Revista Venezolana de Computación
ISSN: 2244-7040
http://www.svc.net.ve/revecom
Vol. 5, No. 2, pp. 31-40, Diciembre 2018
Selección de los Mejores Artículos de CoNCISa 2018

Extensión UML para Clustering Difuso en Data Warehouse

Livia Borjas[1], Rosseline Rodríguez[2], Betzaida Romero[2]
livacaro7@gmail.com, crodrig@usb.ve, betzaidaromero@usb.ve

[1] Instituto Universitario de Tecnología Dr. Federico Rivero Palacios, Caracas, Venezuela
[2] Departamento de Computación, Universidad Simón Bolívar, Caracas, Venezuela

Resumen: La Minería de Datos (MD) aplica métodos que generan modelos inteligibles desde grandes volúmenes de datos, usando técnicas de análisis introspectivo para descubrir patrones y relaciones ocultos. Ésta es una fase importante del proceso conocido como Descubrimiento de Conocimiento en Bases de Datos (KDD). *Clustering* es una técnica de aprendizaje no supervisado, ampliamente utilizada para encontrar "comportamientos" en una larga colección de datos, popularmente usada en KDD. Esta técnica ha sido mejorada aplicando conjuntos difusos, surgiendo los algoritmos de *Clustering* Difuso que permiten descubrir "*clusters*" que se solapan en la frontera. El problema con la aplicación de estas técnicas es que se hace en niveles bajos de abstracción en donde la información es compleja. Sería ideal modelar el proceso de extracción de conocimiento desde niveles altos de abstracción donde los datos son sencillos, inteligible y cuyos modelos no dependen de las herramientas subyacentes para su implementación. Además, se pueden aprovechar los beneficios que ofrecen los modelos multidimensionales que facilitan el modelado del KDD disminuyendo su complejidad de las fases de recopilación e integración de los datos. En el presente artículo se propone una extensión por medio de perfiles UML para la modelación de Minería de Datos, basada en *Clustering* Difuso.

Palabras Clave: Minería de Datos, Clustering Difuso, Data Warehouses, UML, KDD.

Abstract: Data Mining (MD) applies methods that generate intelligible models from large volumes of data, using introspective analysis techniques to discover hidden patterns and relationships. This is an important phase of the process known as Knowledge Discovery in Databases (KDD). Clustering is an unsupervised learning technique, widely used to find "behaviors" in a long data collection, popularly used in KDD. This technique has been improved by applying fuzzy sets, resulting in Fuzzy Clustering algorithms that allow discovering "clusters" that overlap at the border. The problem with the application of these techniques is that it is done at low levels of abstraction where the information is complex. It would be ideal to model the process of extracting knowledge from high levels of abstraction where the data is simple, intelligible and whose models do not depend on the underlying tools for its implementation. In addition, you can take advantage of the benefits offered by the multidimensional models that facilitate the modeling of the KDD, reducing the complexity of the phases of data collection and integration. In the present article, an extension is proposed by means of UML profiles for the modeling of Data Mining, based on Fuzzy Clustering.

Keywords: Data Mining, Fuzzy Clustering, Data Warehouses, UML, KDD.

I. Introducción

Como resultado de la automatización de procesos a toda escala y de los logros alcanzados en tecnologías de información y de almacenamiento de datos, en los últimos años ha aumentado el uso de bases de datos de gran volumen. El análisis de estos extensos volúmenes de datos tiene un gran valor agregado para las organizaciones, brindando conocimiento nuevo de interés para la toma de decisiones estratégicas propias de funciones complejas como la planificación y la predicción, en donde los sistemas de bases de datos tradicionales son insuficientes.

El procesamiento automático de grandes volúmenes de datos a fin de encontrar conocimiento útil para un usuario, es el objetivo principal del proceso de Descubrimiento de Conocimiento en Bases de Datos (*Knowledge Discovery in Databases* o KDD), el cual identifica "patrones comprensibles que se encuentran ocultos en los datos" [1].

Las fases del proceso de KDD son iterativas e incluyen: (1) Recopilación e Integración de los datos en una tabla llamada Atributo-Valor; (2) Selección, Limpieza y Transformación de los Datos para construir la Vista Minable; (3) Minería de Datos (Extracción de Conocimiento) que genera Modelos; (4) Interpretación o Evaluación, para llegar finalmente al producto o Conocimiento; y la fase de (5) Difusión y uso (Divulgación) que lleva a las decisiones estratégicas [2]. Uno de los caminos a seguir en el proceso KDD es implementar un Almacén de Datos (AD), o *Data Warehouse* (DW) durante el pre-procesamiento de los datos. El DW es un repositorio de datos

históricos coleccionado de diversas fuentes bajo un esquema unificado e integrado, frecuentemente modelado en forma multidimensional. En la Figura 1, se observa el esquema del KDD integrado con DW, donde la tabla Atributo-Valor es sustituida por el modelo multidimensional del almacén de datos.

Fayyad [3] describe las fases del KDD integradas con un DW:

1) Recopilación e Integración de los Datos en un DW, que incluye la determinación de fuentes de información que pueden ser útiles y su ubicación; el diseño del esquema DW que unifique de manera operativa toda la información recogida y la implantación del DW que permita la navegación y visualización de sus datos, para discernir aspectos a ser estudiados.

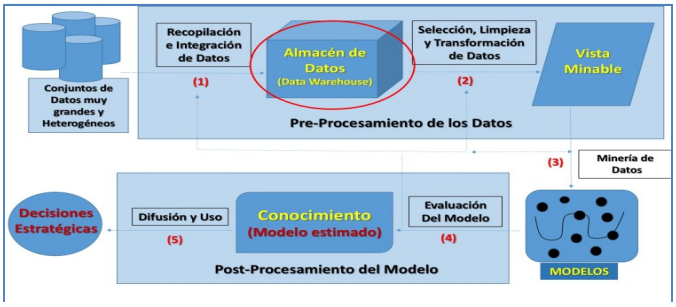

Figura 1: Proceso de KDD Integrado con un DW

1) Recopilación e Integración de los Datos en un DW, que incluye la determinación de fuentes de información que pueden ser útiles y su ubicación; el diseño del esquema DW que unifique de manera operativa toda la información recogida y la implantación del DW que permita la navegación y visualización de sus datos, para discernir aspectos a ser estudiados.

2) Selección, limpieza y transformación de los datos que se van a analizar. Considerando la información disponible relacionada con el dominio de los datos, en esta fase se corrigen o eliminan los datos incorrectos, inconsistentes, ausentes e incompletos. De igual manera se seleccionan los atributos relevantes para el estudio.

3) Minería de Datos (MD). En esta fase se identifica la tarea de MD a realizar, así como el método o la técnica más apropiada para alcanzar los objetivos de análisis planteados.

4) Evaluación del modelo, que incluye la interpretación, transformación y representación de los patrones extraídos; se evalúan los modelos descubiertos con los expertos del dominio del problema y se resuelven posibles inconsistencias o conflictos con el conocimiento disponible.

5) Divulgación y uso del nuevo conocimiento a todos los usuarios, de manera tal que se puedan realizar decisiones estratégicas en la organización interesada en el estudio.

La implantación de un DW como paso previo a la Minería de Datos dentro del KDD resulta útil por dos razones: los pasos de procesamiento e integración de datos para producir el DW genera un repositorio que facilita los objetivos del análisis y el desarrollo del modelo multidimensional del DW favorece la labor de conceptualización de los fenómenos del universo en observación. Un ejemplo de aplicación puede observarse en [4]. Los problemas que se pudieran encontrar durante el proceso KDD son:

- Visualizar el KDD como operaciones simples aisladas en lugar de un proceso integrado lo que produce duplicidad de tiempo y desperdicio de recursos. Por ejemplo, el pre-procesamiento de datos que comparten tanto el KDD como la Minería de Datos.
- Cuando los datos a ser analizados durante la Minería de Datos se encuentran almacenados en archivos planos, sobre todo si se utilizan en un modelo previo al KDD.
- Carencia de mecanismos de modelación del mundo real donde se aplique técnicas de MD.
- Perdida de oportunidad de nuevos conocimientos en cada paso del KDD.

El proceso de modelamiento de almacenes de datos como modelos multidimensionales ha arrojado avances [5][6][7][8] aprovechables en técnicas de minería de datos descriptiva y difusa ampliamente utilizada por su poder expresivo conocida como *Clustering* Difuso. A pesar que existen esfuerzos de propuestas de mecanismos y metodologías que permiten la especificación, modelación o implementación de requisitos difusos [9][10][11][12], no existen propuestas para el modelado de requisitos difusos que involucren técnicas de MD para el tratamiento de grandes volúmenes de datos en espacios multidimensionales.

La disposición de un mecanismo de modelaje de *clustering* difuso en grandes volúmenes de datos que use espacios multidimensionales, contribuiría con el objetivo del modelado conceptual [13]: "captar y enumerar exhaustivamente los requisitos y el dominio de conocimiento, de forma que todos los implicados puedan entenderlos y estar de acuerdo con ellos". El presente trabajo plantea una extensión del mecanismo propuesto por Zubcoff et al. [14] del proceso de KDD en espacio multidimensional, para la modelación de *clustering* difuso como técnica de Minería de Datos, con el fin de resolver los problemas mencionados encontrados en este proceso.

El resto del documento está estructurado de la siguiente forma: la Sección II describe el marco teórico que sustentan este trabajo. En la Sección III, se presenta la propuesta de un perfil UML para diseñar modelos de *clustering* difusos sobre espacios multidimensionales. La Sección IV muestra su aplicación a un caso de estudio. Finalmente, la Sección V presenta las conclusiones y trabajos futuros.

II. Marco Teórico

Para facilitar la comprensión de esta propuesta, se presentan las bases teóricas relacionadas con los mecanismos de extensión de UML, así como, los componentes fundamentales de la arquitectura de Minería de Datos con *Clustering* Difuso en espacios Multidimensionales.

A. Lenguaje Unificado de Modelación (UML)

Un modelo es una representación que describe un sistema o parte de él en un lenguaje con una sintaxis y semántica precisa, que puede ser interpretado y manipulado por un ordenador, de manera que pueda ser comprendido por diferentes diseñadores, independientemente de su implementación. Un artefacto es un

modelo o pieza de información producido en el proceso de desarrollo de software.

UML [15] es un lenguaje gráfico de propósito general, altamente flexible y expresivo, definido para la modelación de sistemas, de amplio uso por los arquitectos de software, cuyo propósito es especificar, construir y documentar los componentes de estos sistemas. Cuando UML resulta insuficiente para modelar dominios muy específicos se restringe o especializa los constructores propios de dicho lenguaje, como son: clases, asociaciones, atributos, operaciones, transiciones, entre otros. Además, UML incluye un mecanismo de extensión que permite definir lenguajes de modelación que son derivados de él [16]. El paquete *Profiles* de UML 2.0 provee mecanismos para extender y adaptar las metaclases de un metamodelo cualquiera a las necesidades concretas de una plataforma o de un dominio de aplicación.

Los perfiles UML están basados en estereotipos, restricciones y valores etiquetados adicionales que son aplicados a los elementos o relaciones de un diagrama. Un perfil se define en un paquete UML, estereotipado «profile», que extiende a un metamodelo o a otro perfil. Para definir perfiles se utilizan tres mecanismos: estereotipos (*stereotypes*), restricciones (constraints), y valores etiquetados (*tagged values*).

Un estereotipo está definido por un nombre y por una serie de elementos del metamodelo sobre los que puede asociarse. Gráficamente, los estereotipos se definen dentro de cajas etiquetadas «stereotype», a las cuales es posible asociarles restricciones, usando el lenguaje OCL o lenguaje natural, que imponen condiciones sobre los elementos del metamodelo. Un valor etiquetado es un metaatributo adicional que se asocia a una metaclase del metamodelo extendido por un perfil. Todo valor etiquetado ha de contar con un nombre y un tipo, y se asocia a un determinado estereotipo.

OCL (*Object Constraint Language*) es un lenguaje formal propuesto por OMG [17], usado para describir expresiones sobre UML que modelan condiciones invariantes que el sistema debe cumplir, así como para modelar pre y post condiciones y consultas sobre los objetos del modelo. Estas restricciones OCL pueden ser omitidas dentro del modelo gráfico, sin embargo son muy útiles en aquellos casos donde el modelo no es suficientemente expresivo y/o se quiere evitar estados indeseables del sistema. Las expresiones OCL son de la forma context TypeName inv Expression, en donde: *context* e *inv* son palabras reservadas del lenguaje; TypeName el nombre de la clase que representa el contexto y Expression la restricción cuyo resultado es un valor booleano. La declaración del contexto es opcional.

Según Fuentes y Vallecillo [16] definir un perfil UML incluye las siguientes consideraciones:

1) Definición del metamodelo de la plataforma o dominio de aplicación a modelar.

2) Definición del perfil dentro del paquete «profile» incluyendo un estereotipo por cada uno de los elementos del metamodelo. Estos estereotipos tendrían el mismo nombre que los elementos del metamodelo, a fin de establecer la relación entre el metamodelo y el perfil.

3) Aplicación de cada estereotipo a la metaclase de UML que se utilizó en el metamodelo del dominio para definir un concepto o una relación.

4) Los elementos del perfil serán los atributos del metamodelo, definidos como valores etiquetados, incluyendo la definición de sus tipos y sus posibles valores iniciales.

5) Las restricciones del dominio serán las restricciones que forman parte del perfil.

En este trabajo se propone extender UML con un perfil que permite representar requisitos difusos para clustering difuso en espacios multidimensionales. El perfil propuesto se basa en estereotipos y en el uso del lenguaje OCL [17] para la especificación formal de tales requisitos.

B. Modelo de Dominio de Clustering Multidimensional

En la presente investigación se reutiliza el perfil UML para el modelo multidimensional de un DW propuesto por Lujan-Mora et al. [18]. Para comprender el perfil UML es necesario comprender primero el modelo multidimensional usado para modelar almacenes de datos así como la técnica de agrupamiento o *clustering* de MD, los cuales se describen a continuación.

Los Almacenes de Datos (*Data Warehouse* o DW) se modelan como espacios multidimensionales donde los datos se organizan en hechos y dimensiones. Los hechos representan colecciones de medidas en forma de datos numéricos, de tal manera que tienen dimensiones asociadas que representan descripciones (datos textuales) que ofrecen un contexto al análisis, formando jerarquías, cuyas medidas pueden ser agregadas a distintos niveles de granularidad [14]. En la Figura 2 [14], se muestra un modelo multidimensional que representa la cantidad de productos adquiridos en una organización, a través cuatro dimensiones: Tiempo, Producto, Cliente y Causa. Cada una de estas dimensiones tiene distintas granularidades o niveles de agregación, como por ejemplo, Fecha, Mes, Año y Todo en la dimensión Tiempo.

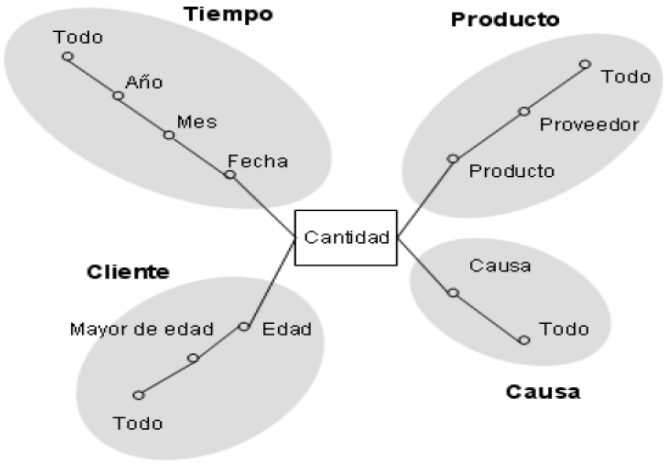

Figura 2: Modelo Multidimensional

Por otro lado, la Minería de Datos (*Data Mining* o DM) es "un mecanismo de explotación, consistente en la búsqueda de información valiosa en grandes volúmenes de datos, con el fin de descubrir patrones, relaciones, reglas, asociaciones o incluso excepciones útiles para la toma de decisiones" [1]. Las técnicas DM se dividen en dos categorías [19]: predictivas y

descriptivas, según su funcionalidad. Las predictivas o supervisadas predicen el valor de un atributo (etiqueta) de un conjunto de datos a partir de datos previamente conocidos, entre ellas están: la clasificación, la regresión y la predicción. Las técnicas descriptivas o no supervisadas descubren patrones y tendencias en los datos, entre las cuales se tienen, el clustering, la asociación y la correlación y dependencia. Dado que el perfil UML aquí presentado se basa en el clustering se dará más detalle de esta técnica.

El *clustering* se basa en la división de los datos en grupos de objetos llamados *clusters* [2]. Consiste en agrupar una colección dada de patrones no etiquetados con el fin de detectar grupos de individuos. También se le denomina clasificación no supervisada pues durante este proceso no hay clases predefinidas ni registros que permitan conocer las relaciones existentes entre los datos. En esta técnica, los grupos se van formando de acuerdo a las características de los datos, maximizando la similitud dentro de los grupos pero a la vez minimizando la similitud entre los distintos grupos [20]. De esta manera, se busca que los objetos que pertenecen a un grupo sean homogéneos entre sí, y que los distintos grupos sean lo más heterogéneos posible (Figura 3).

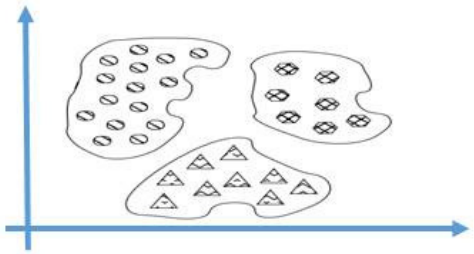

Figura 3: Representación Gráfica del Resultado del *Clustering*

Los algoritmos de *Clustering* intentan minimizar la distancia dentro del grupo de objetos y maximizar la distancia entre grupos, por lo que en ocasiones surge ambigüedad pues no siempre es claro por cuál características agrupar o cuántos grupos hacer. En un modelo de dominio para *clustering* multidimensional, los distintos algoritmos de *clustering* (K-Means, EMI, entre otros) pueden aprovechar los datos estructurados en un modelo multidimensional, de tal manera que los hechos del espacio multidimensional se relacionan directamente con las técnicas de minería de datos, como se observa en la Figura 4 [14].

Los diferentes algoritmos de *clustering* en espacios multidimensionales se caracterizan por:

- Identifican comportamientos comunes en un conjunto de datos cuyos usuarios no podrían derivar a través de la observación casual, aprovechando la potencia de los modelos multidimensionales para descubrir grupos con comportamientos similares.
- La estructura multidimensional facilita la comprensión de los datos, dado que representa el dominio del sistema de una manera muy cercana a la forma de pensar de los analistas.

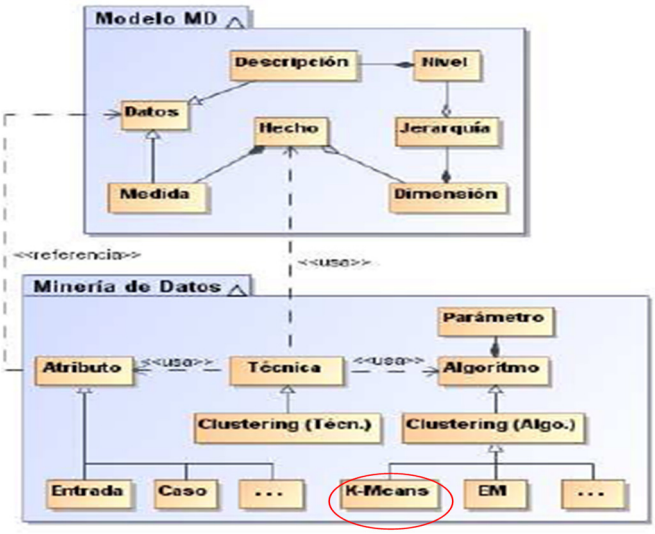

Figura 4: Modelo del Dominio de *Clustering* Multidimensional

- El resultado del análisis es una estructura de partición del conjunto de datos. Por lo que un apropiado modelo multidimensional facilita la representación de los datos en estudio.
- La técnica de *clustering* descubre grupos de objetos similares o con patrones de comportamiento comunes en base al hecho en estudio, considerando las distintas dimensiones en cualquier nivel de detalle en sus jerarquías.
- El modelo multidimensional representa de una manera apropiada y fácil los datos a analizar. Por ejemplo, en la Figura 5 [14] se observa que el hecho bajo análisis (H) contiene medidas (M1 y M2) que son contextualizadas por las dimensiones (D1 a D6). Cada dimensión usada como entrada representará un eje de *clustering* y cada caso corresponderá a particiones de esos ejes (planos C1, C2 y C3). Los *clusters* se muestran como agrupaciones de puntos presentes en las particiones de los ejes de *clustering* que representan las dimensiones usadas como entrada al proceso de análisis.

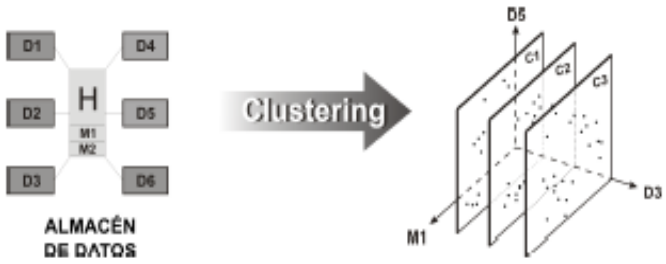

Figura 5: *Clustering* en Minería de Datos

- Los algoritmos de *clustering* tienen como entradas los atributos que utiliza para construir el espacio multidimensional en el cual se miden las similitudes de los datos. La salida de este proceso es un número de

clusters que forman una partición del conjunto de datos en el espacio multidimensional.

C. Clustering Difuso

En el *clustering* clásico, cada patrón pertenece a un único *clúster*. Sin embargo, existen situaciones reales donde los objetos agrupados, debido a su naturaleza, no sólo pertenecen a una partición excluyente e inequívoca, sino que podrían pertenecer a varias particiones que se solapan. Esto genera la necesidad de realizar agrupamientos más flexibles donde la similitud entre los objetos y la pertenencia de un objeto a *clúster* no sea precisa. Una manera de representar esta pertenencia gradual es a través de conjuntos difusos [21], que se caracterizan por una función de membresía cuyo rango está en el intervalo real [0,1]. Cuando el grado de membresía de un elemento es cercano a 1, se dice que está más posiblemente (o certeramente) incluido en el conjunto. Así 0 es la medida de completa exclusión y 1 la de completa inclusión. De esta forma se puede representar cuando un objeto tiene una pertenencia difusa a un grupo. En la Figura 6 [2] se observa el resultado de realizar *clustering* difuso, que produce dos *clústers* F1 y F2, con datos (4, 6, 7) en la intersección. El problema del agrupamiento difuso es encontrar la caracterización de una partición difusa óptima, en base a una relación de similitud entre los objetos.

Son muchos los casos donde resulta útil aplicar un análisis de agrupamiento difuso. Uno de ellos es el agrupamiento de noticias en la web, donde la clasificación de la naturaleza de una noticia es una partición difusa ya que la misma puede pertenecer a diferentes categorías (deportivas, cultural, económica, social, etc.). Otra aplicación en el medio de los negocios es la segmentación de clientes utilizando la agrupación difusa.

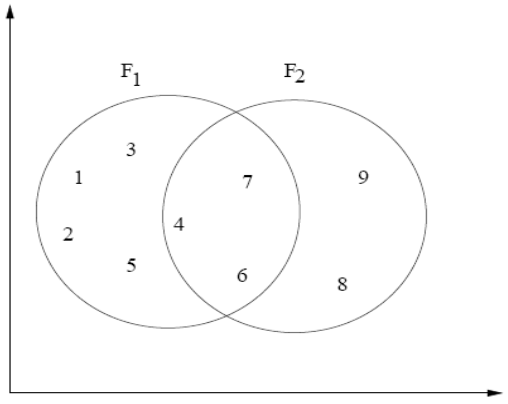

Figura 6: *Clustering* Difuso

D. Algoritmo Fuzzy C-Means

Fuzzy C-Means es un algoritmo de *clustering* difuso de gran difusión, introducido por Ruspini [22], formalizado por Dunn [23] y generalizado por Bezdek [24], cuya idea es obtener particiones difusas óptimas del conjunto de objetos, minimizando una función objetivo que determina los prototipos o centroides de los grupos buscados. Una presentación detallada del *Fuzzy C-Means* y sus versiones se encuentra en Bezdek [25]. Otros algoritmos de clustering difuso pueden encontrarse en [26] así como un análisis comparativo que permite medir el desempeño de éstos sobre diferentes conjuntos de datos. Algunas aplicaciones de *clustering* difuso se describen en [27][28]. Aquí se utilizará una versión general descrita por Hernández [2]. El algoritmo *Fuzzy C-Means* tiene los siguientes pasos [24][28][29]:

1) Dada la matriz de pertenencia $\boldsymbol{\mu}_{n\times k}$ donde un elemento μ_{ij} representa el grado de membresía del objeto i al clúster j, tal que $\mu_{ij} \in [0,1]$, se selecciona una partición difusa inicial de n en k clústers por medio de dicha matriz de pertenencia.

2) Se utiliza $\boldsymbol{\mu}$ para encontrar el valor de la función objetivo de criterio difuso, la cual se explica más adelante. Se reasignan los datos a los clústers para reducir el valor de la función de criterio y se reevalúa $\boldsymbol{\mu}$.

3) Se repite el paso 2 hasta que los valores de $\boldsymbol{\mu}$ no cambien significativamente.

De esta manera, el algoritmo *Fuzzy C-Means* asigna un conjunto de objetos, caracterizados por sus respectivos valores de atributos, a un número *c* determinado de clases (grupos). El resultado del algoritmo *Fuzzy C-Means* se muestra en una tabla donde cada objeto tiene un grado de pertenencia μ_{ij} a cada clase, representada por su centro de clases o grupos construidos, por ello el número de grupos suele ser un parámetro *c* conocido.

Básicamente, el algoritmo *Fuzzy C-Means* requiere de los siguientes parámetros [24][29]:

- Conjunto de Datos: X
- Número de clases o grupos difusos a encontrar: c
- Número de objetos a agrupar: n
- Difusor o grado de difusión: m. Se trata del factor difuso que indica cuánto se quiere que se solapen los grupos. Tiene que cumplirse $m > 1$ ya que la partición se vuelve más difusa conforme se incrementa m y con m=1 la partición dejaría de ser difusa.
- Vector de atributos del objeto j: $y_j, j = 1, \ldots, n$
- Grado de membresía del objeto i a clase j: μ_{ij}, $i = 1,\ldots,n, j = 1,\ldots,c$.

Se han propuesto varios criterios de agrupamiento para obtener la partición difusa óptima. Una de las funciones objetivo de criterio difuso más utilizada es la propuesta por Dunn [23], la cual está asociada con la función de error mínimo cuadrático. Dunn propone minimizar iterativamente la siguiente fórmula $J_m(U,v) = \sum_{k=1}^{n}\sum_{i=1}^{c}(u_{ik})^m d_{ik}^2$. Donde

U: es una matriz de pertenencias, con la *c-partición* difusa de X, contiene el grado de membresía de cada objeto a cada grupo, $U \in M_{fc}$: conjunto de *c-particiones* difusas

$V_i = (v_{i1}, v_{i2}, \ldots, v_{ic})$ es el vector centro del grupo i, es decir, el conjunto de particiones difusas

d_{ik}^2: indica la distancia cuadrada entre los elementos de X= (x_1, x_2, …, x_n), el conjunto de n objetos que es subconjunto del espacio euclidiano de dimensión s con $X \in \mathbb{R}^s$, y los centros de los grupos, en decir distancia cuadrática entre el objeto k el centro V del *clúster* i, calculados con: $d_{ik}^2 = ||x_k - v_i||_A^2 = (x_k - v_i)^T A(x_k - v_i)$ siendo $\| \cdots \|$ la norma inducida por A.

La idea es buscar la partición que produzca la distancia mínima de los objetos al centro de su grupo. Esta distancia está

ponderada por el grado de membresía de cada objeto a un *clúster* y por el factor difuso *m* que indica cuánto se quiere que se solapen los grupos. Para minimizar esta función, se utilizará la propuesta de Bezdek [24], la cual intenta minimizarla de manera iterativa usando el siguiente teorema: "una partición difusa puede ser un mínimo local de la función objetivo *J*, para $m > 1$", cuando se cumplen las siguientes condiciones:

$$u_{ik} = \frac{1}{\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{jk}}\right)^{2/(m-1)}} \text{ y } u_i = \frac{\sum_{k=1}^{n}(u_{ik})^m x_k}{\sum_{k=1}^{n}(u_{ik})^m}$$

III. Modelación Conceptual de Clustering Difuso en Espacios Multidimensionales

En base a los conceptos teóricos mencionados, se presenta a continuación el perfil UML propuesto para modelar *clustering* difuso multidimensional y sus restricciones en OCL.

A. Perfil UML para Clustering Difuso

Con el objetivo de facilitar el análisis de agrupamiento como técnica de minería de datos en un proceso KDD, se propuso en [14] la integración de esta técnica con almacenes de datos, a través de una extensión de *profile* UML cuyo fin es el modelado conceptual de minería de datos con *clustering* en espacios multidimensional. En esta extensión se definen cuatro estereotipos: «clustering» que representa una generalidad del algoritmo con sus parámetros; «entrada» que son los atributos de entrada a la técnica de *clustering* que referencia datos a través de los hechos; «caso» que son los atributos utilizados como caso; y «atributo abstracto» que son aquellos atributos de minería de datos que hacen referencia a los datos multidimensionales. Los ajustes al *clustering* se toman de los parámetros del algoritmo utilizado, por lo que para el caso de *clustering* difuso, se usan los del algoritmo *Fuzzy C-Means*.

En la Figura 7, una adaptación de la propuesta de Zubcoff [14], se muestra el perfil UML para minería de datos con *clustering* difuso, donde en la parte izquierda se observa un extracto de la especificación del perfil UML para modelación multidimensional de DW, propuesto en [18]. Allí se definen las cajas etiquetadas correspondientes a los estereotipos («stereotype») y a las metaclases («metaclass»). Los conceptos del modelo multidimensional (hechos, dimensiones y jerarquías de agregación) son traducidos a la metaclase UML **Class** con los estereotipos **Fact**, **Dimension** y **Base**. Asimismo, los datos multidimensionales: como las medidas (estereotipo **FactAttribute**), las descripciones de los niveles de jerarquía (estereotipo **DimensionAttribute**) y los identificadores de los objetos (estereotipo **OID**). Estos elementos se traducen a la metaclase UML **Property** que típicamente modela atributos de otras metaclases.

En la parte derecha de la Figura 7 se muestra el extracto de la especificación del perfil UML para *clustering* difuso. Aquí se observa el estereotipo **Clustering** que representa la generalización de los algoritmos de *clustering* difuso, que se definen extendiendo la metaclase UML **InstanceSpecification**. La clase **Ajustes** modela los parámetros de los algoritmos de *clustering* difuso, indicando para cada parámetro el dominio y el valor por defecto. Los estereotipos **Entrada**, **Caso** y el **Atributo Abstracto** son tomados de [14] con la misma interpretación, donde la etiqueta referencia permite enlazar con los datos asociados el modelo multidimensional.

B. Restricciones OCL

En cuanto a las restricciones OCL propuestas en [14] para enriquecer la semántica que no puede ser totalmente expresada por el perfil, éstas no pierden vigencia en esta propuesta, por lo que serán reutilizadas también. Entre ellas se destacan las condiciones necesarias para completar el perfil UML, a fin de resolver ambigüedades del dominio de *clustering*: "considerar al menos una entrada para *clustering*", "los parámetros ajustan el *clustering*", "las entradas referencian datos multidimensionales", "los atributos caso pueden referenciar solamente datos multidimensionales descriptivos", "el número de atributos de entrada está limitado en *clustering*". El detalle de estas restricciones puede verse en [14].

Para las restricciones propias de los requisitos de *clustering* difuso se utilizará la extensión de OCL propuesta en [9] que permite incluir términos vagos en las expresiones que representan la semántica formal de tales requisitos. Los términos de la lógica difusa que permite esta extensión abarcan: predicados, modificadores, comparadores, conectores y cuantificadores difusos.

IV. Caso de Estudio

Con el propósito de mostrar la aplicabilidad y viabilidad de la propuesta presentada, a continuación se expone un caso de estudio genérico y simplificado. Club Mercado es un proyecto de desarrollo de una aplicación Web para la compra y *marketing* de productos en línea. Se basa en el consumo colaborativo proveyendo a los clientes de búsquedas para

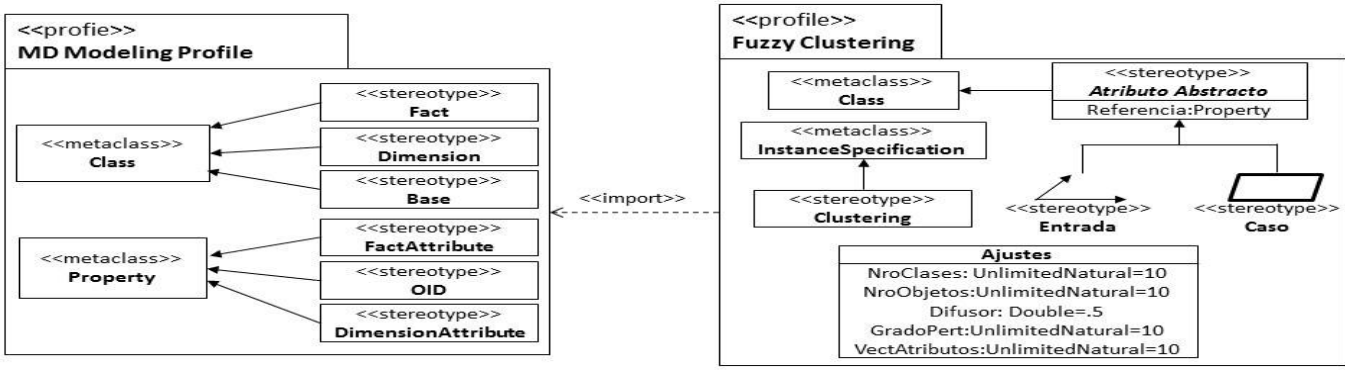

Figura 7: Perfil *Clustering* Difuso

adquirir productos a mejores precios y de superior calidad obtenidos directamente de los productores o distribuidores. La idea es que el sistema sea un apoyo para segmentar los clientes en grupos que tengan conductas similares en cuanto a hábitos de compras, y para agrupar los productos por categorías de mejores precios y/o mejor calidad. De manera especial se espera que el sistema realice el análisis de segmentación de los clientes de forma tal que entregue un detallado conocimiento del perfil de cada cliente a fin que la recomendación de los productos a través del sistema sea consistente con las necesidades y preferencias de éste. El perfil del cliente ha sido segmentado en diferentes grupos, con sus respectivos grados de pertenencia, respetando con mayor fidelidad los intereses reflejados por dichos clientes durante sus visitas y transacciones de compras realizadas históricamente en la aplicación web. Se desea que el sistema sea accesible por medio de dispositivos móviles con una base de datos que permita reducir el uso del internet y en algunos casos prescindir de éste. Las funcionales a proveer son: el registro de usuarios, manejo de carrito de compras, estados de cuenta de las compras realizadas, oferta y recomendación de productos basado en el perfil de pertenencia de cada cliente a través de diversos medios (en comerciales, anuncios de prensa etc.), consultas de productos según preferencias usuario, estadísticas de productos más vendidos a fin de hacer descuentos a los clientes, entre otros.

La interfaz para el registro de usuario solicita los siguientes campos: nombre, apellido, cédula de identidad, edad, número celular, correo electrónico, contraseña, confirmación de contraseña, dirección del usuario, estado, ciudad y municipio a la cual pertenece la dirección especificada. Los datos más relevantes para los productos, incluyen su código, descripción, precio, cantidad, categoría, fechas de inicio y cierre de anuncio publicitario, así como, fotos alusivas. El pago de las compras de los clientes se realizará a través de tarjetas de crédito. Todos los datos son conglomerados en un Almacén de Datos (AD) cuyo modelo multidimensional (MD) se presenta en la Figura 8. Es de notar que el análisis de este caso de estudio está enfocado en un nivel conceptual y de modelamiento, por ellos no se incluyen detalles relacionados a los valores de estos datos (tamaño, dimensionalidad, entre otros). Es decir, el análisis cuantitativo del modelo se escapa de los objetivos del presente trabajo, sin embargo, esta información no afecta la aplicabilidad a nivel conceptual del perfil UML propuesto.

El diagrama presentado en la Figura 8, se obtuvo utilizando el perfil UML para modelado MD [14] descrito en la sección 3.1. Una compra es un hecho de análisis (estereotipado como «fact»), cuya clase se ha identificado como Transacciones, la cual contiene el atributo cantidad de la compra (etiquetado como «FA»). Para el análisis del contexto, se presentan tres instancias de la clase «dimensión»: **Fecha**, **Titular** y **Productos**. Estas dimensiones agregan (flechas con punta de diamante) información a través de jerarquías de agrupación a la compra. Cada nivel de granularidad se indica con la etiqueta «Base». Además, cada nivel de agregación tiene atributos descriptivos. En el caso de la jerarquía definida por la dimensión **Titular** de la tarjeta, con tres niveles: «Base» Usuarios, «Base» Tipo Tarjeta y «Base» Ingresos.

En el primer nivel se observa la cédula de la persona como el identificador de objeto (etiquetado como «OID») y los atributos de la dimensión (etiquetados como «DA»): Nombre, Edad, Teléfono Celular, Sexo, Correo, Contraseña, Confirmación de la Contraseña, Dirección, Usuario, Estado, Ciudad, Municipio. En esta dimensión hay dos jerarquías de agregación que comparten el mismo nivel de granularidad, «Base» Tipo Tarjeta

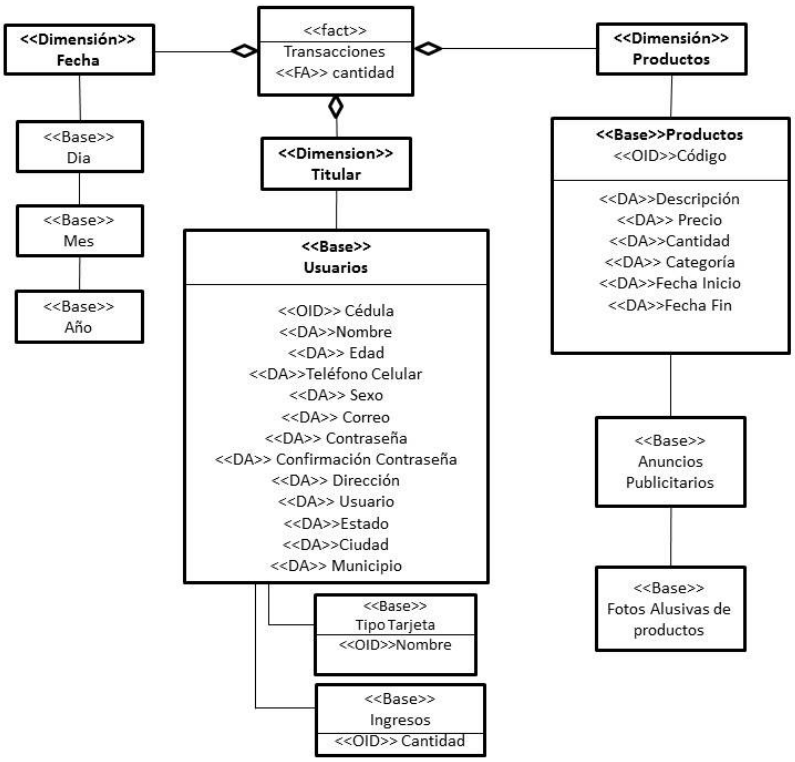

Figura 8: Modelo Multidimensional para Club Mercado

y «Base» Ingresos. Estos niveles de granularidad permiten a los analistas trabajar de manera intuitiva a distintos niveles de detalle, desde las etapas tempranas del desarrollo asegurando la calidad de los datos. En el nivel «Base» Tipo Tarjeta se tiene como identificador de objeto («OID») el nombre del titular de la tarjeta y en el nivel «Base» Ingresos a la cantidad («OID»). En estos niveles no hay más atributos porque corresponden a los datos de entrada necesarios para construir el modelo multidimensional en el cual se miden las similitudes existentes en los datos observados.

Para el análisis del contexto en el caso de la dimensiones **Productos** y **Fecha** de la compra se modelan cada uno con tres niveles de jerarquía. En la dimensión **Productos** se tienen los niveles «Base» Productos, «Base» Anuncios Publicitarios y «Base» Fotos alusivas a productos. En la dimensión **Fecha** los tres niveles de jerarquía son «Base» día, «Base» mes y «Base» año, con la finalidad de permitir diferentes consultas necesarias por los atributos descriptivos.

Para abordar el requisito de segmentación difusa de los clientes almacenados en el MD, se modeló un requisito de análisis de agrupamiento difuso sobre los clientes, en donde el objetivo de la analítica se resume en obtener una estructura de partición (agrupación) difusa (no disjunta) de todos los clientes potenciales del Club Mercado. Se quiere que la partición represente los criterios de preferencias de los clientes demostrada en los históricos de compras por tarjetas de crédito, y así construir un modelo de conocimiento que describa el perfil de *marketing* útil para especializar el sistema recomendador y de ofertas de productos a las necesidades específicas de cada cliente. Es importante resaltar que este modelo no resultaría tan real si se representa con una partición no difusa dentro de un contexto de clientes con diversidad de preferencias, poco sesgadas, donde sus intereses están solapados entre las categorías de la partición.

En este análisis de segmentación difusa de clientes, al aplicar el algoritmo *Fuzzy C-Means*, los clientes son el conjunto de objetos a particionar, caracterizados con los valores de los atributos descritos en la dimensión Titular, analizando las Transacciones que dichos Titulares han realizado. El modelo de agrupamiento difuso obtenido se observa en la Figura 9, que usa la propuesta de Rodríguez y Goncalves [9].

Para esto se usa una instancia de la clase **Ajustes** del perfil de agrupamiento difuso, estereotipada como «clustering». Para este modelo de agrupamiento, se han ajustado los valores de los parámetros como sigue: para NroClases se aplicará el proceso para valores de c desde c=2 hasta y c=10, El NroObjetos indica el número de clientes que en este caso son 25 millones, el factor difusor (mínSoporte) Difusor=10 se ajustó alto para garantizar una partición más difusa, manteniendo el resto de los parámetros con sus valores por defecto, en vista que variar estos valores no aportan diferencias al modelamiento. Las flechas punteadas indican que los atributos son tipos de datos dependientes y la etiqueta *use* que son datos de entrada.

Durante el proceso de *clustering* difuso sobre las transacciones de compras con tarjetas de créditos se utiliza la cantidad comprada como un atributo de entrada, lo cual se indica con la etiqueta [Referencia=Compra:Cantidad]. Los atributos del usuario (como cedula, nombre, edad, sexo, dirección, estados, teléfono, municipio, ciudad, tipo de tarjeta, ingresos, contraseñas, correo, cantidad), aparecen con la etiqueta *use* indicando que son datos de entrada.

En cuanto a los valores de referencia como es el caso de los atributos Tipo de Tarjeta, Cantidad, e Ingresos, corresponden a los atributos de entrada que permitirán la partición difusa de los datos del esquema multidimensional asociados a los clientes basado en el análisis de sus transacciones.

Al aplicar el algoritmo *fuzzy c-means*, se determinan los parámetros de este algoritmo utilizando alguna de las técnicas heurísticas más usadas sobre la base del estudio de casos, para luego modificar el número de agrupaciones. En este caso práctico se propone aplicar el algoritmo *fuzzy c-means* para cada número de agrupaciones entre c = 2 y c = 10. De tal manera que los valores de pertenencia de todos los clientes a

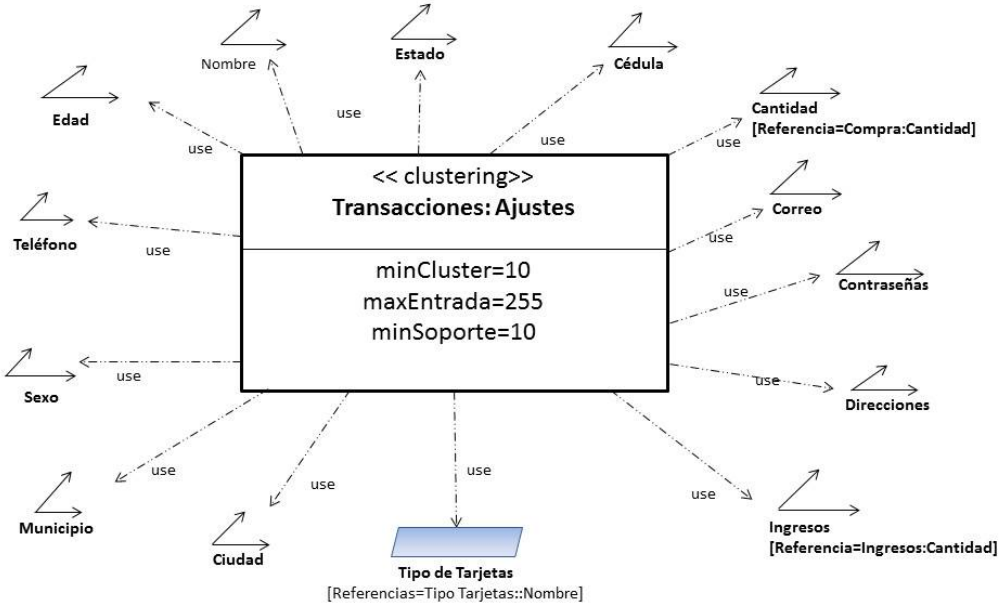

Figura 9: Modelo *Clustering* Difuso MD Compras TC

las agrupaciones 1 a c son calculados y presentados. El algoritmo finalmente muestra una matriz con el centro de todas las agrupaciones (clases) c.

De esta manera el análisis de partición difusa proporciona grados de pertenencia de los objetos a cada una de las agrupaciones. En este sentido, en la segmentación de clientes se discrimina de manera más diferenciada cada cliente en cada clase, en vista que el algoritmo calcula para cada cliente los valores de pertenencia en cada clase. De esta manera los clientes que muestran valores característicos de diferentes segmentos se tratarán con actividades de *marketing* especializadas de acuerdo a sus valores de pertenencia. Como resultado, el recomendador de *marketing* estará en capacidad de ofrecer a los clientes del sistema productos más ajustados a sus intereses, preferencias y necesidades.

Se puede observar que el perfil UML propuesto para *clustering* difuso simplifica el proceso del analista en la labor de modelado, abstrayéndolo hacia niveles altos del proceso de partición difusa separándole de los detalles de implementación que son mucho más complejos. De esta forma, resulta más sencillo configurar el requisito de agrupamiento difuso y modelarlo a través del perfil propuesto, el cual aprovecha las ventajas ofrecidas por el MD. También se facilita el proceso de realizar un análisis de agrupamiento difuso, a través de una notación más intuitiva que se abstrae e independiza de los complejos detalles de su implementación.

V. Conclusiones y Trabajos Futuros

En esta contribución se han propuesto un perfil UML para diseño conceptual de *clustering* difuso en el tope del modelo multidimensional de un almacén de datos. Este ha sido trasladado a una extensión del paquete Profile de UML 2.0, mecanismo de extensión ligera de UML.

Esta propuesta permite diseñar modelos de *clustering* difuso en espacio multidimensional articulado por un almacén de datos, al nivel de abstracción apropiado para concentrarse exclusivamente en los principales conceptos del *clustering* y aprovechando toda la información y el conocimiento del dominio bajo estudio capturado en el modelo multidimensional del almacén de datos. Este modelo facilita la abstracción, flexibilidad y reusabilidad, provee a los usuarios la semántica requerida para la comprensión del sistema modelado, simplificando el diseño de la minería de datos con técnicas difusas de *clustering* en una notación intuitiva. La principal ventaja de esta propuesta es que permite a los analistas llevar a cabo el proceso de KDD estableciendo los objetivos empresariales desde etapas tempranas del desarrollo del proyecto, asegurando así la calidad de los datos.

Los aportes principales de esta propuesta son:

- Facilita el diseño del proceso de minería de datos gracias al uso de modelos conceptuales considerando los objetivos empresariales desde etapas tempranas del proyecto de KDD.
- Incorporar a tempranas etapas del proceso de análisis, construcciones con una notación especializada para modelar la semántica relacionada con requisitos de agrupamiento difuso.
- Provee un Modelo Conceptual para *clustering* difuso, independiente de herramientas y algoritmos específicos, incorporando una nueva notación y semántica para símbolos ya existentes en UML.
- Proporciona una sintaxis y terminología común para el dominio de aplicaciones de agrupamiento difuso, cuyas construcciones actualmente no cuentan con una notación propia.
- Facilita el diseño de minería de datos en espacios multidimensionales.
- Aprovecha las ventajas derivadas de los pasos previos del DW, asegurando la calidad de los datos al integrar los DW al proceso global de KDD con requisitos de *clustering* difuso.
- Provee un camino para el modelado de software de minería de datos difusas guiado por arquitecturas (MDA, *Model Driven Architecture*), mediante la definición y transformación de modelos para este dominio de aplicación de uso y relevancia en la actualidad.

A partir de esta propuesta quedan caminos abiertos, de los cuales se quiere explorar en trabajos futuros, los siguientes:

1. Aplicación completa del perfil UML propuesto en el caso de estudio genérico, culminando el proceso de KDD utilizando datos reales, con el fin de ofrecer los resultados del análisis de segmentación difuso obtenido de manera cuantificada, ofreciendo una comparativa con los resultados alcanzados en la aplicación del caso homólogo preciso.
2. Proponer una metodología para tratamiento de requisitos difusos con técnicas de minería de datos.
3. Extender UML a un perfil de modelación de KDD con diversas técnicas de minería de datos difusa, visto como proceso integrado y como tratamiento de requisitos difusos.
4. Aplicación de la propuesta de extensión de perfiles UML para *clustering* difuso, en diferentes casos de estudio de interés real, tales como: segmentación de pozos petroleros sobre un DW de variables de producción y explotación de la Industria Petrolera Venezolana, segmentación de perfiles de estudiantes que han desertado del sistema educativo formal venezolano, así como segmentación de pacientes con enfermedades metabólicas, proclives a desarrollar enfermedades crónicas y degenerativas. Esto permitirá validar si esta propuesta es un modelo de diseño conceptual novedoso para *clustering* difuso sobre DW.

Proponer un escenario de transformaciones, de los perfiles UML propuestos para minería de datos difusa, que puedan ser automatizadas, de tal manera de realizar un aporte al modelado de software de minería de datos difusa guiado por arquitecturas (MDA, *Model Driven Architecture*).

Agradecimientos

Agradecemos Aquél que nos da fe y valor para emprender proyectos hacia lo desconocido: “Por la fe Abraham, siendo llamado, obedeció para salir al lugar que había de recibir como herencia; y salió sin saber a dónde iba” (Hebreos 11:8).

REFERENCIAS

[1] P. Tan and M. Steinbach, V. Kumar. *Introduction to Data Mining*. Addison Wesley, USA, 2006.

[2] E. Hernández. *Algoritmo de Clustering basado en Entropía para Descubrir Grupos en Atributos de Tipo Mixto*. Tesis para obtener el grado de Maestro en Ciencias en la Especialidad de Ingeniería Eléctrica Opción Computación. Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, México, D.F. 2006. https://www.cs.cinvestav.mx/TesisGraduados/2006/tesisEdnaHernandez.pdf.

[3] U. Fayyad. *Data Mining and Knowledge Discovery: Making Sense out of Data*. IEEE Expert, vol. 11, no. 5, pp. 20-25, October 1996.

[4] O. Moscoso-Zea, A. Sampedro, and S. Luján-Mora. *Datawarehouse Design for Educational Data Mining*. 2016 15th International Conference on Information Technology Based Higher Education and Training (ITHET), pp. 1-6, Istanbul, Turkey, September 2016.

[5] R. Agrawal, A. Gupta, and S. Sarawagi. *Modeling Multidimensional Databases*. Technical Report. IBM. IBM Almoden Research Center. 1995. https://infolab.usc.edu/csci599/Fall2002/paper/I3_agrawal95modeling.pdf.

[6] L. Cabibbo and R. Torlone. *A Logical Approach to Multidimensional Databases*. EDBT ´98 Proceedings of the 6th International Conference on Extending Database Technology, vol. 1, pp. 183-197, Valencia, España, March 1998.

[7] A. Datta and H. Thomas. *The Cube Data Model: a Conceptual Model and Algebra for On-line Analytical Processing in Data Warehouses*. Decision Support Systems, vol. 27, no. 3, pp. 289-301. December 1999.

[8] A. Gosain, S. Sabharwal, and S. Nagpal. *Predicting Quality of Data Warehouse using Fuzzy Logic*. International Journal of Business and Systems Research (IJBSR), vol. 6, no. 3, pp. 255-268. January 2012.

[9] R. Rodríguez y M. Goncalves. *Perfil UML para el Modelado Visual de Requisitos Difusos*. Enl@ce: Revista Venezolana de Información, Tecnología y Conocimiento, vol. 6, no. 3, pp. 29-46, Septiembre 2009.

[10] R. Rodríguez y L. Tineo. *Elementos Gramaticales y Características que Determinan Aplicaciones con Requerimientos Difusos*. Revista Tekhne, vol. 12, pp.50-64, Enero 2009.

[11] R. Rodríguez y M. Goncalves. *Implementación de Requisitos en Sistemas Orientados a Datos con Lenguaje OCL y Lógica Difusa*. Enl@ce Revista Venezolana de Información, Tecnología y Conocimiento, vol. 8, no. 1, pp. 31-54, Enero 2011.

[12] W. Pereira y L. Tineo. *Modelo Orientado a Objetos Difuso*. Acta Científica Venezolana, vol. 51, no. 2, pp. 357, Noviembre 2000.

[13] G. Booch, J. Rumbaugh, and I. Jacobson. *The Unified Modeling Language User Guide*. Addison Wesley. USA. 2005.

[14] J. Zubcoff, J. Pardillo, and J. Trujillo. *Integrating Clustering Data Mining into the Multidimensional Modeling of Data Warehouses with UML Profiles*. In Data Warehousing and Knowledge Discovery. DaWaK 2007. Lecture Notes in Computer Science, 4654:199-208. Springer, Berlin, Heidelberg. Septiembre 2007.

[15] ISO/IEC. *Unified Modeling Language (UML). Version 1.5*. International Standard ISO/IEC 19501.

[16] L. Fuentes y A. Vallecillo. *Una Introducción a los Perfiles UML*. Novática: Revista de la Asociación de Técnicos de Informática, ISSN 0211-2124, no.168, pp. 6-11, Enero 2004.

[17] Object Management Group. *Object Constraint Language Specification, version 2.0*. http://www.omg.org/technology/documents/ formal/ocl.htm.

[18] S. Lujan-Mora, J. Trujillo, and I. Song. *A UML Profile for Multidimensional Modeling in Data Warehouses*. Data & Knowledge Engineering, vol. 59, no. 3, pp. 725-769. December 2006.

[19] S. Mitra and T. Acharya. *Data Mining: Multimedia, Soft Computing and Bioinformatics*. Wiley-InterScience, John Wiley & Sons, Inc., Publication, USA, 2003.

[20] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc, Elseiver, UK, 2006.

[21] L. Zadeh. *Fuzzy Sets. Information and Control*, vol. 8, no. 3, pp. 338-353, 1965.

[22] E. H. Ruspini. *Numerical Methods for Fuzzy Clustering*. Information Sciences, vol. 2, no. 3, pp. 319-350, July 1970.

[23] J. C. Dunn. *A Fuzzy Relative of the ISODATA Process and its Use in Detecting Compact Well-Separated Clusters*. Journal of Cybernetics and Systems, vol. 3, no. 3, pp. 32-57, September 1973.

[24] J. Bezdek, R. Ehrlich, and W. Full. *FCM: The Fuzzy c-Means Clustering Algorithm*. Computers & Geosciences, vol. 10, no. 2-3, pp. 191-203, 1984.

[25] J. Bezdek. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Kluwer Academic Publishers Norwell, MA, USA. 1981.

[26] A. Gosaina and S. Dahiya. *Performance Analysis of Various Fuzzy Clustering Algorithms: A Review*. In Proceedings of the International Conference on Communication, Computing and Virtualization (ICCCV), vol 79. pp. 100-111, Procedia Computer Science. Elseiver. Mumbai, India, Febrary 2016.

[27] W. Meier, R. Weber, and H. Zimmermann. *Fuzzy Data Analysis - Methods and Industrial Applications*. Fuzzy Sets and Systems, vol. 61, no. 1, pp.19-28, January 1994.

[28] J. Strackeljan and R. Weber. *Quality Control and Maintenance*. In: Practical Applications of Fuzzy Technologies. The Handbooks of Fuzzy Sets Series, vol 6. Springer, Boston, MA, pp. 161-184. 1999.

[29] J. Bezdek, M. Pal, J. Keller, and R. Krisnapuram. *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*. Kluwer Academic Publishers Norwell, MA, USA, 1999.