REGLAS DE ASOCIACIÓN

Búsqueda de patrones frecuentes, asociaciones, correlaciones o estructuras causales entre conjuntos de elemento u objetos en bases de datos de transacciones, bases de datos relacionales y otros repositorios de información disponibles

APLICACIONES

Diseño de catálogos

OBJETIVO

El objetivo de la minería de reglas de asociación es encontrar todas las reglas teniendo:
UMBRAL MÍNIMO DE SOPORTE
UMBRAL MÍNIMO DE CONFIANZA

Enfoque de fuerza bruta

- Lista todas las reglas de asociación posibles
- Compruebe el soporte y la confianza para cada regla
- Elimine las reglas que fallan en los umbrales mínimo

TECNICAS: RAM, PRINCIPIO APRIORI, ECLAT.

RAM: ENFOQUE DE DOS PASOS

Generación de elementos frecuentes: Generar todos los conjuntos de elementos cuyo soporte ≥ min sup.

Generación de reglas: Generar reglas de alta confianza a partir de un conjunto de elementos frecuentes. Cada regla es una petición binaria de un conjunto de elementos frecuentes.

Cada conjunto de elementos en la red es un conjunto de elementos frecuente-candidato.

PRINCIPIO APRIORI

Si un conjunto de elementos es frecuente, entonces todos sus sus subconjuntos también deben ser frecuentes. El principio de A priori se mantiene debido a la siguiente propiedad de la medida de soporte:

$$\forall X, Y : (X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

El soporte de un conjunto de elementos nunca excede el soporte de sus subconjuntos. Esto se conoce como la propiedad anti-monótona de soporte.

Algoritmo

Identificar todos los itemsets que ocurren con una frecuencia por encima de un determinado límite (itemsets frecuentes).

Convertir esos itemsets frecuentes en reglas de asociación.

PREDICCIÓN

La predicción es una técnica que se utiliza para proyectar los tipos de datos que se verás en el futuro o predecir el resultado de un evento lo cual te puede ayudar a precisar algo que sucederá en el futuro.

- Existen cuestiones relativas a la relación temporal de las variables de entrada o predictores de la variable objetivo.
- Los valores son generalmente continuos.
- Las predicciones son a menudo (no siempre) sobre el futuro.

Variables independientes – atributos ya conocidos

Variables de respuesta – Lo que queremos saber

Cualquiera de las técnicas utilizadas para la clasificación y la estimación puede ser adaptada para su uso en la predicción mediante el uso de ejemplos de entrenamiento donde el valor de la variable que se predijo que ya es conocido, junto con los datos históricos de esos ejemplos.

APLICACIONES

- Revisar historiales crediticios de las consumidoras y compras hacia un riesgo crediticio.
- Predecir el precio de una propiedad.
- Predecir si va a llover en función de la humedad actual.
- Predecir la puntuación de cualquier equipo durante un partido de fútbol.

TÉCNICAS

1. *Métodos de regresión:*
    a. Regresión lineal: El objetivo del Análisis de regresión es determinar una función matemática sencilla que describa el comportamiento de una variable dados los valores de otra u otras variables.
    b. Regresión lineal multivariante: Permite generar un modelo lineal en el que el valor de la variable dependiente o respuesta y, se determina a partir de un conjunto de variables independientes llamadas predictores $X_1,X_2,\ldots,X_n$; Es una extensión de la regresión lineal simple.
    c. Regresión no lineal multivariante: Es una regresión en la que las variables dependientes o de criterio se modelan como una función no lineal de los parámetros del modelo y una o más variables independientes.
2. *Redes neuronales:* Utiliza los datos para modificar las conexiones ponderadas entre todas sus funciones hasta que sea capaz de predecir los datos con precisión.

REGRESIÓN

Una Regresión es un modelo matemático para determinar el grado de dependencia entre una o más variables, es decir conocer si existe relación entre ellas.

Existen dos tipos de regresión:

Regresión Lineal: cuando una variable independiente ejerce influencia sobre otra variable dependiente.

Regresión Lineal Múltiple: cuando dos o más variables independientes influyen sobre una variable dependiente.

En materia de minería de datos se llama **Predictivo** la cual tiene como objetivo analizar los datos de un conjunto y en base a esto predecir lo que puede ocurrir con ese conjunto de datos en un futuro.

Permite examinar la relación entre dos o más variables e identificar cuáles son las que tienen mayor impacto en un tema de interés.

Variable(s) dependiente(s): Es el factor más importante, el cual se está tratando de entender o predecir.

Variable(s) independiente(s): Es el factor que tú crees que puede impactar en tu variable dependiente.

El análisis de regresión nos permite explicar un fenómeno y predecir cosas acerca del futuro, por lo que nos será de ayuda para tomar decisiones y obtener los mejores resultados.

El objetivo es determinar una gráfica o tendencia recta con la ecuación de la forma:

$$y = mx + b$$

Por lo tanto necesitamos m y b

Entonces

$m = \frac{\sum x \sum y - n\sum(xy)}{(\sum x)^2 - n\sum x^2}$ y $b = \bar{y} - m\bar{x}$

Para determinar el nivel de eficacia del ajuste existen varios parámetros estadísticas, uno de ellos es el siguiente:

$R = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$ *donde;* $\sigma_x = \sqrt{\frac{\sum(x^2)}{n} - \bar{x}^2}$; $\sigma_y = \sqrt{\frac{\sum(y^2)}{n} - \bar{y}^2}$; $\sigma_{xy} = \frac{\sum(xy)}{n} - \bar{x} * \bar{y}$

CLUSTERING

También conocido como agrupamiento, es una de las técnicas de minería de datos, el proceso consiste en la división de los datos en grupos de objetos similares.

Las técnicas de Clustering son las que utilizando algoritmos matemáticos se encargan de agrupar objetos. Usando la información que brindan las variables que pertenecen a cada objeto se mide la similitud entre los mismos, y una vez hecho esto se coloca en clases que son muy similares internamente y a la vez diferente entre los miembros de las diferentes clases.

Cluster es una colección de objetos de datos similares entre sí dentro de un mismo grupo y no similar a objetos de otros grupos.

Análisis de cluster dado un conjunto de puntos de datos tratar de entender su estructura. Encuentra similitudes entre los datos de acuerdo con las características encontradas en los datos. Es un aprendizaje no supervisado ya que no hay clases predefinidas.

APLICACIONES

- *Estudios de terremotos*: los epicentros del terremoto observados deben agruparse a lo largo de fallas continentales.
- *Aseguradoras:* Identificación de grupos de asegurados de seguros de un grupo con un alto costo promedio de reclamos.
- *Planificación de la ciudad:* Identificación de grupos de casas según su tipo de casa, valor, ubicación y geografía.
- *Marketing:* Ayuda a profesionales a descubrir distintos grupos en sus bases de datos en cuanto a clientes.
- *Uso del suelo:* Identificación de áreas de uso similar de la tierra en una base de datos de observación de la tierra.

MÉTODOS DE AGRUPACIÓN

- Asignación jerárquica frente a punto
- Datos numéricos y/o simbólicos
- Determinística vs probabilística
- Exclusivo vs Superpuesto
- Jerárquico
- De arriba-abajo y abajo-arriba

ALGORITMOS

1. **Simple K-Means**
   Este algoritmo debe definir el número de clusters que se desean obtener, así se convierte en un algoritmo voraz para particionar.
2. **X-Means**
   Este algoritmo es una variante mejorada del **K-Means**. Su ventaja fundamental está en haber solucionado una de las mayores deficiencias presentadas en K-Means, el hecho de tener que seleccionar a priori el número de clusters que se deseen obtener.
3. **Cobweb**
   Se caracteriza por la utilización de aprendizaje incremental, esto quiere decir que realiza las agrupaciones instancia a instancia. Durante la ejecución del algoritmo se forma un árbol donde las hojas representan segmentos y el nodo raíz engloba por completo el conjunto de datos.
4. **EM**
   Este tipo de algoritmo se pueden utilizar para segmentar conjuntos de datos. Está clasificado como un método de particionado y recolocación, es decir, **Clustering Probabilístico**. Se trata de obtener la función de densidad probabilística desconocida a la que pertenecen el conjunto completo de datos.

CLASIFICACIÓN

Es una técnica de la minería de datos.

Es el ordenamiento o disposición por clases tomando en cuenta las características de los elementos que contiene.

La clasificación tiene muchos métodos y estos son algunos de ellos:

Análisis discrimente: Método utilizado para encontrar una combinación lineal de rasgos que separan clases de objetos o eventos.

Árboles de decisión: Método analítico que a través de una representación esquemática facilita la toma de decisiones.

Reglas de clasificación: Buscan términos no clasificados de forma periódica, si se encuentra una coincidencia se agrega a los datos de clasificación.

Redes neuronales artificiales: (También conocido como sistema conexionista) es un modelo de unidades conectadas para trasmitir señales.

CARACTERÍSTICAS

- Precisión en la predicción
- Eficiencia
- Robustez
- Escalabilidad
- Interpretabilidad

OUTLIERS

Detección de Outliers: Estudia el comportamiento de valores extremos que difieren del patrón general de una muestra, así como los valores atípicos.

Los valores atípicos son observaciones cuyos valores son muy diferentes a las otras observaciones del mismo grupo de datos, los cuales son ocasionados por:

- Errores de entrada de datos y procedimiento
- Acontecimientos extraordinarios
- Valores extremos y/o faltantes
- Causa no conocidas

Los datos atípicos distorsionan los resultados. Se calculan con dos métodos principales:

a. Métodos univariantes de detección de outliers
b. Métodos multivariantes de detección de outliers

TÉCNICAS

- Prueba de Grubbs
- Prueba de Dixon
- Prueba de Tukey
- Análisis de valores
- Regresión simple

Al detectarse se puede eliminar o sustituir si se corrobora que los datos atípicos se deben a un error de captura o en la medición de la variable.

PATRONES SECUENCIALES

Minería de Datos Secuenciales: Es la extracción de patrones frecuentes relacionados con el tiempo u otro tipo de secuencia. Son eventos que se enlazan con el paso del tiempo,
El orden de acontecimientos es considerado.

Se busca asociaciones de la forma "si sucede de la forma X en el instante de tiempo t entonces sucederá en el evento Y en el instante t+n". El objetivo es poder describir de forma concisa relaciones temporales que existen entre los valores de los atributos del conjunto de ejemplos.

Reglas de asociación secuencial: Expresan patrones secuenciales, esto quiere decir que se dan en instantes distintos en el tiempo.

Características

El orden importa.

- Objetivo: encontrar patrones secuenciales.
- El tamaño de una secuencia es su cantidad de elementos.
- La longitud de la secuencia es la cantidad de ítems.
- El soporte de una secuencia es el porcentaje de secuencias que la contienen en un conjunto de secuencias S.
- Las secuencias frecuentes son las subsecuencias de una secuencia que tiene soporte mínimo.

Tipos de datos

1. ADN y proteínas
2. Recorrido de clientes en un supermercado
3. Registros de accesos a una página web

Aplicaciones

- Medicina: Predecir si un compuesto químico causa cáncer (Agrupamiento de patrones secuenciales).
- Análisis de mercado: Comportamiento de compras (Agrupamiento de patrones secuenciales).
- Web: Reconocimiento de spam de un correo electrónico (Clasificación de datos secuenciales).

VISUALIZACIÓN

La visualización de datos es la presentación de información en formato ilustrado o gráfico. Al utilizar elementos visuales como cuadros, gráficos o mapas, nos proporciona una manera accesible de ver y comprender tendencias, valores atípicos y/o patrones en los datos. Dicho de otra forma, permite a los tomadores de decisiones analizar la información presentada de forma visual de modo que puedan captar conceptos difíciles o identificar nuevos patrones.

- Es importante conocer que existen diferentes tipos de Visualización de datos ya que uno de los grandes retos que enfrentan los usuarios de empresas, es que tipo de elemento visual se debe

utilizar para representar la información de la mejor forma. Aunque existen muchos tipos, mencionaremos los mas comunes:

- **Gráficos**: Este es el tipo más común y conocido, que utilizamos en nuestro día a día con las hojas de cálculo, para representar datos de manera sencilla, como Gráficos Circulares, Líneas, Columnas y Barras aisladas o agrupadas, Burbujas, áreas, Diagramas de Dispersión y Mapas de tipo Árbol.
- **Mapas**: Con la popularización de Google Maps y su conocida API (interfaz de programación para aplicaciones), todos conocemos la visualización de datos en mapas para conocer, por ejemplo, la localización de nuestra flota de vehículos en tiempo real o bien la de las tiendas de un supermercado o los cajeros automáticos de nuestro banco en un mapa.
- **Infografías:** Una infografía es una colección de imágenes, gráficos y texto simple que resume un tema para que se pueda entender fácilmente. son excelentes para ayudarnos a procesar más fácil, la información compleja.
- **Cuadros de mando:** Es una herramienta que permite saber en todo momento el estado de los indicadores del negocio: de ventas, económicos, de producción, de recursos humanos, etc. y que nos dice lo que está pasando en la empresa (idealmente en tiempo real) para poder tomar decisiones adecuadas, ya sean correctivas o de planeación.
-

Aplicaciones

1. **Comprender la información con rapidez**
2. **Identificar relaciones y patrones**
3. **Identifique tendencias emergentes**
4. **Comunique la historia a otras personas**