Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

Alumno: Ricardo Zarek Sánchez Olvera

Matrícula 1795134

Materia: Minería de Datos

REGLAS DE ASOSIACIÓN

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

Sus aplicaciones vendrían siendo:

Análisis de datos de la banca, Corss-marketing, Diseño de catálogos

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

El núcleo del algoritmo de Apriori

•Utilizar los conjuntos frecuentes (k 1 para generar candidatos a k ítems frecuentes

•Utilizar el escaneo de la base de datos y la coincidencia de patrones para recoger los recuentos de los conjuntos de elementos candidatos

Comprimir una gran base de datos en una estructura compacta de árbol de patrones frecuentes (FP tree)
Muy condensado, pero completo para la minería de patrones frecuentes Evita costosos análisis de bases de datos

Utilice un método de minería de patrones frecuentes, basado en el árbol de FP
Una metodología de dividir y conquistar descomponer las tareas de minería en los más pequeños

CLASIFICACIÓN

Tareas predictivas: Predecir el valor de un atributo en particular basándose en los datos recolectados de otros atributos

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

Métodos:

Análisis discriminante: método utilizado para encontrar una combinación lineal de rasgos que separan clases de objetos o eventos

Arboles 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

DETECCIÓN DE OUTLIERS

Estudia el comportamiento de valores extremos que difieren del patrón general de una muestra

Un valor atípico son observaciones cuyos valores son muy diferentes a las otras observaciones del mismo grupo de datos, estos datos distorsionan los resultados de los análisis y por esta razón hay que identificarlos y tratarlos de manera adecuada

Existen distintos tipos de técnicas pero se pueden separar en dos categorías principales:
Metodos univariantes de detección de outliers
Metodos multivariantes de detección de outliers

Una vez detectados los datos atípicos, eliminarlos o sustituirlos puede modificar las inferencias que se realicen debido a que introduce un sesgo, disminuye el tamaño real y puede afectar la distribución o varianzas, la mejor opción es quitarle peso mediante otras técnicas

PATRONES SECUANCIALES

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

- 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.

Fases del método GSP (Generalized Sequential Pattern)
Fase 1:
Recorrer la base de datos para obtener todas las secuencias frecuentes de 1 elemento.
Fase 2:

Generación:
Generar k-secuencias candidatas a partir de las (k-1)-secuencias frecuentes.
Poda:
Podar k-secuencias candidatas que contengan alguna (k-1)-secuencia no frecuente.
Conteo:
Obtener el soporte de las candidatas
Eliminación:
Eliminar las k-secuencias candidatas cuyo soporte real esté por debajo del Umbral de soporte mínimo de frecuencia

PREDICCION

La predicción es una técnica que se utiliza para proyectar los tipos de datos que se verán en el futuro o predecir el resultado de un evento.

En muchos casos, el simple hecho de reconocer y comprender las tendencias históricas es suficiente para trazar una predicción algo precisa de lo que sucederá en el futuro.

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

La mayoría de las técnicas de predicción se basan en modelos
Matemáticos:
- Modelos estadísticos simples como regresión
- Estadísticas no lineales como series de potencias
- Redes neuronales, RBF, etc.
Todo basado en ajustar una curva a través de los datos, es decir,
Encontrar una relación entre los predictores y los pronosticados.

REGRESION

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:
1. Regresión Lineal: cuando una variable independiente ejerce influencia sobre otra variable dependiente.
2. Regresión Lineal Múltiple: cuando dos o más variables independientes influyen sobre una variable dependiente

En Minería de Datos la Regresión se encuentra dentro de la categoría Predictivo. Esta categoría tiene como objetivo analizar los datos de un conjunto y en base a eso, predecir lo que puede ocurrir con ese conjunto de datos en un futuro.

Un análisis de regresión 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.

VISUALIZACION DE DATOS

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.

•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 más comunes:
Gráficos, Mapas, Infografías, Cuadros de mando

A medida que la "era del bigdata" entra en pleno apogeo, la visualización es una herramienta cada vez más importante para darle sentido a los billones de filas de datos que se generan cada día. La visualización de datos ayuda a contar historias seleccionando los datos en una forma más fácil de entender, destacando las tendencias y los valores atípicos. Una buena visualización cuenta una historia, eliminando el ruido de los datos y resaltando la información útil.

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.

Un clúster es una colección de objetos de datos. Similares entre sí dentro del mismo grupo. Disimilar a los objetos en otros grupos.

Análisis de clúster: 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.

Metodos de agrupación:

1- Asignación jerárquica
2- Datos numéricos y/o símbolos
3- Determinística vs probabilística
4- Exclusivo vs superpuesto
5- Jerárquico vs plano
6- De arriba a abajo y de abajo a arriba

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 clúster que se deseen obtener, a X Means se le define un límite inferior K min (número mínimo de clúster) y un límite superior K Max (número máximo de clúster) y este algoritmo es capaz de obtener en ese rango el número óptimo de clúster, dando de esta manera más flexibilidad al usuario