KDD y Data Minig

- Descubrimiento de Conocimiento en Bases de Datos
- Se define como la extracción no trivial de información implícita, desconocida, y potencialmente útil de los datos.
- La palabra descubrimiento se relaciona con la idea de que la información más valiosa es aquella que no es conocida de antemano; sin embargo estas técnicas también pueden servir para confirmar la sospecha de un cierto comportamiento del sistema en una situación particular.

KDD y Data Minig

- Suele buscarse ese conocimiento expresable en forma de reglas porque son más inteligibles para los humanos, y permite entender el modelo del sistema en relación a los datos observados.
- Por otra parte estas reglas se usarán para la predicción de estados futuros.
- A veces sólo se pretende realizar una tarea de clasificación de datos conocidos.
- Se necesitan pues herramientas que automaticen el proceso de descubrimiento de conocimiento.

Aplicaciones empresariales

- **Detección del fraude**: Se puede considerar como una tarea de clasificación. De hecho, cuando el algoritmo analiza una gran cantidad de transacciones las clasificará en legales e ilegales, identificando en estas últimas ciertas características que tengan en común.
- **Análisis de riesgos en créditos**: En este caso ya existen técnicas tradicionales, y el Data Mining ayudaría en la decisión propuesta por estas técnicas.
- **Personalización** en la oferta de productos

Aplicaciones científicas

- **Análisis de datos recogidos por satélites** (meteorológicos, astronómicos, geológicos...)
- **Astronomía**: Detección de cuerpos estelares – se encuadra dentro de lo que se llama reconocimiento de patrones.
  - A este área también pertenecen los OCR, sistemas de visión artificial,...
- **Bioinformática** :genética (identificación de genes)

Aplicaciones en Internet

- Comportamiento de usuarios de sites
- Personalización y recomendaciones on-line
  - En la navegación y en la búsqueda
  - En las compras de tiendas virtuales (Amazon)
- Text Mining: catalogación de textos
- E-learning
- Otras aplicaciones: ¡Espiar a la gente! (Echelon, carnivore, sniffers)

Ejemplo clásico

- Un caso muy popular ocurrió en unos supermercados den EEUU.
- Se usaron técnicas de Data Mining para estudiar el comportamiento de los clientes. Se encontraron relaciones interesantes entre los pañales las cervezas, el género del comprador y el día de la semana.

Ejemplo clásico

- Encontraron que los jueves y los sábados los clientes hombres que compran pañales también compraban cervezas.
- Esta información que no es evidente a primera vista puede servir para colocar los pañales y las cervezas cerca la una de la otra.
- Así, el resultado debe ser analizado después por un experto humano para intentar encontrar una explicación a esa asociación.

Descripción del proceso

1. **Estudio del dominio:** Se debe tener información sobre el domino objetivo. Características, objetivo del proceso de descubrimiento, tipos de patrones, fuentes de datos...
2. **Creación del conjunto de datos:** a partir de la información recolectada se ha de decidir cuál va a ser la fuente de datos que se usará, y seleccionar los atributos que se crean necesarios.

Descripción del proceso

3. **Preprocesado de los datos Filtrado de datos (Data cleaning):** Se ha de estudiar las posibles circunstancias que afecten a la calidad de los datos
   - Elementos extraños
   - Ruido
   - Valores perdidos (como tratarlos)
   - Discretización de valores continuos

Descripción del proceso

4. **Reducción de datos y proyección:** No todos los métodos soportan bien grandes cantidades de datos
   - Selección de atributos mediante técnicas estadísticas (¿Qué es lo realmente relevante?)
   - Selección de instancias (¿Son necesarias todas?)

Descripción del proceso

5. **Elección del objetivo de descubrimiento:** aunque pueda parecer obvio, definir claramente los objetivos que se persiguen, es la clave del éxito del Data Mining
6. **Elección de las metodologías adecuadas:** lo marcarán el tipo de objetivo y las características de los datos.
7. **Aplicación de las metodologías o Data Mining:** se deberán ajustar los parámetros elegidos para obtener los mejores resultados.

Descripción del proceso

8. **Interpretación de los resultados:** a partir del conocimiento sobre el dominio (experto)
9. **Incorporación del nuevo conocimiento:** puesta en práctica de los resultados obtenidos. Se deben revisar los resultados obtenidos y su validez, de forma que se pueda incorporar el conocimiento adquirido a la organización.

Objetivos del proceso de KDD

- **Clasificación**
- **Agrupación o clustering**
- **Predicción**
- **Sintetización (Summarization):** descripción compacta que resuma las características de los de los datos
- **Cambio:** se buscan modelos que permitan descubrir patrones en datos entre los que existe una dependencia temporal o espacial.

Métodos para el KDD

- **Arboles de decisión, reglas de clasificación**
- **Clasificadores, métodos de regresión:** escasa interpretabilidad pero buena predictibilidad
  - **Regresión estadística**
  - **Redes neuronales**
  - **KNN**
- **Particionamiento o agrupamiento:** permiten particionar conjuntos de datos o descubrir grupos característicos
  - K Means, SOM (Teuvo Kohonnen)