Gráfico, Diagrama

Descripción generada automáticamente

Inteligencia de negocios hace referencia a un conjunto de productos y servicios para acceder a los datos, analizarlos y convertirlos en información a partir de Data Warehouses o Data Marts mediante herramientas de extracción, consulta y modelización. La minería de datos es un proceso de extracción automatizada de información sintetizada, relevante y con capacidad predictiva a partir de grandes bases de datos. Abarca diversas técnicas enfocadas a la visualización, análisis y modelización de procesos a partir de grandes cantidades de datos. Su objetivo es identificar patrones significativos que sean válidos, novedosos y potencialmente útiles para el negocio.

Aplicando técnicas de inteligencia artificial desde el punto de vista disciplinar, este proceso puede definirse como multidisciplinar, ya que abarca conceptos y prácticas de estadística aplicando inteligencia artificial sobre repositorios de datos, específicamente bases de datos multidimensionales.

Diagrama

Descripción generada automáticamente

Un proceso típico de minería de datos, entendiéndose como tal un conjunto de etapas sucesivas, consta de los siguientes pasos generales: a partir de los datos registrados en una base transaccional, se ejecuta primero una selección de un conjunto de datos que incluyan las variables que se quiere predecir, calcular o inferir y las que sirven para hacer el cálculo o proceso. Luego se preparan los datos para detectar la presencia de valores atípicos y la ausencia de datos o valores nulos. Se procede a la transformación del conjunto de datos de entrada. A este paso también se lo conoce como preprocesamiento de los datos ya limpios y válidos para almacenarlos en un Data Warehouse. La selección y aplicación de la técnica de minería de datos construye el modelo predictivo o descriptivo, seleccionando una técnica de minería apropiada (las veremos luego en detalle), y finalmente el quinto paso: extracción de conocimiento, interpretación y evaluación de datos. Como resultado de la técnica aplicada, se obtiene un modelo de conocimiento que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. Una vez obtenido el modelo, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno de los modelos alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos.

Texto, Carta

Descripción generada automáticamente

Veamos ahora la diferencia que hay entre consultas que se hagan a la base transaccional y la minería de datos. En los sistemas estándar de gestión de base de datos transaccionales, las consultas se resuelven accediendo a distintos conjuntos de datos almacenados. Por ejemplo, ventas del último mes de un producto. El usuario empieza con una consulta o una suposición, explora los datos y construye un modelo. Los sistemas de Data Mining infieren conocimiento de la base de datos en forma de estructuras y patrones. Por ejemplo, qué clientes son potenciales compradores de un producto. El usuario no propone modelos, el sistema los encuentra y los sugiere al usuario.

Diagrama

Descripción generada automáticamente

Este es un esquema de una arquitectura típica del proceso de minería de datos: la fuente de datos es el repositorio de los datos ya transformados, como lo son los Data Warehouses multidimensionales. El servidor de la base de datos se utiliza en los sistemas en red para gestionar la información. La base de conocimiento guía la búsqueda, evaluando los patrones, metadatos y la obtención de conocimiento previo de los datos, surgidos a partir de la investigación en inteligencia artificial como respuesta a las necesidades que las aplicaciones de esta disciplina planteaban. Contienen elementos de conocimiento, normalmente en la forma de hechos y reglas de inferencia, así como la manera en que estos han de ser utilizados. El algoritmo de minería de datos permite realizar diferentes tipos de análisis como caracterización, asociación, clasificación, análisis de grupos, evolución y análisis de desviaciones. Los veremos luego.

La evaluación de patrones son medidas para conocer la relevancia de los patrones identificados. De no ser significativa, se vuelve a ejecutar el proceso de minería de datos, enfocando en variables diferentes para obtener un nuevo modelo que pueda ser relevante. La interfaz gráfica finalmente permite la interacción con el usuario y la visualización de los patrones hallados bajo distintas formas.

Imagen de la pantalla de un celular con texto

Descripción generada automáticamente con confianza media

Las técnicas de minería de datos crean modelos que pueden ser predictivos o descriptivos. Un modelo predictivo responde preguntas sobre un comportamiento futuro de los datos. Un modelo descriptivo, en cambio, proporciona información sobre las relaciones encontradas entre los datos y sus características para poder entenderlos mejor.

Veamos ahora cada una de las técnicas que utilizan los modelos predictivos.

Texto

Descripción generada automáticamente con confianza baja

Imagen que contiene Diagrama

Descripción generada automáticamente

En base a una clasificación, se evalúa mediante modelos correlacionales las probabilidades de un resultado. Por ejemplo, si un cliente pagará o no pagará de acuerdo a determinados atributos o el tipo de dolencia que puede tener un paciente de acuerdo a los síntomas. En base a una regresión lineal, que es la más utilizada, se encuentran relaciones entre variables; por ejemplo, calcular el tiempo previsible que se empleará en corregir los errores de un desarrollo de software. Los modelos estadísticos son una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales para indicar los diferentes factores que modifican la variable de respuesta. Por ejemplo, análisis de varianza, discriminante, dispersión, etcétera.

La secuenciación intenta predecir el valor de una variable en función de su evolución en el tiempo a partir de este conocimiento y bajo el supuesto de que no van a producirse cambios estructurales. Por ejemplo, estimar la demanda de energía eléctrica para un periodo determinado. Siguiendo con las técnicas predictivas, tenemos el árbol de decisión: sobre los datos se van realizando sucesivas bifurcaciones hasta llegar a un resultado modelo. Se determina la división más discriminante entre los criterios seleccionados, aquella que permite diferenciar mejor a los distintos grupos del criterio base. Se obtiene de este modo la primera segmentación y se realizan nuevas segmentaciones de cada uno de los segmentos resultantes en forma iterativa. Por ejemplo, se supone que se desea conocer qué pasajeros del Titanic tuvieron más probabilidad de sobrevivir a su hundimiento y qué características estuvieron asociadas a la supervivencia del naufragio. Un procedimiento de árbol de decisión selecciona automáticamente los grupos homogéneos con la mayor diferencia en proporción de supervivientes; en este caso, ha encontrado el sexo (hombres y mujeres). El siguiente paso subdivide cada uno de estos grupos en función de otra característica, resultando que los hombres son divididos en adultos y niños, mientras que las mujeres se dividen en grupos basados en la clase en la que viajaban en el barco. Si un pasajero del Titanic es hombre y es adulto, tiene una probabilidad de sobrevivir, según este modelo, del 20%. La probabilidad de supervivencia de cada una de las subdivisiones puede utilizarse con fines predictivos para estimar el grado de probabilidad de supervivencia de los miembros de cada grupo.

Las reglas de asociación encuentran interrelaciones entre datos generando reglas del tipo "qué pasa si". A partir de un proceso de aprendizaje de modelo, por ejemplo, conocer en un hipermercado si un cliente que compra leche muy probablemente compre también pan. Se utilizan los llamados algoritmos de aprendizaje supervisados que se basan en entrenar un modelo por medio de diferentes datos conocidos para poder predecir un dato desconocido a priori. Trabaja sobre dos tablas: la de aprendizaje o entrenamiento y la tabla de verificación, en donde se aplica el modelo generado a partir de la tabla de aprendizaje para evaluar la calidad de este modelo para una correcta predicción.

Las redes neuronales son un paradigma de aprendizaje y procesamiento automatizado inspirado en la forma en que funciona el sistema nervioso humano. Cada centro neuronal recibe múltiples estímulos simultáneos que determinan una única salida. Se trata de un sistema de interconexión de nodos neuronales en una red que colabora para producir un único estímulo de salida. Se estructuran como un grupo numeroso de procesadores conectados entre sí, trabajando en paralelo. Aprenden a través del entrenamiento. Las tareas básicas de las redes neuronales son reconocer, clasificar, agrupar, asociar, almacenar patrones y someter a los datos a un tratamiento mediante algoritmos que dan lugar a resultados de índole predictiva.

Diagrama

Descripción generada automáticamente

En cuanto a las técnicas usadas por los modelos descriptivos, tenemos el agrupamiento o clustering, que clasifica elementos en grupos definidos por el usuario en base a sus características, agrupando los datos según indicadores de frecuencia en segmentos de comportamiento homogéneo. El objetivo es maximizar la similitud de elementos del clúster y las diferencias entre clústers. Cada elemento debe estar asignado a un único clúster. Su finalidad esencial es revelar concentraciones en los datos para su agrupamiento eficiente en estos clústers según su homogeneidad, pudiendo utilizarse variables cualitativas o cuantitativas. El agrupamiento se realiza basándose en la proximidad o lejanía de unos con otros. Es esencial el uso adecuado del concepto de distancia. Por ejemplo, en los análisis de clientes con tarjeta de crédito de tipo Gold por género, hijos, estado civil e ingresos, dio como resultado este ejemplo de agrupamiento. Los métodos simétricos o de aprendizaje no supervisado trabajan sobre elementos clasificados inductivamente en grupos que no son conocidos con anterioridad, a diferencia de la técnica de clustering donde son definidos por el usuario, de modo que todas las variables son tratadas en el mismo nivel de importancia y no hay una hipótesis previa de causalidad.