UNIVERSIDAD AUTÓNOMA DE NUEVO LÉON

FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

MÍNERÍA DE DATOS

RESÚMEN DE PRESENTACIONES GRUPALES

MAESTRA: MAYRA CRISTINA BERRONES REYES

ALUMNA: KARYME MAYELA GAUNA RODRÍGUEZ.

MATRÍCULA: 1819032

GRUPO:003

SÁN NICÓLAS DE LOS GARZA, 25 DE SEPTIEMBRE DEL 2020

REGLAS DE ASOCIACIÓN

Reglas de asociación es una técnica de inteligencia artificial ampliamente utilizada en Data Mining.

Data Mining es el proceso de descubrimiento de tendencias o patrones en grandes bases de dato con el objetivo de guiar futuras decisiones

Las **reglas de asociación** sirven para describir una relación de asociación entre los elementos de un conjunto de datos relevantes

Soporte: fracción de transacciones que contiene un itemset.

1000 transacciones

400 pan

Entonces el soporte es 400, o 40%, la probabilidad que aparezca pan en una transacción

Si soporte mide frecuencia, confianza mide la fortaleza de la frecuencia

Confianza (c): mide que tan frecuente items en Y aparecen en transacciones que contienen X. Probabilidad condicional

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 partición binaria de un conjunto de elementos frecuente.

Principio de Apiorí : si un conjunto de elementos es frecuente, entonces todos sus subconjuntos también deben ser frecuentes.

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

VALORES ATíPICOS

Técnicas para la detección de valores atípicos

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

Programas para la detección

- Excel
- Minitab

Una vez detectados los valores atípicos, se pueden eliminar o sustituir si se corrobora que los datos atípicos se deben a un error de captura o en la medición de la variable.

Aplicaciones de la minería de datos en outliers

- Detección de fraudes financieros
- Tecnología informática y telecomunicaciones
- Nutrición y salud
- Negocios

Outliers tipo:

- Error (error en carga de datos)
- Límites(valores que se escapan del grupo medio)
- Punto de interés (casos anómalos)

REGRESIÓN

Una regresión es un modelo para determinar el grado de dependencia entre una o más variables. Tipos de regresión:

- Regresión lineal (Una variable independiente influye entre otra dependiente)
- Regresión lineal múltiple (dos o más variables independientes influyen sobre una dependiente)

La regresión en la minería de datos tiene como objetivo analizar datos de un conjunto y en base a eso predecir lo que ocurrirá en un futuro.

El análisis de la regresión nos permite analizar la relación entre dos o más variables e identificar cuales son los que tienen mayor impacto, además de ayudarnos a tomar mejores decisiones en base a lo obtenido.

- Variables dependientes (Factor que se trata de predecir)
- Variables independientes (factor que se cree que puede impactar a la v.dependiente)

Regresión lineal

Se tiene como objetivo obtener la ecuación de la forma Y= mx +b

En general se utiliza para ver qué tan bueno es un modelo.

CLAUSTERING

El proceso consiste en la división de los datos en grupos de objetos similares. Las técnicas son las que utilizando algoritmos se encargar de agrupar objetos

Un cluster es una colección de objetos de datos. Similares entre si dentro del mismo grupo.

Análisis de clauster: dado un conjunto de puntos de datos tratar de entender su estructura. De acuerdo con las características presentadas en los datos se encuentran las similitudes de los mismos

Aplicaciones:

- Estudio de terremotos
- Aseguradoras
- Planificación de la ciudad
- Uso de suelo
- Marketing

Simple K-Means

Es un algoritmo que debe definir el número de clusters que se desean obtener, por consiguiente se convierte en un algoritmo voraz para particionar.

Cobweb

Este algoritmo realiza las agrupaciones instancia a instancia. Una vez ejecutado el algoritmo se desarrlla un árbol de clasificación, donde las hojas representan los segmentos y el nodo raíz engloba al conjunto de datos que se tiene.

En el que se toma en cuenta dos parámetros

- Acuity: La utilidad de la categoría está basada en la estimación de la media y la desv. Est. Del valor de un atributo para un nodo en particular
- Cut-off: Se utiliza para evitar el crecimiento descontrolado de la cantidad de segmentos.

EM

Se utiliza para segmentar conjunto de datos . Se trata de obtener la Función de Densidad de Probabilidad (FDP) desconocida a la que pertenecen el conjunto total de datos.

- Expectacion: Utiliza los valores de los parámetros, iniciales o proporcionados por el paso maximization obteniendo diferentes formas de FDP
- Maximization: Obtiene nuevos valores de los parámetros a partir del paso anterior.

PREDICCIÓN

Técnica utilizada para procetar los tipos de datos que se verán en el futuro o realizar una preddicion para el resultado de un evento. Para realizar una predicción es suficiente comprender y reconocer las tendencias históricas.

Variables independientes(atributos ya conocidos)

Variables de respuesta(lo que queremos saber)

Aplicaciones

- Revisar los historiales crediticios
- Predecir el tiempo en una entidad
- Precio de venta de alguna propiedad
- Puntuación de algún partido de futbol

Técnicas

- Modelos estadísticos
- Estadísticas no lineales
- Redes neuronales, RBF

Métodos de regresión

- Regresión lineal: tiene como objetivo determinar una función matemática que describa el comportamiento de una variable dado los valores de otras.
- Regresión líneal multivariante: genera un modelo lineal en el que el valor de la variable dependiente se determina en base al conjunto de las v. indeoendientes.

PATRONES SECUENCIALES

La minería de datos secuenciales, se refere a la extracción de patrones frecuentes relacionados con el tiempo u otro tipo de secuencia.

El objetivo es describir de forma concisa relacione temporales que existen entre los valores de atributos del conjunto.

Las reglas de asociación secuencial expresan patrones secuenciales.

Características:

- Es importante el orden
- La cantidad de elementos representa el tamaño de la secuencia
- Su objetivo es encontrar patrone secuenciales
- La cantidad de ítems es la longitud de la misma

Las ventajas que tienen los patrones secuenciales son la flexibilidad y eficiencia, mientras que sus desventajas son que esta sesgado por los primeros patrones y la utilización.

Aplicaciones

- Comportamiento en las compras (supermercado, ropa, etc)
- Predecir si un compuesto químico puede llegar a provocar cáncer.
- Reconocer el spam en los correos electrónicos

Secuencias

|s| es el número de elementos en una secuencia

Una k-secuencia es una secuencia con k eventos

Subsecuencias

Una subsecuencia también es una secuencia, pero esta se encuentra dentro de otra, con la condición de cumplir ciertas normas.

Análisis de secuencias

1. Bases de datos
2. Secuencia
3. Elemento(transacción)
4. Evento (Ítem)

Método GSP(Generalized sequential pattern)

1. Recorrer la base de datos
2. Generar k – secuencias candidatas a partir de las (k-1) secuencias frecuentes
3. Podar k secuencias candidatas que contengan alguna secuencia no frecuente
4. Obtener el soporte de candidatas
5. Eliminar las k – secuencias candidatas cuyo soporte

VISUALIZACIÓN DE DATOS

La visualización de datos es la presentación de información en formato gráfico, esto nos ayuda a comprender y ver tendencias valores atípicos y/o patrones en los datos.

Tipos:

- Gráficos: gráficos circulare, líneas, columnas, barras aisladas o agrupadas, burbujas, áreas, Diagramas de Dispersiónm etc.
- Mapas
- Infografías: Colección de imágenes gráficos y texto simple que resume un tema para que se pueda entender con facilidad.
- Cuadros de Mando : Herramienta que nos permite conocer el estado de los indicadores de alguno negocio.

Aplicaciones

- Comprender la información con rápidez
- Identificar relaciones y patrones
- Identifique tendencias emergentes
- Comunicar la historia a otras personas

La visualización de datos es importante debido a que es una herramienta cada vez más importante para darle el sentido a los billones de filas de datos que se generan día con día; además de ayudarnos a contar historia seleccionando los datos de manera más fácil de comprender, destacando la tendencias y valores atípicos.

CLASIFICACIÓN

Una clasificación es aquella técnica de la minería de datos encargada de ordenar por clases tomando en cuenta diversas características de sus elementos

D={t1,t2,…..tn) base de datos de tuplas

C=(C1….Cm) clase

Datos de clasificación

- Empareja dato a grupos predefinidos
- Encuentra modelos que describen y distinguen clases

Métodos en la clasificación de datos

- Análisis discriminante: Se usa para encontrar una combinación lineal de rasgos que separan clases de objetos.
- Árboles de decisión: A través de una representación esquemática facilita la toma de decisiones
- Reglas de clasificación
- Buscan términos no clasificados de forma periódica.
- Redes neuronales artificiales: Modelo de unidades conectadas para transmitir señales.

Características

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