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.

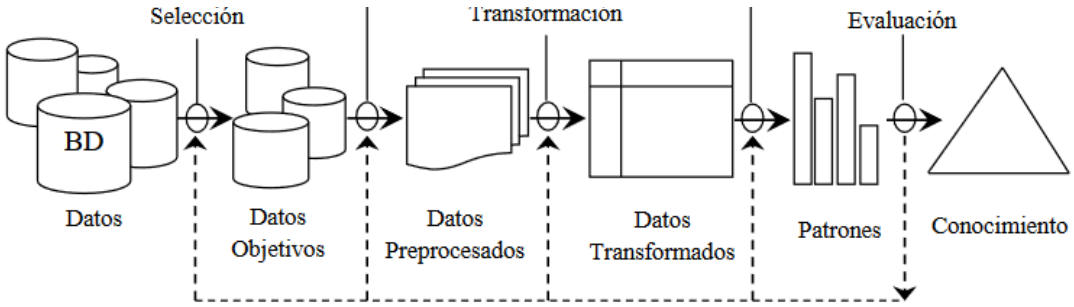
2. **Procesamiento** Esta etapa consiste en la preparación y limpieza de los datos extraídos desde las distintas fuentes de estados en una forma manejable, necesaria para las fases posteriores. En esta etapa se utilizan diversas estrategias para manejar datos faltantes o en blanco, datos inconsistentes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

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

La 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?

Existen diferentes técnicas que posibilitan la exploración de los datos, extrayendo información que no es detectada a simple vista. Una de estas técnicas es la denominada Minería de Datos, la cual combina técnicas semiautomáticas de inteligencia artificial, análisis estadístico, bases de datos y visualización gráfica, para la obtención de información que no esté representada explícitamente en los datos

La minería de datos descubre relaciones, tendencias, desviaciones, comportamientos atípicos, patrones y trayectorias ocultas con el propósito de soportar los procesos de toma de decisiones con mayor conocimiento. La minería de datos se puede ubicar en el nivel más alto de la evolución de los procesos tecnológicos de análisis de datos.

Algunas definiciones de minería de datos pueden ser:

1. Es el proceso de detectar la información procesable de los conjuntos grandes de datos. Utiliza el análisis matemático para deducir patrones y tendencias que existen en los datos. Normalmente estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiados datos.
2. Proceso que permite transformar información en conocimiento útil para el negocio, a través del descubrimiento y cuantificación de relaciones en una gran base de datos
3. Conjunto de técnicas que automatizan la detección de patrones relevantes.

¿Qué no es minería de datos?

Data Mining no es estadística

Comúnmente ambos términos se confunden ya que Data Mining es el sucesor de la estadística tal como se usa actualmente. La estadística y el Data Mining tienen el mismo objetivo el cual es construir modelos compactos y comprensibles que rindan cuenta de las relaciones establecidas entre la descripción de una situación y un resultado con dicha descripción.

La diferencia entre ambas consiste en que las técnicas del Data Mining construyen el modelo de manera automática mientras que las técnicas estadísticas clásicas necesitan ser manejadas y orientadas por un estadístico profesional. Las técnicas de Data Minig permiten ganar tanto en performance como en manejabilidad e incluso en tiempo de trabajo, la posibilidad de realizar uno mismo sus propios modelos sin necesidad de contratar ni ponerse de acuerdo con un estadístico, proporciona gran libertad a los usuarios profesionales.

Data Mining no es OLAP

Las herramientas OLAP permiten navegar rápidamente por los datos, pero no se genera información en el proceso. Se llaman sistemas OLAP (On Line Analytical Processing) a aquellos sistemas que deben ser:

- Soportar requerimientos complejos de análisis
- Analizar datos desde diferentes perspectivas
- Soportar análisis complejos contra un volumen ingente de datos

La funcionalidad de los sistemas OLAP se caracteriza por ser un análisis multidimensional de datos mediante navegación del usuario por los mismos de modo asistido.

Existen dos arquitecturas diferentes para los sistemas OLAP: OLAP multidimensional (MD-OLAP) y OLAP relacionales (ROLAP).

La arquitectura MD-OLAP usa bases de datos multidimensionales, la arquitectura ROLAP implanta OLAP sobre bases de datos relacionales. La arquitectura MDOLAP requiere unos cálculos intensivos de compilación.

La arquitectura ROLAP, accede a los datos almacenados en un Data Warehouse para proporcionar los análisis OLAP. La premisa de los sistemas ROLAP es que las capacidades OLAP se soportan mejor contra las bases de datos relacionales.

Los usuarios finales ejecutan sus análisis multidimensionales a través del motor ROLAP, que transforma dinámicamente sus consultas a consultas SQL. Se ejecutan estas consultas SQL en las bases de datos relacionales, y sus resultados se relacionan mediante tablas cruzadas y conjuntos multidimensionales para devolver los resultados a los usuarios. ROLAP es una arquitectura flexible y general que crece para dar soporte a amplios requerimientos OLAP. El MOLAP es una solución particular, adecuada para soluciones departamentales con unos volúmenes de información y número de dimensiones más modestos

Proceso de minería de datos

El proceso de Data Mining se inicia con la identificación de los datos. Para ello hay que imaginar qué datos se necesitan, dónde se pueden encontrar (en una o varias bases de datos; en papel, etc.) y cómo conseguirlos. Una vez que se dispone de los datos, se deben preparar, poniéndolos en bases de datos en un formato adecuado o construir una warehouse. Esta es una de las tareas más difíciles del Data Mining. Una vez que se tiene los datos en el formato adecuado hay que realizar una selección de los datos esenciales y eliminación de los innecesarios.

Antes de proceder al análisis de los datos por Data Mining, conviene tener una idea de qué es lo que interesa averiguar, qué herramientas se necesitan y cómo proceder. Tras aplicar la herramienta elegida o construida por nosotros mismos hay que saber interpretar los resultados o patrones obtenidos para saber los que son significativos y cómo podarlos para extraer únicamente los resultados útiles. Tras examinar los resultados útiles hay que identificar las acciones que deben de ser tomadas, discutirlas y pensar en los procedimientos para llevarlas a cabo e implementarlas.

Una vez implementadas hay que evaluarlas para ello hay que observar los resultados, los beneficios y el coste para poder reevaluar el procedimiento completo. Para entonces los datos pueden haber cambiado, nuevas herramientas pueden estar disponibles y probablemente habrá que planificar el siguiente ciclo de minería.

A continuación, se proporciona una visión general del proceso de minería de datos y una breve descripción de cada estado:

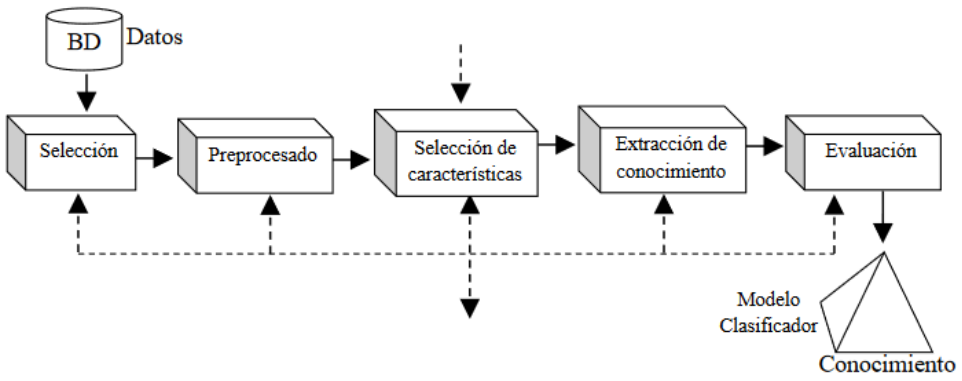


Ilustración 3Proceso de la Minería de Datos

Procesado de los datos

El formato de los datos contenidos en la fuente de datos (base de datos, Data Warehouse, etc.) nunca es el idóneo, y la mayoría de las veces no es posible ni siquiera utilizar ningún algoritmo de minería sobre los datos en bruto. Mediante el preprocesado, se filtran los datos (de forma que se eliminan valores incorrectos, no válidos, desconocidos; según las necesidades y el algoritmo a usar), se obtienen muestras de los mismos (en busca de una mayor velocidad de respuesta del proceso), o se reducen el número de valores posibles (mediante redondeo, clustering, etc.).

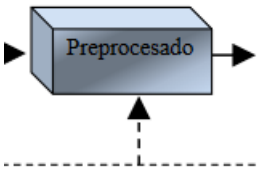


Ilustración 4Preprocesado

Selección de características

Después de haber sido preprocesados, en la mayoría de los casos se tiene una cantidad ingente de datos. La selección de características reduce el tamaño de los datos eligiendo las variables más influyentes en el problema, sin apenas sacrificar la calidad del modelo de conocimiento obtenido del proceso de minería. Los métodos para la selección de características son básicamente dos:

1. Aquellos basados en la elección de los mejores atributos del problema,
2. Aquellos que buscan variables independientes mediante test de sensibilidad, algoritmos de distancia o heurísticos.

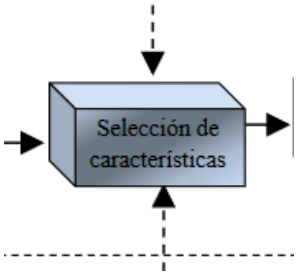


Ilustración 5 Selección de características

Algoritmo de aprendizaje

Mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un preprocesado diferente de los datos.

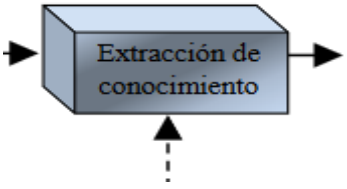


Ilustración 6 Extracción de conocimiento

Evaluación y validación

Una vez obtenido el modelo, se debe proceder a su validación, comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

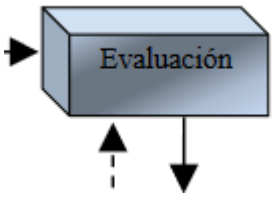


Ilustración 7 Evaluación

La minería de datos se puede definir como un proceso analítico diseñado para explorar grandes cantidades de datos (generalmente datos de negocio y mercado) con el objetivo de detectar patrones de comportamiento consistentes o relaciones entre las diferentes variables para aplicarlos a nuevos conjuntos de datos.

Capítulo III Métodos y técnicas de minería de datos

Métodos de minería de datos

Los métodos de minería de datos tienen como metas primarias (en un alto nivel) la predicción de datos desconocidos y la descripción de patrones. Pueden emplearse diferentes criterios para clasificar los sistemas de minería de datos y, en general, los sistemas de aprendizaje inductivo en computadoras:

- Dependiendo del objetivo para el que se realiza el aprendizaje, pueden distinguirse sistemas para: clasificación, regresión, agrupamiento de conceptos, compactación, modelado de dependencias, detección de desviaciones, etc.

- Dependiendo de la tendencia con que se aborde el problema, se pueden distinguir tres grandes líneas de investigación o paradigmas: sistemas conexionistas (redes neuronales), sistemas evolucionistas (algoritmos genéticos) y sistemas simbólicos.

- Dependiendo del lenguaje utilizado para representar del conocimiento, se pueden distinguir: representaciones basadas en la lógica de proposiciones, representaciones basadas en lógica de predicados de primer orden, representaciones estructuradas, representaciones a través de ejemplos y representaciones no simbólicas como las redes neuronales.

Agrupamiento

También llamada Segmentación, esta herramienta permite la identificación de tipologías o grupos donde los elementos guardan similitud entre sí y diferencias aquellos de otros grupos.

Para alcanzar las distintas tipologías o grupos existentes en una base de datos, estas herramientas requieren, como entrada, información sobre el colectivo a segmentar. Esta información corresponderá a los valores concretos, para cada elemento en un momento del tiempo, de una serie de variables ("Segmentación estática") o a través del comportamiento en el tiempo de cada uno de los elementos del colectivo ("Segmentación dinámica").

Asociación

Este tipo de herramientas establece las posibles relaciones o correlaciones entre distintas acciones o sucesos aparentemente independientes, pudiendo reconocer como la ocurrencia de un suceso o acción puede inducir o generar la aparición de otros. Normalmente este tipo de herramientas se fundamenta en técnicas estadísticas como los análisis de correlación y de variación.

Secuenciamiento

Esta herramienta permite identificar como, en el tiempo, la ocurrencia de una acción desencadena otras posteriormente. Es muy similar a la anteriormente analizada si bien, en este caso, el tiempo es una variable crítica e imprescindible a introducir en la información a analizar.

Reconocimiento de patrones

Estas herramientas son usadas por elementos que son tan habituales como un procesador de texto o un despertador. Los patrones pueden ser cualquier elemento de información que deseemos. En el ámbito particular del DM estas herramientas pueden ayudamos en la identificación de problemas e incidencias y de sus posibles soluciones toda vez que dispongamos de la base de información necesaria en la cual buscar.

Previsión

La Previsión establece el comportamiento futuro más probable dependiendo de la evolución pasada y presente. Esta herramienta tiene su uso fundamental en el tratamiento de Series Temporales y las técnicas asociadas disponen de una importante madurez.

Simulación

Las herramientas de Simulación forman parte también del conjunto de herramientas veteranas de la investigación científica. Como ejemplo están las herramientas de diseño y producción asistidas por ordenador, "CAD" -"CAM", en las cuales se revisan los diseños sometiéndoles a una amplísima serie de condiciones reales normales y extremas. Ello permite no sólo ajustar y adaptar el diseño sino posteriormente establecer márgenes y límites de funcionamiento.

Optimización

Al igual que la Previsión y la Simulación, las herramientas de Optimización tienen una amplia tradición de uso. La optimización ha sido y es extensivamente usada en la resolución de los problemas asociados a la logística de distribución y a la gestión de

"Stocks" en los negocios y en la determinación de parámetros teóricos a partir de los experimentos en la investigación científica.

La optimización resuelve el problema de la minimización o maximización de una función que depende de una serie de variables, encontrando los valores de éstas que satisfacen esa condición de máximo, típicamente beneficios, o mínimo, normalmente costes

Clasificación

La clasificación agrupa todas aquellas herramientas que permiten asignar a un elemento la pertenencia a un grupo o clase. Ello se instrumenta a través de la dependencia de la pertenencia a las clases en los valores de una serie de atributos o variables.

A través del análisis de un colectivo de elementos, o casos de los cuales conocemos la clase a la que pertenecen, se establece un mecanismo que establece la pertenencia a tales clases en función de los valores de las distintas variables y nos permite establecer el grado de discriminación o influencia de éstas.

También se utiliza para estas herramientas la denominación de Predicción o Evaluación para aquellos casos donde se aplican técnicas, normalmente numéricas, que establecen para cada elemento un valor dependiente de los valores que tengan las variables en tal elemento.

Técnicas de minería de datos

La minería de datos ha dado lugar a una paulatina sustitución del análisis de datos dirigido a la verificación por un enfoque de análisis de datos dirigido al descubrimiento del conocimiento. La principal diferencia entre ambos se encuentra en que en el último se descubre información sin necesidad de formular previamente una hipótesis. La aplicación automatizada de algoritmos de minería de datos permite detectar fácilmente patrones en los datos, razón por la cual esta técnica es mucho más eficiente que el análisis dirigido a la verificación cuando se intenta explorar datos procedentes de repositorios de gran tamaño y complejidad elevada.

Dichas técnicas emergentes se encuentran en continua evolución como resultado de la colaboración entre campos de investigación tales como bases de datos, reconocimiento de patrones, inteligencia artificial, sistemas expertos, estadística, visualización, recuperación de información, y computación de altas prestaciones. los algoritmos de minería de datos se clasifican en dos grandes categorías: supervisados o predictivos y no supervisados o de descubrimiento del conocimiento.

SUPERVISADOS	NO SUPERVISADOS
Árboles de decisión	Detección de Desviaciones
Inducción neuronal	Segmentación
Regresión	Agrupamiento (Clustering)
Series temporales	Reglas de Asociación
	Patrones Secuenciales

Ilustración 8Tecnicas de minera de datos

Método estadístico

La estadística es tradicionalmente la técnica que se ha usado para el tratamiento de grandes volúmenes de datos numéricos y nadie pone en duda su efectividad al poseer un amplísimo conjunto de modelos de análisis para cubrir el tratamiento de todo tipo de poblaciones y series de datos. Estos son algunos de los métodos estadísticos más utilizados:

1. **ANOVA**: Análisis de la Varianza, contrasta si existen diferencias significativas entre las medidas de una o más variables continuas en grupos de población distintos.
2. **Análisis de clusters**: Permite clasificar una población en un número determinado de grupos, sobre la base de semejanzas y diferencias de perfiles existentes entre los diferentes componentes de dicha población.
3. Análisis discriminante: Método de clasificación de individuos en grupos que previamente se han establecido, y que permite encontrar la regla de clasificación de los elementos de estos grupos, y por tanto identificar cuáles son las variables que mejor definan la pertenencia al grupo.

Métodos basados en arboles de decisión

Son herramientas analíticas empleadas para el descubrimiento de reglas y relaciones mediante la ruptura y subdivisión sistemática de la información contenida en el conjunto de datos. El árbol de decisión se construye partiendo el conjunto de datos en dos (CART) o más (CHAID) subconjuntos de observaciones a partir de los valores que toman las variables predictoras. Cada uno de estos subconjuntos vuelve después a ser particionado utilizando el mismo algoritmo.

Este proceso continúa hasta que no se encuentran diferencias significativas en la influencia de las variables de predicción de uno de estos grupos hacia el valor de la variable de respuesta.

Reglas de asociación

Derivan de un tipo de análisis que extrae información por coincidencias. Este análisis a veces llamado "cesta de la compra" permite descubrir correlaciones o co-ocurrencias en los sucesos de la base de datos a analizar y se formaliza en la obtención de reglas de tipo; SI ... ENTONCES...

Redes neuronales

Las Redes Neuronales constituyen una técnica inspirada en los trabajos de investigación, iniciados en 1930, que pretendían modelar computacionalmente el aprendizaje humano llevado a cabo a través de las neuronas en el cerebro.

Las redes neuronales son una nueva forma de analizar la información con una diferencia fundamental con respecto a las técnicas tradicionales: son capaces de detectar y aprender patrones y características dentro de los datos. Se construyen estructurando en una serie de niveles o capas compuesta por nodos o "neuronas". Poseen dos formas de aprendizaje derivadas del tipo de paradigma que usan: el supervisado y el no supervisado.

Algoritmos genéricos

Los Algoritmos Genéticos son otra técnica que debe su inspiración, de nuevo, a la Biología como las Redes Neuronales.

Estos algoritmos representan la modelización matemática de como los cromosomas en un marco evolucionista alcanzan la estructura y composición más óptima en aras dela supervivencia. Entendiendo la evolución como un proceso de búsqueda y optimización de la adaptación de las especies que se plasma en mutaciones y cambios en los genes o cromosomas.

Los Algoritmos Genéticos hacen uso de las técnicas biológicas de reproducción (mutación y cruce) para ser utilizadas en todo tipo de problemas de búsqueda y optimización.

Algoritmos matemáticos

Sin llegar a ser técnicas que den soporte a unas necesidades concretas como las anteriores, existe una amplia gama de algoritmos matemáticos que son especialmente útiles y eficaces en la resolución y tratamiento de problemas muy específicos y. puntuales y que, normalmente, son incorporados en alguna de aquellas técnicas con el objeto de mejorarlas.

Referencias

Gestion.org. (s.f.). *Gestion.org*. Obtenido de https://www.gestion.org/tecnicas-de-mineria-de-datos/

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

Micrsosoft. (2017). *Conceptos de mineria de datos*. Obtenido de https://docs.microsoft.com/es-es/sql/analysis-services/data-mining/data-mining-concepts?view=sql-server-2017

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

Roman, J. V. (s.f.). *Mineria de datos.* Obtenido de http://ocw.uc3m.es/ingenieria-telematica/inteligencia-en-redes-de-comunicaciones/material-de-clase-1/07-mineria-de-datos

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