Instituto Tecnológico de Tlaxcala

Ingeniería en sistemas computacionales

Minería de datos

Ricardo Zamora Picazo

Índice

Índice de figuras

Introducción

Hoy en día la tecnología ha hecho que la información digitalizada sea más fácil de capturar, procesar, almacenar, distribuir y transmitir. Gracias a su progreso en informática y su constante uso en diferentes aspectos de la vida se continúa recogiendo y almacenando en bases de datos inmensas cantidades de información.

El avance de la tecnología para la gestión de bases de datos hace posible integrar diferentes tipos de datos tales como imágenes, videos, texto y otro tipo de información en una base de datos sencilla por lo cual los métodos tradicionales de técnicas estadísticas y herramientas de gestión de datos no son eficientes para analizar esta vasta colección de datos.

Actualmente se estima que el suministro de datos del mundo se duplica cada 20 meses, lo que implica un crecimiento excesivo en el volumen de datos que se manejan en sectores productivos como la economía que sobrepasa la capacidad humana de analizar, resumir y extraer conocimientos a tales cantidades de información lo cual hace necesaria una nueva herramienta capaz de automatizar el análisis de los datos almacenados. El conjunto de estas herramientas lo estudia un nuevo campo de investigación llamado minería de datos.

La minería de datos se ha convertido en una herramienta estratégica para la toma de decisiones de mercadeo, producción, organización y otros factores de las empresas que de cierta manera las hace más competitivas, pero aún es muy común que algunas grandes empresas no implementen este tipo de tecnología, sin embargo, todo apunta a que en algún futuro no muy lejano la minería de datos sea usada por la sociedad al menos con la mismo peso que actualmente tiene la estadística.

En la investigación se hablará de ¿Qué es una minería de datos?, ¿Qué beneficios aporta? ¿Cómo puede influir esta tecnología en la resolución de los problemas diarios? ¿Qué tecnologías están detrás de la minería de datos? ¿Cómo la minería de datos una técnica ampliamente relacionada con la investigación de operacionales incide en el diseño de estrategias de mercadeo?

Estas cuestiones se aclararán mediante una introducción a la minería de datos como: definiciones, ejemplos de la vida cotidiana, técnicas usadas y sus principales tendencias.

Capítulo I Descubrimiento del conocimiento en base de datos (KDD)

¿Qué es KDD?

El término de Descubrimiento de Conocimiento en Bases de Datos (Knowledge Discovery in Databases) o por sus siglas KDD empezó a utilizarse en 1989 para referirse al proceso no trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos de algún repositorio de información.

El proceso no es automático, es un proceso iterativo que exhaustivamente explora volúmenes muy grandes de datos para determinar relaciones, este extrae información de calidad que puede usarse para dibujar conclusiones basadas en las relaciones o modelos de datos.

KDD implica la evaluación e interpretación de patrones y modelos para tomar decisiones con respecto a lo que constituye conocimiento y lo que no lo es. Por lo tanto, KDD requiere de un amplio y profundo conocimiento sobre el área de estudio. Por otra parte, la minería de datos no requiere de tanto conocimiento del tema, si no de más conocimiento técnico. La minería de datos es un paso que forma parte del KDD e implica el análisis de grandes cantidades de datos observacionales para encontrar relaciones sospechosas.

El campo del Descubrimiento de Conocimiento en Bases de Datos es la convergencia del aprendizaje automático, la estadística, el reconocimiento de patrones, la inteligencia artificial, las bases de datos, la visualización de los datos, los sistemas para el apoyo a la toma de decisiones, la recuperación de información y otros muchos campos.

Procesos de KDD

Los principales pasos dentro del proceso interactivo e iterativo del KDD son los siguientes:

1. **Selección de los datos** En esta etapa se determinan las fuentes de datos y el tipo de información a utilizar. Es la etapa donde los datos relevantes para el análisis son extraídos desde la o las fuentes de datos.

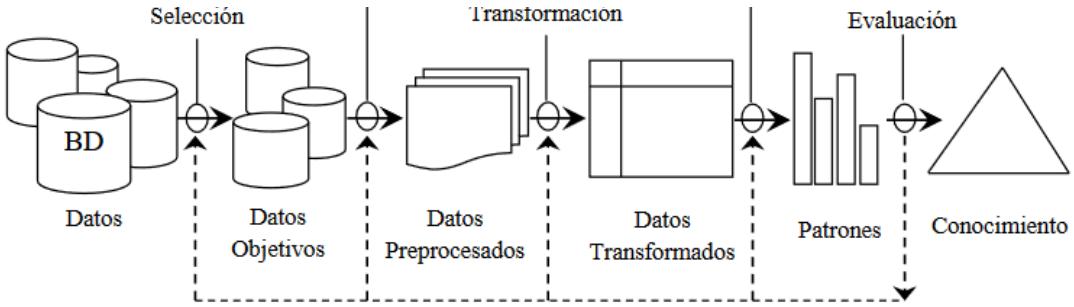
2. **Procesamiento** Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de estados en una forma manejable, necesaria para las fases posteriores. En esta etapa se utilizan diversas estrategias para manejar datos faltantes o en blanco, datos inconsistentes o que están fuera del rango.

3. **Transformación** Consiste en el tratamiento preliminar de los datos, transformación y generalización de nuevas variables a partir de las ya existentes con una estructura de datos apropiada. Aquí se realizan operaciones de agregación o normalización consolidando los datos para la fase siguiente.

4. **Minería de datos** Es la fase de modelamiento en donde métodos inteligentes son aplicados con el objetivo de extraer patrones previamente desconocidos, validos, nuevos, potencialmente útiles comprensibles que contenidos u ocultos en los datos.

5. **Interpretación y evaluación** Se identifican los patrones obtenidos y que son realmente interesantes, basándose en algunas medidas y se realizan una evaluación de los resultados obtenidos.

Ilustración 1Procesos del KDD



Muchas veces los pasos que construyen el proceso de KDD no están tan claramente diferenciados. Las interacciones entre las decisiones tomadas en diferentes pasos, así como los parámetros de métodos utilizados y la forma de representar el problema suelen ser extremadamente complejos.

Históricamente el desarrollo de la estadística no ha proporcionado métodos para analizar datos, encontrar correcciones y dependencias entre ellos. Sin embargo, el análisis de datos a cambiado recientemente y ha adquirido mayor importancia debido a tres factores:

- Incremento de la potencia de las computadoras. Aunque la mayoría de los métodos fueron desarrollados durante los años 60 y 70, las potencias de cálculo de las grandes computadoras de aquella restringían su aplicación a pequeños ejemplos de los cuales resultaban ser demasiado pobres.

- Incremento del ritmo de adquisición de datos. El crecimiento de la cantidad de datos almacenados se ve favorecido no solo por los sistemas de almacenamiento masivo, sino también por la automatización de muchos experimentos y técnicas de recolecciones datos.

- Nuevos métodos. Han surgido principalmente de aprendizaje y representación de conocimiento desarrollados por la comunidad de inteligencia artificial. Estos métodos complementan a las tradicionales técnicas de estadística en el sentido de que son capaces de inducir relaciones cualitativas generales.

Estos nuevos métodos matemáticos y técnicas de software, para el análisis inteligente de datos se denominan actualmente como técnicas de minería de datos, la minería de datos ha permitido el rápido desarrollo de lo que se le conoce como descubrimiento de conocimiento en bases de datos.

Las técnicas de minería de datos han surgido a partir de sistemas de aprendizaje inductivo en computadoras, siendo la principal diferencia entre ellos los datos sobre los datos sobre los que se realiza la búsqueda de nuevo conocimiento. En el caso tradicional de aprendizaje en computadoras (Machine learning), se usa un conjunto de datos pequeños y cuidadosamente seleccionado para entrenar al sistema. Por lo contrario, en la minería de datos se parte de una base de datos generalmente grande en la que los datos han sido generados y almacenados para propósitos diferentes del aprendizaje con los mismos.

Fases del KDD

Existen diferentes técnicas de las diferentes disciplinas que se utilizan en las diversas fases. Si bien los términos de Minería de datos y descubrimiento de conocimiento en bases de datos son usados como sinónimos, el termino de KDD describe el proceso completo de extracción de conocimiento a partir de los datos. Mientras que Data Mining, se refiere exclusivamente al estado descubrimiento de un proceso general KDD.

En el proceso KDD es posible definir al menos 6 estados: Recolección de datos, Selección, Limpieza y Transformación de datos, Minería de datos, Evaluación y Validación, Interpretación y difusión, Actualización y Monitorización.

Recolección de Datos

En este paso es cuando reconocemos las fuentes de información más importantes y quienes tienen el control sobre ellas. La primera fase del KDD determina que las fases sucesivas sean capaces de extraer conocimiento valido y útil a partir de la información original. Generalmente, la información que se requiere investigar sobre dominio de la organización se encuentra:

- En bases de datos.
- Muchas de estas fuentes son las que se utilizan para el trabajo transaccional.

Selección, Limpieza y Transformación de Datos.

Se debe eliminar el mayor número de posibles errores o inconsistentes e irrelevantes.

- Histogramas (Detención de datos anómalos).
- Selección de datos (muestreo, ya sea verticalmente, eliminando atributos u horizontalmente, eliminando tuplas).
- Redefinición de atributos (agrupación o separación).

Acciones ante datos anómalos (outliers):

- Ignorar: algunos algoritmos son robustos a datos anómalos (por ejemplo: arboles)
- Filtrar la columna (eliminar o remplazar): solución extrema, pero a veces existe otra columna dependiente con mayor calidad.

- Filtrar la fila: muchas veces las causas de un dato erróneo están relacionadas con casos o tipos especiales.
- Remplazar el valor: por el valor nulo si el algoritmo lo trata bien o por máximos o mínimos, dependiendo por donde es el outlier. A veces se puede predecir a partir de otros datos, utilizando cualquier técnica de ML.
- Discretizar; transformar un valor continuo en uno discreto (por ejemplo: muy alto, alto, medio, bajo, muy bajo) hace que los outliers caigan en muy alto o muy bajo sin problemas.

Acciones ante datos faltantes:

- Remplazar el valor por medidas: A veces se puede predecir a partir de otros datos utilizando cualquier técnica de ML.
- Segmentar: se segmentan las tuplas por los valores que tienen disponibles. Se obtienen modelos diferentes para cada segmento y luego se combinan.
- Modificar la política de calidad de los datos y esperar hasta que los datos faltantes estén disponibles.

Razones sobre datos faltantes:

- Algunos valores expresan características relevantes. Por ejemplo: la falta de teléfono puede representar en muchos casos un deseo de que no se ke moleste a la persona en cuestión, o un cambio de domicilio reciente.
- Valores no existentes: muchos valores faltantes existen en la realidad, pero otros no. Por ejemplo: un cliente que se acaba de dar de alta no tiene consumo medio de los últimos 12 meses.
- Datos incompletos: si los datos vienen de fuentes diferentes al combinarlos se puede hacer la unión y no la inserción de campos, con lo que muchos datos faltantes representan que esas tuplas vienen de una fuente diferente.

Minería de datos.

Características especiales de la minería de datos.

Aparte del gran volumen de los datos ¿Por qué las técnicas de aprendizaje automático y estadística no son directamente aplicables?

- Los datos residen en el disco. No se pueden escanear múltiples veces.
- Algunas técnicas de muestreo no son compatibles con algoritmos no incrementales.
- Muy alta dimensionalidad (muchos campos).
- Evidencia positiva.
- Datos imperfectos.

Aunque algunos se aplican casi directamente, el interés en la minería de datos está en su adaptación.

Patrones a descubrir:

- Una vez recolectados los datos de interés, un explorador puede decidir qué tipo de patrón quiere descubrir.
- El tipo de conocimiento que se desea va a marcar caramente la técnica de minería de datos a utilizar.
- Según como sea la búsqueda del conocimiento se puede distinguir entre:
    - Directed data mining: se sabe claramente lo que se busca, generalmente predecir unos ciertos datos o clases.
    - Undirected data mining: no se sabe lo que se busca, se trabaja con los datos.

En el primer caso, los propios sistemas de minería de datos se encargan generalmente de elegir el algoritmo más idóneo entre los disponibles para un determinado tipo de patrón a buscar.

Evaluación y validación.

La fase anterior procede una o más hipótesis de modelos. Para seleccionar y validar estos modelos es necesario el uso de criterios de evaluación de hipótesis. Por ejemplo:

1ª Fase: comprobación de la precisión del modelo en un banco de ejemplos independiente del que se ha utilizado para aprender el modelo.

2ª Fase: se puede realizar una experiencia piloto con ese modelo. Por ejemplo: si el modelo encontrado se quería utilizar para predecir la respuesta de los clientes a un nuevo producto, se puede enviar un nailon aun subconjunto de clientes y evaluar la fiabilidad del modelo.

Interpretación y difusión.

El despliegue del modelo a veces es trivial pero otras veces requiere un proceso de implementación o interpretación:

- El modelo puede requerir implementación, por ejemplo: Tiempo real de detección de tarjetas fraudulentas.
- El modelo es descriptivo y requiere interpretación, por ejemplo: Una caracterización de zonas geográficas según la distribución de los productos vendidos.
- El modelo puede tener muchos usuarios y necesita difusión, puede requerir ser expresado de una manera comprensible para ser distribuido en la organización, por ejemplo: Las cervezas y los productos congelados se compran frecuentemente en conjunto.

Actualización y monitorización.

Los procesos derivan en un mantenimiento:

- Actualización: un modelo valido puede dejar de serlo, cambia de contexto (económico, competencia, fuentes de datos, etc.).

- Monitorización: consiste en ir revalidando el modelo con cierta frecuencia sobre nuevos datos, con el objetivo de detectar si el modelo requiere una actualización.

Producen retroalimentaciones en el proceso KDD.

Capítulo II Introducción a la Minería de datos

Conceptos e Historia.

Las raíces de la Data Mining (DM) se remonta a los años 50. Los departamentos de informática preparaban resúmenes de la informática, principalmente de tipo comercial que se encontraba en los ficheros del ordenador central, con el propósito de facilitar la labor directiva. Así nacieron los sistemas de información para la dirección, que, sin embargo, eran voluminosos, poco flexibles, y difíciles de leer para los no informáticos. En los 60 nacen los sistemas gestores de base de datos que aún se mostraban rígidos y carecían de flexibilidad para realizar consultas. Luego aparecieron los motores relacionales resolviendo estos problemas, aunque los informes resultaban muy laboriosos de preparar y depurar, perdiéndose relevancia por su bajo nivel de actualización. Otro grave problema era la diversidad de bases de datos no integradas establecidas por los diferentes departamentos de una organización.

El Data Warehouse (DW) viene a solucionar este problema en los finales de los 80. La existencia de DW ha estimulado el desarrollo de los enfoques de DM, en los que las tareas de análisis se automatizan y dan un paso más al posibilitar la extracción de conocimiento inductivo.

Etapa	Cuestión planteada	Tecnologías	Características
Recolección de datos {Años 60}	'Dime mis beneficios totales en los últimos 4 años.'	Ordenadores, cintas, discos.	Retrospectivo, datos estáticos.
Acceso a los datos. {Años 80}	'Ventas en Cataluña durante las últimas Navidades'	Bases de Datos Relacionales (SQL) ODBC	Retrospectivo, datos dinámicos a nivel de registro.
Data Warehouse y soporte a la toma de decisiones. {Años 90}	'Ventas en Andalucia detalle por delegación y descender a nivel tienda.	{OLAP}, bases de datos multi-dimensionales, data warehouse	Retrospectivo, obtención dinámica de datos a múltiples niveles.
Data Mining	Justifica la tendencia de venta en Castilla para el próximo año	Algoritmos avanzados, ordenadores, multiprocesadores, bases de datos masivas.	Prospectivo, obtención proactiva de información.

Ilustración 2Evolucion de las Tecnologías relacionadas con DM

El data mining es una tecnología compuesta por etapas que integra varias áreas y que no se debe confundir con un gran software. Durante el desarrollo de un proyecto de este tipo se usan diferentes aplicaciones software en cada etapa que pueden ser estadísticas, de visualización de datos o de inteligencia artificial, principalmente. Actualmente existen aplicaciones o herramientas comerciales de data mining muy poderosas que contienen un sinfín de utilerías que facilitan el desarrollo de un proyecto. Sin embargo, casi siempre acaban complementándose con otra herramienta.

La data mining es la etapa de descubrimiento en el proceso de KDD: Paso consistente en el uso de algoritmos concretos que generan una enumeración de patrones a partir de los datos preprocesados (Fayyad et al., 1996). Aunque se suelen usar indistintamente los términos KDD y Minería de Datos.

¿Qué es la minería de datos?

¿Qué no es minería de datos?

Proceso de minería de datos

Capítulo III Métodos y técnicas de minería de datos

Referencias

MArtinez, B. B. (s.f.). *Notas MD.* Obtenido de http://bbeltran.cs.buap.mx/NotasMD.pdf

Mining, M. D. (s.f.). *Minerva Data Mining*. Obtenido de https://mnrva.io/kdd-platform.html

WebMining. (s.f.). *WebMining*. Obtenido de http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/