

Introducción

En vista de los comentarios y sugerencias que nos hicieron, via mail y por chat, sobre la posibilidad de la creación de nuevo conocimiento, he creído conveniente introducir el tema Data Mining (DM) como una posibilidad de creación de conocimiento en las organizaciones. Luego de esto entraremos de lleno al desarrollo metodológico de nuestra solución de inteligencia de negocios.

Panorama Actual

"Segmentamos a nuestros clientes usando Data Mining..", "Data Mining incrementa la satisfacción de nuestros clientes..", "Nuestros competidores están usando DM para incrementar su cuota de mercado, necesitamos levantarnos! ..". Son algunos de los comentarios en las organizaciones que se pueden percibir.

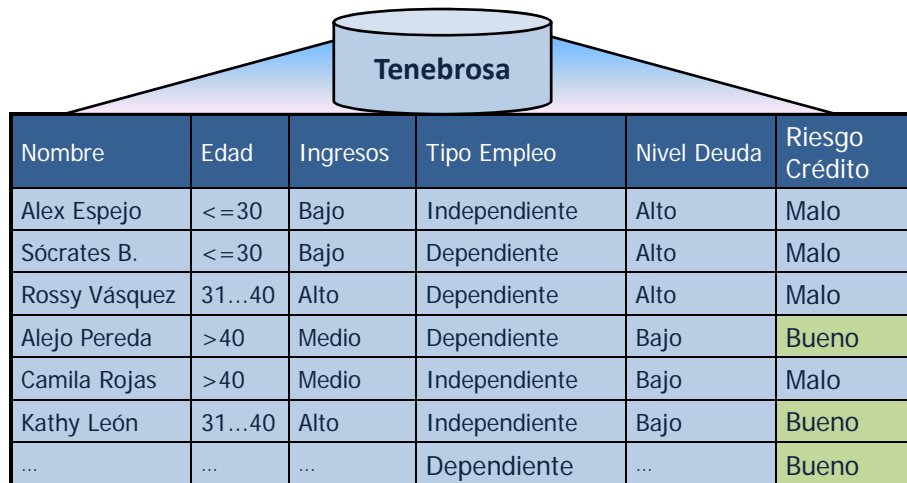
Pero que es el Data Mining? Como esta tecnología puede resolver los problemas diarios de las organizaciones? Cuál es el ciclo de vida de un DM?

Qué es Data Mining?

Data Mining constituye un miembro clave del Business Intelligence (BI) y permite analizar datos, hallando patrones escondidos, de manera automática o semi-automática. En lo que va del tiempo muchas empresas han acumulado una gran cantidad de datos en sus bases de datos, el resultado de esta colección de datos es que las organizaciones tienen "datos ricos" pero "pobre conocimiento".

El propósito principal del DM es extraer de los datos patrones, incrementar su valor intrínseco y transformar la data en conocimiento.

Imagine los datos de una tabla relacional, como se muestran en la fig. 1 conteniendo información de clientes.



Nombre	Edad	Ingresos	Tipo Empleo	Nivel Deuda	Riesgo Crédito
Alex Espejo	<=30	Bajo	Independiente	Alto	Malo
Sócrates B.	<=30	Bajo	Dependiente	Alto	Malo
Rosy Vásquez	31...40	Alto	Dependiente	Alto	Malo
Alejo Pereda	>40	Medio	Dependiente	Bajo	Bueno
Camila Rojas	>40	Medio	Independiente	Bajo	Malo
Kathy León	31...40	Alto	Independiente	Bajo	Bueno
...	Dependiente	...	Bueno

Fig. 1 Tabla de Clientes

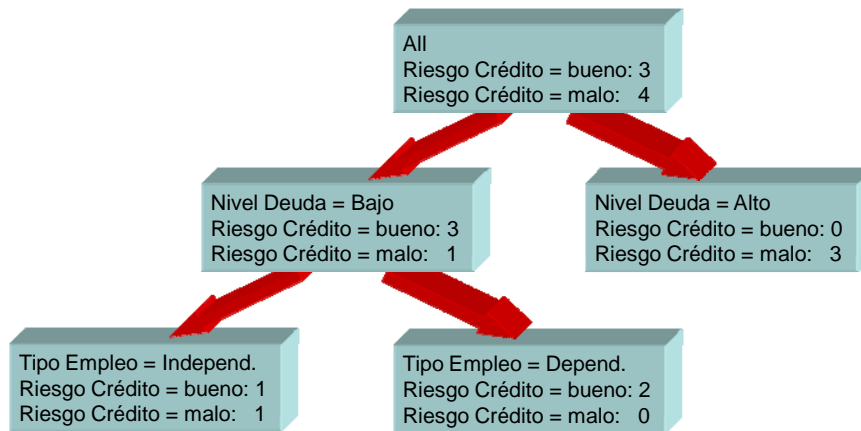
Una de las metas a encontrar podría ser: **¿A qué cliente o grupo de clientes le puedo dar un préstamo con un nivel de riesgo Bueno?**

Podríamos escribir una consulta para buscar cuantos con tipo de Empleo dependiente hay y cuantos no. El impacto de la edad seria otra variable a tener en cuenta o tal vez en función a su nivel de ingresos o deuda que tenía y seguramente tendríamos que escribir miles de consultas cuando queramos combinarlas algunas o todas a la vez, imagine si existiera mas columnas a usar y algunas columnas sean valores numéricos como los ingresos mensuales de un cliente.

En contraste el DM hace un acercamiento más simple ha esta pregunta. Todo lo que tiene que hacer es seleccionar el Algoritmo correcto de DM y especificar el uso las columnas a usar, el significado de las columnas de entrada y las columnas predictivas. En el caso anterior las columnas: edad, ingresos, tipo de empleo, nivel de deuda serian las de entrada. La columna **Riesgo Crédito** seria la columna predictiva. Un modelo de decisión de árbol podría ayudarnos a responder esa preguntar,

El algoritmo revisa la data y analiza el impacto de cada atributo ingresado

Árboles de Decisión (*Decision Trees*)



Volvamos a la pregunta original **¿A qué cliente o grupo de clientes le puedo dar un préstamo con un nivel de riesgo Bueno?**

Se imagina llegar a la respuesta de: los clientes con tipo de empleado **Dependiente** que tengan un nivel de **deuda bajo** y que tengan **más de 40 años** son los que representan **menos riesgo de deuda**.

El DM proporciona un enorme valor a las organizaciones. En estos tiempos el DM puede implementarse con más transparencia:

- **Gran cantidad de data disponible:** las organizaciones llegaron a implementar sistemas transaccionales (ventas, almacenes, producción, personal, contabilidad, etc) y estos en el tiempo han ido almacenando información aunado a la baja de los costos de almacenamiento han acumulado grandes volúmenes de datos.
- **Alto nivel de competencia:** la competencia actualmente es alta como resultado de marketing moderno y canales de distribución como internet y comunicaciones, así como la participación de corporaciones nacionales y extranjeras en el mercado. En este 2008 en nuestra ciudad Trujillo estamos siendo testigos de la aparición de 2 malls con una infraestructura bastante atractiva para los clientes, por mencionar un ejemplo de competencia.
- **Tecnología Lista:** el DM anteriormente era mayormente una solución de laboratorio, ahora ya es una tecnología madura y está lista para ser aplicada en las organizaciones. Los algoritmos y el equipamiento existente son más eficientes para trabajar con data complicada si fuera

el caso. Las API del DM están estandarizándose cada vez mas amplitud y esto permite a los desarrolladores construir aplicaciones

Realidad!

Hace poco conversaba con un Gerente de una empresa comercializadora, de gran presencia en el mercado regional y me comentaba entre otras cosas que, **sino contara con un sistema de información, no podría estar competido con estas corporaciones** – *cuenta con gran cantidad de datos y competencia de primera-* y que justo había invertido en un servidor con una configuración de primera – *Tecnología Lista* - .

Piense la ventaja de conocer la información que descubriría un DM

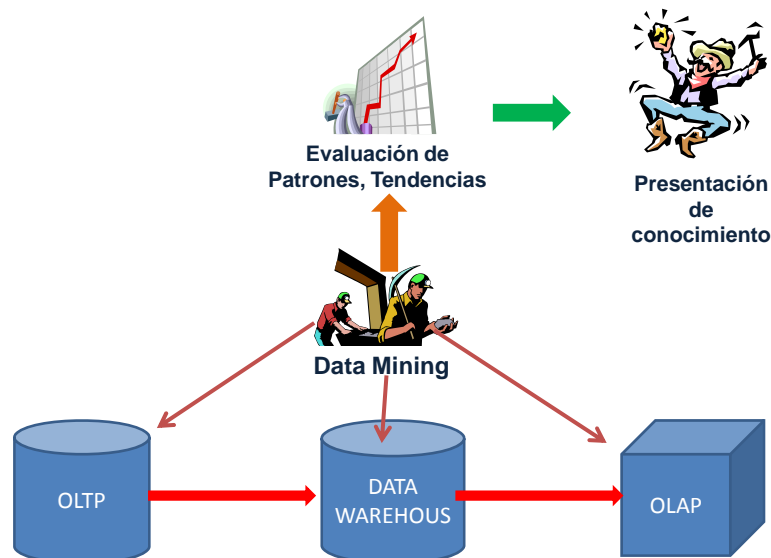
- Imagine una organización de retails en donde identifique los grupos de clientes que adquieren ciertos grupos de productos.
- En un casino de juegos las personas mayores de 55 de género femenino que permanecen 20 minutos “jugando” prefieren ciertos juegos rentables.
- Conocer que clientes son los que continuamente cambian de operador. En algunos países la inversión de un operador telefónica por cliente representa un promedio de 200 dólares, de ahí la importancia de retenerlo
- Conocer el perfil de los clientes que constantemente realizan reclamos en una empresa de servicios.
- Tener la posibilidad de plantear un conjunto de promociones a un determinado grupo de clientes.
- Disminuir el riesgo de proporcionar un préstamo a un solicitante en una entidad financiera.

Qué datos usa Data Mining?

Si su organización cuenta con un Data Warehouse o Data Mart, que es donde mayormente se aplica DM, donde la data ya se encuentra “limpia”. En pequeñas organizaciones es posible que no exista un Data Warehouse por lo que se podría “minar” directamente en las tablas transaccionales. En este sentido se recomienda tener una BD a parte con los datos necesarios y validados.

También es posible aplicarlo directamente en un Cubo OLAP, que como vimos en capítulos posteriores es una BD Multidimensional compuesta por Medidas y Dimensiones.

En general el DM busca descubrir y evaluar patrones y tendencias con miras a presentar un nuevo conocimiento de la organización.



Que Datos usa Data Mining

Ciclo de un Proyecto en DM

Seguramente se estarán preguntando cuales son los pasos para construir un proyecto de DM, aquí van!

Paso 1: Colección de Datos

Los datos del negocio podrían estar en muchos sistemas. Para tener una idea, en Microsoft, existen cientos de Base de Datos y algo de 70 Data Warehouse (1)

Paso 2: Limpieza de Datos y Transformación

La data limpia y transformada es el insumo vital para el DM, por lo que solo considerar la data relevante.

Paso 3: Construir un Modelo

Una vez que la data está limpia y las variables a usar transformadas, podemos empezar a construir modelos comprendiendo la meta que percibe el proyecto de Data Mining para luego ejecutar el tipo de tarea de DM. La idea es entender a los analistas del negocio que conocimiento intentan descubrir. En el caso de postulantes a la universidad por ejemplo: quienes serán los que tendrán más éxito en su vida universitaria.

(1)Data Mining con SQL Server 2005 . ZhaoHui Tang

Esta etapa es clave, conociendo el tipo de análisis a realizar es relativamente sencillo elegir el algoritmo a aplicar. Seguramente serán varios escenarios a desarrollar.

Paso 4: Modelo Preparado

Aplicados los algoritmos necesarios con sus respectivos parámetros. La idea es evaluar e identificar el significado de los patrones encontradas y elegir el modelo a seguir.

Paso 5: Reportear

Entregar reportes de lo encontrado a los usuarios finales para su conformidad respectiva

Paso 6: Predicción

En algunos proyectos el entregar los patrones descubiertos es una media mitad del trabajo, la otra corresponde a crear modelos predictivos incorporando nuevos escenarios

Paso 7: Integración de Aplicación

Es necesario crear una aplicación para integrarla al negocio. Por ejemplo en el caso del CRM la segmentación de mercado es un tema muy aplicable con DM o en el caso de un ERP o Sistemas Desarrollados el tema de los presupuestos cobran más exactitud al aplicarse DM

Paso 8: Administración del Modelo

En el caso de que exista variación con los modelos diseñados es necesario mantenerse vigilante, lo cual obligaría a crear nuevas versiones del DM.

Hasta el próximo artículo en donde tocaremos la Planificación del Proyecto de BI/DW basado en 3 puntos:

- Documento Visión del Producto
- Equipo del Proyecto
- Cronograma del Proyecto

Bibliografía Utilizada:

(1) Data Mining con SQL Server 2005 . ZhaoHui Tang - 2005. USA

(2) Curso de Postgrado en IT-ESAN - Nov 2007. Trujillo-Perú

(3) Experiencia Personal - Abril 2008 . Trujillo-Perú