KDD (Knowledge Discovery in Databases): Un proceso centrado en el usuario

Héctor Oscar Nigro, Daniel Xodo, Gabriel Corti, Damián Terren
INCA/INTIA - Departamento de Computación y Sistemas
Facultad de Ciencias Exactas - UNICEN – Tandil
Campus Universitario - Paraje Arroyo Seco s/n
TE: +54-2293-432466 – FAX: +54-2293-444431
e-mail: onigro@exa.unicen.edu.ar; dxodo@exa.unicen.edu.ar, gabrielcorti@fastmail.fm,

Resumen

"La tecnología KDD está basada en un bien definido proceso KDD de múltiples pasos, para el descubrimiento de conocimiento en grandes colecciones de datos. El proceso KDD es iterativo por naturaleza, y depende de la interacción para la toma de decisiones, de manera dinámica" [GUP]

Por otro lado el *´data stream´*, es una manera tradicional con la que se ha trabajado en los principales paquetes comerciales de Data mining donde creemos encontrar algunos inconvenientes y falta de flexibilidad.

A partir de estas premisas, nuestra intención es elaborar un sistema de descubrimiento de conocimiento en grandes bases de datos con una interfaz más simple y fácil de utilizar a los efectos que el usuario de Data mining obtenga una herramienta que posea la misma funcionalidad de aquellas analizadas, con las mismas ventajas, pero que solucione las debilidades detectadas en la utilización del *´data stream´*.

Se tomará como paquete de software de referencia el Clementine V.5.00, y a partir de sus fortalezas, es decir todas las funcionalidades que ofrece, se establecerá como objetivo superar sus debilidades a través de una nueva organización de sus elementos y módulos en una arquitectura flexible para incorporar nuevos componentes y una interfaz simplificada que ayude al usuario cumpliendo con un número importante de características, como así también hemos pensado la incorporación de un agente inteligente que asista en todas las fases del proceso de descubrimiento de conocimiento.

Finalmente se llevarán a cabo pruebas del sistema en bases de datos reales

1.1 Knowledge discovery in databases

En la literatura actual se puede encontrar un gran número de definiciones acerca del descubrimiento de conocimiento en bases de datos, término que resumiremos con las siglas KDD (Knowledge discovery in databases).

Una de las definiciones más completas, es la siguiente:

"El descubrimiento de conocimiento en bases de datos es un campo de la inteligencia artificial de rápido crecimiento, que combina técnicas del aprendizaje de máquina, reconocimiento de patrones, estadística, bases de datos, y visualización para automáticamente extraer conocimiento (o información), de un nivel bajo de datos (bases de datos)" [FAYY97].

El KDD es un área que está tomando importancia dado el crecimiento actual de las bases de datos (incluyendo bases de datos relacionales, bases de datos de objetos, bases de datos *spatial-time* y otras), y de la capacidad del hardware disponible para procesar estos datos.

Hay que tener en cuenta que el KDD no es un producto de software, sino un proceso compuesto de varias etapas. En estos tiempos están surgiendo herramientas con una gran cantidad de librerías que ayudan en cada uno de estos pasos.

1.2 Proceso KDD centrado en el usuario

El KDD es un proceso centrado en el usuario, que tiene la propiedad de ser altamente interactivo, y que debe ser guiado por las decisiones que toma el usuario, o también por un agente inteligente. La naturaleza centrada en el usuario del proceso KDD posee varias cuestiones actualmente en investigación. Una de ella es, como asistir al usuario en la correcta selección de herramientas y técnicas apropiadas, para lograr los objetivos del usuario.

Es un desafío real, dar al sistema, la inteligencia necesaria para obtener conocimiento e impartir el mismo, en el momento de decidir las herramientas apropiadas, para *que* tipo de problemas y *cuando*. Particularmente en KDD esto es un problema importante de abarcar, aún si el usuario es un investigador que desarrollo técnicas específicas, ya que se necesita al sistema completo para resolver un problema.

Mientras que hay desarrollos de *Knowledge discovery workbenches* y sistemas integrados que incluyen más de un paso del proceso KDD, ellos no se involucran demasiado en las cuestiones de un sistema amigable para el usuario, para ser utilizado por un analista. Un analista, no es usualmente un experto en KDD, pero sí alguien que tiene la responsabilidad de sacar el significado de los datos usando técnicas de KDD disponibles. Para que un sistema cualquiera de KDD sea exitoso, necesita integrarse bien dentro de un ambiente existente para proveer una completa solución a un analista.

Por lo tanto, un reto para los investigadores y practicantes del KDD es poner más énfasis en el proceso general de KDD y en las herramientas para soportar sus varios pasos. Se le debería prestar mayor atención a la interacción con el humano y menos a la automatización total, con el fin de soportar tanto a expertos como a usuarios novatos. El desarrollo de herramientas apropiadas de visualización, interpretación y análisis de descubrimiento de patrones son de particular importancia.

Tales ambientes interactivos, a través de la reducción del tiempo para comprender la complejidad de los datos, habilitan la posibilidad de obtener soluciones prácticas a muchos de los problemas de la vida real, de una manera más rápida que lo hace el ser humano o una computadora operando independientemente.

1.3 Proceso asistido por un agente inteligente

Todas estas metodologías se preocupan en como convertir los datos en información, como llevar a cabo un trabajo organizado, y como diseminar la información de una manera en la cual los *stakeholders*[1] puedan fácilmente convertirla en recursos para la toma de decisiones.

Algunas de las herramientas que se encuentran actualmente en el mercado fueron específicamente diseñadas y documentadas para caber dentro de uno de estos marcos específicos.

Como dijimos anteriormente el proceso de KDD es interactivo e iterativo por naturaleza, e involucra una serie de pasos, en los que se incluyen decisiones tomadas por el usuario. En general, involucrando más, o menos pasos, la estructura general sigue la siguiente forma:

[1] Stakeholders: Son las personas afectadas por el proyecto o que pueden influenciarlo, pero no están directamente involucradas con el proyecto. Por ejemplo los gerentes afectados por el proyecto, los propietarios del proceso, las personas que trabajan con el proyecto a prueba, los departamentos internos que apoyan el proceso, clientes y departamento comercial.

1) Entender el dominio de aplicación, cuál es el problema a resolver, y cuales son los objetivos.

2) Seleccionar del conjunto de datos originales, un subconjunto apropiado, para el problema que deseamos resolver. Eliminando por ejemplo variables irrelevantes.

3) En la etapa de limpieza y pre procesamiento se deberían tomar decisiones con respecto a valores faltantes, atípicos, erróneos, etc. (ruido). También se podría necesitar normalizar los valores de las variables o llevar a cabo otras tareas similares. La etapa de preparación y limpieza es a veces una etapa descuidada pero de suma importancia en este proceso, dado que grandes cantidades de datos son recolectados por medio de métodos automáticos (ej. vía web). A veces el método por el cuál los datos fueron obtenidos no fue cuidadosamente controlado, y así los datos podrían contener valores fuera de rango (ej. edades negativas), combinaciones incorrectas de datos (Sexo: masculino; embarazada: si), y otros. Ingresar estos datos erróneos a los algoritmos de data mining solo lleva a entorpecer su proceso de aprendizaje, o a conseguir resultados alejados del comportamiento real.

4) Encontrar características útiles para representar a los datos dependiendo de los objetivos. Reducción de dimensiones (ej. Kohonen) para llevar adelante el trabajo con un número reducido de variables. Eliminar columnas que varían juntas, como por ejemplo fecha de nacimiento y edad, simplemente tabular, *aggregation* (calcular estadísticos descriptivos), o técnicas más sofisticadas como clustering o análisis de componentes principales.

5) Elegir las herramientas de data mining adecuadas al problema a resolver, teniendo en cuenta el objetivo (predecir, explicar, clasificar, agrupar, etc). También se debe en esta etapa, establecer los parámetros de las redes utilizadas (arquitectura de la red, datos de entrenamiento, de validación y de testeo, etc). Una vez realizada la tarea, se procede con el descubrimiento de patrones y relaciones en los datos, para presentárselos al usuario de una manera adecuada (gráficos, árboles, reglas, etc)

6) Interpretación de los datos, llevada a cabo por el analista.

7) En esta etapa se debería consolidar el conocimiento ganado, probando los modelos creados contra los resultados obtenidos de la aplicación de estos modelos en el mundo real.

Como dijimos anteriormente existen herramientas que fueron desarrolladas para caber en una de éstas metodologías, pero ninguna puso aún, real énfasis en el propósito de asistir al usuario durante el seguimiento de estos pasos en busca del conocimiento. Por tal motivo el sistema que desarrollaremos buscará adaptarse a la metodología general anteriormente mostrada, pero incorporando algunas características importantes como lo son:

- ✓ La posibilidad de realizar un ciclo entre cualesquiera dos pasos, ya que a veces el conocimiento descubierto puede ser directamente aplicable, y otras veces puede guiar al refinamiento de los objetivos de la minería.

- ✓ No hay un progreso determinístico asumido desde un paso a otro. También, los pasos interpretativos y evaluativos, pueden involucrar retrocesos a cualesquiera de los pasos anteriores, cualquier número de veces.

- ✓ La incorporación de un agente inteligente, encargado de monitorear las actividades del usuario y brindarle a este asistencia a lo largo de todo el proceso de KDD.

1.4 Objetivos

A lo largo de este trabajo se va a desarrollar una herramienta capaz de asistir a un usuario a través de todas las etapas del proceso de descubrimiento de conocimiento en bases de datos. También se realizará una comparación con aquellas herramientas consideradas líderes en el mercado para demostrar cuán acertados estábamos a la hora de pensar en una innovación en este tipo de sistemas.

Un sistema como el que estamos pensando permitirá agilizar el proceso de descubrimiento de conocimiento en bases de datos, ya que el mismo puede proveer soporte en todo el proceso, sin la necesidad de buscar otras herramientas complementarias. La manera de trabajar con el sistema será así bastante sencilla gracias a la incorporación de varios componentes visuales. El soporte y la ayuda en cada una de las etapas y principalmente en la preparación de los datos la cual es la etapa que más tiempo consume, proveyendo indicadores estadísticos de cada uno de los atributos, diferentes gráficos de distribuciones de los mismos, posibilidad para manejar valores faltantes o extremos, variar dinámicamente la variable a predecir y además, modificar también de forma dinámica la muestra con la que se seguirá adelante en el proceso, son características muy importantes en estas clases de sistemas.

También la inclusión de un agente inteligente, cubrirá la necesidad de asistencia al usuario, área de gran importancia en la cual hoy día se siguen llevando a cabo investigaciones. La incorporación de este agente ayuda a que el sistema sea de utilidad para los diferentes grupos de usuarios, llevando adelante tareas que el usuario podría olvidar o bien no saber que sería conveniente que las llevara a cabo (como por ejemplo eliminar un valor extremo, el cual podría incorporar ruido y hacer que el modelo realice posteriormente predicciones erróneas).

El agente, encargado entre otras cosas de mostrar los mensajes de error, lo debería hacer de una manera amigable y brindando algunas ideas de cómo solucionar los inconvenientes y recuperarse ante los errores. Esto estimula a que el usuario siga adelante con su trabajo. Por consiguiente un objetivo importante de este trabajo será mostrar que el sistema a desarrollar poseerá buenas aptitudes y una interfaz capaz de adaptarse al nivel de usuario que se encuentre en ese momento utilizando el sistema, y un conjunto de características competentes para las diferentes áreas en las que se desee aplicar el KDD.

Por otro lado, el sistema tendrá que mostrar una arquitectura flexible a la hora de incorporar nuevos componentes tales como diferentes algoritmos de redes neuronales, componentes estadísticos u otros. Esto se podrá lograr gracias a diferentes abstracciones que se llevaron a cabo y al diseño orientado a objetos del sistema. La modificación del algoritmo interno que utiliza algún componente no alterará en absoluto el resto del comportamiento de la aplicación. Esta característica brinda la posibilidad de ir incorporando nuevos algoritmos o mejoras a los ya existentes con un mínimo de esfuerzo.

El diseño de la arquitectura general del sistema fue pensado con la idea de poder incorporar en trabajos posteriores nuevos componentes que permitan la expansión del sistema a nuevas áreas como pueden ser OLAP, spatial data mining, nuevos algoritmos matemáticos, etc. Obteniendo por último una herramienta con diferentes módulos capaz de abarcar una amplia cantidad de tareas que se pueden llevar a cabo al utilizar los datos almacenados en una base de datos. También será

necesario brindar flexibilidad en cuanto a la base de datos a utilizar ya que esta debería poder ser sustituida sin necesidad de modificar toda la aplicación.

En conclusión, intentaremos demostrar los avances con respecto a los sistemas que hoy día se encuentran al frente en el mercado a través de nuestra investigación y desarrollando una nueva aplicación que evite las deficiencias encontradas en los sistemas tradicionales.

1.5. Referencias Bibliográficas

[FAYY97] Usama Fayyad y Evangelos Simoudis. (1997). *Data Mining and Knowledge Discovery in Databases*.

[GUP] SK Gupta; Vasudha Bhatnagar; SK Wasan. *A proposal for Data Mining Management System*

1.6 Bibliografía a utilizar

Carlos A. I. Fernández – *Fundamentos de los Agentes Inteligentes* – Informe Técnico UPM/ DIT/ GSI 16/ 97. Departamento de Sistemas de Ingeniería Telemáticos – Universidad Politécnica de Madrid

Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) y Rudiger Wirth (DaimlerChrysler). *CRISP-DM 1.0. Step-by-step data mining guide.*

Martín y Serrano, C. (1993): "Self Organizing Neural Networks for the Analysis and Representation of Data: some Financial Cases", Neural Computing & Applications, Vol. 1, nº 2, diciembre, pp. 193-206

Borgelt, Christian y Kruse, Rudolf. *Induction of Association Rules: A priori Implementation*. Department of Knowledge Processing and Language Engineering School of Computer Science. Otto-von-Guericke-University of Magdeburg. Universit¨atsplatz 2, D-39106 Magdeburg, Germany.

Quintana Truyenque, Michel Alain. *Modelo híbrido para los procesos de Data Mining en el apoyo a la toma de decisiones basados en tecnologías inteligentes conexionistas y difusas.*. Escuela Profesional de Ingeniería de Sistemas. Universidad Nacional de San Agustín. Arequipa – Perú. Julio del 2002.

Molina Félix, Luis Carlos. *Data Mining: Torturando los datos hasta que confiesen.* Coordinador del programa de *Data mining* (UOC). 2002

Fayyad, U.M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. (ed.) (1996). *Advances in knowledge and data mining*. Cambridge (Massachussets): AAAI/MIT Press.

SK Gupta; Vasudha Bhatnagar; SK Wasan. *A proposal for Data Mining Management System*