Preprocesamiento de datos estructurados

Structured Data Preprocessing

[*]Claudia L. Hernández G.

[**]Jorge E. Rodríguez R.

Fecha de recepción: 13 de marzo de 2008
Fecha de aceptación: 20 de abril de 2008

Resumen

El propósito del preprocesamiento de datos es principalmente corregir las inconsistencias de los datos que serán la base de análisis en procesos de minería de datos. En el caso de las fuentes de datos estructuradas, el propósito no es distinto y pueden ser aplicadas diversas técnicas estadísticas y de aprendizaje computacional. Con el preprocesamiento de datos se pretende que los datos que van a ser utilizados en tareas de análisis o descubrimiento de conocimiento conserven su coherencia. A lo largo del presente artículo, se realizará la descripción de diferentes técnicas existentes junto con algunos algoritmos asociados a tareas destacadas de preprocesamiento de datos estructurados como limpieza y transformación. Luego, se hace una revisión de algunos algoritmos asociados a las técnicas utilizadas con más frecuencia, lo cual podrá permitir la comparación de su efectividad dependiendo del conjunto de datos utilizado, en trabajos futuros.

Palabras clave: preprocesamiento, discretización, minería OLAP, normalización, limpieza, integración, transformación, reducción de la dimensionalidad.

* Ingeniera de Sistemas. Candidata a Magíster en Ciencias de la Computación y las Comunicaciones, Universidad Distrital Francisco José de Caldas. Correo electrónico: clhernandez@gmail.com

** Ingeniero de Sistemas. Especialista en Telemática. Especialista en Ingeniería de Software. Magíster en Ingeniería de Sistemas. Docente de la Universidad Distrital Francisco José de Caldas. Correo electrónico: jrodri@udistrital.edu.co

Abstract

The purpose of data preprocessing is mostly correct the inconsistencies in the data analysis will be based on processes of data mining. For structured data sources, the purpose is not different and can be applied several statistical techniques and machine learning. Really wanted the data to be used in analysis tasks or knowledge discovery, are very close to reality and keep their consistency. In this paper, we will be description about data difficulties, different techniques with some existing algorithms outstanding tasks associated with data preprocessing such as data cleaning and transformation. Next is to review in detail the algorithms associated with the techniques used most frequently.

Key words: data preprocessing, discretization, OLAP data mining, normalization, cleaning, integration, transformation, attribute selection.

1. Introducción

La gran cantidad de datos que actualmente manejan las organizaciones ha generado la necesidad de tener sistemas en los cuales confluya toda la información que es recopilada en fuentes de datos estructuradas como las bases de datos transaccionales.

En estas condiciones, la implementación de data warehouse y aplicaciones con tecnología OLAP (Procesamiento Analítico en Línea) se ha incrementado, para lo cual es necesario establecer procesos que permitan agilizar la creación y actualización de sistemas de este estilo. Diversos factores están haciendo que las organizaciones dirijan la mirada a la tecnología OLAP como alternativa para el manejo de los datos estructurados orientados al apoyo de la toma de decisiones, y más aún cuando Internet está promoviendo formas diferentes en que las compañías acceden y extraen información. La calidad de los datos juega un papel muy importante, ya que este tipo de aplicaciones permiten básicamente visualizar los grandes volúmenes de datos almacenados en un data warehouse y a través de diferentes operaciones es posible analizar los datos y convertirlos en soporte para la toma de decisiones.

El objetivo principal de cualquier análisis de datos, antes que el descubrimiento de conocimiento, es utilizar los resultados para resolver problemas o para la toma de decisiones. En la mayoría de los casos, las imperfecciones con los datos sólo son notadas cuando se inicia el análisis de los datos. Para disminuir tiempo y costos es importante preparar los datos para dicho análisis; en esta línea, ya existen diversas técnicas que están orientadas a apoyar el proceso de minería de datos. Sin embargo, recientemente se han estudiado mecanismos alternos que aplican de manera apropiada al modelo OLAP.

En este artículo se presenta una visión general de la teoría y algunas las técnicas utilizadas en el preprocesamiento de datos, y cómo se están abordando con la tecnología OLAP,

a la vez que se revisan trabajos realizados en este campo y que aplican, en primera instancia y específicamente para OLAP. El artículo está estructurado en tres secciones: en la primera, se hace una descripción de ciertas causas por las cuales los datos se deben someter al preprocesamiento y técnicas asociadas, en la segunda sección, se aborda la descripción general de las aplicaciones OLAP y cómo se relacionan con el preprocesamiento, técnicas empleadas y trabajos relacionados, la última sección, presenta las conclusiones de esta parte preliminar de verificación del estado del arte en el marco de una investigación futura en relación con la comparación de eficiencia de las técnicas utilizadas en el preprocesamiento de datos y sus posibles usos en tecnología OLAP.

2. Preprocesamiento de datos

El preprocesamiento es una tarea necesaria para la preparación de los datos que serán utilizados para data warehouse o en análisis de datos. La justificación de este proceso preliminar al análisis de datos, generalmente, radica en que los datos vienen con ruido por diferentes razones, entre las cuales se encuentran [12]:

- Datos incompletos: valores faltantes para algunos atributos o sólo se tienen los datos agregados y no se cuenta con el detalle.
- Ruido: errores en los datos. Por ejemplo, manejar valores negativos para un atributo que maneja salarios.
- Inconsistencias: contiene discrepancias en los datos. Por ejemplo, edad de un empleado = 30 y fecha de nacimiento = 03/07/1998.

Los algoritmos de aprendizaje computacional (*machine learning*) suelen ser empleados para llevar a cabo ciertas tareas en el proceso de análisis de datos. Con frecuencia, el preprocesamiento de los datos tiene un impacto significativo en el desempeño general de los algoritmos de aprendizaje supervisado. Aplicar algunas técnicas de preprocesamiento permite que los algoritmos de aprendizaje sean más eficientes, por ejemplo, si se reduce la dimensionalidad, los algoritmos de aprendizaje podrían actuar de forma más rápida y su efectividad podría mejorar [14].

A continuación, se describen algunos de los problemas más frecuentes con los cuales se puede enfrentar un equipo de trabajo en el momento de realizar un proceso de análisis de datos. Luego, se describen las tareas de preprocesamiento y las técnicas más comúnmente utilizadas en la solución de los inconvenientes con los datos.

2.1 Problemas con los datos

En muchas ocasiones, la naturaleza y severidad de los problemas dependen del control de los operarios humanos de las aplicaciones que nutren las fuentes originales de datos. Debido a los efectos de estos problemas en los resultados del análisis de datos, se ha establecido como meta rectificarlos o en el peor de los casos sólo reconocer los efectos que existen sobre los resultados [9][12].

2.1.1 Datos con ruido

El ruido en los datos puede estar atribuido a errores en la medida, transmisión de datos, características inherentes a los sistemas de los cuales se obtienen los datos, etc. [9].

2.1.2 Extracción de atributos

En aplicaciones de análisis complejo puede existir la posibilidad de que en el momento de registrar la información haya datos que no fueron incluidos, simplemente porque no fueron considerados importantes durante el registro [9].

2.1.3 Datos irrelevantes

Muchas aplicaciones de análisis requieren extracción de datos significativos, a partir de un conjunto de datos. Cuando los humanos realizan un proceso similar, ellos seleccionan los datos relevantes enfocándose en las piezas clave de la información y algunas veces utilizan el resto sólo para confirmar o aclarar ambigüedades. La complejidad puede ser reducida si los datos irrelevantes son eliminados y sólo la mayoría de los atributos relevantes son utilizados para el análisis de los datos. La reducción de dimensionalidad, a través de la eliminación de valores irrelevantes, puede también mejorar el rendimiento de la herramienta de análisis de datos [9][12].

2.1.4 Volúmenes de datos demasiado grandes

La cantidad de datos algunas veces excede la capacidad disponible de hardware y software usado para el análisis. El volumen de datos y la rata a la cual son producidos puede ser un factor limitante en el análisis de datos [9].

2.1.5 Datos numéricos y simbólicos

Cuando los datos están organizados para el análisis, generalmente se tienen datos numéricos que resultan de parámetros medidos, los cuales pueden ser representados por un número y en este caso los datos pueden ser discretos o numéricos. Los datos simbólicos o categóricos resultan de procesos de medición o características de sistema; esta clase de datos es usualmente cualitativa. Analizar datos que involucran parámetros simbólicos y numéricos es una tarea compleja que requiere atención durante el preprocesamiento de datos y uso apropiado de herramientas de análisis de datos [9].

2.1.6 Atributos faltantes

Atributos faltantes o insuficientes son ejemplos de problemas de datos que pueden complicar las tareas de análisis de datos. Por ejemplo, en el caso del aprendizaje, estos datos insuficientes limitan el algoritmo de aprendizaje o las herramientas estadísticas para recolectar los datos [9][12].

2.1.7 Valores de atributos faltantes

En este caso, los datos asociados a los registros no están completos, algunos contienen valores faltantes para los atributos. Estos registros no pueden ser eliminados, porque la cantidad de datos podría no ser suficiente y porque los datos remanentes podrían contener información útil para el análisis. Tradicionalmente si más del 20% de los valores de los atributos son faltantes, el registro entero puede ser eliminado [9][12].

2.1.8 Poca cantidad de datos

En este caso, todos los atributos están disponibles; el principal problema es que la cantidad total de datos no es suficiente para todas las clases de análisis de datos. Por ejemplo, la mayoría de algoritmos de análisis de datos requieren cerca de 100 ejemplos de datos de entrenamiento. Los conceptos aprendidos o reglas pueden no ser suficientes si los ejemplos disponibles no son suficientes [9].

2.1.9 Múltiples fuentes de datos

En la mayoría de los casos los datos son adquiridos y mantenidos usando diferentes sistemas de software, en este caso es cuando se presentan inconvenientes con la unificación de los datos debido a que el análisis de datos se realiza a partir de varias fuentes de datos [9][12].

2.1.10 Datos desde múltiples niveles de granularidad

En algunas aplicaciones los datos provienen de más de un nivel de granularidad, lo cual haría difícil la comparación y el análisis de los datos [12].

2.2 Tareas de preprocesamiento

Los datos reales tienden a tener ruido, ser incompletos e inconsistentes. Las tareas y técnicas de preprocesamiento de datos pueden mejorar la calidad de los datos, ayudando a mejorar la precisión y eficiencia de los procesos de análisis de datos, de ahí que el preprocesamiento de datos se convierta en un paso preliminar importante. Detectando anomalías, corrigiéndolas a tiempo y reduciendo los datos que serán analizados se puede ayudar para que la toma de decisiones sea mucho más eficaz. La figura 1 resume las tareas de preprocesamiento que se describen a continuación.

2.2.1 Limpieza de datos [12][1]

Esta tarea consiste en llenar los valores faltantes, suavizar los datos erróneos, identificar o remover los datos inconsistentes. La importancia de la limpieza de los datos es reconocida en diversa literatura como una de las tareas más importantes y exhaustiva del preprocesamiento. "La limpieza de datos es uno de los tres problemas más grandes del data warehousing" – Ralph Kimball.

2.2.2 Integración de datos

Combina datos desde múltiples fuentes y maneja la integración de esquemas de datos a través de la combinación de los metadatos.

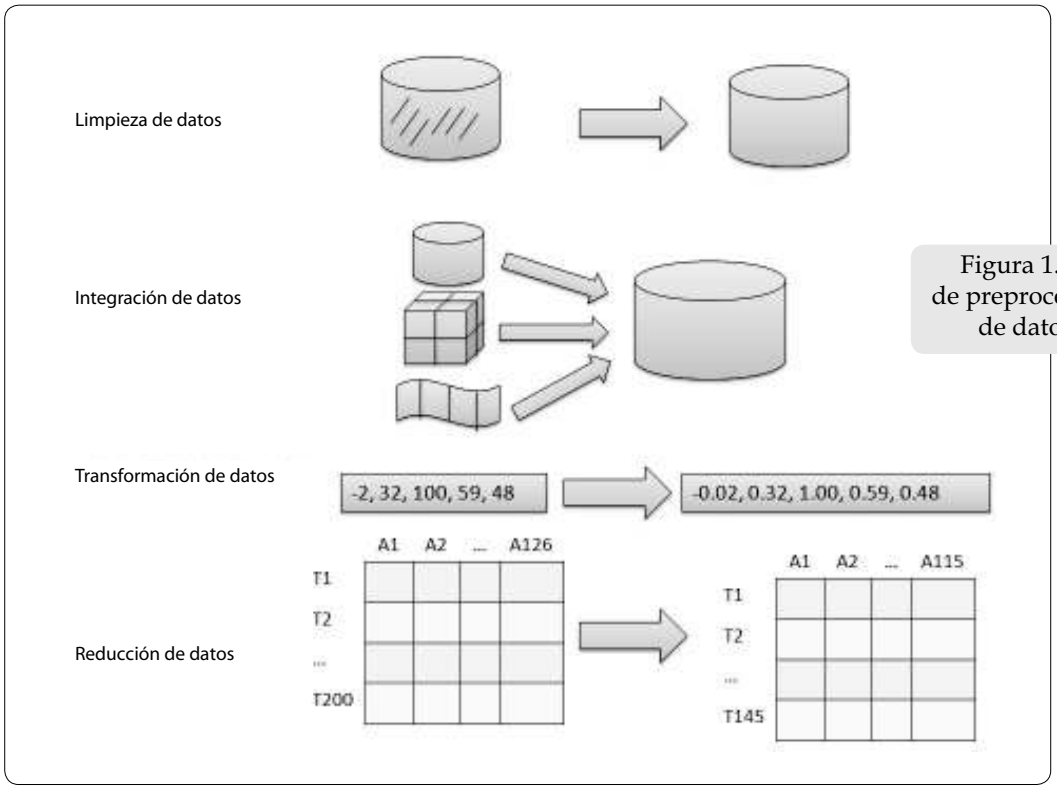


Figura 1. Tareas de preprocesamiento de datos [12]

Se pueden presentar problema en la identificación de las entidades y los atributos para establecer las relaciones o los mapeos correspondientes. Otro de los inconvenientes que se experimentan en esta fase del proceso, es la detección y solución de los valores en conflicto, ya que atributos de diferentes fuentes de datos pueden significar lo mismo, pero denominarse diferente o llamarse igual y tener un significado completamente distinto. Las razones por las cuales se presentan estas discrepancias en los datos aluden a diferentes representaciones, diferentes escalas, entre otras. Es por esto que la redundancia de datos es frecuente cuando se deben integrar los datos de varias fuentes de datos.

2.2.3 Transformación de datos[12]

La transformación de datos involucra lo siguiente:

- Normalización: donde los atributos son escalados dentro de un rango pequeño de valores como entre -1 y 1 o entre 0 y 1.
- Suavizado: el cual es utilizado para remover el ruido de los datos.
- Agregación: donde las operaciones de síntesis o agregación son aplicadas a los datos. Por ejemplo, las ventas diarias pueden ser agregadas en ventas mensuales o ventas anuales.
- Generalización: los datos de bajo nivel o primitivos son reemplazados por conceptos de más alto nivel, haciendo uso del concepto de jerarquía. Por ejemplo, para atributos categóricos como el caso de calles puede ser generalizado al concepto de nivel más alto como ciudad. De forma similar, con los atributos numéricos como la edad puede establecerse correspondencia con conceptos de nivel superior como joven, adulto, anciano.

2.3.4 Reducción de datos

Obtener representación reducida en volumen, pero produciendo los mismos resultados o similares en el análisis [7]. La discretización es una parte de la reducción de datos, pero con importancia particular, especialmente para datos numéricos [7].

2.3 Técnicas para preprocesamiento

2.3.1 Limpieza de datos [12][1]

Las tareas de limpieza de datos involucran llenado de los datos faltantes, suavizar los errores de los datos, corregir los datos inconsistentes y resolver la redundancia causada por la integración de los datos. Los datos no siempre están disponibles, esto debido a diferentes causas como mal funcionamiento de los equipos, inconsistencia con otros registros de datos que son borrados, datos no ingresados; debido al no entendimiento, ciertos datos pueden no ser considerados importantes en el momento del registro, no registrar historia o cambios de los datos. Otros problemas de datos que requieren limpieza se refieren a registros duplicados o datos incompletos.

Así, para el manejo de detección de valores anómalos existen algunas técnicas como el algoritmo de agrupación jerárquica (CURE-Clustering Using Representatives) y el algoritmo DBSCAN, ambos basados en clustering, pero cada uno manejándolo de forma diferente.

Las opciones que se tienen en el manejo de los datos faltantes son:

- Ignorar la tupla: usualmente, se hace cuando falta la etiqueta de la clase, no es efectiva cuando el porcentaje de valores faltantes por atributo varía considerablemente.

- Llenar los valores faltantes manualmente: es una tarea tediosa.
- Llenar los valores faltantes automáticamente con: constantes globales, la media del atributo, el valor más probable (basado en la inferencia como el método bayesiano o el árbol de decisión).

Existen diversos métodos para dar soporte a cada una de las tareas en preprocesamiento. Una de las opciones para el manejo de los datos con ruido es el método Binning, que permite reducir la numerosidad y en el cual primero se ordenan los datos y se realiza la partición en bins del mismo tamaño o bins de la media o bins de los extremos.

Para el relleno de los datos faltantes también se puede utilizar el algoritmo K-Medias (K-Means) que es un método de agrupamiento por vecindad en el que se parte de un número determinado de prototipos y de un conjunto de ejemplos por agrupar. K-Medias es uno de los algoritmos de clustering utilizados con más frecuencia. La "K" se refiere al hecho de que el algoritmo funciona para un número fijo de clústeres, los cuales son definidos en términos de la proximidad entre los puntos de datos [3][1].

El procedimiento es el siguiente [19]:

- Se calcula, para cada ejemplo X_k el prototipo más próximo A_g, y se incluye en la lista de ejemplos de dicho prototipo.

$$A_g = \arg \quad \min A_i \quad \{d(X_k, A_i)\} \quad \forall i = 1...n$$

(Ecuación 1)

- Después de haber introducido todos los ejemplos, cada prototipo A_k tendrá un conjunto de ejemplos a los que representa:

$$L(A_k) = \{X_{k_1}, X_{k_2}, ..., X_{k_n}\}$$

(Ecuación 2)

- Se desplaza el prototipo hacia el centro de masas de su conjunto de ejemplos.

$$A_k = \frac{\sum_{l=1}^{m} X_{k_l}}{m}$$

(Ecuación 3)

- Se repite el procedimiento hasta que ya no se desplazan los prototipos.

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 ésta. Dichos centros se determinan con el objetivo de minimizar las distancias cuadráticas euclídeas entre los patrones de entrada y el centro más cercano, es decir, minimizando el valor *J:*

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

(Ecuación 4)

Tabla 1. Ejemplo de utilización de la técnica de Bins

Datos ordenados: 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

2. Bins del mismo tamaño	3. Bins de media	4. Bins de los extremos
Bin 1: 4, 8, 9, 15 Bin 2: 21, 21, 24, 25 Bin 3: 26, 28, 29, 34	Bin 1: 9, 9, 9, 9 Bin 2: 23, 23, 23, 23 Bin 3: 29, 29, 29, 29	Bin 1: 4, 4, 4, 15 Bin 2: 21, 21, 25, 25 Bin 3: 26, 26, 26, 34

Donde m es el conjunto de patrones; d_{EUCL} es la distancia euclídea; 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.

A continuación se presenta un ejemplo en el cual se aplica el algoritmo K medias:

Tabla 2. Datos iniciales ejemplo algoritmo K-Medias

ATRIBUTO	CLASE
5	+
2.2	-
1.8	-
4	+
2	+
3	-
5	+
?	-
?	+

Se utilizan todos los datos para aplicar el algoritmo:

- En primer lugar, se especifican cuántos clúster se van a crear. Este número corresponde a la cantidad de valores que puede tomar el atributo clase, en este ejemplo, 2 (+ y -). Para obtener el valor de cada centro se halla la media de los valores que pertenezcan a cada valor tomado por el atributo clase. En este caso, el primer centro corresponde al valor + del atributo clase que tendrá un valor inicial de 4.0; el segundo centro corresponde al valor – del atributo clase que tendrá un valor inicial de 2.33

- A continuación cada una de las instancias, ejemplos, es asignada al centro del clúster más cercano de acuerdo con la distancia Euclidiana que le separa de él.

- Después de haber introducido todos los ejemplos, cada centro tendrá un conjunto de ejemplos a los que representa. En este caso los centros quedaron con los siguientes conjuntos de datos, centro 1 (+): 5, 4, 5; centro 2 (-): 2.2, 1.8, 2,3.

    - Se desplaza el prototipo hacia el centro de masas de su conjunto de ejemplos. Es decir, se hallan los nuevos centros calculando la media de las distancias que pertenecen a cada centro. En este caso los nuevos centros serán:

Tabla 3. Resultados obtenidos en la primera iteración del algoritmo K medias

Instancias	Distancia a centro 1	Distancia a centro 2	Clúster más cercano según distancia
5	1	2.67	1
2.2	1.8	0.13	2
1.8	2.2	0.53	2
4	0	1.67	1
2	2	0.33	2
3	1	0.67	2
5	1	2.67	1

centro1 (+) = (1+0+1)/3 = 0,666
centro2 (-) = (0,13+0,53+0,33+0,67)/4 = 0,415

- Se repite el procedimiento hasta que ya no se desplacen los centros. El resultado final del algoritmo es que para valores con clase + el valor con el cual se rellenará es 4.666 y para valores con clase – el valor con el cual se rellenará es 2,25.

Una extensión de este algoritmo es el denominado K-Modas en el cual se sustituye la media por la moda, para aplicarlo a datos categóricos, ya que K-Medias está orientado a datos numéricos.

2.3.2 Integración de datos [12]

La redundancia de datos puede ser detectada por el análisis correlacional [12]. Por ejemplo, dados dos atributos, la correlación entre los atributos puede ser medida por:

$$\frac{P(A \wedge B)}{P(A)P(B)}$$

(Ecuación 5)

Si el resultado es mayor que 1 entonces A y B están positivamente correlacionados. Cuanto más alto sea el valor mayor implicación habrá entre uno y otro. Por tanto, un valor alto puede indicar que alguno de los dos puede ser removido como redundancia.

Si el valor es igual a 1 indica que los dos atributos son independientes y que no existe correlación entre ellos. Si el valor es menor que 1 entonces A y B están correlacionados negativamente. Esto quiere decir que cada atributo disuade del otro.

Una integración de datos cuidadosa puede ayudar a reducir o prevenir las redundancias e inconsistencias y mejorar la calidad y velocidad de la obtención de resultados del análisis de datos.

2.3.3 Transformación de datos[12]

Normalizacion: Algunas técnicas de normalización son las siguientes:

- Normalización Min-Max: Ejecuta una transformación lineal de los datos originales. Con base en los valores mínimo y máximo de un atributo, se calcula un valor de normalización v' con base en el valor v de acuerdo con la sig*u*iente expresión:

$$v' = \frac{v - \min_A}{\max_A - \min_A}(nuevo\_\max_A - nuevo\_\min_A) + nuevo\_\min_A$$

(Ecuación 6)

Este método conserva las relaciones entre los datos originales.

Ejemplo: suponiendo que el valor mínimo y máximo de un atributo son $12 y $98 respectivamente, se requiere mapear los valores en un rango entre 0 y 1. Entonces tomando un valor de 73 por normalización min-max es transformado en:

$$\frac{73-12}{98-12}(1-0)+0 = \frac{61}{86} = 0,7093$$

- Normalización z-core
  Los valores para un atributo A son normalizados basados en la media y la desviación estándar de A. Un valor v de A es normalizado a v' con el cálculo de la siguiente expresión:

$$v' = \frac{v - media_A}{des\_est_A}$$

(Ecuación 7)

Este método es utilizado cuando el máximo y el mínimo del atributo A son desconocidos o cuando hay valores anómalos que predominan al utilizar la normalización min-max.

Ejemplo: suponiendo que la media y la desviación estándar de un atributo son $54 y $16 respectivamente. Con la normalización z-core un valor de 73 se transformaría en:

$$v' = \frac{73 - 54}{16} = \frac{19}{16} = 1{,}1875$$

- Normalización de escala decimal: Normaliza moviendo los puntos decimales de los valores del atributo A. El número de puntos decimales movidos depende del máximo valor absoluto de A. Un valor v de A es normalizado a v' con el cálculo de la siguiente *e*xpresión:

$$v' = \frac{v}{10^j}$$

(Ecuación 8)

Donde j es el entero más pequeño de Max($|v'|$) < 1.

Ejemplo: suponer que el rango de valores de los registros del atributo A es de -986 a 917. El máximo valor absoluto de A es 986 para normalizar por escala decimal se debe dividir cada valor por 1000 (j=3) entonces -986 es normalizado como -0,986.

Es de notar, que la normalización puede cambiar los datos originales un poco, especialmente los dos últimos métodos mencionados. También es necesario guardar los parámetros como la media o desviación estándar para uso futuro y que se pueda normalizar de manera uniforme.

Suavizado: las técnicas incluidas aquí son binning, clustering y regresión.

Agregación: este paso es generalmente usado en la construcción de los cubos de datos para el análisis de datos en diferente granularidad.

2.3.4 Reducción de datos

Las técnicas de reducción de datos pueden ser aplicadas para obtener una representación reducida de los datos manteniendo la integridad de los datos originales. Las estrategias para la reducción son las siguientes [12]:

- Agregación de cubos de datos: donde las operaciones de agregación son aplicadas a los datos en la construcción de los cubos.

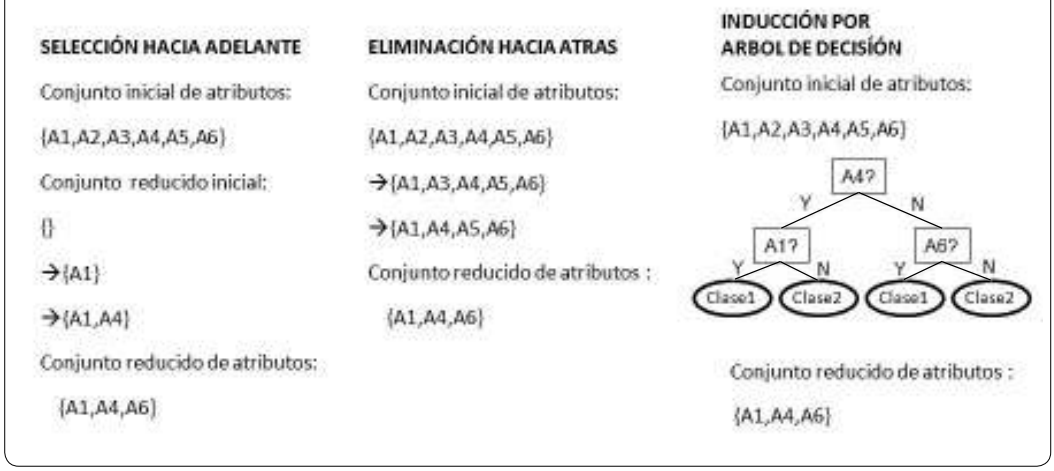
- Reducción de dimensión: donde pueden ser detectados y eliminados atributos o dimensiones poco relevantes o redundantes. La figura 2 muestra algunos métodos de selección de atributos.

También es utilizada la técnica de selección de atributos relevantes basada en Bootstraping.

- Compresión de datos: donde son usados mecanismos de codificación para reducir el tamaño del conjunto de datos. En este caso las técnicas utilizadas son la transformada de wavelet (DWT) o análisis de componentes principales (PCA).

- Reducción de numerosidad: donde todos los datos son reemplazados o estimados por representaciones de datos pequeños

Figura 2. Métodos heurísticos básicos para selección de atributos[12]



como modelos paramétricos, de los cuales sólo se guardan los parámetros y no los datos, o los no paramétricos como el clustering, el muestreo o el uso de histogramas.

- Discretización y generación del concepto de jerarquía: donde los valores son reemplazados por rangos o por datos de niveles conceptuales superiores.

Las técnicas de discretización pueden ser usadas para reducir el número de valores de un atributo continuo, dividiendo el rango del atributo en intervalos. Las etiquetas de los intervalos pueden ser usadas para reemplazar los valores actuales de datos. El concepto de jerarquías organiza los valores de los atributos o dimensiones en niveles graduales de abstracción. Ellos son una forma de discretización [12][18].

La generación automática del concepto de jerarquías para datos categóricos debe estar basada en el número de valores distintos de los atributos definidos en la jerarquía. Para datos numéricos, las técnicas que pueden ser usadas son: segmentación por reglas de partición, análisis de histogramas y análisis de clustering. El Chi-Merge también es un algoritmo de discretización automático que analiza la calidad de múltiples intervalos utilizando el estadístico Chi Cuadrado (χ^2).

3. OLAP y minería de datos

3.1 OLAP

OLAP es típicamente ejecutado para la validación de hipótesis de usuarios. Las funcionalidades OLAP incluyen dril-down, roll-up, slice, dice y operaciones de pivoteo para manejo flexible y transformación de datos. [16]. En el caso particular de esta revisión, no sólo se quiere aplicar OLAP para entender o visualizar datos, sino también para generar nuevos datos que puedan ser usados para producir nuevas hipótesis de aplicación de algoritmos de descubrimiento de conocimiento. En la tecnología OLAP se considera que los datos deben ser integrados en un DW o en un datamart como prerrequisito 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 datamart [16]. Así, todo lo que ha dado origen a OLAP hace que se refiera a una nueva forma para manejar la información importante especialmente para la toma de decisiones [11].

Los data warehouse y OLAP necesitan datos históricos, sumarizados y consolidados. Los datos están recopilados desde ciertas bases de datos operacionales y otras fuentes externas. Esto requiere no sólo consultas profundas y complejas, joins y agregaciones, sino también muchos GB o TB de capacidad de almacenamiento y una arquitectura típica como la que se muestra en la figura 3 [11].

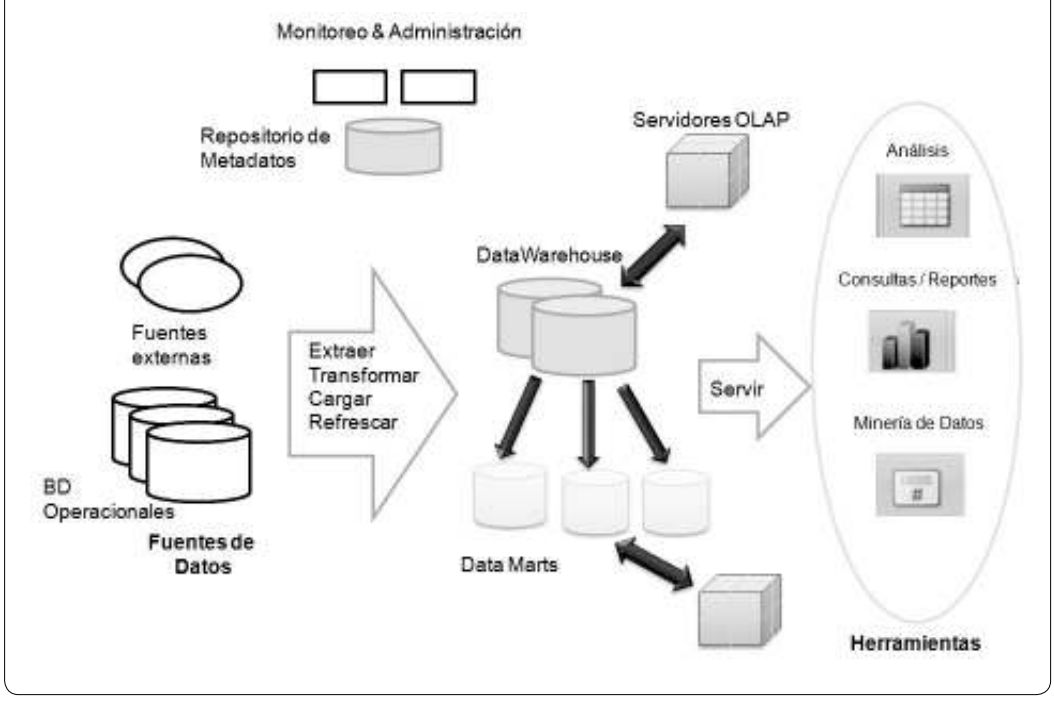
Algunos temas importantes en el área que involucra tecnología OLAP incluyen limpieza de datos, selección de índices, particionamiento de datos, vistas materializadas y administración del datawarehouse. El análisis de datos requiere extracción de datos relevantes del datawarehouse, agregando datos y analizando los resultados [6].

3.2 Minería OLAP (OLAP Mining)

Minería OLAP es un mecanismo que integra OLAP con minería de datos [11]. De esta forma se considera una exploración sana de cómo puede ser integrada la visualización de los datos con otras técnicas como aprendizaje inductivo y clustering jerárquico [11].

Minería OLAP se basa en las siguientes razones:

Figura 3. Arquitectura de un sistema típico de data warehouse [6]



- Las herramientas de minería de datos necesitan trabajar con datos integrados, consistentes y sin ruido, que requieren de los pasos de preprocesamiento en los cuales se hacen la limpieza e integración de datos. Un data warehouse construido por servidores de preprocesamiento, es tan valioso para OLAP como para minería de datos por la fuente de datos limpia e integrada.

- Minería OLAP facilita el análisis interactivo de datos exploratorios. Es proporcionado como una herramienta para las funciones OLAP de cualquier conjunto de datos en los cubos para análisis de datos en diferentes niveles de abstracción y para flexibilidad de interacción con motor de minería basado en resultados de minería intermedios.

- Por la integración de OLAP con múltiples módulos de minería de datos, minería OLAP proporciona flexibilidad para seleccionar las funciones de minería deseadas e intercambiar dinámicamente tareas de minería de datos.

Ejecutar Minería OLAP en cooperación con las funciones de minería de datos [11]:

- OLAP basado en caracterización y comparación: sumariza y caracteriza un conjunto de datos obtenidos de las tareas relevantes basados en la generalización de datos. Para la minería de conocimiento a múltiples niveles el drill-down y el roll-up son técnicas que pueden ser utilizadas.

  En el caso de que se quiera integrar caracterización y comparación multinivel, en cada paso del drill-down o del roll-up la caracterización y la comparación producen un cuboid, con la misma estructura de datos. Entonces, cualquier módulo de minería puede tratar el resultado de la caracterización o comparación como un cubo de datos y ejecutarse directamente sobre estos datos.

- OLAP basado en asociación:
  - Asociación intra-atributo: es la asociación formada entre uno o un grupo de atributos formado por la agrupación de otro conjunto de atributos en una relación.
  - Asociación inter-atributo: es la asociación entre un conjunto de atributos en una relación.

- OLAP basado en clasificación: han sido muchos los métodos estudiados para clasificación, incluidos los métodos de árboles de decisión como ID3 o C4.5, métodos estadísticos, redes neuronales, etc.

  La relevancia de los atributos está basada en el análisis de una medida incierta, una medida que determina qué tan relevante es un atributo en una clase. En el proceso de clasificación, el clasificador adopta un método de inducción basado en árbol de decisión que integra tecnología de cubos OLAP y luego árboles de decisión. En este caso, primero ejecuta la mínima generalización en los datos de entrenamiento y luego ejecuta el árbol de decisión sobre los datos generalizados. Para manejar el ruido y datos excepcionales y facilitar el análisis estadístico se introducen dos umbrales, el umbral de clasificación y el umbral de excepción.

- OLAP, basado en la predicción: minería OLAP puede ser integrada con la predicción, si cualquier clase de predicción puede ser identificada por un criterio de selección de la clase y esta característica puede ser mostrada. Entonces, las operaciones de cubos pueden ser ejecutadas sobre un cuboid seleccionado.

- OLAP basado en análisis de clustering: clustering de datos, conocido como "aprendizaje no supervisado", es un proceso de particionamiento de un conjunto de datos en un conjunto de clases llamadas clústeres, con los miembros de cada clúster compartiendo alguna propiedad común. Un buen método de clúster producirá clústeres de alta calidad, de tal forma que la similitud intraclase es alta y la similitud interclase es baja.

  El analizador de clúster está basado en el paradigma k-means. Comparado con otros métodos éste es prometedor por su eficiencia en procesamiento de gran volumen de datos. Sin embargo, como ya se había mencionado, su utilización se limita generalmente al uso de datos numéricos.

3.3 Preprocesamiento OLAP

En la práctica, la tarea de preprocesamiento de datos consume tiempo y tiene una importante influencia en la calidad de los modelos generados [16]. Diversas experiencias muestran que, por lo menos, los tres cuartos de tiempo deben ser empleados en la transformación de los datos para ser manejados en un formato apropiado para el aprendizaje y que este proceso ha influido significativamente en los modelos finales generados [8].

El preprocesamiento de datos formalmente se define como el conjunto de acciones tomadas antes de que inicie el proceso de análisis de datos. Esto es esencialmente una transformación T a los vectores de datos en nuevos vectores de datos [9].

$$Y_{ij} = T(X_{ik})$$
(Ecuación 9)

Donde Y_{ij} conserva la información relevante de X_{ik}, Y_{ij} elimina al menos uno de los problemas de X_{ik}. Con i como el número de objetos, j el número de atributos antes del preprocesamiento y k el número de atributos después del preprocesamiento. En general, j es diferente de k.

La ejecución de tareas de preprocesamiento tiene su justificación básicamente en las siguientes razones [9]:

- Solucionar problemas de datos para prevenir la obtención de resultados erróneos en el análisis de datos.
- Entender la naturaleza de los datos y realizar un análisis de datos más significativo
- Extraer el conocimiento más significativo de un conjunto de datos.

En la mayoría de aplicaciones es necesario aplicar más de una técnica de preprocesamiento, por esto una tarea crucial es la identificación del tipo de preprocesamiento [9]. Un ejemplo realista del preprocesamiento de datos puede ser hallado en la tecnología de base de datos, lo que generalmente es llamado como data warehouse. El resultado de consultas específicas, generalmente es guardado en vista que son independientes de la BD y por lo mismo permiten respuestas mucho más rápidas [5].

En cuanto a la tarea de discretización de variables numéricas, ésta utiliza diferentes aproximaciones como discretización dentro de un número de categorías, utilizando los puntos de corte equidistantes o discretización basada en la media y la desviación estándar. La discretización en el algoritmo CN4 está basada en la entropía o el estimado de LaPlace. Todos estos sistemas discretizan atributos numéricos "on-line", es decir, durante el aprendizaje. Los algoritmos que se denominan "off-line" indican que

discretizan antes de que la máquina inicie el paso de aprendizaje [2].

- Algoritmo para discretización [2]:
  Se trata cada atributo numérico por separado. La idea básica es crear intervalos para que la distribución de clases a posteriori P(C|intervalo) sea diferente de la distribución de clases a priori P(C) en el mismo conjunto de datos de entrenamiento. El número de intervalos resultantes es "controlado" por la definición de un umbral para el mínimo número de objetos dentro de un intervalo, los intervalos menos frecuentes son denominados como desconocidos.

  El algoritmo es descrito de la siguiente manera:

- Algoritmo para agrupamiento [2]:La agrupación de los valores nominales empieza a ser importante si el número de estos valores es bastante largo (por ejemplo: cientos de códigos ZIP o códigos de profesión).

  El algoritmo de agrupamiento está basado en la misma idea del algoritmo de discretización. Nuevamente se crean los grupos de valores, de tal forma que la distribución a posteriori P(C|grupo) difiera significativamente de la distribución a priori P(C). La principal diferencia con el algoritmo de discretización es que se crea un grupo para cada valor del atributo de la clase y un grupo adicional llamado "desconocido".

Tabla 4. Descripción del algoritmo de discretización

CICLO PRINCIPAL:
1. Crear lista ordenada de los valores de los atributos
2. Para cada valor:
   a. Calcular las frecuencias de ocurrencia de los objetos con respecto a cada clase.
   b. Asignar el nombre de la clase a todos los valores usando el procedimiento ASIGNAR.
Fin
3. Crear los intervalos a partir de los valores usando el procedimiento INTERVALO

ASIGNAR:
Si para todos los valores dados de los objetos estos pertenecen a la misma clase entonces asignar el valor de la clase
Si no, si para los valores dados de la distribución de objetos existe mucha diferencia con los miembros de la clase entonces el valor asignado es el de la clase más frecuente.
Si no asignar el valor "desconocido".

INTERVALO:
a. Si una secuencia de valores pertenece a la misma clase entonces crear el intervalo
INT_i = [$Limiteinferior_i$, $Limitesuperior_i$] a partir de estos valores
b. Si el intervalo INT_i pertenece a la clase "desconocido" entonces
Si los intervalos vecinos INT_{i-1}, INT_{i+1} pertenece a la misma clase entonces crear el intervalo INT_{i-1} U INT U INT_{i+1}
Si no, crear el intervalo INT_{i-1} U INT o INT U INT_{i+1} de acuerdo con los criterios dados.
c. Crear cubrimiento continuo de los atributos asignando $Limiteinferior_i$ = ($Limiteinferior_i$ + $Limitesuperior_{i-1}$)/2 y $Limitesuperior_i$ = $Limiteinferior_i$

El algoritmo es descrito de la siguiente forma:

- Evaluación de los algoritmos [2]: durante la discretización o agrupamiento de los datos se puede perder información oculta en los datos. Se puede medir la pérdida por el número de contradicciones antes y después del preprocesamiento. Contradicciones significa que los objetos descritos por el mismo valor de atributo pertenezcan a clases distintas. Cualquier algoritmo de aprendizaje clasificará estos objetos como objetos pertenecientes a la misma clase y objetos pertenecientes a otras clases serán clasificados erróneamente. Se cuentan los errores y se establece el máximo de exactitud de acuerdo con:

$$1 - \frac{No.Errores}{No.Objetos}$$

(Ecuación 10)

Como efecto del procesamiento "off-line", el número de objetos en los conjuntos de datos y el número de atributos puede ser reducido. Esta reducción puede dar alguna información acerca de patrones regulares en los datos.

Otro uso de la información es el número de intervalos o número de grupos. Si en el preprocesamiento resulta un solo grupo o un solo intervalo con alta frecuencia en los datos, se puede ignorar el atributo correspondiente en el siguiente paso de la máquina de aprendizaje.

- Selección de atributos basado en análisis de conflictos [15]:
  - Una buena aproximación para hallar la asociación entre dos variables es el análisis de la tabla de contingencia. Tradicionalmente, la tabla de contingencia se utiliza para variables nominales, variables cuyos valores provienen de un conjunto desordenado. Para la selección de atributos, se puede construir una tabla de contingencia de tal forma que las filas

Tabla 5. Descripción del algoritmo de agrupamiento

CICLO PRINCIPAL:
1. Para cada valor:
a. Calcular las frecuencias de ocurrencia de los objetos con respecto a cada clase.
b. Asignar el nombre de la clase a todos los valores usando el procedimiento ASIGNAR.
Fin.
2. Crea los grupos a partir de los valores usando el procedimiento AGRUPAR
ASIGNAR:
Si para todos los valores dados de los objetos pertenecen a la misma clase entonces asignar el valor de la clase.
Si no, si para los valores dados de la distribución de objetos existe mucha diferencia con los miembros de la clase entonces el valor asignado es el de la clase más frecuente.
Si no asignar el valor "desconocido".

AGRUPAR:
Crear grupos para valores con el mismo nombre de la clase.

Tabla 6. Datos ejemplo 1 de selección de atributos basado en el análisis de conflictos

	Rango		
Sexo	Profesor asistente	Profesor asociado	Profesor
Masculino	20	10	5
Femenino	5	3	2

representan valores distintos de un registro y las columnas son etiquetadas por las clases. Las entradas de la tabla son enteros no negativos, dando el número de tuplas en la clase, representadas por la columna y con un valor de atributo particular, representado por la fila.

- Ejemplo 1: si un registro de datos de una universidad tiene el atributo *sexo* y otro atributo *rango* como la etiqueta de la clase, la tabla de contingencia podría ser como la tabla 6.

Al analizar este caso e independientemente de la distribución que se seleccione para el análisis, que generalmente es χ^2, se puede ver que el sexo no es perfecto asociado con el rango, porque contiene más de una fila cuya entrada es diferente de cero. En cuyo caso, dando el valor del atributo no se puede determinar únicamente su etiqueta de clase. En este caso se denominan *entradas conflcitivas*.

- Ejemplo 2: la tabla de contingencia que relaciona *rangos de salario* como atributo con *rango* es representada con la tabla 7.

Aquí se puede observar que no hay conflicto de entradas en la tabla. Dado un valor de salario, el rango puede ser determinado de forma única. La selección de atributos para problemas de clasificación se puede reducir al problema de hallar un conjunto de atributos que no estén en conflicto en las filas de entrada de la tabla.

La selección de atributos, basada en el análisis de conflicto de la tabla de contingencia, se espera que trabaje mejor con atributos nominales. La principal razón es que usualmente hay números pequeños de valores distintos para cada

Tabla 7. Datos ejemplo 2 de selección de atributos basado en el análisis de conflictos

	Rango		
Rangos de salario	Profesor asistente	Profesor asociado	Profesor
[3000-5000]	25	0	0
[5000-8000]	0	13	0
[8000-12000]	0	0	7

atributo. Si el número de muestras y el conjunto de entrenamiento es suficientemente grande, habrá suficientes repeticiones de los mismos valores de atributos en el conjunto de entrenamiento.

El segundo aspecto en el análisis de conflictos es aquél en el que los atributos son escogidos para construir la contingencia. Así, es frecuente el caso en el que un conjunto de atributos colectivamente determinan la etiqueta de clase de una tupla.

La mayoría de datos de aplicaciones reales contienen ruido. Para problemas de clasificación como el ruido puede dar lugar a etiquetas inconsistentes de tuplas e introducir entradas conflictivas en la tabla de contingencia. Una solución sencilla es usar un valor umbral, llamado tolerancia, para tolerar un pequeño porcentaje de ruido.

3.4 Trabajos relacionados

Existe una colección de técnicas denominadas Soft Computing que tiene en cuenta el preprocesamiento de datos, el cual constituye uno de los procesos fundamentales para el manejo de los datos, de forma que se puedan seleccionar los atributos adecuados para que el proceso de minería de datos se pueda realizar con mayor información. En una forma similar, como ocurre con los diferentes pasos de la minería de datos, entre los cuales se tiene la limpieza de datos y el preprocesamiento, dentro del contexto de modelamiento y más concretamente en el modelamiento difuso, el objetivo es hallar un conjunto de relaciones que describan el comportamiento con los datos actuales de acuerdo con un conjunto de patrones o con reglas difusas del tipo if-then [10].

De la misma forma, DBMiner fue desarrollado como apoyo a Minería OLAP; además de las funciones de caracterización, incluye otras funciones de minería de datos como la asociación, clasificación, predicción, clustering y secuenciación [11][1].

3.4.1 Reducción de datos con diseño de experimentos para preprocesamiento de datos [17]

Mientras hoy en día muchas técnicas son usadas para la reducción de datos y pueden incrementar la velocidad en el proceso analítico de un data warehouse, siempre hay una opción de borrar un conjunto de características variables que podrían no tener contribución en el pasado, pero que podrían ser significativas después. El análisis siempre ha sido influenciado por la selección de la técnica; en gran medida, el diseño de experimentos puede ayudar en la reducción de datos durante el preprocesamiento de los datos, indicando un ahorro significativo en costos y tiempo.

El diseño de experimentos ha sido una disciplina con una amplia aplicación a través de las ciencias de la ingeniería. El uso de técnicas como factorial fraccional, arreglos ortogonales y especialmente las técnicas taguchi han sido muy investigadas y utilizadas ampliamente en control de calidad estadístico, pero no mucho en data warehouse y minería de datos. El enfoque para hallar los factores que afectan un producto en el diseño de experimentos, puede reducir drásticamente el número de pruebas requeridas para reunir los datos necesarios.

3.4.2 Algoritmos de "asignación" para OLAP sobre datos imprecisos [4]

Mientras hay mucho trabajo en la representación y consulta sobre datos ambiguos, también es importante definir algunos criterios para el manejo de datos ambiguos manejados

1 En este articulo se examinan los principios de minería OLAP y se estudian las técnicas de implementación con el sistema DBMiner como ejemplo de implementación

por aplicaciones OLAP. Uno de los criterios es la coherencia, la cual representa las relaciones entre las consultas similares generadas de relacionar nodos en un orden de jerarquía que satisfagan la intuición de los usuarios, en la cual ellos pueden navegar a través de la jerarquía.

El segundo criterio es llamado la fidelidad, captura la intuición de datos más precisos, que dirigen a mejores resultados. El tercer criterio llamado correlación-preservación esencialmente requiere que las propiedades estadísticas de los datos deberían no afectarse por la asignación de registros de datos ambiguos.

Recientes trabajos han propuesto extender el modelo OLAP para soportar la ambigüedad de los datos, específicamente imprecisión e incertidumbre. Un proceso llamado *asignación* fue propuesto para transformar una tabla de hechos imprecisa en una forma llamada la base de datos extendida que puede ser leída para responder consultas de agregaciones OLAP. Se realiza una extensión del modelo relacional para OLAP y, de esta forma, poder manejar los datos imprecisos y con incertidumbre. Los atributos en el modelo OLAP estándar son de dos clases, dimensiones y medidas. En el momento de extender el modelo, se soporta la incertidumbre en los valores de las medidas y la imprecisión en los valores de las dimensiones.

3.4.3 M4 - un metamodelo para preprocesamiento de datos [20]

Éste es el metamodelo utilizado por Mining Mart, un sistema para soportar preprocesamiento para minería de datos. En particular, los metadatos pueden ser cualquier información relacionada con las definiciones de esquema y configuración de especificaciones, almacenamiento físico, derechos de acceso, conceptos de negocio, terminología y detalles acerca de los reportes de usuario.

El propósito de tecnologías avanzadas, como minería de datos y data warehouse es la extracción de información y conocimiento a partir de los datos. La minería de datos busca detectar patrones desconocidos en datos que son usados para soporte de análisis de negocios y predicción. Las operaciones de preprocesamiento incluyen transformaciones de datos, agregación, discretización, segmentación y muestreo. Las experiencias prácticas han demostrado que entre el 50 y el 80 % de los esfuerzos de descubrimiento del conocimiento son gastados en el preprocesamiento de datos, lo cual no sólo consume tiempo, sino que también requiere conocimiento profundo del conocimiento del negocio, la minería de datos y las bases de datos.

M4 combina ideas desde dos estándares existentes para representación de metadatos e intercambio en el área de data warehouse. Ellos son drásticamente simplificados y extendidos con la minería de datos y los elementos del preprocesamiento para crear el metamodelo de dominio específico.

En este caso, es usado el contexto de Mining Mart para representar un ejemplo típico de herramientas orientadas a metadatos y su correspondiente metamodelo. Los repositorios son integrados o interoperan rastreando su esquema de metadatos con el metamodelo común. Un paso en esta dirección ha sido el estándar para representación e intercambio propuesto por la OMG, llamado Common Warehouse Metamodel (CWM).

3.4.4 discretización de dimensiones con valores continuos en cubos de datos OLAP [17]

Lo que se busca con el desarrollo de este proyecto es otra forma de integrar OLAP con

minería de datos, 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. Actualmente, existen diversos algoritmos que están automatizando las tareas asociadas a la discretización, pero las herramientas OLAP aún no los incorporan en sus aplicaciones. Se encuentran determinando cómo aplicar la discretización automática en la definición de cubos OLAP que permitan simplificar los datos con la menor pérdida posible de información.

4. Trabajos futuros

Las diversas técnicas de preprocesamiento de datos y las tendencias que se presentan en las compañías, en relación con el manejo y acceso de su información para la toma de decisiones, generan inquietudes que extienden el espectro de posibilidades en las cuales se pueden orientar nuevas investigaciones. Generalmente, la calidad de los datos es una prioridad en el momento de visualizar o analizar datos, más aún cuando éstos son estructurados. Un enfoque interesante en este sentido, podría ser la realización de la revisión de las técnicas utilizadas con mayor frecuencia para el preprocesamiento de datos y sus algoritmos asociados, en particular, lo relacionado con el manejo de datos faltantes, de tal forma que sea posible la comparación de su efectividad para el análisis, dependiendo del conjunto de datos utilizado y encaminados a aplicaciones OLAP.

5. Conclusiones

Durante el preprocesamiento, los datos son principalmente orientados en alguna de las siguientes tres direcciones:

- Limpieza de datos: tratamiento del ruido, valores faltantes, valores anómalos, redundancia, entre otros.
- Alterar la dimensionalidad de los datos: generación de atributos, filtrado, transformaciones, etc.
- Alterando la cantidad de datos: por selección, muestreo y balanceando los registros de datos disponibles.
- El preprocesamiento de datos es una etapa en la cual los cambios hechos a un conjunto de datos pueden brindar una pronta solución a un problema de descubrimiento de conocimiento.
- Al igual que diferentes funciones conocidas para minería de datos, las tareas de preprocesamiento pueden ser aplicables en entornos OLAP con la recién denominada minería OLAP. En este sentido, se pueden generar proyectos de investigación válidos en los cuales se puede analizar la pertinencia en aplicaciones OLAP de cada una de las técnicas utilizadas en minería de datos.
- Las experiencias relacionadas con el preprocesamiento de datos permiten determinar que las técnicas de manejo de datos faltantes son ampliamente utilizadas y pueden ser objetivo de investigación para determinar su efectividad en datos estructurados y en particular en aplicaciones de tecnología OLAP.
- Algunas de las técnicas de preprocesamiento están siendo automatizadas y ya son aplicables así en procesos de minería de datos, tal es el caso de la discretización. Se han iniciado avances y ambientes de prueba para realizar la revisión de posibilidades de automatizar esta y otras tareas de preprocesamiento de datos aplicables a los cubos OLAP y el modelo dimensional.
- La calidad de los datos juega un papel muy importante en el proceso analítico, a partir de cuyos resultados se podrían establecer reglas o patrones para la toma de decisiones. Generalmente, los datos que se obtienen de aplicaciones de

producción de las organizaciones, por su dinamismo, son generados con ciertas inconsistencias, las cuales, sin duda, afectan significativamente los resultados obtenidos de cualquier procedimiento metódico que sobre ellos se realice. Allí es donde radica la importancia del preprocesamiento y la necesidad de analizar cuál puede ser la mejor técnica que se va a utilizar en cada una de las tareas, para que los cubos de datos OLAP sean construidos con datos de alta calidad, mejorando el desempeño de las aplicaciones y permitiendo el ahorro de tiempo y costos durante la etapa de diseño e implementación.

Referencias bibliográficas

[1] Barrera, H., Correa, J., y Rodríguez, J. Prototipo de software para el preprocesamiento de datos - UDClear". IV Simposio Internacional de Sistemas de Información e Ingeniería de Software en la Sociedad del Conocimiento, libro de actas volumen 1, ISBN 84-690-0258-9.

[2] Berka, Petr y Bruha Ivan. *Discretization and Grouping: Preprocessing Steps for Data Mining*. 1998.

[3] Berry, Michael J.A., Linoff Gordon S. *Data Mining Techniques*. Wiley Publishing, Inc. 2004.

[4] Burdick, Doug, et al. *Efficient Allocation Algorithms for OLAP over Imprecise Data*. VLDB 06, September 12-15, 2006, Seoul, Korea. Copyright 2006 VLDB Endowment, ACM 1-59593-385-9/06/09.

[5] Cadoli, Marco, Donini Francesco, Liberatore Paolo y Shaerf Marco. *Preprocessing of Intractable Problems*. Dipartimento de Informatica e Sistemistica, Universita di Roma "La Sapienza", Italy. Technical Report. 1997.

[6] Cheung Pui Ling Pauline, et al. *Data Warehousing and OLAP*. 2000.

[7] Clifton, Chris. *Introduction to Data Mining*. Purdue University, 2004.

[8] Engels, Robert y Theusinger Christiane. *Using a Data Metric for Preprocessing Advice for Data Mining Applications*. ECAI 98, 13th European Conference on Artificial Intelligence. Jhon Willey & Sons, 1998.

[9] Famili, A., Shen Wei-Min, Weber Richard y Simoudis Evangelos. Data Preprocessing and Intelligent Data Analysis. *Submitted to Intelligent Data Analysis Journal*, 1997.

[10] Gómez-Skarmeta, Antonio, Jiménez Fernando e Ibañez Jesus. *Data Preprocessing in Knowledge Discovery with Fuzzy-Evolutionary Algorithms*. Departamento de Informática, Inteligencia Artificial y Electrónica, Universidad de Murcia. 1998.

[11] Han, Jiawei. Olap *Mining: An Integration of OLAP with Data Mining*. Intelligent Database Systems Research Laboratory. 1997.

[12] Han, Jiawei y Kamber Micheline. *Data mining, Concepts and Techniques*. Segunda edición. 2006.

[13] Hing-Yan, Lee y Hwee-Leng Ong. A New V*isualisation Technique for Knowledge Discovery in OLAP*. Japan-Singapore AI Centre Information Technology Institute. Singapore. 2000.

[14] Kotsiantis, S. B., Kanellopoulos D. y Pintelas P. E. Data Preprocessing for Supervised Learning. *International Journal of Computer Science*, Vol. 1 No. 2 2006.

[15] Lu, Hongjun, Sung Sam Yuan y Lu Ying. *On Preprocessing Data for Effective Classification*. Department of Information Systems and Computer Science, National University of Singapore. 1996.

[16] Maedche, Alexander, Hotho Andreas y Markus Wiese. *Enhancing Preprocessing in Data-Intensive Domains using Online-Analytical Processing*. 2000.

[17] Maliakal, Jose. Data Reduction with Design of Experiments (DoE) for Data Mining Pre-Processing. *Proceedings of World Academy of Science, Engineering and Technology,* Vol. 26, December 2007.

[18] Palaniappan, Sellappan y Hong Tan Kim. Discretization of Continuos Valued Dimensions in OLAP Data Cubes. *IJCSNS International Journal of Computer Science and Network Security*, Vol.8 No.11. November 2008.

[19] Rodríguez, Nestor y Sánchez Wilson. Proyecto de grado: Software para preprocesamiento de datos UDCLEAR versión 2.0. Universidad Francisco José de Caldas, Facultad Tecnológica. 2008.

[20] Vaduva, Anca, Kietz JörgUwe y Zücker Regina. *M4 - A Metamodel for Data Preprocessing*. 2001.