Minería y almacenamiento de los datos

Big data e inteligencia artificial

tech

CONTENIDO

Objetivos

- Entender el concepto de minería de datos.
- Entender el tratamiento y ejemplo de aplicación de minería de datos.

Conceptos básicos de la minería de datos

¿Qué es el *datamining*?

El *datamining* (minería de datos), es el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

Básicamente, el *datamining* surge para intentar ayudar a comprender el contenido de un repositorio de datos. Con este fin, hace uso de prácticas estadísticas y, en algunos casos, de algoritmos de búsqueda próximos a la inteligencia artificial y a las redes neuronales.

De forma general, los datos son la materia prima bruta. En el momento que el usuario les atribuye algún significado especial pasan a convertirse en información. Cuando los especialistas elaboran o encuentran un modelo, haciendo que la interpretación que surge entre la información y ese modelo represente un valor agregado, entonces se hace referencia al conocimiento.

Etapas del *datamining*

Aunque en *datamining* cada caso concreto puede ser radicalmente distinto al anterior, el proceso común a todos ellos se suele componer de cuatro etapas principales:

- **Selección del conjunto de datos:** para remover el ruido, datos inconsistentes y seleccionar aquellos que sean útiles para el análisis.
- **Análisis de las propiedades de los datos:** para caracterizar variables, mediante histogramas u otras herramientas.
- **Transformación del conjunto de datos de entrada:** también conocida como consolidación de datos, es una fase en la que a los datos seleccionados se les da una estructura apropiada para el proceso de minería.
- **Seleccionar y aplicar la técnica de minería de datos:** donde se aplican métodos inteligentes para extraer los patrones información.
- **Extracción de conocimiento:** esto se logra identificando patrones relevantes, basados en variables de interés.
- **Interpretación y evaluación de datos:** mediante tablas y gráficas que representan un insumo valioso para la toma de decisiones.

Big data y *datamining*

El *big data* se centra en analizar los grandes volúmenes de datos que superan la capacidad de los procesamientos informáticos habituales. Su objetivo es el de analizar en el menor tiempo posible y de forma eficaz toda la información. Es por ese motivo que hace uso de *software*, que le permite definir las características a nivel cliente y usuario.

En cambio, **la minería de datos o** *data mining* analiza los grandes volúmenes de datos. Sintetiza e identifica y agrupa patrones de comportamiento entre los datos. Generalmente los datos que analiza pertenecen a clientes y consumidores. Como ejemplo para el uso de *data mining*, podría aplicarse al caso de necesitar patrones de conducta de clientes, periodos de contratación de un servicio determinado o periodos de compra, fuga a otras compañías o, incluso, riesgos de estafas a partir de patrones sospechosos o inusuales.

En resumen, lo comentado anteriormente, significa que *data mining* consiste en el **conjunto de técnicas** para la extracción de la información y que *big data* es la **tecnología** capaz de capturar, gestionar y procesar en un tiempo razonable y de forma veraz estos datos.

Preprocesamiento, limpieza y normalización

Los datos del mundo real y los datos en sus etapas más tempranas suelen estar sucios. Pueden ser incompletos, inconsistentes y estar llenos de errores. Una de las formas más exitosas de salvaguardar datos concisos para su análisis es normalizar datos y preprocesarlos.

El procesamiento de datos comprende cuatro técnicas que si se usan correctamente dan como resultado unos datos perfectamente transformados.

Las técnicas de procesamiento de datos son las siguientes:

- ***Data cleaning:*** la limpieza de datos elimina ruido y resuelve las inconsistencias en los datos.
- ***Data integration:*** con la Integración de datos se migran datos de varias fuentes a una fuente coherente como un data warehouse.
- ***Data transformation:*** la transformación de datos sirve para normalizar datos de cualquier tipo.
- ***Data reduction:*** la reducción de datos reduce el tamaño de los datos agregándolos.

Todas estas técnicas pueden trabajar juntas o individualmente para crear un fuerte conjunto de datos. Una gran parte del preprocesamiento de datos es el aspecto de la transformación. Cuando se trata de datos sin procesar nunca se sabe lo que se va a obtener. Por lo tanto, normalizar datos a través del proceso de transformación, es una de las maneras más rápidas y eficientes para alcanzar el objetivo final de unos datos limpios y utilizables.

- **¿Qué es normalizar datos?**

  Normalizar datos es una técnica que se aplica a un conjunto de datos para reducir su redundancia. El objetivo principal de esta técnica es asociar formas similares a los mismos datos en una única forma de datos. Esto es, en cierto modo, cogiendo datos específicos como "número", "num.", "nro.", "N.º" o "#" y normalizándolo a "Número" en todos los casos.

Extracción de información, traducción automática, análisis de sentimientos, etc.

Tareas del *datamining*

Las tareas de *data mining* se suelen dividir en dos grandes categorías:

- **Tareas predictivas:** cuyo objetivo es predecir el valor de un atributo (característica) en particular, basándose en los valores de otros atributos.
  - **Clasificación y predicción:** son dos tipos de análisis de datos, aquellos que pueden ser usados para clasificar datos y los que se usan para predecir tendencias. La clasificación de datos predice clases de etiquetas mientras la predicción de datos predice funciones de valores continuos. Aplicaciones típicas incluyen análisis de riesgo para préstamos y predicciones de crecimiento. Algunas técnicas para clasificación de datos incluyen: clasificaciones bayesianas, K-Nearest Neighbor, algoritmos genéticos, entre otros.
  - **Arboles de decisión:** definen un conjunto de clases, asignando a cada dato de entrada una clase y determina la probabilidad de que ese registro pertenezca a la clase.

    Se pueden distinguir dos tipos de árboles:

    » El primero es el árbol de decisión de clasificación, donde cada registro a clasificar fluye por una rama del árbol. La rama por seguir es determinada por una serie de preguntas definidas por los nodos de la rama. Cuando el registro llega a un nodo hoja, se le asigna a la clase del nodo hoja.

    » El segundo es el árbol de decisión de regresión, cuando el registro llega a un nodo hoja, a la variable de salida de ese nodo, se le asigna el promedio de los valores de la variable de salida de los registros que cayeron en ese nodo hoja durante el proceso de entrenamiento.

- **Redes neuronales:** son modelos predictivos no lineales que aprenden a través del entrenamiento. Existen diferentes tipos de redes neuronales, las más conocidas son las simples y multicapas. Las tareas básicas de las redes neuronales son reconocer, clasificar, agrupar, asociar, almacenar patrones, aproximación de funciones, sistemas (predicción, control, entre otros) y optimización de transacciones comerciales y reconocimiento de patrones.

- **Tareas descriptivas:** cuyo objetivo es obtener patrones que representen las relaciones subyacentes existentes en los datos.
  - **Descripción de clases:** hay tres formas de ver este punto.
    - » La primera se denomina caracterización de los datos (*data caracterization*), el cuál realiza un resumen de las características generales de una clase particular de datos; los resultados suelen representarse en términos de reglas de caracterización.
    - » La segunda es la discriminación de datos (*data discrimination*), que es una comparación entre las características generales de los objetos de una clase respecto a las de otro conjunto contrastante.
    - » Finalmente, también se puede aplicar una combinación de ambas.
  - **Análisis de asociación:** es el descubrimiento de reglas de asociación que muestran condiciones del tipo atributo-valor que ocurre con frecuencia dentro de un conjunto de datos.

    La minería mediante reglas de asociación es el proceso de búsqueda interesante de correlaciones entre un conjunto grande de datos. El descubrimiento de reglas de asociación en grandes volúmenes de transacciones de negocios puede facilitar el proceso de toma de decisiones.
  - **Análisis de clústeres:** aquí se analizan objetos sin consultar clases conocidas.

    El proceso trabaja agrupando objetos según el principio de "maximizar la similitud dentro de una clase y minimizar la similitud entre clases". Un clúster es una colección de objetos de datos mutuamente similares. *Clustering* es el proceso de agrupamiento de objetos.

Ejemplo de aplicaciones del *DATAMINIG*

Ahora se puede entender el presente para anticiparse al futuro. Estos son algunos ejemplos de *data mining* en la industria actual:

- ***Marketing:*** la minería de datos se utiliza para explorar bases de datos cada vez mayores y mejorar la segmentación del mercado. Analizando las relaciones entre parámetros como edad de los clientes, género, gustos, etc., es posible adivinar su comportamiento para dirigir campañas personalizadas de fidelización o captación. El d*ata mining* en *marketing* predice también qué usuarios pueden darse de baja de un servicio, qué les interesa según sus búsquedas o qué debe incluir una lista de correo para lograr una tasa de respuesta mayor.

- **Comercio minorista:** los supermercados, por ejemplo, emplean los patrones de compra conjunta para identificar asociaciones de productos y decidir cómo situarlos en los diferentes pasillos y estanterías de los lineales. El *data mining* detecta, además, qué ofertas son las más valoradas por los clientes o incrementa la venta en la cola de caja.
- **Banca:** los bancos recurren a la minería de datos para entender mejor los riesgos del mercado. Es habitual que se aplique a la calificación crediticia (*rating*) y a sistemas inteligentes antifraude para analizar transacciones, movimientos de tarjetas, patrones de compra y datos financieros de los clientes. El *data mining* también permite a la banca conocer más sobre las preferencias o hábitos de sus clientes en internet para optimizar el retorno de sus campañas de *marketing*, estudiar el rendimiento de los canales de venta o gestionar las obligaciones de cumplimiento de las regulaciones.
- **Medicina:** la minería de datos favorece diagnósticos más precisos. Al contar con toda la información del paciente (historial, examen físico y patrones de terapias anteriores) se pueden prescribir tratamientos más efectivos. También posibilita una gestión más eficaz, eficiente y económica de los recursos sanitarios al identificar riesgos, predecir enfermedades en ciertos segmentos de la población o pronosticar la duración del ingreso hospitalario. Detectar fraudes e irregularidades y estrechar vínculos con los pacientes al ahondar en el conocimiento de sus necesidades son también ventajas de emplear el *data mining* en medicina.
- **Televisión y radio:** hay cadenas que aplican la minería de datos en tiempo real a sus registros de audiencia en televisión *online* (IPTV) y radio. Estos sistemas recaban y analizan sobre la marcha información anónima de las visualizaciones, las retransmisiones y la programación de los canales. Gracias al *data mining* se pueden emitir recomendaciones personalizadas a los radioyentes y telespectadores, conocer en directo sus intereses y su actividad y entender mejor su conducta. Las cadenas obtienen, además, conocimiento muy valioso para sus anunciantes, que aprovechan estos datos para llegar con más precisión a sus clientes potenciales.

Minería de textos

Se trata de una rama de la minería de datos que analiza la información de tipo textual. Es una disciplina transversal y de creciente interés, cuyas aplicaciones son múltiples. Entre otras: indexación de documentos, traducción automática, resumen automático de textos, reconocimiento de voz o identificación de la autoría de textos.

El análisis de sentimientos, también conocido como minería de opinión, se trata de una tarea de clasificación masiva de documentos de manera automática, que se centra en catalogar los documentos en función de la connotación positiva o negativa del lenguaje ocupado en el mismo.

Con las redes sociales, los usuarios tienen hoy en día todo tipo de facilidades para mostrar sus opiniones sobre cualquier tema que deseen. Tener constancia sobre las opiniones referentes a una marca o producto y medir su impacto es actualmente de vital importancia para todas las empresas, ya que es la imagen lo que está en juego.

A toda la información que se recopila de esta forma se le denomina minería de opinión (*opinion mining*) y gracias a ella, las empresas tienen una inmediata disponibilidad de la información deseada. Además, la minería de opinión no solo permite responder ¿qué opinan los internautas sobre su propia marca o producto?, sino que facilita, mediante los medios adecuados, obtener ventajas competitivas en diferentes ámbitos.

Tipos de almacenamiento de datos

El objetivo importante debe ser encontrar un producto que permita recopilar datos de manera más efectiva. Algunos de los problemas clave involucrados en la evaluación y selección de productos de almacenamiento de datos incluyen los siguientes:

- La plataforma de almacenamiento debe ofrecer alto rendimiento y escalabilidad y administrar los costos de manera efectiva.
- El rendimiento debe abarcar tanto el alto rendimiento como la baja latencia.
- Producir buenos modelos de IA significa recopilar muchos *terabytes* o *petabytes* de datos, lo que puede ser costoso. Las organizaciones deben ser conscientes del costo general de administrar una plataforma de aprendizaje automático e IA.

Hay bases de datos relacionales, como *MySQL, SQL Server y Oracle*. Como su nombre lo indica utilizan el modelo relacional y siempre es mejor usarlas cuando los datos son consistentes y ya se tiene algo planificado.

También existen las no relacionales, como *MongoDB* y *Redis*, conocidas como *NO-SQL* (*Not Only SQL*). Estas son más flexibles en cuanto a consistencia de datos y se han convertido en una opción que intenta solucionar algunas limitaciones que tiene el modelo relacional.

Además, hay otras BBDD no tan tradicionales, como las basadas en grafos o aquellas que tienen información cartográfica, que pueden servir, por ejemplo, si estás creando un *e-commerce* para encontrar relaciones entre los productos y las preferencias de los usuarios.

Tener un buen diseño de base de datos desde el comienzo te puede ayudar a ahorrar tiempo. Las bases de datos relacionales y no relacionales se organizan de formas diferentes y trabajan con tipos de datos distintos, así que es importante entender cómo se diseña cada una. En el caso de bases de datos relacionales se trabaja con el estándar *SQL*, que se usa para actualizar o recuperar datos.

Un ejemplo claro para entender las bases de datos es el funcionamiento de una aerolínea. Todas las aerolíneas tienen aplicaciones para hacer reservas y trabajan con diferentes rutas. Todas las personas que van a viajar tienen que proveer datos para hacer una reservación de una ruta específica, por ejemplo, la fecha en que van a viajar. Para esto se hace necesario un repositorio en el que se pueda almacenar esta información y se pueda cruzar, además, con una cantidad de sillas por avión. Una base de datos permite no redundar en los datos.

Bibliografía

[1] O., Zaiane. CMPUT690. *Principles in knowledge discovery in databases. chapter i: introduction to data mining*. University of Alberta, 1999.

[2] H., Jiawei, K., Micheline, P., Jian, *Data mining: concepts and techniques*. Morgan Kaufmann, Elsevier. 2012.