la investigación se concluye:

En primer lugar, la técnica híbrida KMediasSom puede ser usada con garantía de resultados adecuados, en tareas de relleno de valores faltantes en aplicaciones OLAP.

En segundo lugar, se comprueba que la combinación de estas dos técnicas permite generar en promedio resultados aceptables.

AGRADECIMIENTOS

Agradecemos a la Universidad Distrital “Francisco José de Caldas” y a los expertos que apoyaron el proyecto de investigación e hicieron posible mostrar los resultados expuestos.

REFERENCIAS

- [1] R. Hernández, C. Fernández y P. Baptista, “Metodología de la investigación”. Tercera edición ed., Mc Graw Hill. 2003.
- [2] J. Han y M. Kamber, “Data mining, Concepts and techniques”. Segunda ed. San Francisco, CA, USA. Morgan Kaufmann, pp. 61-65, 110-127. 2006.
- [3] D. Pyle. “Data preparation for data mining”. San Francisco, CA, USA. Morgan Kaufmann Publishers. 1999.
- [4] S. Zhang, C. Zhang and Q. Yang. “Data preparation for data mining”. Applied Artificial Intelligence. Vol. 17, pp. 375-381. 2003.
- [5] J. Villalobos. “Calidad de datos en las organizaciones”, pp. 28-68. Fecha: 2008. Fecha de consulta: junio de 2012. URL: http://www.acis.org.co/fileadmin/Base_de_Conocimiento/XXVIII_Salon_de_Informatica/ConferenciaJorgeVillalobosAlvarado.pdf
- [6] A. Maedche, A. Hotho and M. Wiese. “Enhancing Preprocessing in Data-Intensive Domains using Online-Analytical Processing”. DaWaK 2000 Proceedings of the Second International Conference on Data Warehousing and Knowledge Discovery. London, UK. Springer-Verlag. 2000.
- [7] C. Clifton. “Data Mining. Course Overview”, pp. 9-64. January, 2006. Fecha de consulta: enero de 2012. URL: <http://www.cs.purdue.edu/homes/clifton/cs590d/Intro.ppt>
- [8] L.P. Cheung, K.R. Lau, W.A. Lee, C.L. Tsoi and K.F. Yip. “Data Warehousing and OLAP”, pp. 1-15. 1997. Fecha de consulta: 2011. URL: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.31.1896&rep=rep1&type=pdf>
- [9] S. Palaniappan and T. K. Hong. “Discretization of Continuous Valued Dimensions in OLAP Data Cubes”. IJCSNS International Journal of Computer Science and Network Security. Vol. 8 N° 11, pp. 116-126. November, 2008.
- [10] J. Han. “Olap Mining: An Integration of OLAP with Data Mining”. Proceedings of the 7th IFIP 2.6 Working Conference on Database Semantics (DS-7). Leysin, Switzerland. 1997.
- [11] H.-Y. Lee and H.-L. Ong. “A New Visualisation Technique for Knowledge Discovery in OLAP”. Proceedings of the DOOD’95 Post-Conference Workshops on Integration of Knowledge Discovery in Databases with Deductive and Object-Oriented Databases (KDOOD) and Temporal Reasoning in Deductive and Object-Oriented Databases (TDOOD). Singapore, Singapore. 1995.
- [12] R.S. Somasundaram and R. Nedunchezian. “Evaluation of three Simple Imputation Methods for Enhancing Preprocessing of Data with Missing Values”. International Journal of Computer Applications. Vol. 21 N° 10, pp. 14-19, mayo 2011.
- [13] D. Li, J. Deogun, W. Spaulding y B. Shuart. “Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method”. Rough Sets and Current Trends in Computing. Lecture Notes in Computer Science. Vol. 3066, pp. 573-579. Uppsala, Sweden, Springer, 2004.
- [14] R. Pressman. “Ingeniería del Software. Un enfoque práctico”. Tercera edición ed., Madrid, España: McGraw-Hill, 1993.
- [15] Y. Ayala y O.O. Melo. “Estimación de datos faltantes en medidas repetidas con respuesta binaria”. Revista Colombiana de Estadística. Vol. 30 N° 2, pp. 265-285, diciembre 2007.
- [16] C. Bishop, M. Svensén and C. Williams. “GTM: The generative topographic mapping”.

- Neural Computation. Vol. 10 N° 1, pp. 215-234. 1998.
- [17] C. Bishop, M. Svensén and C. Williams. "Magnification Factors for the GTM Algorithm". Proceedings IEEE Fifth International Conference on Artificial Neural Networks. Cambridge, UK. 1997.
- [18] S. Kaski. "Data Exploration Using Self-Organizing Maps". Mathematics, Computing and Management in Engineering Series. N° 82, pp. 57. 1997.
- [19] J. Pérez, M.F. Henriques, R. Pazos, L. Cruz, G. Reyes, J. Salinas y A. Mexicano. "Mejora al algoritmo de agrupamiento K-means mediante un nuevo criterio de convergencia y su aplicación a bases de datos poblacionales de cáncer". 2° Taller Latino Iberoamericano de Investigación de Operaciones. Acapulco, México, 2007.
- [20] F. Fernández y A. Murillo. "Clasificación automática simbólica por medio de algoritmos genéticos". Revista de Matemática: Teoría y Aplicaciones. Vol. 16 N° 2, pp. 283-292. 2009.
- [21] E. Yolís, P. Britos, G. Perichisky y R. García. "Algoritmos genéticos aplicados a la categorización automática de documentos", pp. 1470. 2003. Fecha de consulta: 2011. URL: http://sedici.unlp.edu.ar/bitstream/handle/10915/22639/Documento_completo.pdf?sequence=1
- [22] G. Chen, S. Jaradat, N. Banerjee, T. Tanaka, M. Ko and M. Zhang. "Evaluation and Comparison of Clustering Algorithms in Analyzing ES Cell Gene Expression Data". Statistica Sinica. Vol. 12, pp. 241-262. 2002.
- [23] B. Peralta, J. Ortega y H. Benítez. "Un Sistema de Patrones de Software para Redes Neuronales Artificiales". Universidad Autónoma de México - UNAM, Departamento de Matemáticas, Departamento de Ingeniería de Sistemas Computacionales y Automatización, 2005.
- [24] I. Amón Uribe. "Guía metodológica para la selección de técnicas de depuración de datos". Tesis maestría, Universidad Nacional de Colombia - Medellín, Antioquia, Medellín, 2010.
- [25] F. Medina y M. Galvan. "Imputación de datos: Teoría y práctica: patrones de comportamiento de los datos omitidos". Serie de estudios estadísticos y prospectivos CEPAL, pp. 26-27, julio 2007.
- [26] J. Hernández, J. Ramírez y C. Ferri. "Introducción a la Minería de Datos". Pearson. Prentice Hall, 2004.
- [27] X. Wu, V. Kumar, R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. McLachlan, A. Ng, B. Liu, P. Yu, Z.-H. Zhou, M. Steinbach, D. Hand y D. Steinberg. "Top 10 algorithms in data mining". Knowledge and Information Systems. Vol. 14 N° 1, pp. 1-37, diciembre 2007.
- [28] H. Barrera, J. Correa y J. Rodríguez. "Prototipo de Software para el preprocesamiento de datos. "UD-Clear". IV Simposio Internacional de Sistemas de Información e Ingeniería de Software en la Sociedad del Conocimiento SISOFIT, Cartagena, Colombia, 2006.
- [29] M. Berry and G.S. Linoff. "Data Mining Techniques". Indianapolis, Indiana. Wiley. 2004.
- [30] N. Rodríguez y W. Sánchez. "Software para preprocesamiento de datos udclear versión 2.0". Bogotá, Colombia, 2008.
- [31] J.R. Hilera y V.J. Martínez. "Redes neuronales artificiales: Fundamentos, modelos y aplicaciones". Addison-Wesley Iberoamericana Ra-Ma. Madrid, España. 1995.
- [32] J. Vesanto, J. Himberg, E. Alhoniemi and J. Parhankangas. "Self-Organizing Map in Matlab: the SOM Toolbox" de Proceedings of the Matlab DSP Conference. 2000.
- [33] J.D. Martín. "Determinación de tendencias en un portal web utilizando técnicas no supervisadas. Aplicación a sistemas de recomendaciones basados en filtrado colaborativo". Valencia, España. 2004.
- [34] J. Vesanto. "Neural Network Tool for Data Mining: SOM Toolbox". Proceedings of Symposium on Tool Environments and Development Methods for Intelligent Systems (TOOLMET2000). Oulu, Finland. 2000.
- [35] J. Navarro y J. Losilla. "Análisis de datos faltantes mediante redes neuronales artificiales". Psichotema. Vol. 12 N° 3, pp. 503-510. 2000.
- [36] M. Dash, H. Liu and J. Yao. "Dimensionality Reduction for Unsupervised Data". de Ninth IEEE International Conference on Tools with AI, ICTAI'97. 1997.

- [37] L.C. Cagnina. “Optimización mono y multiobjetivo a través de una heurística de inteligencia colectiva”. San Luis, Argentina. 2010.
- [38] M.L. Ramírez. “El método de jerarquías analíticas de Saaty en la ponderación de variables. Aplicación al nivel de mortalidad y morbilidad en la provincia del chaco”. Comunicaciones científicas y tecnológicas. Universidad Nacional del Nordeste, pp. 4. 2004.
- [39] A. Pérez, G. García, J. Medina, J. Martínez y J. Carrasco. “Algoritmo de agrupamiento para colecciones de documentos”. Reporte Técnico de Minería de Datos. Centro de Aplicaciones de Tecnologías de Avanzada (CENATAV). La Habana, Cuba. 2008.