UANL
UNIVERSIDAD AUTÓNOMA DE NUEVO LEÓN

FCFM
FACULTAD DE CIENCIAS FÍSICO MATEMÁTICAS

MINERIA DE DATOS

Resumen

GRUPO 03
CHRISTIAN SERVANDO GARZA GONZALEZ
1505813

Reglas de Asociación

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.

Algunas de sus aplicaciones son en los catálogos.

Las reglas de asociación de minera tienen un enfoque en dos pasos:

1. generación de elementos frecuentes: aquí todos los conjuntos de elementos cuyo soporte es mayor o igual al soporte mínimo.
2. generación de reglas: Se generan reglas de alta confianza a partir de un conjunto de elementos frecuentes. Cada regla es una partición binaria de un conjunto de elementos frecuente.

El principio de a priori habla de si un conjunto de elementos es frecuente, entonces todos sus subconjuntos también deben ser frecuentes. El algoritmo a priori fue uno de los primeros algoritmos desarrollados para búsqueda de reglas de asociación y sigue siendo uno de los mas empleados, tiene dos etapas:

1. Identificar todos los ítems sets que ocurren con una frecuencia por encima de un determinado limite (itemsets frecuentes)
2. Convertir esos itemsets frecuentes en reglas de asociación.

Detección de outliers

Estudia el comportamiento de valores extremos que difieren del patrón general de una muestra, o dicho de otra manera, identifica el valor atípico.

Hablar de valor atípico es referirse a observaciones cuyos valores son muy diferentes a las otras observaciones del mismo grupo de datos. Estos datos atípicos son ocasionados por:

- Errores de entrada de datos y procedimiento
- Acontecimientos extraordinarios
- Valores extremos y/o faltantes.

Los valores atípicos se calculan usando diferentes técnicas (métodos univariantes y multivariantes). Dentro de algunas técnicas para la detección de valores atípicos, se encuentran:

1. Prueba de grubbs
2. Prueba de Dixon
3. Prueba de Tukey (diagrama de caja)
4. análisis de valores atípicos de mahalanobis

5. regresión Simple (regresión por mínimos cuadrados)

Después de detectar los outliers, 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. En caso de que no se deba a un error, lo mejor sería quitarles peso a esas observaciones atípicas mediante tecinas robustas, ya que pueden introducir un sesgo.

Dentro de sus aplicaciones en la minería de datos, se encuentra:

- La detección de fraudes financieros
- Tecnología informática y telecomunicaciones
- nutrición y salud
- Negocios

Regresión lineal

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

Existen dos tipos:

1. regresión lineal: cuando una variable independiente ejerce influencia sobre otra variable dependiente
2. regresión lineal múltiple: cuando dos o mas variables independientes influyen sobre una variable dependiente.

En la minería de datos, la regresión se encuentra dentro de la categoría predictivo.

El análisis de regresión permite examinar la relación entre dos o mas variables e identificar cuales son las que tienen mayor impacto en un tema de interés. También 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.

Clustering

Este método es conocido como un agrupamiento donde 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.

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

Algunos ejemplos de donde se usa el clustering son:

- Aseguradoras
- Uso del suelo
- Marketing

- Planificación de la ciudad
- Estudios de terremotos

Los algoritmos de cluster más comunes son:

1. Simple K-means: este algoritmo debe definir el número de clusters que se desean obtener, asi se convierte en un algoritmo voraz para particionar.
2. X-means es una variante del k means
3. Cobweb: Pertenece a la familia de algoritmos jerárquicos. Se caracteriza por la utilización de aprendizaje incremental, esto quiere decir, que realiza las agrupaciones a instancia.
4. EM: Este algoritmo pertenece a una familia de modelos que se conocen como finite Mixture MOdels, los cuales se pueden utilizar para segmentar conjuntos de datos.

Predicción

Es una técnica que se utiliza para para proyectar los tipos de datos que se verán en el futuro o predecir el resultado de un evento. En algunos casos, el simple hecho de conocer y comprender las tendencias históricas es suficiente para trazar una predicción de lo que sucederá en el futuro.

Existen cuestiones relativas a la relación temporal de las variables de entrada o predictores de la variable objetivo. Esta técnica tiene relación con otras técnicas.

Dentro de sus aplicaciones:

1. Se puede revisar los historiales crediticios de los consumidores y las compras pasadas para predecir si serán un riesgo en el futuro.
2. Predecir si lloverá
3. Predecir el precio de venta de una propiedad

La mayoría de las técnicas de predicción se basan en modelos matemáticos:

- Modelos estadísticos simples como regresión

Dentro de los tipos de métodos de regresión:

1. 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.
2. Regresión lineal multivariante: permite generar un modelo lineal…
3. regresión no lineal uni variable y multivariable: Método para encontrar un modelo no lineal para la relación entre la variable dependiente y un conjunto de variables independientes.

4. Redes neuronales: utiliza los datos para modificar las conexiones ponderadas entre todas sus funciones hasta que sea capaz de predecir los datos con precisión.

Patrones secuenciales

Características:

- El orden importa
- Su objetivo es encontrar patrones secuenciales
- El tamaño de una secuencia es su cantidad de elementos
- La longitud de la secuencia es la cantidad de ítems

Una de las ventajas de los patrones secuenciales es que son flexibles y que son eficientes, ya que solo basta correr una vez para obtener un patrón mientras que una de las desventajas es que los primeros patrones pueden estar sesgados.

De los tipos más importantes:

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

Dentro de las aplicaciones, se separan en agrupamiento de patrones secuenciales como en medicina o análisis de mercado y en clasificación con datos secuenciales como el reconocimiento de spam de un correo electrónico en una pagina web.

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.

Métodos de clasificación:

- Análisis discriminante: método utilizado para encontrar una combinación lineal de rasgos que separan clases de objetos o eventos
- Reglas de clasificación: buscan términos no clasificados de forma periódica, si se encuentra una coincidencia se agrega a los datos de clasificación
- Arboles de decisión: método analítico que a través de una representación esquemática facilita la toma de decisiones
- Redes neuronales artificiales (también conocido como sistema conexionista) es un modelo de unidades conectadas para trasmitir señales

Las características finales que podemos ver en clasificación son:

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

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.

Dentro de los tipos de visualización se encuentran:

- Gráficos
- Mapas
- Infografías
- Cuadros de mando

Hay algunos software que utilizar los diferentes tipos de visualización de datos para representar la información, por ejemplo uno muy utilizado es Excel o por ejemplo se encuentra también R que es un software mas sofisticado.

Dentro de las aplicaciones:

- Comprender la información con rapidez Mediante el uso de representaciones gráficas de información de negocios, las empresas pueden ver grandes cantidades de datos de formas claras y cohesivas y sacar conclusiones a partir de esa información.
- Identificar relaciones y patrones. Incluso muy grandes cantidades de datos complicados comienzan a tener sentido cuando se presentan de manera gráfica; las empresas pueden reconocer parámetros con una correlación muy estrecha.
- Identifique tendencias emergentes. El uso de la visualización de datos para descubrir tendencias en los negocios y en el mercado puede dar a las empresas una ventaja sobre la competencia, y eventualmente tener un impacto en la base de operación.