Extracción y Visualización de Conocimiento de Bases de Datos Médicas

José Hernández Orallo
M.Carmen Juan Lizandra
Neus Minaya Collado
Carlos Monserrat Aranda

El aumento del volumen y variedad de información que se encuentra informatizada en bases de datos digitales ha crecido espectacularmente en la última década. Gran parte de esta información es histórica, es decir, representa transacciones o situaciones que se han producido en el pasado y que se han ido registrando a través de la interacción con la ubicuidad de sistemas informáticos que nos rodean.

Aparte de su función de "memoria de la organización" a la cual pertenecen, esta información está empezando a tener un valor mucho más importante que la de un simple registro: la información histórica es útil para predecir la información futura.

Aunque a primera vista parece que predecir el futuro es algo bastante mágico y poco sistematizable, las ciencias inductivas, especialmente en la segunda mitad del siglo XX, han establecido con rotundidad lo que puede ser predicho a partir de experiencias anteriores. Además, se conoce aproximadamente el esfuerzo computacional necesario y se puede obtener una estimación más que razonable de la fiabilidad de la predicción. De hecho, predecir el futuro no tiene nada de mágico, la mayoría de nuestros actos cotidianos se basan en predicciones. Uno coge una línea de autobús para ir al trabajo porque predice que ese autobús le llevará a donde quiere porque durante los últimos meses ha venido haciéndolo así. En cambio decide coger el metro aquellos días que llueve, porque sabe que el autobús encontrará más atasco de lo habitual. Gran parte de nuestros actos se basan en experiencias pasadas y en su extrapolación futura.

La mayoría de decisiones de empresas, organizaciones e instituciones se basan también en información de experiencias pasadas extraídas de fuentes muy diversas. A diferencia de las decisiones personales, las decisiones colectivas suelen tener consecuencias mucho más graves, especialmente económicas, y, recientemente, se deben basar en volúmenes de datos que desbordan la capacidad humana.

A partir de aquí se puede explicar porqué el área de la extracción semi-automática de conocimiento de bases de datos ha adquirido recientemente una importancia científica y económica inusual. Como hemos dicho, existe una vasta cantidad de información de manera digital. Esto supone, en primer lugar, que las personas ya no son capaces de analizar toda esta información en un tiempo razonable. Por tanto, las decisiones se ven abrumadas por

millones de datos que deben analizarse. En segundo lugar, afortunadamente, el hecho de que la información esté en formato digital permite que ésta pueda ser analizada directamente por sistemas informáticos.

Llegados a este punto, la pregunta obvia es qué tiene esto de novedoso, cuando el estudio estadístico de datos de distinta índole se viene realizando desde hace décadas. La respuesta a esta pregunta depende de varios factores, pero especialmente al hecho de que las técnicas estadísticas son principalmente de validación, es decir, requieren que el usuario proponga un modelo y sean las herramientas estadísticas las que lo parametricen y le den un grado de plausibilidad respecto a unos determinados datos. Más aún, para proponer estos modelos e interpretar los resultados de los distintos tests estadísticos, el usuario debe tener ciertos conocimientos de estadística.

Sin embargo, las necesidades de toma de decisión requieren el descubrimiento de *nuevos* modelos no esperados o imposibles de descubrir manualmente a partir de tal magnitud de datos. Este descubrimiento, además, debe requerir de la mínima pericia posible por parte del usuario. Nace el "Descubrimiento de Conocimiento a partir de Bases de Datos" (KDD, del inglés *Knowledge Discovery from Databases*). Fayyad et al. [2] lo definen como el "*proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos*".

Es importante resaltar que el KDD puede utilizarse a la vez para descubrir y verificar una hipótesis. Pero además, estos modelos deben ser novedosos (no previamente esperados o conocidos) y además comprensibles. Para ello se necesitan lenguajes de representación para las hipótesis más inteligibles que los modelos estadísticos y también se hacen necesarias nuevas técnicas de visualización, para ayudar a entender los patrones que ha extraído el sistema, y convertirlos en conocimiento consciente y, por tanto, útil.

Finalmente, existía un término similar a KDD, denominado "Análisis Inteligente de Datos" (IDA, del inglés Intelligent Data Analysis) que correspondía con el uso de técnicas de inteligencia artificial en el análisis de los datos, menos especializado, y que últimamente ha caído en desuso debido a que solapa en gran medida con KDD.

En este artículo comentaremos algunas técnicas de extracción de conocimiento sobre bases de datos y de su visualización, centrándonos en bases de datos médicas, debido a que este área ilustra de una manera más clara este cambio entre el uso de herramientas estadísticas tradicionales y la minería de datos.

EL PROCESO DEL KDD

Muchas veces se confunde el KDD con la minería (o prospección) de datos (DM, del inglés *Data Mining*). La DM se encarga de generar y contrastar hipótesis y es, por tanto, sólo una parte del proceso complejo del KDD. El KDD engloba una serie de fases:

1. Determinar las fuentes de información que pueden ser útiles y dónde conseguirlas.

2. Diseñar el esquema de un almacén de datos (Data Warehouse) que consiga unificar de manera operativa toda la información recogida.

3. Implantación del almacén de datos que permita la "navegación" y visualización previa de sus datos, para discernir qué aspectos puede interesar que sean estudiados.

4. Selección, limpieza y transformación de los datos que se van a analizar. La selección incluye tanto una criba o

fusión horizontal (filas) como vertical (atributos).

5. Seleccionar el método de minería de datos apropiado y aplicarlo.
6. Interpretación, transformación y representación de los patrones extraídos.
7. Difusión y uso del nuevo conocimiento.

El proceso resultante es el que se ilustra simplificadamente en la figura 1:

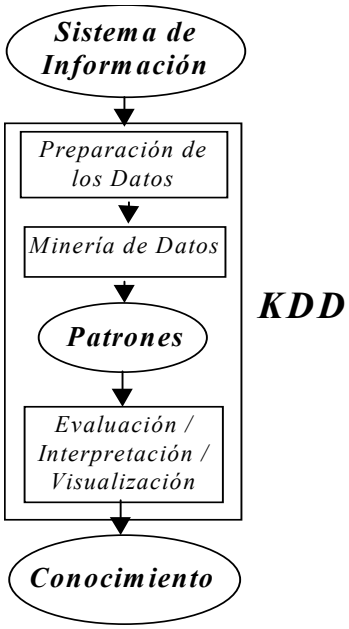

Fig. 1. Proceso de Extracción de Conocimiento

Vulgarmente se dice que el KDD transforma datos (información) en conocimiento.

RECOGIDA Y PREPROCESADO DE DATOS

Las primeras fases del KDD han cobrado gran relevancia en el área de sistemas de información, y también es una de las que determinan que las fases sucesivas sean capaces de extraer conocimiento válido y útil a partir de la información original.

Generalmente, la información que se quiere investigar sobre un cierto dominio de la organización se encuentra en bases de datos y otras fuentes muy diversas, tanto internas como externas. Para poder operar eficientemente con esos datos y debido a que los costes de almacenamiento masivo y conectividad se han reducido drásticamente en los últimos años, parece razonable recoger (copiar) los datos en un sistema unificado. Aparte de facilitar el acceso a los datos en tiempo real (conocido como OLAP, *On-Line Analytical Processing*), esta separación de los datos con respecto a las fuentes permite que el trabajo transaccional diario de los sistemas de información originales (conocido como OLTP, *On-Line Transactional Processing*) no se vea interferido por el proceso de minería de datos, generalmente muy exigente, sobre los sistemas de gestión de bases de datos que mantienen la información.

Aunque es un área todavía abierta, los esquemas de almacenes de datos más comunes son los denominados en estrella simple y estrella jerárquica (copo de nieve). Esta estructura permite la sumarización, la visualización y la navegación de la información según las dimensiones de la estrella (mediante el pliegue y despliegue de ramas de la estrella).

Esta estructura está especialmente indicada para un tipo de usuarios, denominados 'picapedreros' (también conocidos como granjeros). Estos usuarios se dedican fundamentalmente a realizar informes periódicos, ver la evolución de determinados parámetros, controlar valores anómalos, etc. En el ámbito empresarial estudiaran fundamentalmente la evolución de ventas, costes fijos y variables, nuevos clientes, fidelización de los mismos, etc. Sin embargo, la estructura en estrella también facilita la tarea de otro tipo de usuarios, denominados 'exploradores' que van a ser los encargados de encontrar nuevos patrones significativos utilizando técnicas de minería de datos.

Finalmente, un punto importante a destacar, especialmente en el caso de almacenes de datos médicos, es la privacidad. Muchas veces los registros corresponden con datos de historiales de pacientes. Existen dos soluciones para su privacidad: un acceso restringido a toda la información, que sólo permitiría analizar los datos a muy pocas personas autorizadas, o un acceso más general pero sin posibilidad de obtener información individual, sólo de grupos.

MINERÍA DE DATOS

Una vez recogidos los datos de interés en un almacén de datos, un explorador puede decidir qué tipos de patrón quiere descubrir. Es importante destacar que la elección del tipo de conocimiento que se desea extraer va a marcar claramente la técnica de minería de datos a utilizar. Afortunadamente, los propios sistemas de minería de datos se encargan generalmente de elegir la técnica más idónea entre las disponibles para un determinado tipo de patrón a buscar, con lo que el explorador sólo debe determinar el tipo de patrón. Veamos cuáles son estas tipologías:

- Asociaciones: Una asociación entre dos atributos ocurre cuando la frecuencia de que se den dos valores determinados de cada uno conjuntamente es relativamente alta. Es uno de los patrones con más interés comercial, sobre todo en el análisis de hábitos de los clientes, donde, por ejemplo, en un supermercado se analiza si los pañales y los potitos de bebé se compran conjuntamente.

- Dependencias: Una dependencia fundamental (aproximada o absoluta) es un patrón en el que se establece que uno o más atributos determinan el valor de otro. Uno de los mayores problemas de la búsqueda de dependencias es que suelen existir muchas dependencias nada interesantes y en las que la causalidad es justamente la inversa. Por ejemplo el hecho de que un paciente haya sido ingresado en maternidad determina su sexo.

- Clasificación: Una clasificación se puede ver como el esclarecimiento de una dependencia, en la que el atributo dependiente puede tomar un valor entre varias clases, ya conocidas. Muchas veces se conoce como aprendizaje supervisado. Por ejemplo, si se sabe por un estudio de dependencias que los atributos edad, número de miopías y astigmatismo han determinado los pacientes para los que su operación de cirugía ocular ha sido satisfactoria, podemos intentar determinar las reglas exactas que clasifican un caso como positivo o negativo a partir de esos atributos.

- Segmentación: La segmentación (o clustering) es la detección de grupos de individuos. Se diferencia de la clasificación en el que no se conocen ni las clases ni su número (aprendizaje no supervisado), con lo que el objetivo es determinar grupos o racimos (clusters) diferenciados del resto.

- Tendencias: El objetivo es predecir los valores de una variable continua a partir de la evolución de otra variable continua, generalmente el tiempo. Por ejemplo, se intenta predecir el número de clientes o pacientes, los ingresos, llamadas, ganancias, costes, etc. a partir de los resultados de semanas, meses o años anteriores.

- Reglas Generales: Evidentemente muchos patrones no se ajustan a los tipos anteriores. Recientemente los sistemas incorporan capacidad para establecer otros patrones más generales.

Una vez determinado el tipo de patrón a buscar, el sistema (u opcionalmente el usuario) puede elegir la técnica más apropiada. Según la manera de representar los patrones, podemos distinguir entre técnicas no simbólicas y técnicas simbólicas. Las más numerosas y tradicionales son las técnicas no simbólicas, generalmente más apropiadas para variables continuas y con un conocimiento más claro de lo que se quiere buscar.

- Técnicas estadísticas: Parte de las múltiples técnicas estadísticas se pueden utilizar para confirmar asociaciones y dependencias, y para realizar segmentaciones. Una técnica muy importante es el uso de regresión lineal (y no lineal) y redes de regresión para establecer tendencias. También son originariamente estadísticos los árboles de regresión, aunque los comentaremos más adelante.

- Vecinos más próximos y razonamiento por casos. Se aplican generalmente a clasificación y segmentación, basándose en medidas de distancias o similitud con el prototipo o los miembros de los grupos.

- Redes Neuronales Artificiales, Lógica Difusa, Algoritmos Genéticos y combinaciones entre ellos. Son técnicas tradicionales de aprendizaje automático con aplicaciones especialmente en clasificación y segmentación. Su mayor inconveniente es la mala inteligibilidad de los resultados, aunque algunas nuevas combinaciones y técnicas permiten extraer reglas a partir de los modelos inducidos.
- Algoritmos específicos: algoritmos eficientes para la búsqueda de asociaciones o dependencias. Por ejemplo, en un supermercado con miles de artículos, no se puede evaluar estadísticamente cada uno de los posibles pares o tripletes de combinaciones entre productos. Así, existen algoritmos específicos que permiten buscar todas las asociaciones significativas existentes eficientemente (véase p.ej. el capítulo de Agrawal et al. en [2]).

El mayor inconveniente de las técnicas no simbólicas es su poca (o nula) inteligibilidad. En el caso del razonamiento por casos o las redes neuronales, el resultado del proceso es una caja negra que sirve para predecir o clasificar nuevos casos, pero no se sabe cómo y, por tanto, no se ha obtenido conocimiento. Por el contrario, las técnicas simbólicas generan un modelo "legible" y además aceptan mayor variedad de variables y mayor riqueza en la estructura de los datos. Entre las técnicas simbólicas, podemos destacar:

- Árboles de Decisión: Utilizados fundamentalmente para clasificación y segmentación, consisten en una serie de tests que van separando el problema, siguiendo la técnica del divide y vencerás, hasta llegar a las hojas del árbol que determinan la clase o grupo a la que pertenece el registro o individuo. La fig. 2 muestra un árbol de decisión. Existen muchísimas técnicas para inducir árboles de decisión, siendo el más famoso el algoritmo C4.5 de Quinlan [8]. Los árboles de regresión son similares a los árboles de decisión pero basados en técnicas estadísticas.

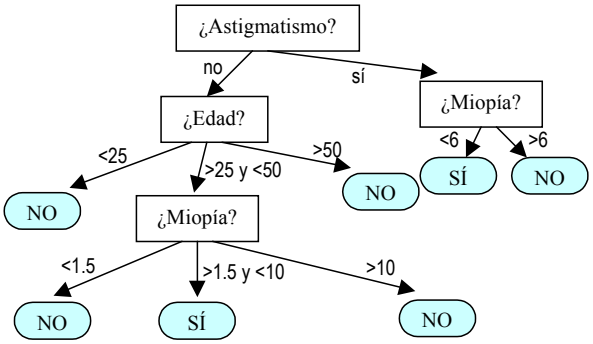

Fig. 2. Árbol de Decisión para Determinar Recomendación o No de Cirugía Ocular.

- Programación Inductiva y Otras Técnicas de Inducción Simbólica de Alto Nivel: fundamentalmente se usan para obtener patrones de tipo general, que se pueden establecer entre varios individuos o son intrínsecamente estructurales. Aunque existen algunas aproximaciones basadas en reglas simples, es la programación lógica inductiva (ILP) el área que ha experimentado un mayor avance en la década de los noventa [7]. ILP se basa en utilizar la lógica de primer orden para expresar los datos, el conocimiento previo y las hipótesis. Como la mayoría de bases de datos actuales siguen el modelo relacional, ILP puede trabajar directamente con la estructura de la misma, ya que una base de datos relacional se puede ver como una teoría lógica. Aparte de esta naturalidad que puede evitar o simplificar la fase de preprocesado, ILP permite representar hipótesis o patrones relacionales, aprovechando y descubriendo nuevas relaciones entre individuos. Por ejemplo no tiene sentido enviarle propaganda de piscinas a una persona si ésta convive con otra que ya se instaló una piscina recientemente. Estos patrones son imposibles de expresar con representaciones clásicas. Nótese que un árbol de decisión siempre se puede convertir fácilmente en un conjunto de reglas (como de la fig. 2 a la 3) pero no viceversa.

¿Operación?

SI Astig.=No **Y** 25<Edad<25 **Y** 1.5<Miopía<10 **ENTONCES** SÍ

SI Astig.=Sí **Y** Miopía<6 **ENTONCES** SÍ

EN OTRO CASO Operación=NO

Fig. 3. Reglas correspondientes al árbol de la fig. 2.

En definitiva, existen multitud de técnicas, combinaciones y nuevas variantes que aparecen recientemente, debido al interés del campo. Así, los sistemas de KDD se afanan por incorporar la mayor cantidad de técnicas, así como ciertas heurísticas para determinar o asesorar al usuario sobre qué métodos son mejores para distintos problemas.

KDD EN MEDICINA

Durante los ochenta, aparte del campo de imagen médica, ya tratado en un número anterior de esta revista [5], las aplicaciones de inteligencia artificial y de técnicas avanzadas de computación a la medicina se basaban en el desarrollo de sistemas expertos de apoyo al diagnóstico. La mayoría de estos sistemas no llegaron a calar en la práctica habitual de muchos profesionales, especialmente en nuestro país, donde pocas consultas en aquella época disponían incluso de ordenador personal y pocos hospitales tenían (y tienen) completamente informatizados los historiales. En los sistemas expertos, los modelos eran introducidos y validados por expertos manualmente. A partir de estos modelos y de los datos introducidos por el médico tras examinar el paciente y su historial, el sistema automáticamente sugería un diagnóstico que, evidentemente, sólo debía servir de apoyo al diagnóstico final que debía realizar el médico.

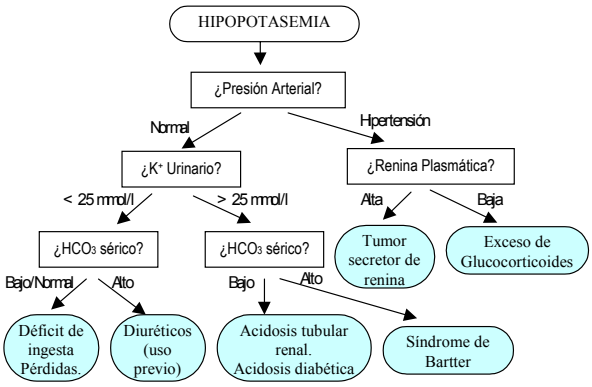

Fig.4. Árbol de Decisión para el Diagnóstico Diferencial de las causas de Hipopotasemia (adaptado de [3])

Generalmente, su utilidad se basaba en diagnóstico diferencial, donde complejos árboles de decisión como el de la figura 4, eran mucho más fáciles de aplicar automáticamente que siguiendo manualmente las flechas del árbol que aparecen en muchos manuales de medicina.

Sin embargo, estos sistemas tenían algunas carencias importantes: generalmente el experto simplificaba el conocimiento que introducía en el sistema. La razón no estaba en una falta de rigor sino en que muchas veces los expertos eran incapaces de expresar todo su conocimiento con reglas, ya que muchos expertos basan su decisión en procesos no completamente conscientes. Esto es similar a los procesos que utilizan los labradores (o sus perros) cuando predicen que va a llover mañana. Esto no quiere decir que el experto (o el labrador, o incluso el perro) no tenga un modelo, sino que este modelo es neuronal, basado en interconexiones, pesos y refuerzos de un circuito imposible de mostrar conscientemente. Nótese que esto no quiere decir que no haya modelos comprensibles.

Por último, muchos de estos sistemas de los ochenta quedaban obsoletos con el tiempo y no se particularizaban a las características propias del área donde el sistema iba a ser aplicado. No es lo mismo un diagnóstico diferencial con síntomas de vómitos y diarreas agudas en una zona no tropical que tropical.

En definitiva, en esta última década el objetivo es la adquisición, el descubrimiento y el mantenimiento de gran parte del conocimiento de una manera automática. La inclusión manual del mismo a partir de expertos u otras fuentes debería minimizarse, siempre que hubiera alternativa automática. Para ello, como hemos visto, surge el KDD. Los nuevos objetivos del KDD en medicina son [6]:

- La interpretación comprehensiva de los datos de los pacientes de una manera contextual y la presentación de tales interpretaciones de una manera visual o simbólica.

- La extracción (descubrimiento) de conocimiento médico a partir del diagnóstico, revisiones médicas, seguimientos, terapias o tareas globales de gestión de los pacientes.

Dentro de la extracción de conocimiento se incluye el uso de conocimiento previo, ya sea para su refinamiento o particularización, o para ayudar al descubrimiento de nuevos patrones o modelos.

Otra característica importante es que los usuarios de estos nuevos sistemas son profesionales de la medicina que, aunque con ciertos conocimientos de estadística obligatorios en su formación, no tienen conocimientos de aprendizaje de modelos ni de la mayoría de técnicas presentadas en el punto anterior. Por tanto, los sistemas deben ser sencillos de manejar, los patrones descubiertos deben ser fáciles de entender (ya sean simbólicos o visuales) y la interrelación con el resto de sistemas informáticos de adquisición de datos, visualización y gestión de los centros asistenciales debe ser transparente para el usuario.

A partir de aquí los nuevos sistemas deben permitir de una manera cómoda y eficaz:

- Asociación de síntomas y clasificación diferencial de patologías.
- Estudio de factores (genéticos, precedentes, hábitos, alimenticios, etc.) de riesgo/salud en distintas patologías.
- Segmentación de pacientes para una atención más inteligente según su grupo.
- Predicciones temporales de los centros asistenciales para el mejor uso de recursos, consultas, salas y habitaciones.
- Estudios epidemiológicos, análisis de rendimientos de campañas de información, prevención, sustitución de fármacos, etc.

Como se puede comprobar, el abanico de aplicaciones es muy amplio, las repercusiones tanto en calidad de servicio y disminución de costes pueden ser altamente significativas. Sin embargo, el uso de estas técnicas todavía es reducido, especialmente en nuestro país.

Por citar algunos ejemplos más concretos del uso de estas técnicas, Abbott [1] ha utilizado un sistema de KDD para determinar los principales factores que motivan el cambio de atención de largo plazo (crónica) a atención aguda. Existían 44.000 registros (individuos) de diferentes fuentes, cada uno con 1.095 variables. Tras el preprocesado y limpieza, el número de registros potenciales para el análisis se redujo a 14.000. Además se redujo el número de variables a 135. Se utilizó una red neuronal artificial para clasificar los pacientes en dos clases: atención a largo plazo y atención aguda. El modelo aprendido mostró una efectividad de clasificación global del 94%. Se esperaba que algunos factores como dispnea, cadera rota, ataque cardiaco o sepsis aparecerían, y así se verificó por el sistema. Sin embargo, algunas variables que no se esperaban ayudaron a refinar el modelo. Finalmente, hasta 23 factores se mostraron significativos, algunos esperados y otros descubiertos.

Otra aplicación similar es la desarrollada en el Hospital Central de Karlstad en Suecia, donde se analiza la recurrencia de nuevos tumores en los cinco años posteriores a una operación de cáncer de mama. El tipo de seguimiento del paciente se decide por un conjunto de reglas obtenidas automáticamente a partir de los datos recogidos en el hospital durante años.

Finalmente, desde un punto de vista de gestión hospitalaria, el sistema descrito por Matheus et al [2] realiza informes en lenguaje natural sobre distintos aspectos de gestión de un hospital a partir de datos de admisiones, materiales, pagos, servicios, recursos humanos, etc., con el fin de estudiar las causas de aumento (o disminución) de costes, y corregirlos en aquellos casos que sea posible.

VISUALIZACIÓN DE LOS DATOS

Aunque parte de los procesos de descubrimiento ya se realizan de manera completamente automática, muchos de ellos requieren todavía de intervención humana o, por el contrario, pueden mejorar la capacidad humana de extraer y comprender patrones. Para potenciar estas capacidades se suelen utilizar las técnicas de visualización de datos.

Las técnicas de visualización de datos se utilizan fundamentalmente con dos objetivos: en primer lugar aprovechar la gran capacidad humana de extraer patrones a partir de imágenes (todavía no se han inventado reconocedores de caracteres mejores que el ser humano) y, en segundo lugar, ayudar al usuario a comprender más rápidamente patrones descubiertos automáticamente por un sistema de KDD.

Aunque las técnicas no son excluyentes, estos dos objetivos marcan dos momentos diferentes del uso de la visualización de los

datos: visualización *previa* (a veces conocida como Visual Data Mining [10]) y visualización *posterior* al proceso de minería de datos.

La visualización *previa* se utiliza para entender mejor los datos y sugerir posibles patrones o qué tipo de herramienta de KDD utilizar. Se utiliza frecuentemente por picapedreros, para ver tendencias y resúmenes de los datos, y por exploradores, para ver 'filones' que investigar.

Un ejemplo típico de estas visualizaciones previas es la segmentación mediante funciones de densidad, generalmente representadas tridimensionalmente, donde los seres humanos ven claramente los segmentos (clusters) que aparecen con distintos parámetros (en este caso altura, figuras 5-10):

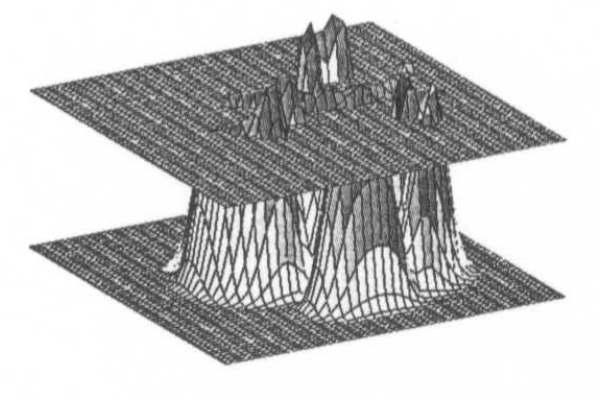

Fig.5. Gráfico de Densidad con Corte a Altura 4 (de [10]).

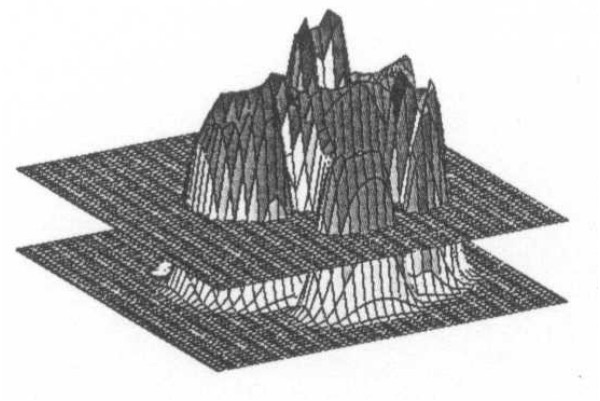

Fig.6. Gráfico de Densidad con Corte a Altura 2 (de [10]).

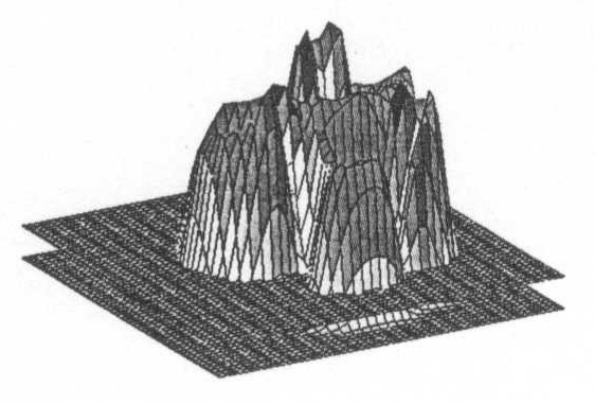

Fig.7. Gráfico de Densidad con Corte a Altura 1 (de [10]).

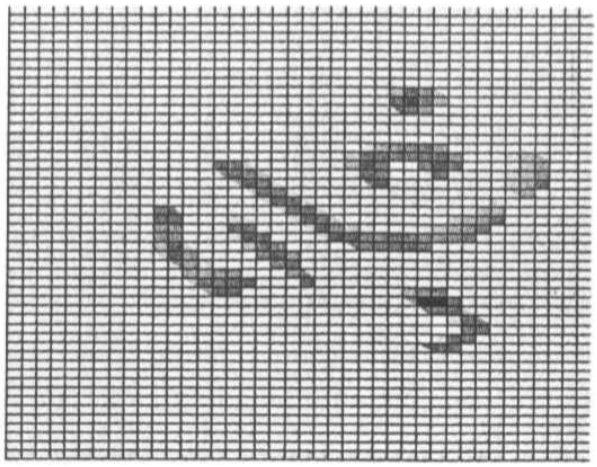

Fig.8. Corte y Segmentación a Altura 4 (de [10]).

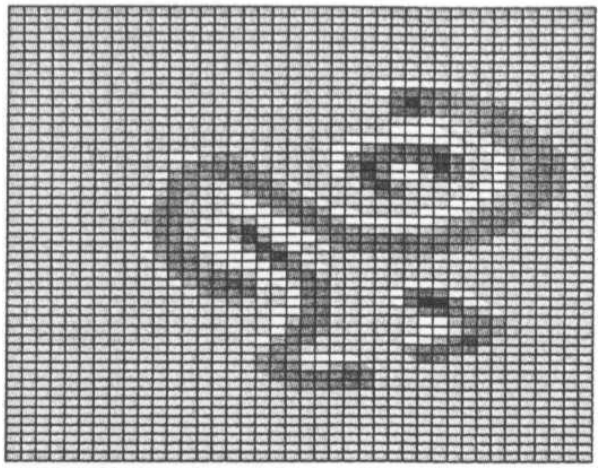

Fig.9. Corte y Segmentación a Altura 2 (de [10]).

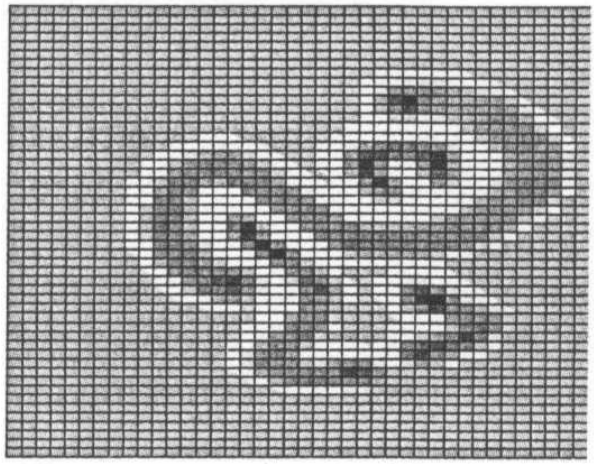

Fig.10. Corte y Segmentación a Altura 1 (de [10]).

También es muy común para encontrar asociaciones el uso de gráficos bidimensionales donde en las abscisas y ordenadas están todos los factores y en la intersección se muestran las frecuencias (utilizando color, puntos gordos o una tercera dimensión) de cada par de factores. Esta técnica se puede utilizar hasta con asociaciones de tres factores.

Para detectar valores extraños y anomalías se utilizan gráficos especializados, algunos de ellos tradicionales en estadística: *outliers* (valores extremos), *scatterplots* (diagramas de dispersión) o *alarming* (resalte de anomalías).

Sin embargo, el mayor problema de la visualización de datos es que la información de almacenes de datos suele ser multidimensional, siendo además el número de dimensiones mucho mayor que 3. El objetivo es por tanto conseguir proyectar las dimensiones en una representación en 2 (ó 3 simuladas) dimensiones, que son las únicas que pueden ser representadas en la pantalla de un ordenador o en papel.

La proyección geométrica que se ha popularizado más en la década de los noventa es la técnica de visualización de coordenadas paralelas [4]. La idea consiste en mapear el espacio k-dimensional en dos dimensiones mediante el uso de *k* ejes de ordenadas (escalados linealmente) por uno de abscisas. Cada punto en el espacio *k*-dimensional se hace corresponder con una línea poligonal (polígono abierto), donde cada vértice de la línea poligonal intersecta los *k* ejes en el valor para la dimensión. Cuando hay pocos datos cada línea se dibuja de un color. Cuando hay muchos datos se utiliza una tercera dimensión para los casos.

Por ejemplo, dados los siguientes datos de pacientes (tabaquismo, colesterol, tensión, obesidad, alcoholismo, precedentes, estrés) y su riesgo (muy bajo, bajo, medio, alto, muy alto) de enfermedades coronarias:

Tbco.	Clstrl	Tnsn	Obsd	Alcl	Prcd	Strs	Rsg
Med	Alto	8	No	Sí	Sí	No	Alto
Bajo	Med	9	Sí	No	No	No	Bajo
Alto	Bajo	8,5	No	No	No	No	Med
Bajo	Med	7	No	No	No	No	Bajo
Bajo	Bajo	8,5	No	Sí	Sí	Sí	Med
Bajo	Med	9	No	No	Sí	No	Med
Med	Bajo	9	No	No	Sí	No	Med
Alto	Med	11	No	No	No	No	Alto
Alto	Alto	13	Sí	No	Sí	No	M.A.
Bajo	Bajo	7	No	No	No	No	M.B.
Bajo	Alto	12	Sí	Sí	Sí	Sí	M.A.
Alto	Med	11	No	No	No	Sí	Alto
Alto	Med	8	No	No	No	No	Med

se pueden representar utilizando coordenadas paralelas de la siguiente manera:

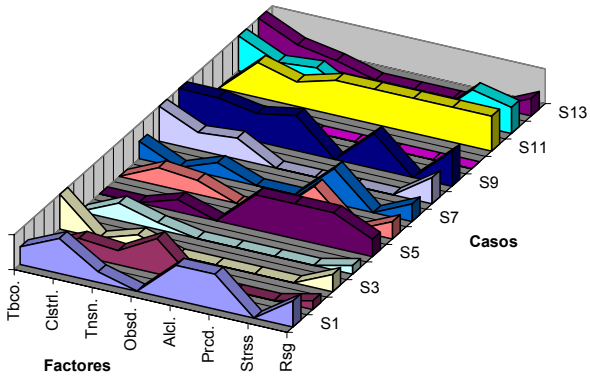

Fig.11. Representación de Coordenadas Paralelas.

El mayor problema de estas representaciones (y de otras muchas) es que no acomodan bien las variables discretas.

En este sentido, existen otro tipo de técnicas que sí permiten combinar atributos continuos y discretos, mediante el uso de transformaciones menos estándar y el uso de iconos. Se utilizan rasgos compatibles y diferenciados para distintas dimensiones, como son círculos, estrellas, puntos, etc., con la ventaja de que se pueden combinar más convenientemente valores discretos y continuos.

Para el ejemplo anterior, podríamos tener el siguiente gráfico:

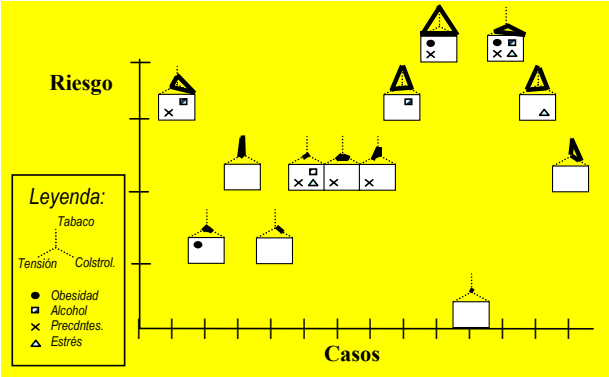

Fig.12. Representación Icónica.

donde los valores continuos (tabaco, colesterol y tensión) aparecen en miniejes tridimensionales, mientras que los discretos aparecen como iconos (círculo, cuadrado, cruz y triángulo para obsesidad, alcoholismo, precedentes y estrés, respectivamente).

Otras aproximaciones más sofisticadas se basan en estructuras jerárquicas, como por ejemplo, los Cone Trees [9].

Por otro lado, la visualización *posterior* se utiliza para mostrar los patrones y entenderlos mejor. Un árbol de decisión es un ejemplo de visualización posterior. Nótese la diferencia entre los datos de la tabla anterior o su visualización en las figuras 11 y 12, y un árbol de decisión en concreto generado a partir de dicha tabla. De hecho, existen muchos árboles de decisión posibles para unos datos dados.

Otros gráficos de visualización posterior de patrones son los que muestran una determinada segmentación de los datos, una asociación, una determinada clasificación, utilizando para ello gráficos de visualización *previa* en los que además se señala el patrón. Por ejemplo, la figura 13 muestra una segmentación lineal realizada para el corte y segmentación de la figura 9.

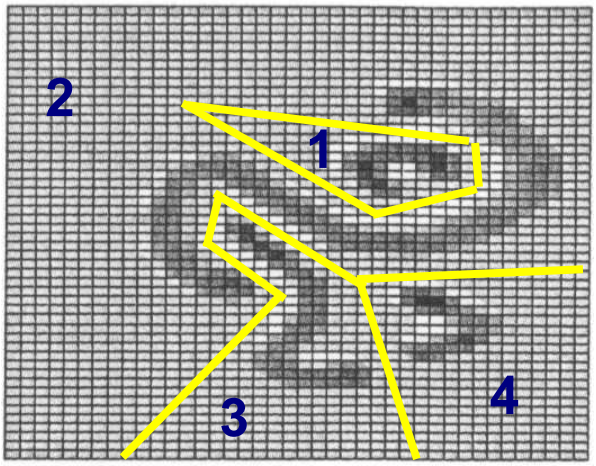

Fig.13. Segmentación sugerida a partir de la fig. 9.

En nuestra opinión, las herramientas gráficas requieren mayor experiencia para seleccionar qué gráfico nos interesa utilizar entre los cientos de gráficas que proporcionan los sistemas actuales.

CONCLUSIONES

En este artículo se han mostrado las diferencias, y por tanto las nuevas posibilidades, del descubrimiento de conocimiento a partir de bases de datos, en comparación con otras aproximaciones más clásicas. En particular, se diferencia de los sistemas expertos en que, en éstos, los modelos se introducen manualmente y el sistema se dedica a aplicarlos, mientras que en el KDD se descubren estos modelos. El KDD se diferencia de las herramientas de análisis de datos estadísticas en que éstas

verifican un modelo propuesto por el usuario (o a lo sumo lo parametrizan) mientras que los sistemas de KDD descubren patrones novedosos (que por sus propios criterios de selección en la generación están ya verificados).

Las nuevas técnicas de visualización son mucho más complejas que las tradicionales. Muchas de ellas se basan en complejas representaciones tridimensionales o icónicas que ayudan al usuario a sugerirle patrones que analizar o a entender un patrón descubierto por el sistema.

Hemos mostrado algunos ejemplos de la utilidad y el gran abanico de posibilidades del KDD en el área de la salud, donde se ilustra con mayor crudeza los posibles beneficios que se están desaprovechando, por el desconocimiento de unas técnicas que están, por tanto, seriamente infrautilizadas.

Afortunadamente, esta tendencia está cambiando. El descubrimiento de conocimiento a partir de bases de datos es un área incipiente que va a cobrar cada vez mayor importancia y ubicuidad a medida que los posibles usuarios (cualquier persona o entidad que tenga que tomar decisiones importantes) vayan familiarizándose con las nuevas herramientas y sistemas que, bien seguro, serán cada día más potentes y fáciles de manejar.

REFERENCIAS

1. Abbot, P. *Predicting Long-Term Care Admissions: A Connectionist Approach to Knowledge Discovery in the Minimum Data Set,* Tesis Doctoral, Universidad de Maryland, Baltimore, 1999.

2. Fayyad U., Piatesky-Shapiro G., Smyth P. (eds.) *Advances in Knowledge Discovery and Data Mining*, MIT Press, 1996.

3. Harrison et al. (eds.) *Principios de Medicina Interna, 12ª Edición*, McGraw Hill, 1991.

4. Inselberg, A.; Dimsdale, B. "Parallel Coordinates: A Tool for Visualizing Multidimensional Geometry" *Proc. Visualization 1990,* IEEE CS Press, Los Alamitos, Calif. 1990, pp. 361-370.

5. Juan Lizandra, M.Carmen; Monserrat Aranda, Carlos; Hernández-Orallo, José: Imagen Médica. Síntesis de Imagen Médica, ACTA, Vol. 11, pp.:44-55, 1999.

6. Lavrac, N. "Selected Techniques for Data Mining in Medicine" *Artificial Intelligence in Medicine,* (16) 1: 3-24, 1999.

7. Muggleton, S.; De Raedt, L. "Inductive Logic Programming: Theory and Methods" *Journal of Logic Programming,* (19-20: 629-679), 1994.

8. Quinlan, J.R. *C4.5 Programs for Machine Learning*, Morgan Kaufmann, 1993.

9. Robertson, G.; Card, S.; Mackinlay, J. "Cone Trees: Animated 3D Visualisations of Hierarchical Information", Proc. ACM CHI Intl. Conf. On Human Factors in Computing, ACM Press, New York, 1991, pp. 189-194.

10. Wong, P.C. "Visual Data Mining", Special Issue of *IEEE Computer Graphics and Applications*, Sep/Oct 1999, pp. 20-46.