05

Fecha de presentación: septiembre, 2016
Fecha de aceptación: noviembre, 2016
Fecha de publicación: Diciembre, 2016

ADECUACIÓN A METODOLOGÍA

DE MINERÍA DE DATOS PARA APLICAR A PROBLEMAS NO SUPERVISADOS TIPO ATRIBUTO-VALOR

ADAPTATION TO A METHODOLOGY OF DATA MINING FOR APPLYING TO UN-SUPERVISED PROBLEMS TYPE ATTRIBUTE-VALUE

Lic. Ciro Rodríguez León[1]
E-mail:crleon@ucf.edu.cu
Dra. C María Matilde García Lorenzo[2]
E-mail: mmgarcía@uclv.edu.cu
[1]Universidad de Cienfuegos. Cuba.
[2]Universidad Central de Las Villas. Santa Clara. Cuba.

¿Cómo referenciar este artículo?

Rodríguez León, C., &García Lorenzo, M. M. (2016).Adecuación a metodología de minería de datos para aplicar a problemas no supervisados tipo atributo-valor. Universidad y Sociedad [seriada en línea], 8 (4). pp. 43-53. Recuperado de http://rus.ucf.edu.cu/

RESUMEN

Debido a que la cantidad de datos almacenados, de todo tipo, van en aumento exponencial, existe la necesidad de tener mecanismos eficientes para manipularlos y extraer conocimientos de ellos. La minería de datos es de las principales encargadas de este tipo de proceso y para hacer menos complejos sus procedimientos se han diseñado metodologías que los guíen. Debido a que estas metodologías son de propósito general en ellas no se describen cuestiones importantes como técnicas y algoritmos a usar en cada etapa. En la presente investigación, luego de un estudio comparativo, se escoge la metodología CRISP-DM para realizar su adecuación a problemas no supervisados tipo atributo-valor. De esta forma, reduciendo el dominio de aplicación, se logra llegar a un nivel de especificación más profundo en cada una de las seis fases que son propuestas originalmente, se ahorra así tiempo a los especialistas que se propongan realizar este tipo de actividad. Para demostrar el uso de esta adecuación y sus resultados acertados, es aplicada a un caso de estudio real, consistente en un grupo de pacientes diabéticos tipo 2, se obtienen resultados satisfactorios luego de hacer un análisis independiente por sexo. Los grupos encontrados representan diferentes niveles de riesgo en la evolución de la enfermedad, los que mejoran su proceso de prevención y diagnóstico.

Palabras clave: Minería de datos, CRISP-DM, agrupamiento, índices de validación, diabetes.

ABSTRACT

The amount of any kind of stored data is going in an exponential increment. That is why it is needed to create efficient procedures to manipulate this data and extract knowledge from them. Data mining is in charge of this type of process and to make their procedures less complex. Methodologies have been designed to guide them. As these methodologies are general they do not describe important issues as techniques and algorithms to be used in each period. In the present research, after a comparative study, CRISP-DM methodology is selected to be adapted to un-supervised problems type attribute-value. In this way, by reducing the application domain, it is achieved a deeper specification level in each of the six phases which were originally proposed, so time of specialists with the purpose of doing this kind of activity, is saved. To demonstrate the use of this adaptation and its successful results, it is applied to a real case study, consisting in a group of type 2 diabetic patients in which satisfactory results are achieved after an independent analysis by sex. The groups found represent different levels of risk factors in the disease evolution who improve their prevention process and diagnosis.

Keywords: Data mining, CRISP-DM, clustering, validation index, diabetes.

INTRODUCCIÓN

La cantidad de información continúa creciendo; sin embargo, la habilidad de los humanos para procesarla y asimilarla permanece constante. Además, la información en sí misma tiene pocas ventajas, su sistematización, incorporación y utilización son los elementos que aportan su valor añadido: el conocimiento. Es necesario crear sistemas que generen conocimiento, para asegurar el uso productivo de la información y guiar una toma de decisiones óptima.

Es obvia la incapacidad del hombre de procesar y extraer nueva información de grandes cantidades de datos, por lo que surge un importante campo la Minería de Datos. La Minería de Datos (MD) o KDD (***KnowledgeDiscovery in Databases***) como se le comienza a llamar a inicios del año 1996, se define como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos (Gorunescu, 2011).

Los procesos de MD tienen muchas veces implícitos técnicas de aprendizaje que, de acuerdo con la definición dada por Michalski en 1986, es la habilidad de adquirir nuevo conocimiento, desarrollar habilidades para analizar y evaluar problemas mediante métodos y técnicas, así como también por medio de la experiencia propia; se requiere del aprendizaje entendible para un hombre.

Dependiendo del esfuerzo requerido por el aprendiz (o número de inferencias que necesita sobre la información que tiene disponible) han sido identificadas varias estrategias. Las más estudiadas y conocidas de estás clasificaciones son (Russell & Novig, 2009): aprendizaje por instrucción, aprendizaje por deducción, aprendizaje por inducción. Del último hay dos tipos principales: aprendizaje con ejemplos o supervisado y aprendizaje por observación y descubrimiento o no supervisado.

Se entiende por aprendizaje supervisado: cuando un algoritmo produce una función que establece una correspondencia entre las entradas y las salidas deseadas del sistema. Por tanto, las bases de estos sistemas están formadas por las características (rasgos) de los ejemplos y la clasificación (clases o categoría) a la que estos pertenecen.

Por otro lado, aprendizaje no supervisado: es cuando todo el proceso de modelado se lleva a cabo sobre un conjunto de ejemplos formado tan solo por entradas al sistema, es decir, sus rasgos. No se tiene información sobre las categorías o clasificación de esos ejemplos. Constituye un tipo de aprendizaje por observación y descubrimiento, donde el sistema de aprendizaje analiza una serie de entidades y determina que algunas tienen características comunes, por lo que pueden ser agrupadas formando un concepto. Luego se pueden utilizar los datos ya clasificados o etiquetados para confeccionar herramientas o modelos que sin esta catalogación no tienen sentido. O bien la propia división en clases de los ejemplos puede representar un resultado sustancial, debido a que pueden llevar a conclusiones a especialistas en el tema.

El proceso de MD involucra numerosos pasos e incluye muchas decisiones que deben ser tomadas por el usuario. Entre ellas, se dice que la adecuación de los datos para la utilización de las técnicas de descubrimiento demanda el 70% del esfuerzo. Para organizar este proceso han surgido varias metodologías o procedimientos que lo guían. Dentro de las más usadas se encuentran: Proceso KDD, CRISP-DM y SEMMA.

En el presente trabajo, luego de un estudio comparativo entre las metodologías estudiadas, se escoge CRISP-DM para adecuarla para la categorización de conjuntos de datos no supervisados tipo atributo – valor, debido a que las metodologías y procedimientos de MD de la literatura estudiados son todas de propósito general; no contienen en ellas especificaciones como las técnicas y algoritmos a utilizar en cada problema específico y esto hace compleja su utilización.

DESARROLLO

Para tomar la decisión de cual metodología de MD utilizar para adecuarla a solucionar los problemas enunciados en el presente trabajo, se describen y comparan a continuación tres de estas. Primeramente el proceso KDD (Fayyad, Piatetsky-Shapiro & Smyth, 1996), este es iterativo e interactivo. Se dice iterativo ya que la salida de algunas fases puede retornar a pasos anteriores y porque a menudo son necesarias varias iteraciones para extraer conocimiento de alta calidad. Se habla de interactivo porque el usuario, o generalmente el experto del dominio del problema, deben ayudar en la preparación de los datos, validación del conocimiento extraído, entre otros.

El proceso se organiza en torno a cinco fases, en la primera, ***selección***, es donde se determinan las fuentes de información que pueden ser útiles y donde conseguirlas. Dado que los datos provienen de diferentes fuentes, pueden contener valores erróneos o faltantes. Estas situaciones se tratan en la fase de ***pre-procesamiento***, en la que se eliminan o corrigen los datos incorrectos y se decide la estrategia a seguir con los datos incompletos. En la siguiente, , se realizan trasformaciones a los datos usando métodos de transformación o reducción de dimensiones.

En la fase ***minería de datos*** se decide cuál es la tarea a realizar (clasificar, agrupar, etc.) y se elige el método que se va a utilizar para buscar los patrones de interés. En la fase de ***evaluación e interpretación*** se valoran los patrones y se analizan por los expertos, y si es necesario se vuelve a las fases anteriores para una nueva iteración. Esto incluye resolver posibles conflictos con el conocimiento que se disponía anteriormente.

Por otro lado la metodología CRISP-DM (***Cross-Industry Estándar Processfor Data Mining:*** Procedimiento Industrial Estándar para realizar Minería de Datos), es creada en el 2000 por el grupo de empresas SPSS, NCR y Daimer Chrysler. Es de distribución libre lo que le permite estar en constante desarrollo por la comunidad internacional. Además resulta independiente de la herramienta que se utilice para llevar a cabo el proceso de MD. Es ampliamente usado por los miembros de la industria. El modelo consiste en seis fases definidas de manera cíclica(Chapman et al., 2000): análisis del problema, comprensión de datos, preparación de datos, modelado, evaluación y despliegue.

Las seis fases no son rígidas en procedimiento, de hecho muchas veces existe retroalimentación entre diferentes fases. Depende en gran medida la salida de una fase para saber que cual etapa o tarea de la etapa se va a seguir a continuación. Además analistas experimentados no necesitan aplicar cada una de las etapas al problema en cuestión.

Por último se analiza SEMMA que, comenzando con una representación estadística de los datos, pretende facilitar la exploración estadística, las técnicas de visualización, seleccionar y transformar las variables más significativas en la predicción, modelar las variables para predecir salidas y finalmente confirmar la precisión del modelo. Evaluando la salida de cada etapa en este proceso se puede determinar cómo modelar nuevas interrogantes levantadas por los resultados anteriores. Por eso, puede procederse a la fase de exploración para un refinamiento adicional de los datos. Sus fases son: muestro, exploración, manipulación, modelación y evaluación(Olson & Delen, 2008).

Al hacer un análisis del uso, a nivel mundial, de las metodologías de MD se encuentra con que CRISP-DM es la más seguida y referenciada de todas. Esto se corrobora por una encuesta realizada por el portal para análisis de datos KDnuggets en el 2014 (“Data Mining Community’s Top Resource,” 2014). Compara particularmente las etapas del proceso KDD con las de SEMMA se puede, inicialmente, afirmar que son equivalentes y se examina más profundamente. Las cinco etapas de SEMMA pueden ser vistas como una implementación práctica de las cinco fases del proceso KDD, lo que en este caso está directamente ligado al software de SAS Enterprise Miner(Azevedo & Santos, 2008).

Se puede decir, por último, que CRISP-DM y SEMMA pueden ser vistas como implementaciones del proceso KDD. Con un primer acercamiento podemos pensar que CRISP-DM es más completa que SEMMA, pero mediante un análisis más profundo se puede ver que las etapas que están presentes en la primera pueden ser vistas como fases implícitas dentro de la metodología SEMMA.

Se concluye, entonces, que todas las metodologías analizadas son de propósito general para realizar procesos de MD a toda la gama de problemas que pueden presentarse, pues no especifican en ninguna de sus fases, etapas o tareas como métodos, algoritmos o técnicas usar en dependencia de las situaciones particulares que se presenten. Sin embargo, se considera a CRISP-DM más oportuna para realizar su adecuación a problemas no supervisados tipo tributo - valor pues es de libre distribución y no es dependiente de la herramienta que se utiliza; además es ampliamente la más usada dentro de la comunidad científica.

Adecuación de la metodología CRISP-DM

Se ve a continuación la adecuación de la metodología CRISP-DM para aplicarla a problemas no supervisados tipo atributo-valor.Esta consta de seis fases: análisis del problema, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue; aunque es importante recordar que las fases no son rígidas y que pueden existir retroalimentación entre ellas.

En Chapman et al. (2000), manual de usuario de la metodología CRISP-DM, se describe esta paso por paso. En ese documento se aprecia que la guía está enfocada a toda la amplia gama de problemas de MD. Por tanto, en cada fase los autores solo describen los procedimientos a realizar de manera muy general, pueden usarlas en problemas supervisados, no supervisados, de minería de texto, etcétera. Se desarrolla a continuación el contenido de cada una de las fases particularizadas para los problemas que se ocupan en la presente investigación, se llega a un grado de especificidad más profundo.

En la fase I, ***análisis del problema***, se trata de comprender el negocio, donde se identifica la expectativa del cliente con el proceso de MD. Se determinan los objetivos y la producción del plan del proyecto. Para determinar los objetivos se debe distinguir los beneficios que se brindaran a la institución, organismo, empresa o cliente de forma general, que desea obtener información útil a partir de

datos no supervisados. Es por ello, que esta parte del proceso es imprescindible el diálogo con los expertos del área del conocimiento en que se encuentra, se ofrece explicación de lo que se necesita para lograr lo que ellos esperan; así como las potencialidades de los resultados a obtener. Es necesario precisar desde el comienzo si se tiene alguna idea de cómo quedan conformados los grupos o su cantidad, por hipótesis de variables que son relevantes, problemas similares que se han solucionado, etcétera; para entonces encausar el proceso hacia ese objetivo. Pero en muchas ocasiones solo se tiene la ***materia prima*** pero no se sabe que puede sacarse de ella, se tienen en este caso metas mucho más amplias.

Es provechoso definir en este momento cómo es utilizado el conjunto de datos una vez realizado todo el proceso y categorizadas las instancias. Si con el solo hecho de haberlos dividido en grupos es ya suficiente como para brindar información relevante o esto, por sí solo, no es suficiente y es necesario utilizarlos para crear algún modelo computacional con técnicas estadísticas o de aprendizaje automático, por ejemplo. También este momento se puede tomar la decisión de que herramientas de MD son utilizadas en todo el proceso.

Durante la fase II, ***comprensión de los datos***, se obtiene una visión más realista del conjunto de casos del que se extrae el conocimiento. En Chapman et al. (2000), se orienta mediante la ejecución de tareas como: la recolección de los datos iniciales, describir y explorar los datos y verificar su calidad. Se analiza a continuación como realizar estas actividades en los problemas no supervisados tipo atributo-valor.

Una vez identificadas las líneas a seguir el primer paso es obtener los datos. Si se encuentren sin digitalizar, hay que transformarlos a este formato, o si se encuentran en algún formato digital pero que no sea en forma de tabla (cada columna un atributo y cada fila un caso) se debe llevar a este estilo.El segundo paso es describir los datos de modo que se pueda identificar de manera general sus características. Es recomendable para ello la creación de una estructura como la mostrada en la tabla 1, donde de conjunto con los especialistas del dominio de aplicación se pueda realizar una primera selección de las variables importantes para el problema en cuestión. Atributos muchas veces presentes como identificadores, nombres propios, direcciones, fechas, etcétera, en la mayoría de los casos carecen de relevancia para el problema. Se dejan solamente las variables que evidentemente pueden influir en los patrones deseados a reconocer. De igual forma el tipo de las variables es importante para el tipo de técnica que se puede aplicar a la base de conocimiento que se está creando y además para saber las transformaciones que pueden hacerse más adelante a estos datos.

Tabla 1. Ejemplo de la descripción de las variables.

Identificador	Descripción	Tipo	Relevancia
Variable 1	Descripción 1	Nominal	Relevante
Variable 2	Descripción 2	Numérica	No relevante
⋮	⋮	⋮	⋮
Variable n	Descripción n	Nominal	Relevante

Durante esta tarea se realizan análisis de visualización mediante representaciones espaciales, se analizan gráficos de dispersión, histogramas, entre otros, de algunas variables. También es esta una vía de comprobar si el conjunto de datos sobre el que se trabaja concuerda con la teoría del dominio y no posee errores. Con el análisis de simples estadísticas como media y moda de las variables declaradas como más relevantes por los especialistas pueden indicar el camino correcto.

Es en la fase III, ***preparación de los datos***, en la que se seleccionan, limpian, construyen, integran y da la forma final a los datos. Los criterios para esta selección incluyen: la importancia para el cumplimiento de los objetivos de la MD, la calidad, y las restricciones técnicas como límites sobre el volumen de datos o los tipos de datos. Este proceso incluye no solo la selección de las variables sino también de los casos.

Primeramente es pertinente realizar un análisis sobre los valores perdidos en la muestra, existen varias opciones y tratamientos. En casos donde en una misma variable haya un porciento muy elevado de estos puede valorarse el eliminar esta variable de la muestra; el mismo tratamiento puede hacerse en el caso de las instancias; esta es llamada: variante de eliminación (***case deletion, CD***). También puede reemplazarse estos por la media, en el caso de variables continuas, o por la moda, en el caso de las nominales; llamada imputación de la media (***Mean Imputation, MI***)(Cios, Pedrycz, Swiniarski & Kurgan, 2007). Muy parecida a esta última es la imputación de la mediana (***Median Imputation, MDI***) se sustituye por la mediana el ausente (Acuna & Rodriguez, 2004). Se hallan además métodos que estiman estos valores perdidos como la imputación de KNN (***KNN Imputation, KNI***) (Batista & Monard, 2002) y el ***EMImputation*** descrito en Schafer (1997).

Es el momento ahora de aplicar filtros que modifiquen la muestra al tratar que la obtenida pueda ser mejor generalizada por los métodos de agrupamiento. Los análisis para trabajar con las instancias pueden incluir buscar en

la muestra valores atípicos y extremos. Estos valores pueden aparecer de cinco formas diferentes:

- Tipo A (Punto extremo): forma más simples.
- Tipo B (Extremo contextual): si una instancia de los datos es una rara ocurrencia respecto a un contexto específico de los datos y es normal respecto a otro.
- Tipo C (Extremo colectivo): es el caso en el que una instancia de los datos individualmente no es anómala pero de conjunto con la totalidad de los datos sí es un extremo.
- Tipo D (Extremo real): estas son las observaciones ruidosas que son de interés en el sistema que se analice. Por tanto, no se deben tratar bajo el concepto de ruido sino el de extremos reales.
- Tipo E: (Extremo de error): es el caso en el que alguna observación es denominada incorrectamente como extremo, debido a algo inherente al problema en cuestión o fallo.

Existen dos partes en el proceso de lidiar con los datos ruidosos: encontrarlos y tratarlos.Para encontrarlos enMalik, Sadawarti & Singh (2014), se describen varias estrategias: basadas en proximidad, paramétricas, no paramétricas, entre otras. Luego para el tratamiento de estos el procedimiento es muy parecido al caso de los valores ausentes, se pueden: ignorar, filtrar (eliminar o reemplazar, las filas o columnas),reemplazar el valor o hacerlo discreto.

Además, los datos a menudo son representados por una gran cantidad de atributos en muchas áreas, el problema que nos ocupa no es una excepción. En la práctica no todos los rasgos son relevantes e importantes para la tarea de aprendizaje, muchos de ellos son a menudo redundantes, correlacionados e incluso ruido en ocasiones, lo que puede traer consigo efectos negativos como sobre entrenamiento, baja eficiencia y desempeños. Existen métodos de selección y extracción de rasgos específicos para los problemas de aprendizaje no supervisados, algunos de ellos son: Laplacian Score(He, Cai & Niyogi, 2005), selección de rasgos multi-grupo (Cai, Zhang, & He, 2010), selección de rasgos discriminativa no supervisada (Yang, Shen, Ma, Huang & Zhou, 2011), Selección de rasgos discriminativa no negativa (Li, Yang, Liu, Zhou & Lu, 2012) y Selección de rasgos robusta no supervisada (Qian & Zhai, 2013).

Es válido también nivelar el comportamiento de las variables estandarizando los datos, en caso de que determinadas variables pueden tener un peso mayor que otras, simplemente porque la unidad de medida en que aparecen dan lugar a puntuaciones con valores relativamente altos en comparación con los de las otras. Pero cuando los rasgos vengan en la misma escala, tales como, puntuaciones de ítem en un cuestionario o porcentajes, no es aconsejable la estandarización. Es apropiado igualmente darle un orden aleatorio a los ejemplos de la muestra pues hay algoritmos de agrupamiento que son sensibles al orden de entrada de los datos.

Se sugiere para el caso de variables nominales cuyos dominios tienen múltiples valores, modificar esta situación codificando valores y creando nuevas variables, siempre que en el problema no se necesiten todos estos valores e incluso pueda oscurecer lo que realmente se desea representar. Por ejemplo, se puede tener una variable nominal con cinco posibles valores, uno es que la característica no está presente y los demás cuatro son diferentes atenuaciones de esa característica cuando está presente; es posible que solo se necesite saber si el rasgo está presente o no. Entonces es el momento de codificar esta variable como una nueva con solo dos valores: presente la cualidad y no presente.

Existen casos donde debe realizarse recodificación de variables continuas, por ejemplo, es posible que varios atributos estén presentes en los datos solo porque son utilizados para calcular un valor que es realmente el centro de atención. Es recomendable entonces calcular ese valor para cada caso, añadirlo como una columna y eliminar las demás menos representativas.

Es muy recomendable que para la siguiente fase pasen varios archivos de datos con cada uno de los filtros que se hayan aplicado e incluyendo en estos uno con todas las variables clasificadas como relevantes; puesto que en ocasiones el filtrado, selección y extracción de casos y variables puede traer consecuencias negativas en lugar de positivas debido a pérdida de información. Experimentando luego con toda la gama de datos se está garantizando una mayor variabilidad en los resultados finales y por ende mayor calidad en la solución encontrada.

La fase IV, ***modelado***, constituye la cúspide en el proceso de descubrir el conocimiento. Es aquí donde los algoritmos ***cavarán*** y se adentran en los casos preparados y extraen el preciado producto: la información. Se cumple esta función al realizar tareas como: seleccionar las técnicas de modelado, generar la prueba de diseño, construir y evaluar el modelo (Chapman et al., 2000).

La selección de los métodos de agrupamiento que se van a utilizar en el proceso de MD es de las partes más difíciles del proceso y de las más determinantes también. Dependen en gran medida de la experiencia del ingeniero del conocimiento y de lo bien que se conozcan los datos; por ello si hasta el momento no se tiene idea alguna de

cómo se comportan estos últimos, los análisis como graficado, análisis de distribuciones estadísticas y de cuáles variables son más influyentes son recomendables. Es válido aclarar que no hay una forma determinista de elegir las técnicas a utilizar, la variedad en la experimentación y la correcta evaluación es la que dice la última palabra. No obstante en esta sección se dan algunas sugerencias de cómo escoger estos métodos:

- Si no se tiene idea de la cantidad de grupos que se quieren obtener es recomendable utilizar las técnicas que los estiman como: ***mountainclustering***(Yager & Filev, 1994), ***Expectation Maximization***(Dempster, Laird, & Rubin, 1977) o el Conglomerado en dos fases(Bacher, Wenzig, & Vogler, 2004). La mayoría de estos métodos lo que hacen es ir variando el número de grupos, dentro de un rango lógico o dado por el usuario, y evaluando el agrupamiento con un índice de validación interno en cada caso. Luego es escogido el que mejor evaluado haya sido. Por lo que, si se dispone de la implementación de medidas de evaluación internas del agrupamiento puede lograrse con cualquier algoritmo el mismo comportamiento de manera manual, o la implementación para que este funcione de esta forma no debe ser compleja. Si las predicciones de estas técnicas no coinciden en el mismo número de grupos puede escogerse el rango de posibilidades entre la menor estimación y la mayor para comenzar a experimentar.

- Si el volumen de los datos es elevado se puede escoger métodos como el K-medias(MacQueen, 1967), BIRCH(Zhang, Ramakrishnan & Livny, 1996) y los ***Self-OrganizingMaps***(Kohonen, 1995).

- Si con los anteriores los resultados no son buenos atendiendo a las medidas de evaluación internas una selección o extracción de rasgos más profunda puede mejorar los resultados.

- Escoger la función de distancia para los métodos de agrupamiento que mejor se adecue, con la característica de los datos y los resultados que se desean alcanzar con el agrupamiento, es clave para cumplir los objetivos de la MD. En Deza & Deza (2006), se expone un compendio detallado de una gran cantidad con sus definiciones matemáticas, semántica, entre otros detalles. Si luego de una amplia experimentación si se utilizan varias funciones de distancia y similitud los resultados no son satisfactorios, puede valorarse entonces la opción de aprender una nueva función de distancia que se ajuste mejor a los datos con los que se trabaja, en Brown, Liu, Brodley & Chang (2012) se confecciona un procedimiento para lograr este objetivo.

- Si con los primeros acercamientos al agrupamiento los resultados evaluados con las técnicas pertinentes, son convincentes, es el momento de escoger uno o varios agrupamientos con el visto bueno de los expertos. Pero si no obtiene resultados promisorios se debe probar una mayor variabilidad en los parámetros y de métodos de forma general hasta alcanzar un resultado satisfactorio.

Luego de realizar la experimentación con las técnicas seleccionadas es necesario ***generar las pruebas del diseño***, pues los resultados son, desde un acercamiento inicial, grupos arbitrarios. Algunos de ellos están mejor formados que otros. Una manera de determinarlo es aplicar a los conglomerados formados los índices de validación interna tales como: Ball&Hall (1965);Calinski & Harabasz(1974); Dunn(1974); Hartigan(1975);Davies & Bouldin (1979);y Halkidi & Vazirgiannis (2001), Estos permiten verificar si la estructura de los grupos producido por un algoritmo colocan adecuadamente los datos. La función que ejecuta los índices internos tiene como parámetros principales: una matriz de datos (donde cada columna representa un rasgo y las filas las instancias) y un vector con las etiquetas de las clases conformadas por los algoritmos. La idea más sencilla es escoger el agrupamiento de mayor cantidad de índices con mejores valores; pero puede también hacerse un ordenamiento por el promedio del ranking que le da cada índice de validación a cada resultado. La idea es obtener, de alguna forma, una muestra prometedora, del total de resultados, para mostrársela a los expertos del negocio.

Luego solo queda experimentar con las técnicas seleccionadas y los datos; es importante hacerlo con cada algoritmo lo más que se pueda dentro de los parámetros lógicos que el problema muestre. Esto se logra básicamente variando los parámetros de cada una de las técnicas, se pueden obtener varios agrupamientos con una misma elección de algoritmo. Luego seleccionar las muestras más prometedoras, como se explica anteriormente, con la utilización de los índices de validación internos.

Es en la fase V, ***evaluación***, en los que se analizan los agrupamientos seleccionados como más promisorios en la etapa anterior, no debe ser únicamente el que queda en primer lugar luego del ordenamiento según los índices de validación, sino una muestra de estos. De conjunto con los especialistas en el dominio se selecciona el que mejor soluciona las necesidades existentes y por tanto se comprueba el más útil de la modelación realizada. Para ello pueden realizarse análisis de los centroides de los grupos, comportamientos de las variables dentro de un mismo conglomerado, distribución de las clases, entre otros. En este punto pueden descubrirse cuestiones importantes que pueden hacer que el proceso de MD

regrese a la fase de modelado, o incluso anteriores, con un enfoque diferente.

Por ejemplo, puede que luego de todos los análisis realizados se tiene una idea más clara de la cantidad de grupos a construir. O que los agrupamientos encontrados no brindan información novedosa o útil al cliente; en puntos como este último, en el que el agrupamiento es realizado por la predominancia de una variable y los grupos creados son muy triviales. Puede pensarse repetir el proceso, ahora se toman los conglomerados construidos como conjuntos de datos independientes. También puede considerarse, en caso de tener clases desbalanceadas, el unir dos o más de estas para conformar un agrupamiento con más sentido.

Por último se tiene la fase VI, ***despliegue***, con el objetivo de desplegar los resultados de la MD, se concluye que estrategia es el momento de documentar todo lo realizado y unir todos los reportes que el equipo de trabajo ha realizado independientemente. Todo esto es necesario para trazar una estrategia de monitoreo y mantenimiento del proceso, pues pasado un determinado tiempo pueden comenzarse a cometer errores con los resultados de la minería pues los comportamientos datos puede variar.

Caso de estudio

Con el objetivo de comprobar el desempeño y guía de la adecuación realizada a la metodología CRISP-DM a problemas no supervisados tipo atributo-valor, descrita anteriormente, se decide aplicar esta a un caso de estudio, el cual consta de en un grupo de pacientes diabéticos tipo 2 de la provincia de Cienfuegos. Se muestra a continuación cómo se procede en cada una de las fases se toma como guía lo descrito en la sección anterior.

Luego del análisis con los expertos se extrae como resultado de la ***fase I*** el objetivo de la minería: mejorar el proceso de diagnóstico de la diabetes mellitus tipo 2 (DM), permite que este se haga en etapas tempranas de la enfermedad o aún antes de su debut, en las áreas de atención primaria de Cienfuegos. Por tanto la meta que rige el proceso de MD es identificar grupos existentes dentro de los datos, para luego de categorizados estos puedan obtener comportamientos comunes de los pacientes en cada conglomerado. Además con el nuevo conjunto de datos, ya clasificado, pueden construirse modelos de predicción utilizando el propio resultado del agrupamiento o técnicas supervisadas. Otro punto importante es la herramienta a utilizar en el proceso de MD, en este caso es elegida WEKA, debido a que es de distribución libre desarrollada bajo licencia GPL, lo cual ha impulsado que sea una de las plataformas más utilizadas en el área en los últimos años. Además de su facilidad para añadir extensiones y modificar métodos gracias a su filosofía de paquetería para lograr esto sin la necesidad de modificar el núcleo de la aplicación.

Como resultado de aplicar la ***fase II***, comprensión de datos, se cuenta con una idea clara de la información contenida y de las dificultades que estos presentan. Se hace posible la corrección de algunas de estas dificultades, con el propósito de mejorar su calidad. Por ejemplo, de 77 características iniciales que se poseían de cada caso (paciente) fueron identificadas como relevantes solo 37. Es comprobada la correspondencia entre el comportamiento en los datos y la teoría de las ciencias médicas, verificadas mediante análisis de moda de la variable sexo, promedio a la de edad, entre otras de las variables consideradas significativas en estos pacientes.

Durante la ***fase III***, se realizaron tareas como la de selección de los datos, en la que se decide usar las 37 variables descritas como relevantes de la fase anterior. Por otro lado de los 1951 pacientes iniciales que se dispone luego de realizar un análisis de valores atípicos, se utiliza el filtro ***Interquartile Range***, con la eliminación de 18 casos de la muestra. Además en la muestra solo el 30% de los casos tienen todos sus atributos sin valores perdidos, por lo que son utilizados los filtros no supervisados de atributos que ayudan a mejorar esta situación. Para el caso de las variables numéricas fue el ***EMImputation*** que utiliza el algoritmo de maximización esperada para sustituir los valores perdidos por los que se considera más adecuados. Para los atributos discretos se usa el ***Replace Missing Values*** que sustituye los valores ausentes por la moda. Por último se pasa a estandarizar las variables mediante el filtro ***Standardize***, cuyo resultado es que todos los atributos numéricos pasan a tener media cero y desviación estándar uno.

Lo primero a realizar en la ***fase IV*** es la selección de las técnicas de modelado, que permite definir qué algoritmos de la herramienta WEKA se utilizan para determinar los grupos en los diabéticos tipo 2. La aplicación cuenta con varias técnicas implementadas, para un primer acercamiento al problema (si los resultados no son satisfactorios se realizarán análisis más profundos) se utilizarán:

- ***EM*** (Simple Expectation Maximization): para tener una estimación inicial de la cantidad de grupos a conformar.
- ***Cascade SimpleKMeans***: también estima la cantidad óptima de grupos realiza varios agrupamientos (utiliza como base el algoritmo K-medias) y escoge el mejor al aplicarle la medida de validación interna Calinski y Harabasz.
- ***SimpleK Means***: implementación del algoritmo K-medias.

- *Farthes tFirst*: variantedel k-medias.
- *LVQ*: implementación de *Learning Vector Quantization*.
- *Self OrganizingMap*: implementación de los mapas auto-organizados de Kohonen.

El resultado de las técnicas seleccionas son grupos de pacientes, algunos de los cuales están mejor conformados que otros, por lo que en este punto se generan pruebas para estos resultados. Para ello se utiliza una nueva funcionalidad añadida a WEKA: ***ClusterValidation***, la que tiene implementada cinco índices de validación interna: Ball & Hall (1965); Calinski & Harabasz(1974); Dunn(1974); Hartigan (1975);Davies-Bouldin(1979), y; debido a la no existencia de estos en la aplicación. De esta forma se evalúan los índices a cada uno de los resultados. Luego se realiza un ranking de cada agrupamiento por cada una de las medidas y serán escogidos, para el análisis de los expertos, los que mejor promedio de este ranking hayan tenido.

El próximo paso es ejecutar las técnicas, para luego determinar la calidad de los resultados desde una mirada técnica. Para ello se realiza primeramente una experimentación con el total de los casos. Se aplica el algoritmo EM con la opción que estima la cantidad de grupos y el ***Casca de SimpleK Means***que posee esta misma funcionalidad. Con la utilización de estos resultados se pasa a ejecutar las demás técnicas, a las que hay que especificarle manualmente la cantidad de grupos, se varía este valor entre dos y cuatro, que son los resultados de las dos anteriores.

Lo próximo es pasar a la fase V, ***evaluación***, en la que al hacer un análisis de los agrupamientos que quedaron en primer y segundo lugar, resultan en poner en un conglomerado a los pacientes que son hombres y en otro a las mujeres. Incluso con el análisis de los resultados que se encuentran en el lugar tercero y cuarto, en los que los agrupamientos fueron de tres y cuatro grupos respectivamente, se observa que hay un gran desbalance de las clases obtenidas; la mayoría de los elementos se agrupan en dos grupos y estos están determinados en su mayoría por el sexo igualmente.

Al presentar los resultados a los expertos se determina que este agrupamiento no aporta información útil para el diagnóstico temprano de la diabetes. Entonces, mediante la adecuación de la metodología realizada, que dice si puede haber retroalimentación entre una fase posterior con una anterior, se determina dividir el conjunto de datos en dos, un conjunto de hombres y uno de mujeres. A estos nuevos conjuntos se le aplican las fases IV y V, respectivamente, para tratar de obtener clases útiles para el diagnóstico temprano de la DM.

Durante la ***fase IV***, en este caso con el conjunto dividido por sexo, las técnicas seleccionadas, las pruebas de diseño para validar el agrupamiento y la experimentación con los datos se realizan de la misma forma que la pasada vez de la realización de esta etapa.

En la ***fase V*** nuevamente se comprueba si son útiles o no los modelos resultantes para el diagnóstico temprano de la DM tipo 2, se verifica el cumplimiento de los objetivos de la MD. Atendiendo a que el análisis es por sexo, en este caso los resultados se analizan de manera independiente en los conjuntos.

En el caso de los hombres el resultado logrado por el algoritmo ***EM*** con tres grupos es el que propone la conglomeración más adecuada para el diagnóstico de la DM, de acuerdo al criterio de expertos. Este conforma tres grupos distribuidos como se muestra en la Figura 1a. Obesidad, 50 años de edad como promedio, antecedentes familiares de diabetes mellitus e hipertensión arterial (HTA) son características comunes de los hombres de estos grupos. No obstante, existen en los pacientes de cada uno de estos conjuntos diferencias significativas.

En el grupo 1 se encuentran los hombres con un peso promedio de 86.27 Kg y un índice de masa corporal (IMC) de 30.17, lo que indica un alto nivel de obesidad. Además tienen el hábito tóxico de tomar café. Niveles de glicemia bajos y resultados de análisis médicos en niveles normales son características distintivas de los hombres de este grupo. El grupo 2 está conformado por pacientes tomadores de café, que con un peso promedio de 80 Kg y un IMC de 28.79, tienen glicemias por encima de los 10mmol/L y niveles de colesterol, triglicéridos y microalbuminuria muy altos. Por último, el grupo 3 se compone de aquellos que con un IMC de 27.32, que implica niveles de obesidad bajos, y sin hábitos tóxicos presentan niveles de glicemia altos pero por debajo de los 10mmol/L. Además de tener el colesterol en niveles de riesgo y los triglicéridos un poco altos.

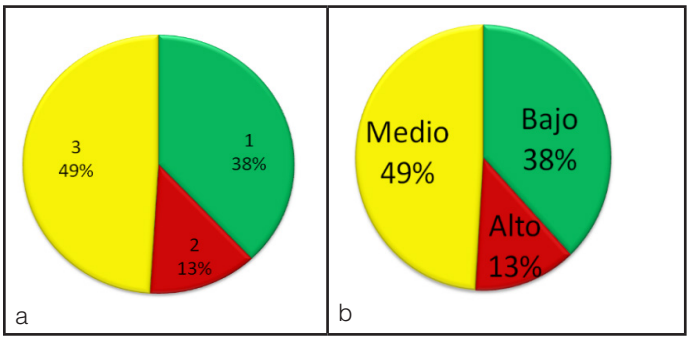

Figura 1. a. Distribución de los hombres en los grupos por el EM con k=3.

b. Etiquetas de las clases en el conjunto de hombres.

Debido a estas diferencias entre los conglomerados pueden identificarse etiquetas en los grupos de pacientes hombres, tres niveles de riesgo: bajo, medio y alto, que coinciden con los grupos 1, 3 y 2 respectivamente (ver Figura1b).Pero lo más interesante es que se puede seguir un comportamiento diferenciado en cada caso por parte de los médicos; se propone actuar de las siguientes formas por los especialistas para cada uno:

- Grupo de riesgo ***bajo***: primeramente realizar una prueba de tolerancia a la glucosa. El segundo paso es modificar los estilos de vida del paciente, para lo que se le asigna una dieta y un plan de ejercicios físicos. Estas personas deben tener seguimiento con el fin de observar el comportamiento de los factores de riesgo, pueden nunca llegar a hacer el debut en la diabetes si siguen las orientaciones médicas.
- Grupo de riesgo ***medio***: estos ya debutaron con diabetes, pero al tener niveles de glucosa por debajo de los 10mmol/L los médicos proponen modificar su estilo de vida y, en caso de ser necesario, indicar un hipoglucemiante oral.
- Grupo de riesgo ***alto***: pacientes con un cuadro clínico tóxico, los médicos no solo le modifican su estilo de vida, mediante una dieta y ejercicios, sino que le indican tratamiento medicamentoso como insulina, durante una primera etapa, luego puede mantenerse con hipoglucemiantes orales.

En el caso de las féminas es el algoritmo ***Self Organizing Map***elque, formando 4 grupos mostrados en la Figura 2a, es escogido por los especialistas en el dominio, ya que se encuentran características interesantes en los diferentes conglomerados. Es interesante que en ellos las glicemias se encuentren entre 7 y 10mmol/L, lo que es considerado por los médicos como glicemias altas, aunque no en niveles críticos. Otra característica común es que las pacientes tienen antecedentes familiares de diabetes.

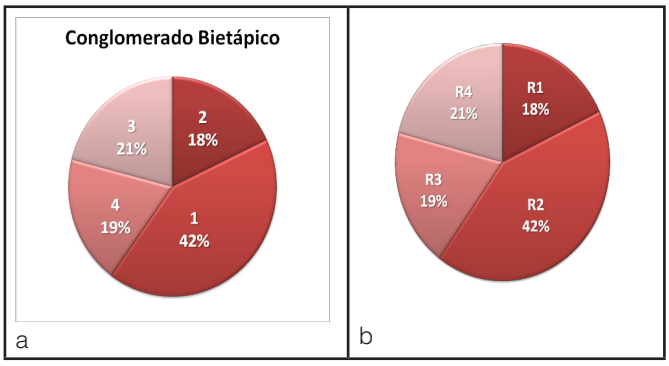

Figura 2 a. Distribución de las mujeres en los grupos por el Self Organizing Map con k=4.

b. Etiquetas de las clases en el conjunto de mujeres.

Pero al igual que en el caso anterior, con los hombres, existen características específicas que diferencian los grupos formados. En el grupo 1 se encuentran las mujeres que con 74.52 Kg de peso tienen un IMC de 30.52, lo que indica niveles de obesidad elevados. Estas mujeres, muy obesas, toman café, son hipertensas y tienen los triglicéridos y el colesterol un poco altos. El segundo grupo está formado por las mujeres de 47 años de edad aproximadamente, con 82.4 Kg de peso promedio y un IMC de 33.48. Lo que indica mujeres jóvenes muy obesas. Otras características interesantes son que padecen HTA y tienen los triglicéridos y la microalbuminuria altos.

El grupo 3 lo constituyen mujeres de 56 años, 64 Kg y un IMC de 26.52. Estas pacientes no son obesas, no tienen hábitos tóxicos, no padecen HTA y sus análisis están normales. El último grupo se compone de mujeres de 63 años, 64.24 Kg y 27.69 de IMC. Estas personas con niveles de obesidad bajos, son tomadoras de café, hipertensas y con los triglicéridos y el colesterol un poco elevados.

Al analizar las características de estos grupos se observa que las diferencias notables no están en los niveles de glicemia, como en el caso de los hombres, sino en la edad y el peso, específicamente los niveles de obesidad. Según los médicos este es un factor agravante del riesgo que tienen los diabéticos de complicarse, por las consecuencias que esta tiene para la salud humana. La edad es otro agente que aumenta la posibilidad de riesgo ya que varias de las complicaciones de la DM aparecen después de padecer la enfermedad por algún tiempo. Además cuando se debuta en edades tempranas es porque se tienen los factores de riesgo en niveles altos. Por lo anterior se pueden identificar en los grupos formados niveles de riesgo de complicación, ya no determinados por la glicemia, sino por estos factores igual de importantes en la prevención de la DM.

Por tanto se proponen como etiquetas, para las clases del conjunto de datos de mujeres, cuatro niveles de riesgo donde uno indica mayor riesgo y cuatro menor riesgo (ver Figura 2b). La clasificación R1 corresponde al grupo 2 conformado por la técnica, ya que estas son las mujeres más jóvenes y con más obesidad. R2 son las mujeres muy obesas pero más avanzadas en edad, agrupadas en el conglomerado 1. El grupo de R3 coincide con el cuarto conglomerado, donde las pacientes son obesas en menor medida. El grupo 3 se considera como R4 ya que estas mujeres no son obesas, ni hipertensas y sus análisis estás en niveles saludables.

Debido a que en todos los grupos las mujeres tienen glicemias entre 7 y 10 mmol/L el procedimiento a seguir es

el mismo para los cuatro. Este consiste en modificar el estilo de vida de la paciente, mediante la dieta y el plan de ejercicios, e indicarle un hipoglucemiante oral, si es necesario. Pero para la indicación de la dieta los médicos tienen en cuenta el peso, la talla, el IMC, la edad, el sexo y la actividad física. Algunos de estos factores influyen también en el tipo de hipoglucemiante oral que le prescriben al paciente. Por lo que, al existir diferencias notables en los grupos conformados de mujeres en estos factores las hay también en el tipo de dieta y el tratamiento de cada una de estas pacientes.

CONCLUSIONES

Se revisaron las metodologías Proceso KDD, CRISP-DM y SEMMA. Se elige CRISP-DM para realizar su adecuación a los problemas nos supervisados tipo atributo-valor por ser de libre distribución, independiente de la herramienta utilizada y la más usada dentro de la comunidad científica.

Se realiza la adecuación de la metodología CRISP-DM a los problemas no supervisados tipo atributo-valor especificando en cada una de sus fases que acciones, procesos y métodos realizar en este tipo particular de dato. Lo que facilita a los investigadores la solución de estos tipos de problemas.

El problema de la diabetes tipo 2 en la ciudad de Cienfuegos se elige como caso de estudio. Luego de aplicar la adecuación de la metodología se decide hacer un análisis independiente por sexo. Se obtienen tres grupos en los hombres y cuatro en las mujeres, en ambos casos fueron interpretados como niveles de riesgo de complicación de la enfermedad.

REFERENCIAS BIBLIOGRÁFICAS

Acuna, E., & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy ***Classification, clustering, and data mining applications*** (pp. 639-647). Berlin: Springer.

Azevedo, A., & Santos, M. F. (2008). ***KDD, SEMMA and CRISP-DM: a parallel overview***. Paper presented at the IADIS European Conf. Data Mining. Recuperado de http://dblp.uni-trier.de/db/conf/iadis/dm2008.html#AzevedoS08

Bacher, J., Wenzig, K., & Vogler, M. (2004). ***SPSS TwoStep Cluster-a first evaluation***: Lehrstuhl für Soziologie Berlin, DE. Amsterdam. Recuperado de http://www.philso.uni-augsburg.de/lehrstuehle/soziologie/sozio2/ehemalige/wenzig/vortraege/2004_spsstwostep.pdf

Ball, G. H., & Hall, D. J. (1965). ISODATA, A novel method of data analysis an pattern classification. In NTIS (Ed.). Menlo Park: Stanford Research Institute.

Batista, G. E., & Monard, M. C. (2002). A Study of K-Nearest Neighbour as an Imputation Method. ***HIS, 87, pp.*** 251-260. Recuperado de http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.14.3558&rep=rep1&type=pdf

Brown, E. T., Liu, J., Brodley, C. E., & Chang, R. (2012). ***Dis-function: Learning distance functions interactively.*** Paper presented at the Visual Analytics Science and Technology (VAST), 2012 IEEE Conference.

Cai, D., Zhang, C., & He, X. (2010). ***Unsupervised feature selection for multi-cluster data.*** Paper presented at the Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining.

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. ***Communications in Statistics - Theory and Methods, 3***(1), pp.1-27. doi: 10.1080/03610927408827101

Cios, K. J., Pedrycz, W., Swiniarski, R. W., & Kurgan, L. A. (2007). ***Data Mining, A Knowledge Discovery Approach***.Berlin: Springer.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide. Recuperado de http://www.iidia.com.ar/rgm/CD-TIpEI/TEI-2-CRISP-DM-GdP-material.pdf

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. ***Pattern Analysis and Machine Intelligence, IEEE Transactions on***(2), pp. 224-227. recuperado de

Dempster, A.P, Laird, N. M., & Rubin, D. (1977). Maximum - likelihoodfromincomplete data via the EM algorithm. ***Journal of the Royal Statistical Society, Series B, 39*** (1), pp. 1 – 38. Recuperado de http://web.mit.edu/6.435/www/Dempster77.pdf

Deza, E., & Deza, M.-M. (2006). ***Dictionary of Distances***. Ámsterdam: Elsevier Science.

Dunn, J. (1974). Well separated clusters and optimal fuzzy partitions. ***Journal on Cybernetics, 4***, pp. 95-104.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). The KDD Process for Extracting Useful Knowledge from Volumes of Data. ***Communications of the ACM, 39***(11), pp. 27-34. doi: 0002-0782

Gorunescu, F. (2011). ***Data Mining - Concepts, Models and Techniques*** (12). Berlin: Springer.

Halkidi, M., & Vazirgiannis, M. (2001). ***Clustering validity assessment: Finding the optimal partitioning of a data set.*** Paper presented at the Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference.

Hartigan, J. A. (1975). Clustering Algorithms, New York: John Willey and Sons. ***Inc. Pages113129***.

He, X., Cai, D., & Niyogi, P. (2005). ***Laplacian score for feature selection.*** Paper presented at the Advances in neural information processing systems.

Kohonen, T. (1995). Self-Organising Maps. Berlin:***Springer.***

Li, Z., Yang, Y., Liu, J., Zhou, X., & Lu, H. (2012). ***Unsupervised Feature Selection Using Nonnegative Spectral Analysis.*** Paper presented at the AAAI.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations ***5th Berkeley Symp, 1***, pp. 281–297. Recuperado de MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations ***5th Berkeley Symp, 1***, pp. 281–297. Recuperado de https://pdfs.semanticscholar.org/a718/b85520bea702533ca9a5954c33576fd162b0.pdf

Malik, K., Sadawarti, H., & Singh, K. G. (2014). Comparative Analysis of Outlier Detection Techniques. ***International Journal of Computer Applications, 97***(8), pp. 12-21. Recuperado de https://www.researchgate.net/profile/Kamal_Malik2/publication/269802647_Comparative_Analysis_of_Outlier_Detection_Techniques/links/54f036890cf25f74d7243135.pdf

Olson, D. L., & Delen, D. (2008). ***Advanced Data Mining Techniques***. Berlin: Springer. Recuperado de http://lib.mdp.ac.id/ebook/Karya%20Umum/Advanced-Data-Mining-Techniques.pdf

Qian, M., & Zhai, C. (2013). ***Robust Unsupervised Feature Selection.*** Paper presented at the IJCAI.

Russell, S., & Novig, P. (2009). ***Artificial Intelligence A modern Approach***. New York: Pearson Education.

Schafer, J. L. (1997). ***Analysis of incomplete multivariate data***. London: Chapman & Hall.

Yager, R. R., & Filev, D. P. (1994). Approximate clustering via the mountain method. ***IEEE Transactions on Systems Man and Cybernetics, 24***(8).

Yang, Y., Shen, H. T., Ma, Z., Huang, Z., & Zhou, X. (2011). ***l2, 1-norm regularized discriminative feature selection for unsupervised learning.*** Paper presented at the IJCAI Proceedings-International Joint Conference on Artificial Intelligence.

Zhang, T., Ramakrishnan, R., & Livny, M. (1996). ***BIRCH: An Efficient Data Clustering Method for Very Large Databases.*** Paper presented at the 1996 ACM SIGMOD International Conference on Magnament of Data. Montreal.