UNIVERSIDAD AUTONOMA DE NUEVO LEON

FACULTAD DE CIENCIAS FISICO MATEMATICAS

MINERIA DE DATOS

GRUPO: 003

Maestra: Mayra Cristina Berrones Reyes

Resumenes de Tecnicas

Alumna: Marlene Calderón Rangel

Matricula: 1811330

Reglas de Asociación

Técnica que se utiliza en la inteligencia artificial. Describe una relación de asociación entre los elementos de un conjunto de datos relevantes.

Existen diferentes aplicaciones: -Análisis de datos de la banca - Cross marketing -Diseño de catálogos.

Tiene como objetivo encontrar todas las reglas teniendo umbral mínimo de soporte y umbral mínimo de confianza. Esto se refiere a que debido a lo que conocemos y estamos acostumbrados es la manera de asociar los datos y es así como se forman las reglas de asociación.

- Conjunto de elementos: una colección de uno o más artículos, o un conjunto de elementos que contienen k elementos.
- Recuento de soporte: frecuencia de ocurrencia de un ítem set.
- Confianza ©: mide que tan frecuente ítems en y a pertenecen en transacciones X

Enfoque de 2 pasos

Generación de elementos frecuentes: generar todos los conjuntos de elementos cuyo soporte sea mayor o igual al soporte mínimo

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 frecuentes.

Cada conjunto de elementos en la red es un conjunto de elementos frecuentes candidato. Calcular el soporte de cada candidato escaneando la base de dato.

Principio A priori

Se utiliza para reducir el número de elementos. Si un conjunto tiende a ser frecuente entonces los subconjuntos deben ser frecuentes. El soporte de un conjunto de elementos nunca excede el soporte de sus subconjuntos.

Detección de outliers

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

Los valores atípicos son valores diferentes a las observaciones del mismo grupo de datos. Los datos atípicos ocasionados por:

- Errores de entrada y procedimiento
- Acontecimientos extraordinarios
- Valores extremos

Existen distintos tipos de técnicas para detectarlos y se pueden dividir en dos categorías principales: -Métodos univariantes de detección - Métodos multivariables

Tecnicas para la detección de datos atípicos:

- Prueba de GRUBBS
- Prueba DIXION
- ANALISIS DE VALORES
- Regresión Simple

Al detectar los outliers podemos eliminarlos o sustituir si son valores atípicos que no aportan nada, pero hay que realizarlo con cuidado ya que podemos sesgar la muestra y puede afectar al tamaño de la muestra, podemos afectar a la varianza.

Aplicaciones de outliers.

- Detección de fraudes financieros: cuenta que se abre y no tiene actividad en un gran tiempo y de repente recibe una fuerte cantidad de dinero
- Tecnología informática y telecomunicaciones: detectar una falla del algoritmo que necesitamos procesar
- Nutrición y salud: al tomar un grupo de personas con buena salud y puede ser un valor atípico alguien con presión alta.
- Negocios: no puedes cambiar el giro del negocio con la información de dos outliers.

Regresión Lineal

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

-Regresión lineal es cuando una variable influye a otra

-Regresión lineal múltiple es cuando unas variables influye a otra

En la minería de datos se encuentra en la categoría de predictivo. El análisis de regresión permite examinar la relación entre dos o más variables. Hay dos tipos de variables:

- Variable(s) Dependiente(s): La variable que se intenta predecir
- Variable(s) Independiente(s): Es el factor que influye en tu variable dependiente

Nos ayuda para poder predecir lo que puede pasar en el futuro y mejoramiento de toma de decisiones gracias a este análisis. Nos permite clasificar qué factores impactan más, cómo se relacionan y cuánta seguridad nos brindan estos factores. Al mismo tiempo nos deja visualizar con muchos tipos de gráficos para entender la relación de estas variables.

Este procedimiento nos va dando una serie de factores los cuales son los siguientes:

-**La R** representa el coeficiente de correlación y significa el nivel de asociación entre las variables.

-**La R^2** representa el coeficiente de determinación, indica porcentualmente el cambio de la dependiente respecto a la independiente.

Se necesita saber si esta regresión es significativa para tener idea si existe estas relaciones entre cada uno. Para saber si lo es, se usa la prueba de significancia y que la R^2 ajustada sea muy alta.

Predicción

Es una técnica que se suele usar para proyectar los tipos de datos, para predecir el resultado de un evento. Con el objetivo de comprender las tendencias históricas es suficiente para trazar una predicción un poco precisa de lo que podría ocurrir en el futuro.

Se tienen ciertas cuestiones relativas a la relación temporal de las variables de entrada o predictoras de la variable objetivo:

- Los valores son generalmente continuos.
- Las predicciones suelen ser sobre el futuro.
- Las variables independientes corresponden a los atributos ya conocidos.
- Las variables de respuesta corresponden a lo que queremos saber.

Aplicaciones

- Banca: Revisar los historiales crediticios de los consumidores y las compras pasadas para predecir si serán un riesgo crediticio en el futuro.
- Clima: Predecir si va a llover en función de la humedad actual.
- Deportes: Predecir la puntuación de cualquier equipo durante un partido de fútbol.
- Inmobiliaria: Predecir el precio de venta de una propiedad.

Técnicas

Prácticamente las técnicas de predicción están basadas en modelos matemáticos y principalmente basados en ajustar una curva a través de los datos, esto se refiere a encontrar una relación entre los predictores y los pronosticados.

Las más comunes son: -Modelos estadísticos simples como regresión -Estadísticas no lineales como series de potencias -Redes neuronales, etc.

Clasificación

Es una técnica de la minería de datos, también es el ordenamiento o disposición por clases tomando en cuenta las características de los elementos que contiene.

Datos de la clasificación

- Empareja datos a grupos predefinidos, junta dependiendo del patrón que siguen los datos.
- Encuentra modelos que describen y distinguen clases o conceptos para futuras predicciones.
- La clasificación se considera como la técnica más sencilla y utilizada.

Métodos utilizados

Análisis discriminante: se utiliza para encontrar una combinación lineal de rasgos que separan clases de objetos.

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

Árboles de decisión*:* esté a través de una representación esquemática facilita la toma de decisiones. Solo puede tener un camino al cual seguir.

Redes neuronales artificiales: modelo de unidades conectadas para transmitir señales. Diferente a árbol de decisión tienes diversas respuestas.

Características de los métodos

1. Precisión en la predicción: capacidad de predecir correctamente, grado de cercanía entre la precisión y el valor real.
2. Eficiencia: realizar adecuadamente una función.
3. Robustez: habilidad de funcionar con ausencia de ciertos valores.
4. Escalabilidad: habilidad para trabajar con grandes cantidades de datos.
5. Interpretabilidad: entendimiento que brinda.

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.

El objetivo y una de las características principales es encontrar los patrones secuenciales.

Ventajas

Flexibilidad: Su comportamiento puede ajustarse gracias a su amplio conjunto de parámetros.

• Eficiencia: Cálculos muy sencillos, basta con recorrer una vez el conjunto de datos.

Desventajas:

• Utilización: Los valores adecuados para los parámetros son difíciles de establecer a priori, por lo que se suele emplear un proceso de prueba y error.

• Sesgado por los primeros patrones: Los resultados obtenidos dependen del orden de presentación de los patrones.

Los patrones secuenciales pueden ser aplicados en diferentes áreas, las principales son en la medicina al predecir si un compuesto químico causa cáncer, en análisis de mercado para analizar el comportamiento de compras y estos serían en agrupamiento de patrones secuenciales. Y en el internet para el reconocimiento de spam de un correo electrónico y este se realiza por medio de clasificación de datos secuenciales.

Clustering

Una técnica utilizada en la minería de datos, su proceso consiste en la división de los datos en grupos de objetos similares, esta técnica es la más utilizada en algoritmos matemáticos se encargan de agrupar objetos.

Esta se puede dividir en dos conceptos:

- Cluster: colección de objetos de datos similares entre si dentro del mismo grupo

- Análisis de cluster: Dado un conjunto de puntos de datos trata de entender su estructura, encontrar similitudes entre los datos.

Aplicaciones:

• Estudio de terremotos

• Planificación de la ciudad

• Marketing

• Aseguradoras

• Uso del suelo

Algoritmos de clustering Simple k-means: para utilizarlo debemos tener definido el número de clusters que se desean obtener. Asumir de forma aleatoria los centros de cada cluster, el algoritmo hará los siguientes pasos

1. Determina coordenadas del centroide
2. Determina la distancia de cada objeto a los centroides
3. Agrupa los objetos basados en la menor distancia

Visualización

Esta representa los datos en un formato ilustrados. Esto nos proporciona una manera accesible de comprender los datos de manera visual o gráfica.

Tipos de Visualización de Datos

- Gráficos
- Mapas
- Infografías
- Cuadros de mando

Aplicaciones

- Comprender la Información
- Identifica Relaciones y Patrones
- Identificar Tendencias Emergentes