UNIVERSIDAD AUTONÓMA DE NUEVO LEÓN

FCFM

Minería de Datos

Resumen Técnicas de Minería de Datos

M.C. Mayra Cristina Berrones Reyes

Arleth Alanis Aguirre
1801925
Grupo 3
Séptimo Semestre

27 septiembre 2020

Reglas de Asociación

Las reglas de asociación es una técnica que se utiliza en la inteligencia artificial en el data meeting lo que hace es describir una regla como su nombre lo indica de asociación entre los conjuntos de datos relevantes. Es la búsqueda de patrones frecuentes, asociaciones, correlaciones o estructuras causales entre conjuntos de elementos u objetos en bases de datos de transacciones, relacionales y otros repositorios de información disponible.

Los conceptos que se utilizan en las reglas de asociación son los siguientes:

- Conjunto de elementos: es la colección de uno o más artículos
- Recuento de soporte: es la frecuencia de ocurrencia de un ítemset.
- Confianza: mide que tan frecuentes son los ítems y aparecen en transacciones, que se puede ver como una probabilidad condicional entre los datos y su asociación

El objetivo es encontrar todas las reglas teniendo un umbral mínimo de soporte y umbral mínimo de confianza.

Reglas de la Asociación en la Minería enfocada en dos pasos

1. Generación de elementos frecuentes: es generar todos los conjuntos de elementos con soporte 2 mínimo superior.

2. Generación de reglas: generar las reglas de alta confianza a partir de un conjunto de elementos frecuentes. Cada regla es una partición binaria de un conjunto de elementos frecuentes.

Reglas de la Asociación Principio "Apriori"

Para las reglas de asociación utilizamos el Principio A priori que dice que, si un conjunto de elementos es frecuente, entonces todos sus subconjuntos también deben de ser frecuentes.

Clasificación

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

Métodos de la clasificación:

- Análisis discriminante: 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 transmitir señales.

Características de los métodos de clasificación:

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

Outliers

La detección de outliers estudia el comportamiento de valores extremos que difieren del patrón general de una muestra.

De acuerdo con la definición anterior, se puede decir que son los valores atípicos en un conjunto de datos, porque son observaciones cuyos valores son muy diferentes a las otras observaciones del mismo grupo de datos, por lo que distorsionan los resultados de los análisis y por esta razón hay que identificarlos y tratarlos de manera adecuada.

Técnicas para la detección de valores atípicos

- Prueba de Grubbs
- Prueba de Dixon
- Prueba de Tukey (Diagrama de caja)
- Análisis de valores atípicos de Mahalanobis
- Regresión simple

¿Qué hacer cuando se han detectado los outliers?

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

Si no es un error, eliminarlo o sustituirlo puede modficicar las inferencias que se realicen a partir de esa información, debido a que se introduce un sesgo, disminuye el tamaño muestral y/o puede afectar la distribución y varianzas. Por lo tanto, la mejor opción es quitarles peso a esas observaciones atípicas mediante técnicas robustas.

Aplicación de la minería de datos en outliers

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

Patrones secuenciales

Conceptos

Minería de datos secuenciales: es la extracción de patrones frecuentes de un conjunto de datos relacionados con el tiempo u otro tipo de secuencia, se enlazan con el paso del tiempo y, el orden de los acontecimientos es muy importante.

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

Características

- El orden importa.
- El 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.
- El soporte de una secuencia es el porcentaje de las secuencias que la contienen en un conjunto de secuencias S.
- Las secuencias frecuentes son las subsecuencias de una secuencia que tiene un soporte mínimo.

Ventajas:

- Flexibilidad
- Eficiencia

Desventajas:

- Utilización: es a prueba y error
- Sesgado por los primeros patrones.

Tipos de datos

- ADN y proteínas en la biotecnología y medicina.
- Recorrido de clientes en un supermercado en la mercadotecnia.
- Registros de acceso a una página web se utiliza en informática e internet.

Aplicaciones

- Agrupamiento de patrones secuenciales: objetos de un grupo similares se agrupa para las características entre sí y diferentes a otros grupos.
- Clasificación con datos secuenciales: funciona como un algoritmo de patrones que se repiten.

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 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.

En esta técnica existen cuestiones relativas a la relación temporal de las variables de entrada o predictores de la variable objetivo, los valores son generalmente continuos y como se mencionó anteriormente, las predicciones son a menudo sobre el futuro.

Las variables independientes son los atributos ya conocidos y las variables de respuesta son lo que queremos conocer.

Aplicaciones

- Revisar los historiales crediticios de los consumidores y las compras pasadas para predecir si son una opción rentable en el futuro.
- Predecir el precio de venta de una propiedad.
- Predecir si lloverá en función de la humedad actual.
- Predecir la puntuación de cualquier equipo durante un partido de futbol.

Técnicas

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

- Modelos estadísticos como regresión.
    - Regresión lineal
    - Regresión lineal multivariante
    - Regresión no lineal
    - Regresión no lineal multivariante.
- Estadísticas no lineales como series de potencias
- Redes neuronales y RBF.

La predicción de basa en ajustar una curva a estos modelos a partir los datos y, así encontrar una relación entre los predictores y los pronosticados.

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:

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, el objetivo es analizar los datos de un conjunto y con base a eso, predecir lo que puede ocurrir en un futuro.

Análisis de regresión

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

- Variables dependientes: es el factor más importante, el cuál se esta tratando de entender o predecir.
- Variables independientes: es el factor que tú crees que puede impactar en tu variable dependiente.

Además, 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.

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.

Tipos de visualización

- Gráficos
- Mapas
- Infografías
- Cuadros de mando

Aplicaciones

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

La visualización de datos es más importante a medida que la era del big data entra en pleno apogeo, la visualización es una herramienta cada vez más importante para darle sentido a los billones de datos que se generan cada día h ayudar a contar datos en una forma más fácil de entender, destacando las tendencias y los valores atípicos,

Clustering

Es una técnica donde le el proceso consiste en la división de los datos en grupos de objetos similares. Se encarga 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 colocan en clases que son muy similares internamente, y a la vez diferente entre los miembros de las diferentes clases.

Un cluster es una colección de objetos de datos similares entres si dentro del mismo grupo. Su objetivo es disimilar a los objetos en otros grupos, es decir que los datos tienen que tener algo parecido pero no iguales

El análisis del cluster dado con conjunto de puntos de datos es para tratar de entender su estructura, encuentra similitudes entre los datos de acuerdo a las características encontradas en los datos.

Aplicaciones del cluster (ejemplos)

Estudio de terremotos: los epicentros del terremoto observados luego de fallas continentales

Planificación de una ciudad: identificación de grupos de casas según su tipo de casa, de valor y ubicación geográfica.

Marketing: descubrir distintos grupos en sus grupos de clientes dado a sus preferencias.

Aseguradoras: identificar grupos de asegurados con un alto costo promedio de reclamo.

Uso del suelo: identificación de áreas de uso similar de la tierra.

Métodos de agrupación

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

Algoritmos del cluster

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.

Pasos:

1. Se determina la cantidad de clusters con los que se quiere agrupar la información.
2. Se asume de forma aleatoria los centros por cada clusters, después se determina las coordenadas del centroide, la distancia de cada objeto a los centroides y agrupa los objetos basados en la menor distancia.
3. Quedarán agrupados por clusters, los grupos de simulaciones según la cantidad de clusters que el investigador definió en el momento de ejecutar el algoritmo.

2. *X- Means*

Este algoritmo es una variante mejorado del K-Means, su ventaja fundamental está en haber solucionado una de las mayores diferencias presentadas en el anterior, en el X-Means se le define un límite inferior y un límite superior de clusters, y así se obtiene el número óptimo de clusters, dando más flexibilidad.

3. *EM*

Funciona para segmentar conjunto de datos, está clasificado como un método de particionado y recoleociación, es decir, clustering probabilístico, se trata de obtener la función de densidad de probabilidad desconocida a la que pertenecen el conjunto completo de datos.

4. *Cobweb*

Este algoritmo realiza las agrupaciones de instancia a instancia, durante su ejecución se forma un árbol de clasificación donde las hojas representan los segmentos y la raíz engloba por completo el conjunto de datos.